

Ricco Rakotomalala

Pratique de la Régression Linéaire Multiple

Diagnostic et sélection de variables

Version 2.1

Université Lumière Lyon 2

Avant-propos

Ce support décrit quelques techniques statistiques destinées à valider et améliorer les résultats fournis par la régression linéaire multiple. Il correspond à la dernière partie des enseignements d'économétrie (je préfère l'appellation *Régression Linéaire Multiple*) en L3-IDS de la Faculté de Sciences Economiques de l'Université Lyon 2 (<http://dis.univ-lyon2.fr/>).

Ce support se veut avant tout opérationnel. Il se concentre sur les principales formules et leur mise en oeuvre pratique avec un tableur. Autant que possible nous ferons le parallèle avec les résultats fournis par les logiciels de statistique. Le bien-fondé des tests, la pertinence des hypothèses à opposer sont peu ou prou discutées. Nous invitons le lecteur désireux d'approfondir les bases de la régression à consulter le document "Économétrie - Régression Linéaire Simple et Multiple" ([18]), accessible sur ma page de fascicules (http://eric.univ-lyon2.fr/~ricco/cours/cours_econometrie.html).

Un document ne vient jamais du néant. Pour élaborer ce support, je me suis appuyé sur différentes références, des ouvrages disais-je plus tôt, mais aussi des ressources en ligne qui sont de plus en plus présents aujourd'hui dans la diffusion de la connaissance.

Les seuls bémols par rapport à ces documents sont (1) le doute que l'on pourrait émettre sur l'exactitude des informations prodiguées, mais la plupart de leurs auteurs sont des enseignants-chercheurs qui font sérieusement leur travail; (2) une disponibilité plus ou moins aléatoire, au gré des migrations des serveurs et de la volonté de leurs auteurs, auquel il est très difficile de remédier¹; (3) les informations sont disparates, avec une absence d'organisation, à la différence des ouvrages qui suivent une ligne pédagogique très structurante.

Néanmoins, ces ressources en ligne renouvellent profondément le panorama des documents disponibles pour les enseignements. La gratuité n'est pas le moindre de leurs atouts.

Ce support n'engage que son auteur. Toutes suggestions ou commentaires qui peuvent en améliorer le contenu sont le bienvenu.

1. Je fais systématiquement des copies... mais je me vois très mal les diffuser moi même via ma page web.

Table des matières

Partie I La régression dans la pratique

1	Étude des résidus	7
1.1	Diagnostic graphique	7
1.1.1	Graphiques des résidus	7
1.1.2	Graphiques des résidus pour les données CONSO	12
1.2	Tester le caractère aléatoire des erreurs	13
1.2.1	Test de Durbin-Watson	15
1.2.2	Test des séquences	17
1.3	Test de normalité	19
1.3.1	Graphique Q-Q plot	20
1.3.2	Test de symétrie de la distribution des résidus	21
1.3.3	Test de Jarque-Bera	22
1.4	Conclusion	25
2	Points aberrants et points influents	27
2.1	Points aberrants : détection univariée	28
2.2	Détection multivariée sur les exogènes : le levier	30
2.3	Résidu standardisé	34
2.4	Résidu studentisé	37
2.5	Autres indicateurs usuels	41
2.5.1	DFFITS	41
2.5.2	Distance de COOK	41
2.5.3	DFBETAS	44
2.5.4	COVRATIO	45
2.6	Bilan et traitement des données atypiques	46
3	Colinéarité et sélection de variables	51
3.1	Détection de la colinéarité	51
3.1.1	Conséquences de la colinéarité	51

3.1.2	Illustration de l'effet nocif de la colinéarité.....	52
3.1.3	Quelques techniques de détection.....	52
3.2	Traitement de la colinéarité - Sélection de variables.....	55
3.2.1	Sélection par optimisation.....	56
3.2.2	Techniques basées sur le F partiel de Fisher.....	62
3.3	Régression stagewise.....	65
3.4	Coefficient de corrélation partielle et sélection de variables.....	67
3.4.1	Coefficient de corrélation brute.....	67
3.4.2	Coefficient de corrélation partielle.....	68
3.4.3	Calcul de la corrélation partielle d'ordre supérieur à 1.....	70
3.4.4	Procédure de sélection fondée sur la corrélation partielle.....	72
3.4.5	Équivalence avec la sélection fondée sur le t de Student de la régression.....	73
3.5	Les régressions partielles.....	74
3.5.1	Principe des régression partielles.....	74
3.5.2	Traitement des données CONSO.....	75
3.6	Régressions croisées.....	77
3.6.1	Principe des régressions croisées.....	77
3.6.2	Régressions croisées sur les données CONSO.....	79
3.7	Conclusion.....	82
4	Régression sur des exogènes qualitatives.....	83
4.1	Analyse de variance à 1 facteur et transposition à la régression.....	83
4.1.1	Un exemple introductif.....	84
4.1.2	ANOVA à 1 facteur.....	84
4.2	Inadéquation du codage disjonctif complet.....	87
4.2.1	Codage disjonctif complet.....	87
4.2.2	Régression sans constante et lecture des coefficients.....	88
4.2.3	Vers des solutions plus générales.....	89
4.3	Codage "Cornered effect" de l'exogène qualitative.....	90
4.3.1	Principe.....	90
4.3.2	Lecture des résultats.....	90
4.3.3	Application aux données LOYER.....	91
4.4	Comparaisons entre groupes.....	92
4.4.1	Comparaisons avec le groupe de référence.....	92
4.4.2	Comparaisons entre deux groupes quelconques.....	96
4.5	Régression avec plusieurs explicatives qualitatives.....	99
4.5.1	Régression sur les indicatrices.....	100
4.5.2	Prise en compte des interactions.....	105
4.5.3	Ajout de nouvelles indicatrices.....	105

4.5.4	Tester la significativité de l'interaction	107
4.5.5	Interprétation des coefficients	108
4.6	Régression avec un mix d'explicatives qualitatives et quantitatives	108
4.6.1	Interprétation des coefficients	109
4.6.2	Prise en compte des interactions	110
4.6.3	Lien avec la comparaison de régressions	112
4.7	Sélection de variables en présence d'exogènes qualitatives	113
4.7.1	Traitement groupé des indicatrices	114
4.7.2	Traitement individuel des indicatrices	118
4.8	Autres stratégies de codage d'une exogène qualitative nominale	120
4.8.1	Le codage "centered effect" d'une exogène qualitative nominale	120
4.8.2	Le codage "contrast effect" d'une exogène qualitative	123
4.9	Codage d'une exogène qualitative ordinale	126
4.9.1	Un exemple introductif	126
4.9.2	(In)adéquation du codage pour variable qualitative nominale	128
4.9.3	Utilisation du codage cumulatif	130
4.9.4	Codage "backward difference"	132
4.9.5	Codage "forward difference"	133
4.9.6	Codage "Helmert"	134
4.10	Codage polynomial orthogonal d'une exogène qualitative ordinale	136
4.10.1	Construction du codage	136
4.10.2	Régression sur les variables recodées	137
4.11	Les erreurs à ne pas commettre	139
4.11.1	Codage numérique d'une variable discrète nominale	139
4.11.2	Codage numérique d'une variable discrète ordinale	139
4.12	Conclusion pour le traitement des exogènes qualitatives	140
5	Rupture de structure	141
5.1	Régression contrainte et régression non-contrainte - Test de Chow	143
5.1.1	Formulation et test statistique	143
5.1.2	Un exemple	145
5.2	Détecter la nature de la rupture	146
5.2.1	Tester la stabilité de la constante	146
5.2.2	Tester la stabilité du coefficient d'une des exogènes	148
5.3	Conclusion	150
6	Détection et traitement de la non linéarité	153
6.1	Non linéarité dans la régression simple	153
6.1.1	Linéarisation par transformation de variables	153
6.1.2	Détecter numériquement la non-linéarité dans la régression simple	155

6.1.3	Tester l'adéquation d'une spécification	160
6.2	Non linéarité dans la régression multiple	162
6.2.1	Lecture des résidus	162
6.2.2	Résidus partiels et résidus partiels augmentés	163
6.2.3	Un exemple "réaliste" : traitement des données "mtcars" sous R.	166
A	Table de Durbin Watson	173
B	Gestion des versions	175
C	Fichiers associés à ce support	177
D	Tutoriels	179
	Littérature	181

La régression dans la pratique

La régression dans la pratique

Le véritable travail du statisticien commence après la première mise en oeuvre de la régression linéaire multiple sur un fichier de données. Après ces calculs, qu'on lance toujours "pour voir", il faut se poser la question de la pertinence des résultats, vérifier le rôle de chaque variable, interpréter les coefficients, etc.

En schématisant, la modélisation statistique passe par plusieurs étapes² : proposer une solution (une configuration de l'équation de régression), estimer les paramètres, diagnostiquer, comprendre les résultats, réfléchir à une formulation concurrente, etc.

Dans ce support, nous mettrons l'accent, sans se limiter à ces points, sur deux aspects de ce processus : le diagnostic de la régression à l'aide de l'analyse des résidus, il peut être réalisé avec des tests statistiques, mais aussi avec des outils graphiques simples ; l'amélioration du modèle à l'aide de la sélection de variables, elle permet entre autres de se dégager du piège de la colinéarité entre les variables exogènes.

Notations

Le point de départ est l'estimation des paramètres d'une régression mettant en jeu une variable endogène Y et p variables exogènes X_j . Nous disposons de n observations.

L'équation de régression s'écrit :

$$y_i = a_0 + a_1x_{i,1} + \dots + a_px_{i,p} + \varepsilon_i \quad (0.1)$$

où y_i est la i -ème observation de la variable Y ; $x_{i,j}$ est la i -ème observation de la j -ème variable ; ε_i est l'erreur du modèle, il résume les informations manquantes qui permettrait d'expliquer linéairement les valeurs de Y à l'aide des p variables X_j .

Nous devons estimer $(p + 1)$ paramètres. En adoptant une écriture matricielle :

$$Y = Xa + \varepsilon \quad (0.2)$$

les dimensions de matrices sont respectivement :

- $Y \rightarrow (n, 1)$
- $X \rightarrow (n, p + 1)$
- $a \rightarrow (p + 1, 1)$
- $\varepsilon \rightarrow (n, 1)$

La matrice X de taille $(n, p + 1)$ contient l'ensemble des observations sur les exogènes, avec une première colonne formée par la valeur 1 indiquant que l'on intègre la constante a_0 dans l'équation.

2. <http://www.itl.nist.gov/div898/handbook/pmd/section4/pmd41.htm>

$$\begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ 1 & x_{2,1} & \cdots & x_{2,p} \\ \vdots & & & \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{pmatrix}$$

Remarque 1 (Régression sans constante). Dans certains problèmes, la régression sans constante peut se justifier. Il y a p paramètres à estimer dans la régression. On peut aussi voir la régression sans constante comme une régression avec la contrainte $a_0 = 0$. Il faut simplement faire attention aux degrés de liberté pour les tests. Il faut noter également que le coefficient de détermination R^2 n'est plus interprétable en termes de décomposition de la variance, il peut prendre des valeurs négatives d'ailleurs.

Données

Autant que possible, nous utiliserons le même fichier de données pour illustrer les différents chapitres de ce support. On veut expliquer la consommation en L/100km de véhicules à partir de $p = 4$ variables : le prix, la cylindrée, la puissance et le poids (Figure 0.1). Nous disposons de $n = 31$ observations. Nous connaissons la marque et le modèle de chaque véhicule, cela nous permettra d'affiner certains commentaires.

i	Modèle Véhicule	x1 (Fr) Prix	x2 (cm3) Cylindrée	x3 (kW) Puissance	x4 (kg) Poids	y (l/100km) Consommation
1	Daihatsu Cuore	11600	846	32	650	5.7
2	Suzuki Swift 1.0 GLS	12490	993	39	790	5.8
3	Fiat Panda Mambo L	10450	899	29	730	6.1
4	VW Polo 1.4 60	17140	1390	44	955	6.5
5	Opel Corsa 1.2i Eco	14825	1195	33	895	6.8
6	Subaru Vivio 4WD	13730	658	32	740	6.8
7	Toyota Corolla	19490	1331	55	1010	7.1
8	Ferrari 456 GT	285000	5474	325	1690	21.3
9	Mercedes S 600	183900	5987	300	2250	18.7
10	Maserati Ghibli GT	92500	2789	209	1485	14.5
11	Opel Astra 1.6i 16V	25000	1597	74	1080	7.4
12	Peugeot 306 XS 108	22350	1761	74	1100	9.0
13	Renault Safrane 2.2 V	36600	2165	101	1500	11.7
14	Seat Ibiza 2.0 GTI	22500	1983	85	1075	9.5
15	VW Golf 2.0 GTI	31580	1984	85	1155	9.5
16	Citroen ZX Volcane	28750	1998	89	1140	8.8
17	Fiat Tempra 1.6 Liberty	22600	1580	65	1080	9.3
18	Fort Escort 1.4i PT	20300	1390	54	1110	8.6
19	Honda Civic Joker 1.4	19900	1396	66	1140	7.7
20	Volvo 850 2.5	39800	2435	106	1370	10.8
21	Ford Fiesta 1.2 Zetec	19740	1242	55	940	6.6
22	Hyundai Sonata 3000	38990	2972	107	1400	11.7
23	Lancia K 3.0 LS	50800	2958	150	1550	11.9
24	Mazda Hachtback V	36200	2497	122	1330	10.8
25	Mitsubishi Galant	31990	1998	66	1300	7.6
26	Opel Omega 2.5i V6	47700	2496	125	1670	11.3
27	Peugeot 806 2.0	36950	1998	89	1560	10.8
28	Nissan Primera 2.0	26950	1997	92	1240	9.2
29	Seat Alhambra 2.0	36400	1984	85	1635	11.6
30	Toyota Previa salon	50900	2438	97	1800	12.8
31	Volvo 960 Kombi aut	49300	2473	125	1570	12.7

Fig. 0.1. Tableau de données CONSO - Consommation des véhicules

Nous effectuons sous TANAGRA une première régression sur l'ensemble des exogènes. Nous en extrayons quelques informations importantes (Figure 0.2) :

- la régression semble de très bonne qualité puisque que nous expliquons $R^2 = 95.45\%$ de la variance de l'endogène ;
- impression confirmée par le test de Fisher, $F = 136.54$ avec une p-value < 0.000001 : le modèle est globalement très significatif ;
- mis à part la variable cylindrée, toutes les variables sont significatives au risque de 10%.

Global results

Endogenous attribute	Consommation
Examples	31
R ²	0.954559
Adjusted-R ²	0.947568
Sigma error	0.817238
F-Test (4,26)	136.5413 (0.000000)

Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	364.7719	4	91.1930	136.5413	0.0000
Residual	17.3648	26	0.6679		
Total	382.1368	30			

Coefficients

Attribute	Coef.	std	t(26)	p-value
Intercept	2.456294	0.626818	3.918671	0.000578
Prix	0.000020	0.000009	2.338943	0.027297
Cylindrée	-0.000501	0.000575	-0.870866	0.391797
Puissance	0.024994	0.009992	2.501486	0.018993
Poids	0.004161	0.000879	4.734462	0.000068

Fig. 0.2. Résultat de la régression sur les données CONSO (cf. Données, figure 0.1)

La même régression sous EXCEL donne exactement les mêmes résultats (Figure 0.3)³. Seul le mode de présentation des résultats est un peu différent. Nous avons calculé dans la foulée la prédiction ponctuelle \hat{y}_i et les résidus $\hat{\varepsilon}_i = y_i - \hat{y}_i$ de la régression.

Remarque 2 (Interprétation des coefficients). D'ores et déjà, sans trop rentrer dans les détails, on note des bizarreries dans le rôle des variables. Que le prix et la consommation soient d'une certaine manière liés, on peut le comprendre. En revanche, imaginer que le prix influe directement sur la consommation paraît étrange. Cela voudrait dire qu'en diminuant artificiellement le prix d'un véhicule, on pourrait diminuer la consommation. Concernant la cylindrée, la taille du moteur, on s'étonne quand même qu'elle ne joue aucun rôle sur la consommation. Cela voudrait dire qu'on peut augmenter indéfiniment la taille du moteur sans que cela ne soit préjudiciable à la consommation de carburant... Nous reviendrons plus en détail sur la sélection des variables et l'interprétation des résultats plus loin.

³. Fonction DROITEREG(...)

i	Modele	Prix	Cylindrée	Puissance	Poids	Consommation	Prédiction	Résidu
1	Daihatsu Cuore	11600	846	32	650	5.7	5.7739	-0.0739
2	Suzuki Swift 1.0 GLS	12490	993	39	790	5.8	6.4759	-0.6759
3	Fiat Panda Mambo L	10450	899	29	730	6.1	5.9817	0.1183
4	VW Polo 1.4 60	17140	1390	44	955	6.5	7.1836	-0.6836
5	Opel Corsa 1.2i Eco	14825	1195	33	895	6.8	6.7094	0.0906
6	Subaru Vivio 4WD	13730	658	32	740	6.8	6.2859	0.5141
7	Toyota Corolla	19490	1331	55	1010	7.1	7.7649	-0.6649
8	Ferrari 456 GT	285000	5474	325	1690	21.3	20.6905	0.6095
9	Mercedes S 600	183900	5987	300	2250	18.7	20.0742	-1.3742
10	Maserati Ghibli GT	92500	2789	209	1485	14.5	14.3514	0.1486
11	Opel Astra 1.6i 16V	25000	1597	74	1080	7.4	8.5104	-1.1104
12	Peugeot 306 XS 108	22350	1761	74	1100	9.0	8.4574	0.5426
13	Renault Safrane 2.2 V	36600	2165	101	1500	11.7	10.8852	0.8148
14	Seat Ibiza 2.0 GTI	22500	1983	85	1075	9.5	8.5202	0.9798
15	VW Golf 2.0 GTI	31580	1984	85	1155	9.5	9.0380	0.4620
16	Citroen ZX Volcane	28750	1998	89	1140	8.8	9.0108	-0.2108
17	Fiat Tempra 1.6 Liberty	22600	1580	65	1080	9.3	8.2449	1.0551
18	Fort Escort 1.4i PT	20300	1390	54	1110	8.6	8.1430	0.4570
19	Honda Civic Joker 1.4	19900	1396	66	1140	7.7	8.5565	-0.8565
20	Volvo 850 2.5	39800	2435	106	1370	10.8	10.3995	0.4005
21	Ford Fiesta 1.2 Zetec	19740	1242	55	940	6.6	7.5233	-0.9233
22	Hyundai Sonata 3000	38990	2972	107	1400	11.7	10.2640	1.4360
23	Lancia K 3.0 LS	50800	2958	150	1550	11.9	12.2110	-0.3110
24	Mazda Hachtback V	36200	2497	122	1330	10.8	10.5284	0.2716
25	Mitsubishi Galant	31990	1998	66	1300	7.6	9.1678	-1.5678
26	Opel Omega 2.5i V6	47700	2496	125	1670	11.3	12.2534	-0.9534
27	Peugeot 806 2.0	36950	1998	89	1560	10.8	10.9257	-0.1257
28	Nissan Primera 2.0	26950	1997	92	1240	9.2	9.4656	-0.2656
29	Seat Alhambra 2.0	36400	1984	85	1635	11.6	11.1335	0.4665
30	Toyota Previa salon	50900	2438	97	1800	12.8	12.1888	0.6112
31	Volvo 960 Kombi aut	49300	2473	125	1570	12.7	11.8815	0.8185

	poids	puissance	cylindrée	prix	constante
coef.	0.004161	0.024994	-0.000501	0.000020	2.456294
e.t.	0.000879	0.009992	0.000575	0.000009	0.626818
R²	0.9546	0.8172	#N/A	#N/A	#N/A
	136.5413	26	#N/A	#N/A	#N/A
	364.7719	17.3648	#N/A	#N/A	#N/A

Fig. 0.3. Résultat de la régression sous EXCEL

Logiciels

Nous utiliserons principalement le tableur EXCEL. Mais à plusieurs reprises nous ferons appel à des logiciels gratuits tels que TANAGRA ⁴, REGRESS ⁵, LAZSTATS/OPENSTAT ⁶ et R ⁷; et à des logiciels commerciaux tels que SPSS ⁸ et STATISTICA ⁹. *Qu'importe le logiciel en réalité, le plus important est de savoir lire correctement les sorties des outils statistiques.*

4. TANAGRA : Un logiciel gratuit de Data Mining pour l'enseignement et la recherche - <http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html>

5. <http://tutoriels-data-mining.blogspot.com/2011/05/regress-dans-la-distribution-sipina.html>

6. <http://tutoriels-data-mining.blogspot.com/2011/05/regression-avec-le-logiciel-lazstats.html>

7. The R Project for Statistical Computing - <http://www.r-project.org/>

8. Pour une lecture détaillée des résultats fournis par SPSS, voir <http://www2.chass.ncsu.edu/garson/PA765/regress.htm>

9. Pour une lecture des résultats de STATISTICA, voir <http://www.statsoft.com/textbook/stmulreg.html>

Étude des résidus

L'inférence statistique relative à la régression (estimation par intervalle des coefficients, tests d'hypothèses, etc.) repose principalement sur les hypothèses liées au terme d'erreur ε qui résume les informations absentes du modèle. Il importe donc que l'on vérifie ces hypothèses afin de pouvoir interpréter les résultats¹.

Rappelons brièvement les hypothèses liées au terme d'erreur :

- sa distribution doit être symétrique, plus précisément elle suit une loi normale ;
- sa variance est constante ;
- les erreurs ε_i ($i = 1, \dots, n$) sont indépendantes.

Pour inspecter ces hypothèses, nous disposons des erreurs observées, les résidus, $\hat{\varepsilon}_i$ produites par la différence entre les valeurs observées de l'endogène y_i et les prédictions ponctuelles de la régression \hat{y}_i

$$\hat{\varepsilon}_i = y_i - \hat{y}_i \tag{1.1}$$

avec $\hat{y}_i = \hat{a}_0 + \hat{a}_1 x_{i,1} + \dots + \hat{a}_p x_{i,p}$

Remarque 3 (Moyenne des résidus). Dans un modèle avec constante, la moyenne des résidus $\bar{\hat{\varepsilon}} = \frac{1}{n} \sum_i \hat{\varepsilon}_i$ est mécaniquement égale à zéro. Ce résultat ne préjuge donc en rien de la pertinence de la régression. En revanche, si elle est différente de 0, cela indique à coup sûr des calculs erronés. Ce commentaire n'a pas lieu d'être pour une régression sans constante.

1.1 Diagnostic graphique

1.1.1 Graphiques des résidus

Aussi simpliste qu'il puisse paraître, le diagnostic graphique est pourtant un outil puissant pour valider une régression. Il fournit un nombre important d'informations que les indicateurs statistiques

1. Voir Dodge, pages 113 à 120.

appréhendent mal. Toute analyse de régression devrait être immédiatement suivie des graphiques des résidus observés... car il y en a plusieurs.

Avant d'énumérer les différents types de graphiques, donnons quelques principes généraux (Figure 1.1) :

- les résidus sont portés en ordonnée ;
- les points doivent être uniformément répartis *au hasard* dans un intervalle, que nous préciserons plus loin², sur l'ordonnée ;
- aucun point ne doit se démarquer ostensiblement des autres ;
- on ne doit pas voir apparaître une forme de régularité dans le nuage de points.

Le type du graphique dépend de l'information que nous portons en abscisse.

Résidus en fonction de l'endogène Y

Ce type de graphique permet de se rendre compte de la qualité de la régression. Les résidus $\hat{\varepsilon}_i$ doivent être répartis aléatoirement autour de la valeur 0, ils ne doivent pas avoir tendance à prendre des valeurs différentes selon les valeurs de Y . On cherche surtout à voir si la prédiction est d'égale qualité sur tout le domaine de valeurs de Y (Figure 1.1). Si pour une valeur ou une plage de valeur de Y , les résidus s'écartent visiblement, il faut s'inquiéter car cela indique que la valeur y_i a été mal reconstituée par le modèle.

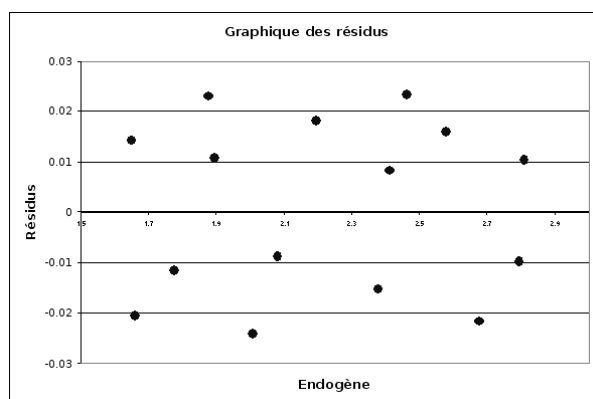


Fig. 1.1. Graphique "normal" des résidus. Endogène vs. Résidus.

Résidus en fonction de chaque exogène X_j

Il doit être produit pour chaque variable exogène. L'idée est de détecter s'il y a une relation quelconque entre le terme d'erreur et les exogènes. Rappelons que les variables exogènes et les erreurs sont indépendantes par hypothèse (covariance nulle), cela doit être confirmé visuellement.

² Voir chapitre 2 sur les points atypiques

Graphique de résidus pour les données longitudinales

Dans le cas particulier des séries temporelles, nous pouvons produire un graphique supplémentaire en portant en abscisse la variable temps. Elle permet d'ordonner les valeurs d'une autre manière. Il est alors possible de détecter une rupture de structure associée à une date particulière (ex. guerre, crise politique, choc économique, etc.).

Cas pathologiques

Il est difficile de prétendre à l'exhaustivité, nous nous contenterons de caractériser quelques situations singulières qui doivent attirer notre attention.

Points atypiques et points influents

Par définition, un *point atypique*, on parle aussi de point aberrant, est une observation qui s'écarte résolument des autres. Cela peut être dû à une erreur de recueil des données, cela peut aussi correspondre à un individu qui n'appartient pas à la population étudiée. Dans le graphique de résidus, il s'agit de points éloignés des autres, que la variable en abscisse soit l'endogène ou une des exogènes (Figure 1.2).

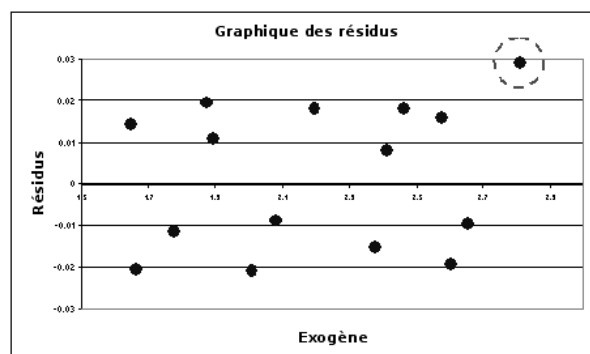


Fig. 1.2. Un point présente une valeur atypique pour une des exogènes. De plus, elle est mal reconstituée par la régression (le résidu est élevé).

Les *points influents* sont des observations qui pèsent exagérément sur les résultats de la régression. On peut les distinguer de plusieurs manières : ils sont "isolés" des autres points, on constate alors que la distribution des résidus est asymétrique (Figure 1.3); ils correspondent à des valeurs extrêmes des variables, en cela ils se rapprochent des points atypiques.

Bien souvent la distinction entre les points atypiques et les points influents est difficile. Elle est assez mal comprise : un point peut être influent sans être atypique, il peut être atypique sans être influent. La meilleure manière de le circonscrire est de recalculer les coefficients de la régression en écartant le point : si les résultats diffèrent significativement, en termes de prédiction ou terme de différence entre les

coefficients estimés, le point est influent. Cela est difficilement discernable dans un graphique des résidus, il est plus approprié de passer par des calculs que nous détaillerons dans le chapitre consacré à la détection des points atypiques et influents (Chapitre 2).

Asymétrie des résidus

Signe que la distribution des résidus ne suit pas la loi normale, cette situation (Figure 1.3) survient

- lorsque certains points se démarquent des autres, ils sont mal reconstitués par la régression. La moyenne des résidus est mécaniquement égale à 0, mais la dispersion est très inégale de part et d'autre de cette valeur.
- lorsque les données sont en réalité formées par plusieurs populations (ex. en médecine, effectuer une régression en mélangeant les hommes et les femmes, sachant qu'ils réagissent de manière différente à la maladie étudiée).
- lorsqu'on est face à un problème de spécification, une variable exogène importante manque.
- etc.

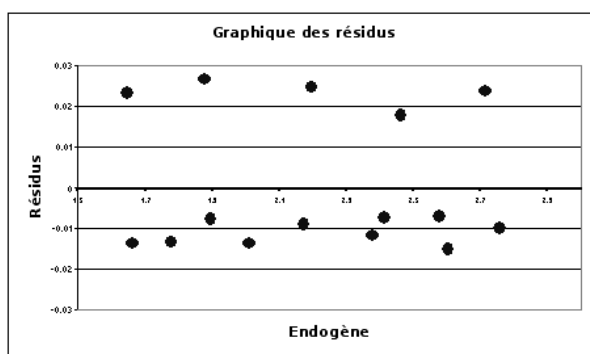


Fig. 1.3. La distribution des résidus est asymétrique.

Non-linéarité

Dans ce cas, la relation étudiée est en réalité non-linéaire, elle ne peut pas être modélisée à l'aide de la régression linéaire multiple. Les résidus apparaissent alors en "blocs" au-dessus (prédiction sous-estimée) ou en-dessous (prédiction sur-estimée) de la valeur 0 (Figure 1.4). On peut y remédier en ajoutant une variable transformée dans le modèle (par ex. en passant une des variables au carré, ou en utilisant une transformation logarithmique, etc.). On peut aussi passer à une régression non-linéaire (ex. réseaux de neurones, etc.).

Rupture de structure

Dans certains cas, il arrive que la relation entre les exogènes et l'endogène ne soit pas la même sur tout le domaine de définition : on parle de rupture de structure. Il y a en réalité deux ou plusieurs

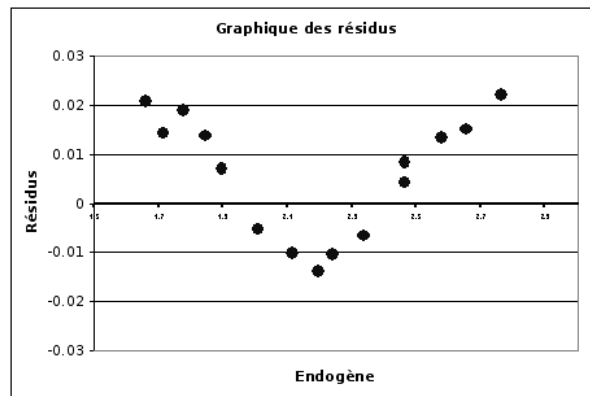


Fig. 1.4. La relation à modéliser est non-linéaire

régressions à mener. Ils peuvent être totalement indépendants. On peut aussi imposer que les coefficients de quelques variables soient identiques d'une régression à l'autre. L'erreur dans ce cas est d'imposer une seule régression pour tous les groupes d'individus. Nous obtenons alors des résidus en "blocs", qui peuvent être assez proches de ce que l'on obtient lorsque les relations sont non-linéaires (Figure 1.4), ils indiquent en tous les cas qu'il y a bien des groupes distincts que l'on ne peut pas modéliser de manière identique dans la population (Figure 1.5).

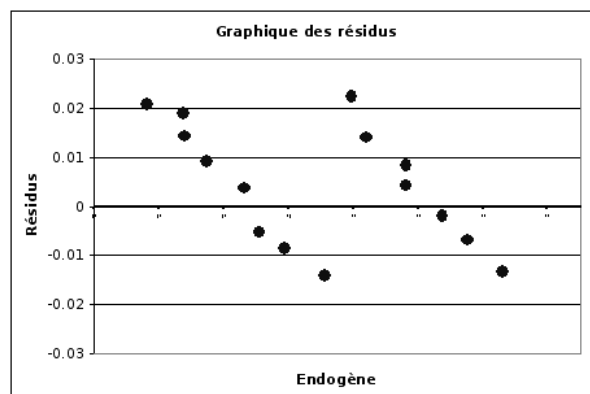


Fig. 1.5. Résidus caractéristiques d'une rupture de structure

Hétéroscédasticité

Souvent associée à une des exogènes en abscisse, ce type de graphique (Figure 1.6) indique que la variance des résidus n'est pas constante, et qu'elle dépend d'une des exogènes. Il existe des tests spécifiques pour détecter l'hétéroscédasticité (Bourbonnais, pages 130 à 143).

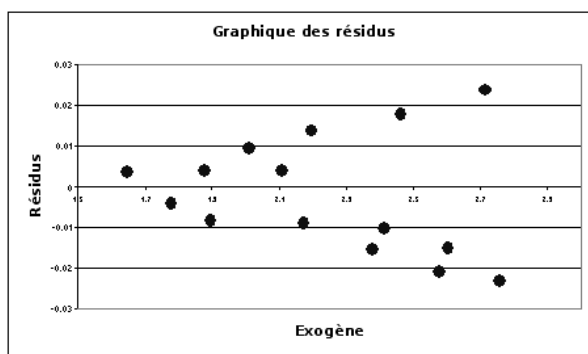


Fig. 1.6. La variance des résidus augmente avec les valeurs d'une des exogènes

Autocorrélation des résidus

Ce problème est spécifique aux données longitudinales. Dans le graphique des résidus, nous plaçons des dates en abscisse, nous essayons de détecter si les erreurs suivent un processus particulier au cours du temps. L'autocorrélation peut être positive (des "blocs" de résidus sont positifs ou négatifs, figure 1.8) ou négative (les résidus sont alternativement positifs et négatifs, figure 1.7).

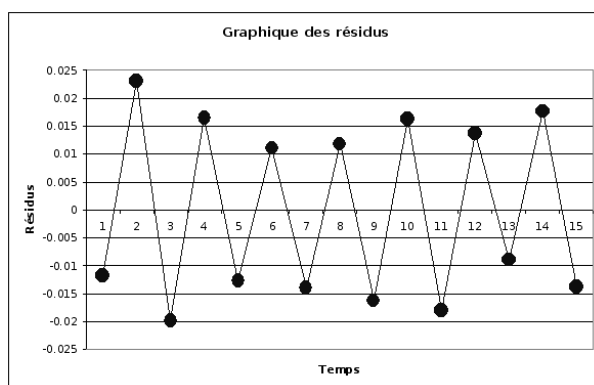


Fig. 1.7. Autocorrélation négative des résidus

1.1.2 Graphiques des résidus pour les données CONSO

Nous avons lancé la régression sur les données CONSO (Figures 0.2 et 0.3). Nous construisons les différents graphiques des résidus en les croisant avec l'endogène et les exogènes (Figure 1.9). Nous avons utilisé le logiciel R.

Une information, essentiellement, saute aux yeux : 2 points semblent se démarquer systématiquement sur l'endogène Y , le prix, la cylindrée et la puissance. Pourtant ils ne semblent pas particulièrement mal restitués par la régression puisque le résidu (erreur de prédiction) ne prend pas des valeurs anormalement

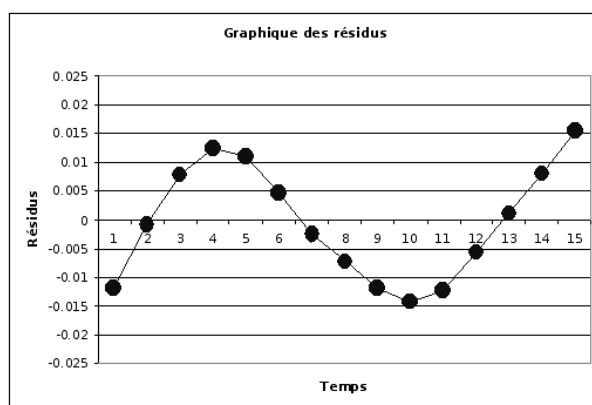


Fig. 1.8. Autocorrélation positive des résidus

élevées (en valeur absolue) sur ces observations. Nous détaillerons l'analyse de ces véhicules dans le chapitre consacré à l'analyse des points atypiques et influents.

1.2 Tester le caractère aléatoire des erreurs

Lorsque nous travaillons avec des données longitudinales, la date définit naturellement l'ordonnancement des observations. Il est important de vérifier que les résidus sont produits de manière totalement aléatoire. Si l'on conclut au rejet de cette hypothèse, les résidus sont produits par un processus quelconque, l'hypothèse d'indépendance des erreurs est rejetée, la méthode des moindres carrés ordinaires n'est plus BLUE³ : elle est certes non-biaisée, mais elle n'est plus à variance minimale, et la matrice de variance covariance n'est plus estimée de manière convergente, les tests de significativité ne sont plus opérants.

La détection de l'autocorrélation des résidus peut s'effectuer visuellement à l'aide du graphique des résidus (Figures 1.8 et 1.7). Elle peut également s'appuyer sur des techniques statistiques. La plus connue est certainement le test de Durbin-Watson qui détecte une forme particulière de l'autocorrélation. Nous pouvons aussi utiliser des tests plus généraux comme le test des séquences de Wald.

Les causes de l'autocorrélation des résidus peuvent être multiples. Elles se rapprochent des problèmes de spécifications à l'origine des violations des hypothèses (Bourbonnais, page 114) : une variable exogène importante est absente de l'équation de régression ; la liaison modélisée n'est pas linéaire ; les données ont été manipulées (ex. moyenne mobile, reconstituée par interpolation, etc.), c'est souvent le cas lorsqu'elles sont produites par des observatoires statistiques.

Remarque 4 (Test l'autocorrélation pour les données transversales). Tester l'autocorrélation des résidus n'a aucun sens sur les données transversales. En effet, il n'y a pas d'ordonnancement naturel des observations. Il sera toujours possible de les mélanger différemment de manière à ce que les résidus ne suivent

3. Best Linear Unbiased Estimator

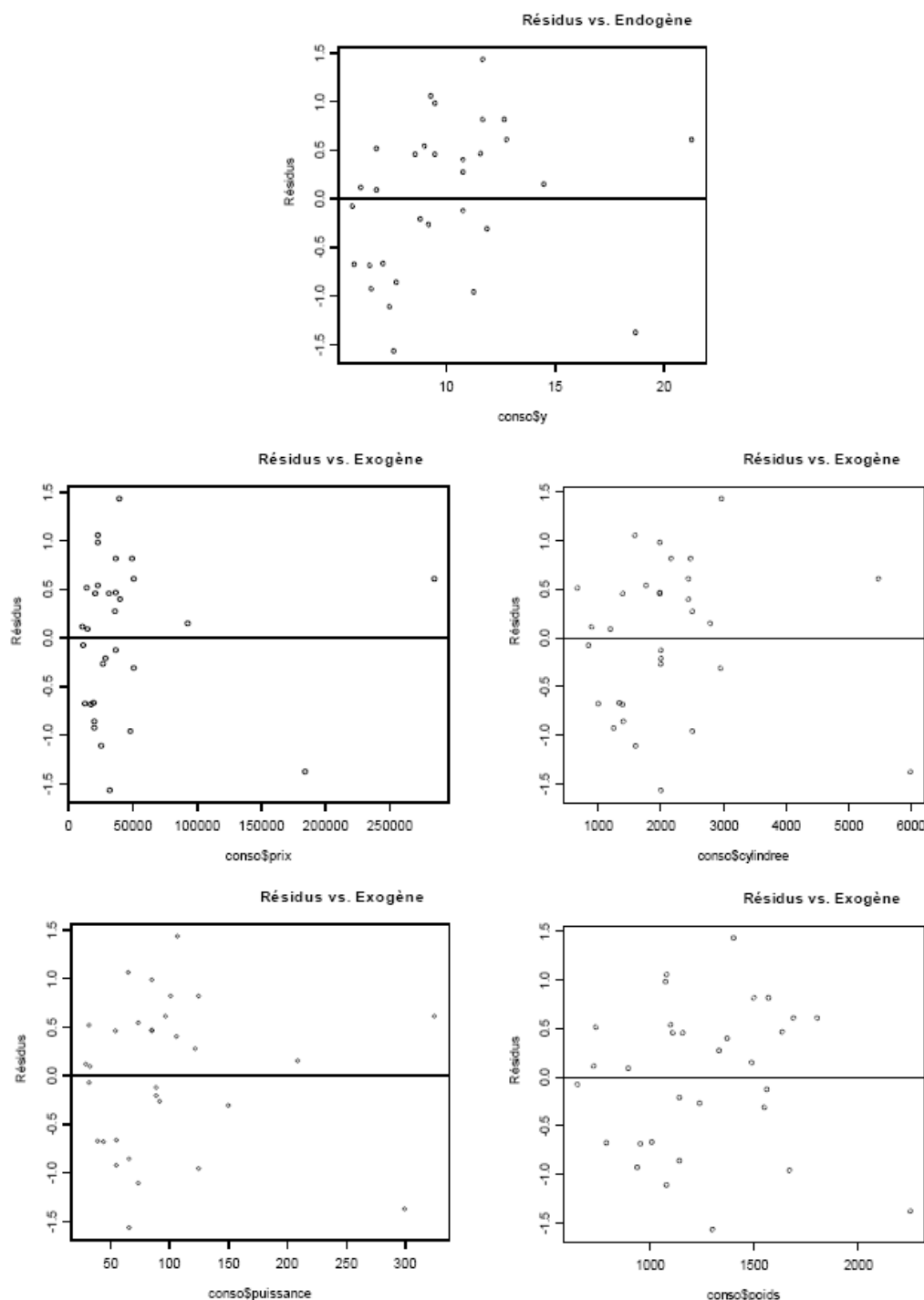


Fig. 1.9. Graphiques des résidus - Données CONSO

aucun processus particulier. Il est néanmoins possible de retrouver un agencement particulier des résidus en les triant selon l'endogène par exemple. Mais il faut rester très prudent par rapport aux tests, le plus sage est de s'appuyer sur les techniques graphiques simples pour détecter d'éventuelles anomalies (ex.

les valeurs négatives des résidus sont regroupés sur les petites valeurs de Y , les valeurs positives sur les grandes valeurs de Y : manifestement il y a un problème dans le modèle...).

1.2.1 Test de Durbin-Watson

Principe

Le test de Durbin-Watson permet de détecter une autocorrélation de la forme :

$$\varepsilon_i = \rho \cdot \varepsilon_{i-1} + \nu_i, \text{ avec } \nu_i \sim \mathcal{N}(0, \sigma_\nu) \quad (1.2)$$

Le test d'hypothèses s'écrit :

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

On utilise la statistique de Durbin-Watson

$$d = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2} \quad (1.3)$$

Par construction, $0 \leq d \leq 4$, $d = 2$ lorsque $\hat{\rho} = 0$. Elle a été tabulée par Durbin et Watson (Annexes A) pour différentes tailles d'échantillon n et de nombre de vraies variables explicatives k (sans compter la constante). La règle de décision n'est pas usuelle, nous pouvons la résumer de la manière suivante pour un test bilatéral (Bourbonnais, pages 115 et 116) :

- Acceptation de H_0 si $d_U < d < 4 - d_U$
- Rejet de H_0 si $d < d_L$ ($\rho > 0$) ou $d > 4 - d_L$ ($\rho < 0$)
- Incertitude si $d_L < d < d_U$ ou $4 - d_U < d < 4 - d_L$

Le test de Durbin-Watson est assez limité. Il ne teste que les autocorrélation des résidus d'ordre 1. De plus, son utilisation est encadrée par des conditions draconiennes (Johnston, page 189) :

- la régression doit comporter un terme constant ;
- les variables X sont certaines (non-stochastiques), en particulier elles ne doivent pas comporter l'endogène retardée⁴.

Remarque 5 (Autres formes d'autocorrélation des résidus). D'autres tests ont été mis au point pour évaluer d'autres formes de relation entre les résidus (ex. processus auto-régressif d'ordre 4 pour les données trimestrielles, etc. – Johnston, pages 180 à 200).

4. On doit utiliser une version modifiée du test de Durbin (Johnston, page 190)

Exemple : Prédiction de la consommation de textile

Pour illustrer la mise en oeuvre du test de Durbin-Watson, nous reprenons un exemple extrait de l'ouvrage de Theil (1971)⁵. L'objectif est de prédire la consommation de textile à partir du revenu par tête des personnes et du prix. Nous disposons d'observations sur 17 années à partir de 1923 (Figure 1.10).

Annee	Conso	Revenu	Prix
1923	99.2	96.7	101
1924	99	98.1	100.1
1925	100	100	100
1926	111.6	104.9	90.6
1927	122.2	104.9	86.5
1928	117.6	109.5	89.7
1929	121.1	110.8	90.6
1930	136	112.3	82.8
1931	154.2	109.3	70.1
1932	153.6	105.3	65.4
1933	158.5	101.7	61.3
1934	140.6	95.4	62.5
1935	136.2	96.4	63.6
1936	168	97.6	52.6
1937	154.3	102.4	59.7
1938	149	101.6	59.5
1939	165.5	103.8	61.3

Fig. 1.10. Données de Theil sur le textile

Annee	Conso	Revenu	Prix	pred(conso)	e	dénominateur	numérateur
1923	99.2	96.7	101	93.692	5.508	30.334	0.000
1924	99	98.1	100.1	96.423	2.577	6.639	8.591
1925	100	100	100	98.579	1.421	2.019	1.335
1926	111.6	104.9	90.6	116.781	-5.181	26.847	43.592
1927	122.2	104.9	86.5	122.452	-0.252	0.063	24.303
1928	117.6	109.5	89.7	122.910	-5.310	28.196	25.587
1929	121.1	110.8	90.6	123.046	-1.946	3.785	11.320
1930	136	112.3	82.8	135.425	0.575	0.330	6.351
1931	154.2	109.3	70.1	149.804	4.396	19.323	14.602
1932	153.6	105.3	65.4	152.057	1.543	2.380	8.141
1933	158.5	101.7	61.3	153.905	4.595	21.110	9.314
1934	140.6	95.4	62.5	145.557	-4.957	24.573	91.234
1935	136.2	96.4	63.6	145.098	-8.898	79.166	15.527
1936	168	97.6	52.6	161.584	6.416	41.160	234.491
1937	154.3	102.4	59.7	156.861	-2.561	6.561	80.587
1938	149	101.6	59.5	156.289	-7.289	53.124	22.347
1939	165.5	103.8	61.3	156.135	9.365	87.703	277.343
Somme						433.31	874.66

d	2.02
dL	1.02
dU	1.54
4-dL	2.98
4-dU	2.46

	prix	revenu	const
coef	-1.38	1.06	130.71
e.t.	0.08	0.27	27.09
R ²	0.95	5.56	#N/A
	136.68	14	#N/A
	8460.94	433.31	#N/A

Fig. 1.11. Test de Durbin-Watson sur les données de Theil

L'équation de régression à mettre en place est

$$y_i = a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \varepsilon_i, \quad i = 1, \dots, 17$$

où y est la consommation en textile, x_1 le prix du textile et x_2 le revenu par habitant.

5. Theil, H., *Principles of Econometrics*, Wiley, 1971. Page 102. L'exemple et la description des résultats du test sont accessibles sur le site <http://shazam.econ.ubc.ca/intro/dwdist.htm>

Les calculs sont organisés comme suit (Figure 1.11) :

1. A l'aide de la fonction DROITEREG() d'EXCEL, nous obtenons les coefficients $a_0 = 130.71$, $a_1 = -1.38$ et $a_2 = 1.06$.
2. Nous formons la prédiction \hat{y}_i avec ces coefficients.
3. Nous calculons l'erreur de prédiction, le résidu de la régression $\hat{\varepsilon}_i = e_i = y_i - \hat{y}_i$.
4. Nous pouvons alors calculer la statistique de Durbin-Watson. En formant le numérateur 874.66 et le dénominateur 433.31, nous obtenons $d = 2.02$.
5. Pour un test bilatéral à 10%, nous récupérons les valeurs critiques dans la table de Durbin-Watson (Annexes A). Pour $n = 17$ et $k = 2$, $d_L = 1.02$ et $d_U = 1.54$.
6. Nous constatons que nous sommes dans la région $d_U < d < 4 - d_U$, l'hypothèse d'absence d'autocorrélation d'ordre 1 des résidus n'est pas contredite par les données ($\rho = 0$).

1.2.2 Test des séquences

Le test des séquences⁶, appelé également *test de Wald-Wolfowitz*, est plus générique que le précédent. Il cherche à détecter toute forme de régularité lorsque les résidus sont ordonnés selon le temps. Il détecte autant les autocorrélations négatives (les résidus sont alternativement négatives et positives) que les autocorrélations positives (des blocs de résidus consécutifs sont positifs ou négatifs). Étant plus générique, il est bien entendu moins puissant pour des formes particulières d'autocorrélation. On lui préférera le test de Durbin-Watson par exemple si on veut vérifier expressément la présence d'un processus auto-régressif d'ordre 1 des résidus.

Principe

Bien entendu, les données doivent être ordonnées pour que le test puisse opérer. Notre référence est la date pour les données longitudinales.

Le test repose sur la détection des séquences de valeurs positives '+' ou négatives '-' des résidus. La statistique du test r est le nombre total de séquences dans la série d'observations.

Exemple 1. Si tous les résidus négatifs sont regroupés sur les petites valeurs de Y , et inversement, les résidus positifs, sur les grandes valeurs de Y , nous aurons simple $r = 2$ séquences. C'est éminemment suspect si l'on se réfère à l'hypothèse H_0 selon laquelle les résidus sont générés aléatoirement.

Posons n_+ (resp. n_-) le nombre de résidus positifs (resp. négatifs) dans la série des résidus. Sous l'hypothèse H_0 le processus de génération des données est aléatoire, la statistique r suit asymptotiquement⁷ une loi normale de paramètres :

6. Voir Siegel, S., Castellan, J., *Nonparametric statistics for the behavioral sciences*, McGraw-Hill, 1988, pages 58 à 64, section "The one-Sample runs test of randomness"

7. Pour les petites valeurs de n_+ et n_- , les valeurs critique de r ont été tabulées. Voir par exemple Siegel-Castellan, Table G, page 331. Curieusement, je n'ai pas pu en trouver en ligne...

$$\mu_r = \frac{2n_+n_-}{n} + 1 \quad (1.4)$$

$$\sigma_r = \sqrt{\frac{(\mu_r - 1)(\mu_r - 2)}{n - 1}} \quad (1.5)$$

Nous pouvons former la statistique centrée et réduite $z = \frac{r - \mu_r}{\sigma_r}$. La région critique du test – rejet de l’hypothèse de génération aléatoire des résidus – s’écrit :

$$R.C. : |z| > u_{1-\frac{\alpha}{2}}$$

où $u_{1-\frac{\alpha}{2}}$ est le fractile d’ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée et réduite $\mathcal{N}(0, 1)$.

Remarque 6 (Le test de séquences est un test bilatéral). Attention, le test des séquences est bien un test bilatéral. Des ‘+’ et ‘-’ alternés (r élevé) sont tout aussi suspects que des blocs de ‘+’ et ‘-’ (r faible). Ce test permet autant de détecter les autocorrélations négatives que positives.

Prédiction de la consommation de textile

Annee	Conso	Revenu	Prix	pred(conso)	e	Sup/Inf	Séquences
1923	99.2	96.7	101	93.692	5.508	+	1
1924	99	98.1	100.1	96.423	2.577	+	
1925	100	100	100	98.579	1.421	+	
1926	111.6	104.9	90.6	116.781	-5.181	-	
1927	122.2	104.9	86.5	122.452	-0.252	-	2
1928	117.6	109.5	89.7	122.910	-5.310	-	
1929	121.1	110.8	90.6	123.046	-1.946	-	
1930	136	112.3	82.8	135.425	0.575	+	
1931	154.2	109.3	70.1	149.804	4.396	+	3
1932	153.6	105.3	65.4	152.057	1.543	+	
1933	158.5	101.7	61.3	153.905	4.595	+	
1934	140.6	95.4	62.5	145.557	-4.957	-	
1935	136.2	96.4	63.6	145.098	-8.898	-	4
1936	168	97.6	52.6	161.584	6.416	+	
1937	154.3	102.4	59.7	156.861	-2.561	-	6
1938	149	101.6	59.5	156.289	-7.289	-	
1939	165.5	103.8	61.3	156.135	9.365	+	7
						r	7

	prix	revenu	const
coef	-1.38	1.06	130.71
e.t.	0.08	0.27	27.09
R ²	0.95	5.56	#N/A
	136.68	14	#N/A
	8460.94	433.31	#N/A

n+	9
n-	8
n	17

Mu	9.47
Sigma	1.99

z	-1.24
---	-------

u(1-alpha/2)	1.64
--------------	------

Fig. 1.12. Test de Wald-Wolfowitz sur les données de Theil

Reprenons l’exemple de la consommation de textile (Theil, 1971), nous reproduisons les calculs à l’aide d’un tableur (Figure 1.12) :

1. A l’aide de la fonction DROITEREG() d’EXCEL, nous obtenons les coefficients $a_0 = 130.71$, $a_1 = -1.38$ et $a_2 = 1.06$.

2. Nous formons la prédiction \hat{y}_i avec ces coefficients.
3. Nous calculons l'erreur de prédiction, le résidu de la régression $\hat{\varepsilon}_i = e_i = y_i - \hat{y}_i$.
4. Nous annotons avec le caractère '+' (resp. '-') les résidus positifs (resp. négatifs).
5. Nous comptons le nombre de valeurs positives et négatives, $n_+ = 9$ et $n_- = 8$, nous vérifions que $n = n_+ + n_- = 17$.
6. Nous pouvons calculer la moyenne et l'écart-type de la statistique de test sous l'hypothèse nulle : $\mu_r = 9.47$ et $\sigma_r = 1.99$.
7. Nous affectons un numéro à chaque séquence de '+' et '-', nous obtenons ainsi le nombre de séquences $r = 7$.
8. Nous calculons enfin la statistique centrée et réduite $z = \frac{7-9.47}{1.99} = -1.24$;
9. Que nous comparons au fractile d'ordre 0.95 (pour un test bilatéral à 10%) de la loi normal centrée et réduite $u_{0.95} = 1.64$.

Nous sommes dans la région d'acceptation de H_0 . Nous pouvons conclure que les résidus sont indépendants, ils sont générés par un processus purement aléatoire.

1.3 Test de normalité

Une grande partie de l'inférence statistique (ex. test de pertinence globale de la régression, prédiction par intervalle, etc.) repose sur l'hypothèse de distribution normale $\mathcal{N}(0, \sigma_\varepsilon)$ du terme d'erreur de l'équation de régression (Équation 0.1). Vérifier cette hypothèse semble incontournable pour obtenir des résultats exacts⁸.

Nous disposons des erreurs observés $\hat{\varepsilon}_i$, les résidus de la régression, pour évaluer les caractéristiques des erreurs théoriques ε_i . Cela n'est pas sans poser des problèmes. En effet, si la variance de l'erreur est constante $V(\varepsilon_i) = \sigma_\varepsilon^2$, la variance du résidu, l'erreur observée, ne l'est pas $V(\hat{\varepsilon}_i) = \sigma_\varepsilon^2(1 - h_{ii})$, où h_{ii} est lue sur la diagonale principale de la *hat matrix* $H = X(X'X)^{-1}X'$. Et surtout, la covariance $cov(\hat{\varepsilon}_i, \hat{\varepsilon}_j) = -\sigma_\varepsilon^2 h_{ij}$ entre deux résidus observés n'est pas nulle en général.

De fait, la loi des statistiques sous H_0 (normalité des erreurs) que l'on pourrait utiliser dans cette section sont modifiés, induisant également une modification des valeurs critiques pour un même risque α . Comment ? Il n'y a pas vraiment de réponses établies. Il semble néanmoins que les tests usuels restent valables, pour peu que l'on ait *suffisamment d'observations* ($n \geq 50$)⁹. Il faut surtout voir les tests comme

8. Pour un tour d'horizon des conséquences des violations des hypothèses dans la régression, nous conseillons l'excellent document de J.Ravet disponible en ligne <http://homepages.ulb.ac.be/~j.ravet/stateco/docs/econometrie.pdf>

9. Cette valeur est vraiment donné comme un ordre d'idées. En réalité, le problème de l'utilisation des résidus pour évaluer la normalité des erreurs est souvent passé sous silence dans la littérature. Le seul ouvrage où cela est posé clairement est celui de Capéraà P., Van Cutsem B., *Méthodes et modèles en statistique non paramétrique - Exposé fondamental*, Dunod, Presse de l'Université de Laval, 1988 ; pages 306 et 307

des indicateurs supplémentaires pour évaluer la régression, il faut réellement s'inquiéter si la distribution empirique des résidus s'écarte *très fortement* de l'hypothèse de normalité c.-à-d. avec des p-value très faibles lorsque les tests sont mis en oeuvre. C'est en ce sens que nous les présentons¹⁰.

1.3.1 Graphique Q-Q plot

Principe

Il ne s'agit pas d'un test au sens statistique du terme. Le graphique *Q-Q plot* (quantile-quantile plot) est un graphique "nuage de points" qui vise à confronter les quantiles de la distribution empirique et les quantiles d'une distribution théorique normale, de moyenne et d'écart type estimés sur les valeurs observées. Si la distribution est compatible avec la loi normale, les points forment une droite. Dans la littérature francophone, ce dispositif est appelé *Droite de Henry*.

Remarque 7. Pour plus de détails, nous conseillons la lecture du document en ligne http://eric.univ-lyon2.fr/~ricco/cours/cours/Test_Normalite.pdf, section 1.5.

Application sur les données CONSO

A partir du descriptif de notre document de référence, nous avons construit la Droite de Henry dans le tableur EXCEL (Figure 1.13). Le détail des calculs est le suivant :

1. Trier les résidus $\hat{\varepsilon}_i$ de manière croissante, ce sont les quantiles observés.
2. Produire la fonction de répartition empirique, lissée en accord avec la loi normale $F_i = \frac{i-0.375}{n+0.25}$
3. Calculer les quantiles théoriques normalisées z_i en utilisant la fonction inverse de la loi normale centrée réduite.
4. En déduire les quantiles théoriques dé-normalisées $\varepsilon_i^* = \tilde{\sigma}_\varepsilon \times z_i$. Si la distribution empirique cadre parfaitement avec la loi normale, les points devraient être alignés sur la diagonale principale. Ici, pour simplifier¹¹, nous prenons $\tilde{\sigma}_\varepsilon = \sqrt{\frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2}$.

Nous constatons que les points sont relativement bien alignés. Il n'y a pas d'incompatibilité manifeste avec une distribution normale.

10. Pour une présentation détaillée des tests d'adéquation à la loi normale d'une distribution empirique, nous conseillons un de nos supports accessibles en ligne http://eric.univ-lyon2.fr/~ricco/cours/cours/Test_Normalite.pdf. Des liens vers d'autres documents et des fichiers exemples sont disponibles sur notre site de supports de cours http://eric.univ-lyon2.fr/~ricco/cours/supports_data_mining.html, section Statistique

11. En toute rigueur, nous devrions utiliser l'estimateur sans biais de l'écart-type de l'erreur ($\hat{\sigma}_\varepsilon$). Mais cette petite entorse ne tire pas à conséquence dans notre procédure. Il s'agit simplement d'un changement d'échelle. Si les résidus sont compatibles avec la distribution normale, les points formeront une droite que l'on utilise l'un ou l'autre des estimateurs.

i	e	F	z	e*
1	-1.5678	0.0200	-2.0537	-1.5371
2	-1.3742	0.0520	-1.6258	-1.2168
3	-1.1104	0.0840	-1.3787	-1.0318
4	-0.9534	0.1160	-1.1952	-0.8945
5	-0.9233	0.1480	-1.0451	-0.7822
6	-0.8565	0.1800	-0.9154	-0.6851
7	-0.6836	0.2120	-0.7995	-0.5984
8	-0.6759	0.2440	-0.6935	-0.5190
9	-0.6649	0.2760	-0.5948	-0.4451
10	-0.3110	0.3080	-0.5015	-0.3754
11	-0.2656	0.3400	-0.4125	-0.3087
12	-0.2108	0.3720	-0.3266	-0.2444
13	-0.1257	0.4040	-0.2430	-0.1819
14	-0.0739	0.4360	-0.1611	-0.1206
15	0.0906	0.4680	-0.0803	-0.0601
16	0.1183	0.5000	0.0000	0.0000
17	0.1486	0.5320	0.0803	0.0601
18	0.2716	0.5640	0.1611	0.1206
19	0.4005	0.5960	0.2430	0.1819
20	0.4570	0.6280	0.3266	0.2444
21	0.4620	0.6600	0.4125	0.3087
22	0.4665	0.6920	0.5015	0.3754
23	0.5141	0.7240	0.5948	0.4451
24	0.5426	0.7560	0.6935	0.5190
25	0.6095	0.7880	0.7995	0.5984
26	0.6112	0.8200	0.9154	0.6851
27	0.8148	0.8520	1.0451	0.7822
28	0.8185	0.8840	1.1952	0.8945
29	0.9798	0.9160	1.3787	1.0318
30	1.0551	0.9480	1.6258	1.2168
31	1.4360	0.9800	2.0537	1.5371

Ecart-type	0.748436
Moyenne	0.000000

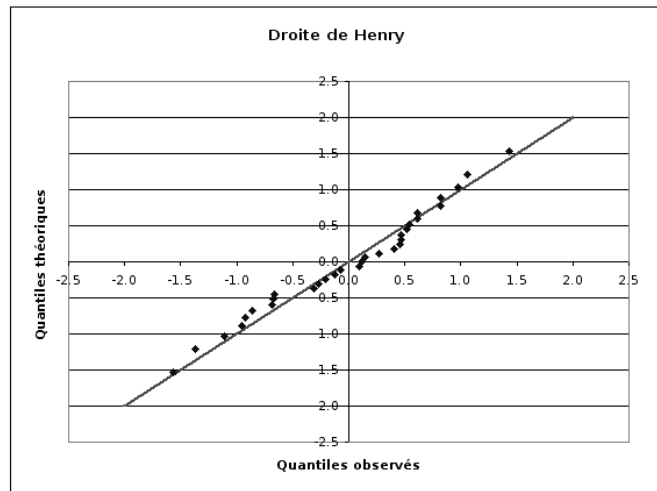


Fig. 1.13. Droite de Henry sur les résidus des MCO – Données CONSO

Bien souvent, on peut se contenter de ce diagnostic. Nous réagissons uniquement si l'écart avec la normalité est très marquée. Néanmoins, pour les puristes, nous pouvons consolider les conclusions en s'appuyant sur la batterie des tests de normalité. Nous nous contenterons de tests asymptotiques simples.

1.3.2 Test de symétrie de la distribution des résidus

Principe du test

Ce test est basé sur le coefficient d'asymétrie

$$\gamma_1 = \frac{\mu_3}{\sigma^3} \quad (1.6)$$

où μ_3 est le moment centré d'ordre 3, et σ l'écart-type.

On sait que γ_1 est égal à 0 si la distribution est normale. Le test d'hypothèses s'écrit de la manière suivante :

$$H_0 : \varepsilon \text{ suit une loi normale, par conséquent } \gamma_1 = 0$$

$$H_1 : \varepsilon \text{ ne suit pas une loi normale, par conséquent } \gamma_1 \neq 0$$

Remarque 8. Attention, les hypothèses ne sont pas symétriques. Si on établit que $\gamma_1 \neq 0$, nous savons que la distribution n'est pas gaussienne. En revanche, conclure $\gamma_1 = 0$ indique que la distribution est seulement *compatible* avec une loi normale.

Statistique du test et région critique

Pour réaliser le test, nous devons définir la statistique du test et sa loi de distribution sous H_0 . Nous utilisons le coefficient d'asymétrie empirique :

$$g_1 = \frac{\frac{1}{n} \sum_i \hat{\varepsilon}_i^3}{\left(\frac{1}{n} \sum_i \hat{\varepsilon}_i^2\right)^{\frac{3}{2}}} \quad (1.7)$$

Sous H_0 , elle suit asymptotiquement une loi normale d'espérance et d'écart-type¹²

$$\begin{aligned} \mu_1 &\approx 0 \\ \sigma_1 &\approx \sqrt{\frac{6}{n}} \end{aligned}$$

Nous formons le rapport $c_1 = \frac{g_1}{\sigma_1}$. Pour un test bilatéral au risque α , la région critique est définie par

$$R.C. : |c_1| \geq u_{1-\frac{\alpha}{2}}$$

où $u_{1-\frac{\alpha}{2}}$ est le fractile d'ordre $1 - \frac{\alpha}{2}$ de la loi normale centrée réduite.

Application sur les données CONSO

Nous construisons le test ci-dessus sur les résidus des MCO sur nos données CONSO. Voici les principales étapes (Figure 1.14) :

1. Nous récupérons la colonne des résidus $\hat{\varepsilon}_i$.
2. Nous calculons les colonnes de $\hat{\varepsilon}_i^2$ et $\hat{\varepsilon}_i^3$.
3. Nous calculons les sommes et formons $g_1 = \frac{-0.1220}{0.5602^{3/2}} = -0.2909$.
4. Nous calculons l'écart-type $\sigma_1 = \sqrt{\frac{6}{31}} = 0.4399$, et le rapport $|c_1| = 0.6612$.
5. Nous observons que $|c_1| < 1.6449 = u_{0.95}$, pour un test bilatéral à 10%. Nous ne sommes pas dans la région critique.

Si l'on se réfère au résultats du test, l'hypothèse de compatibilité avec la normale ne peut pas être rejetée.

1.3.3 Test de Jarque-Bera

Principe

Ce test complète le précédent en intégrant le coefficient d'aplatissement $\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$ dans la procédure. Les hypothèses deviennent :

12. Une formulation plus précise de l'écart-type est disponible dans http://eric.univ-lyon2.fr/~ricco/cours/cours/Test_Normalite.pdf

i	Résidu	e^2	e^3	e^4
1	-0.0739	0.0055	-0.0004	0.0000
2	-0.6759	0.4568	-0.3088	0.2087
3	0.1183	0.0140	0.0017	0.0002
4	-0.6836	0.4673	-0.3194	0.2184
5	0.0906	0.0082	0.0007	0.0001
6	0.5141	0.2643	0.1359	0.0698
7	-0.6649	0.4421	-0.2939	0.1954
8	0.6095	0.3715	0.2264	0.1380
9	-1.3742	1.8885	-2.5953	3.5665
10	0.1486	0.0221	0.0033	0.0005
11	-1.1104	1.2329	-1.3690	1.5202
12	0.5426	0.2944	0.1598	0.0867
13	0.8148	0.6639	0.5409	0.4407
14	0.9798	0.9599	0.9405	0.9215
15	0.4620	0.2134	0.0986	0.0456
16	-0.2108	0.0444	-0.0094	0.0020
17	1.0551	1.1132	1.1745	1.2391
18	0.4570	0.2089	0.0955	0.0436
19	-0.8565	0.7337	-0.6284	0.5382
20	0.4005	0.1604	0.0642	0.0257
21	-0.9233	0.8525	-0.7871	0.7267
22	1.4360	2.0622	2.9615	4.2528
23	-0.3110	0.0967	-0.0301	0.0094
24	0.2716	0.0737	0.0200	0.0054
25	-1.5678	2.4579	-3.8533	6.0410
26	-0.9534	0.9089	-0.8665	0.8261
27	-0.1257	0.0158	-0.0020	0.0002
28	-0.2656	0.0705	-0.0187	0.0050
29	0.4665	0.2176	0.1015	0.0474
30	0.6112	0.3736	0.2284	0.1396
31	0.8185	0.6700	0.5484	0.4489
Somme		17.3648	-3.7806	21.7634
1/n*somme		0.5602	-0.1220	0.7020

g1	-0.2909
sigma1	0.4399
abs(g1/sigma1)	0.6612
u(1-alpha/2)	1.6449

Fig. 1.14. Test de normalité des résidus fondé sur le coefficient de symétrie sur les données CONSO

H_0 : ε suit une loi normale, par conséquent $\gamma_1 = 0$ et $\gamma_2 = 0$

H_1 : ε ne suit pas une loi normale, par conséquent $\gamma_1 \neq 0$ ou $\gamma_2 \neq 0$

où μ_4 est le moment centré d'ordre 4, σ est l'écart-type.

Remarque 9 (Rejet de l'hypothèse de normalité). Ici également, le test n'est pas symétrique. Si la distribution est compatible avec la loi normale, γ_1 et γ_2 sont simultanément à zéro. En revanche, il suffit que l'un des deux soient différents de zéro pour que l'hypothèse de normalité soit rejetée. Autre point important, on conjecture que les statistiques associées à chaque coefficient sont indépendants (asymptotiquement).

Statistique du test et région critique

Estimateur de γ_2

Nous devons déterminer la statistique et la distribution sous H_0 du coefficient d'aplatissement. Le plus simple est d'utiliser l'estimation triviale déduite de la définition du coefficient γ_2 :

$$g_2 = \frac{\frac{1}{n} \sum_i \hat{\varepsilon}_i^4}{\left(\frac{1}{n} \sum_i \hat{\varepsilon}_i^2\right)^2} - 3 \quad (1.8)$$

Sous H_0 , l'espérance et l'écart-type de g_2 sont :

$$\begin{aligned}\mu_2 &\approx 0 \\ \sigma_2 &\approx \sqrt{\frac{24}{n}}\end{aligned}$$

La statistique standardisée suit une loi normale : $c_2 = \frac{g_2}{\sigma_2} \sim \mathcal{N}(0, 1)$.

Statistique de Jarque-Bera

Maintenant, il faut trouver une manière de combiner les deux statistiques g_1 et g_2 . Puisqu'ils sont indépendants (asymptotiquement), le plus simple est de proposer la statistique de Jarque-Bera¹³ :

$$T = \frac{(n-p-1)}{6} \left(g_1^2 + \frac{g_2^2}{4} \right) \quad (1.9)$$

Remarque 10 (Degré de liberté). La valeur $(n-p-1)$ représente le degré de liberté : nous disposons d'un échantillon de taille n , il y a $(p+1)$ coefficients à estimer dans la régression avec constante. Cette prise en compte des degrés de liberté entraîne une correction des résultats fournis par les logiciels (ex. la fonction `jarqueberaTest(.)` du package `fBasics` de R) d'autant plus importante que le nombre de variables vraies p est grand et que la taille de l'échantillon n est faible.

Sous H_0 , la statistique T suit une loi du χ^2 à 2 degrés de liberté. La région critique du test, au risque α , s'écrit :

$$R.C. : T > \chi_{1-\alpha}^2(2)$$

Il s'agit d'un test unilatéral, $\chi_{1-\alpha}^2(2)$ correspond au fractile d'ordre $1-\alpha$ de la loi du χ^2 à 2 degrés de liberté.

Application sur les données CONSO

Nous complétons le test fondé sur le coefficient d'asymétrie en utilisant les résidus de la régression sur les données CONSO. Voici les principales étapes (Figure 1.15) :

1. Nous récupérons la colonne des résidus $\hat{\varepsilon}_i$.
2. Nous calculons les colonnes de $\hat{\varepsilon}_i^2$, $\hat{\varepsilon}_i^3$ et $\hat{\varepsilon}_i^4$.
3. Nous calculons les sommes et formons $g_1 = \frac{-0.1220}{0.5602^{3/2}} = -0.2909$.
4. Nous formons $g_2 = \frac{0.7020}{0.5602^2} - 3 = -0.7626$.
5. Reste à la calculer la statistique de Jarque-Bera : $T = \frac{31-4-1}{6} \left[(-0.2909)^2 + \frac{(-0.7626)^2}{4} \right] = 0.9967$.
6. Que l'on compare avec le seuil critique $\chi_{0.90}^2(2) = 4.6052$.

Au risque de $\alpha = 10\%$, nous ne pouvons pas rejeter l'hypothèse d'une distribution gaussienne des résidus.

13. http://fr.wikipedia.org/wiki/Test_de_Jarque_Bera

i	Résidu	e ²	e ³	e ⁴
1	-0.0739	0.0055	-0.0004	0.0000
2	-0.6759	0.4568	-0.3088	0.2087
3	0.1183	0.0140	0.0017	0.0002
4	-0.6836	0.4673	-0.3194	0.2184
5	0.0906	0.0082	0.0007	0.0001
6	0.5141	0.2643	0.1359	0.0698
7	-0.6649	0.4421	-0.2939	0.1954
8	0.6095	0.3715	0.2264	0.1380
9	-1.3742	1.8885	-2.5953	3.5665
10	0.1486	0.0221	0.0033	0.0005
11	-1.1104	1.2329	-1.3690	1.5202
12	0.5426	0.2944	0.1598	0.0867
13	0.8148	0.6639	0.5409	0.4407
14	0.9798	0.9599	0.9405	0.9215
15	0.4620	0.2134	0.0986	0.0456
16	-0.2108	0.0444	-0.0094	0.0020
17	1.0551	1.1132	1.1745	1.2391
18	0.4570	0.2089	0.0955	0.0436
19	-0.8565	0.7337	-0.6284	0.5382
20	0.4005	0.1604	0.0642	0.0257
21	-0.9233	0.8525	-0.7871	0.7267
22	1.4360	2.0622	2.9615	4.2528
23	-0.3110	0.0967	-0.0301	0.0094
24	0.2716	0.0737	0.0200	0.0054
25	-1.5678	2.4579	-3.8533	6.0410
26	-0.9534	0.9089	-0.8665	0.8261
27	-0.1257	0.0158	-0.0020	0.0002
28	-0.2656	0.0705	-0.0187	0.0050
29	0.4665	0.2176	0.1015	0.0474
30	0.6112	0.3736	0.2284	0.1396
31	0.8185	0.6700	0.5484	0.4489
Somme		17.3648	-3.7806	21.7634
1/n*somme		0.5602	-0.1220	0.7020

g1	-0.2909
g2	-0.7626
T	0.9967
chi2_{1-alpha}(2)	4.6052

Fig. 1.15. Test de Jarque-Bera pour vérifier la normalité des résidus sur les données CONSO

1.4 Conclusion

Examiner les résidus est un des moyens les plus sûrs d'évaluer la qualité d'une régression. Nous avons présenté dans ce chapitre quelques outils, plus ou moins sophistiqués, pour apprécier correctement les informations qu'ils peuvent nous apporter. Dans la majorité des cas, les écueils qui peuvent invalider une régression sont :

- la liaison étudiée est non-linéaire ;
- un problème de spécification, par ex. une variable exogène importante manque ;
- l'existence de points atypiques ou exagérément influents ;
- les erreurs ne sont pas indépendants et/ou dépendent d'une des exogènes ;
- il y a une rupture de structure dans la relation ou les données sont organisées en blocs non homogènes,...

Malgré la puissance des procédures numériques avancées, les techniques graphiques très simples sont à privilégier, au moins dans un premier temps : leurs conditions d'applications sont universelles, elles proposent un diagnostic nuancé de situations qui peuvent s'avérer complexes. Rien ne nous empêche par la suite de compléter le diagnostic visuel à l'aide des tests statistiques.

Détection des points aberrants et des points influents

L'objectif de la détection des points aberrants et influents est de repérer des points qui jouent un rôle anormal dans la régression, jusqu'à en fausser les résultats. Il faut s'entendre sur le terme *anormal*, nous pourrions en résumer les différentes tournures de la manière suivante :

- L'observation prend une valeur inhabituelle sur une des variables. Nous parlons alors de détection univariée car nous étudions les variables individuellement. Par exemple, un des véhicules a une puissance 700 cv, nous avons intégré une Formule 1 dans notre fichier de véhicules.
- Une combinaison de valeurs chez les exogènes est inhabituelle. Par exemple, une voiture très légère et très puissante : le poids pris individuellement ne se démarque pas, la puissance non plus, mais leur concomitance est surprenante (Figure 2.1).
- L'observation est très mal reconstituée par la régression, n'obéissant pas de manière ostensible à la relation modélisée entre les exogènes et l'endogène. Dans ce cas, le résidu observé est trop élevé.
- L'observation pèse de manière exagérée dans la régression, au point que les résultats obtenus (prédiction, coefficient, ...) sont *très différents* selon que nous l'intégrons ou non dans la régression.

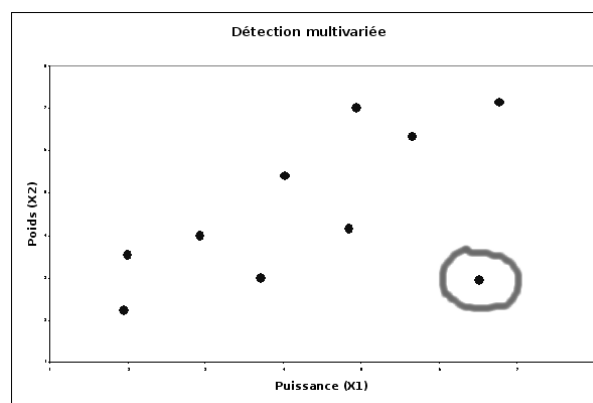


Fig. 2.1. Le point entouré est suspect car la combinaison de valeurs est inhabituelle

Outre les ouvrages énumérés en bibliographie, deux références en ligne complètent à merveille ce chapitre : le document de J. Confais et M. Le Guen [12], section 4.3, pages 307 à 311 ; et la présentation de

A.Gueguen, *La régression linéaires - Outils diagnostics*, <http://ifr69.vjf.inserm.fr/~webifr/ppt/outilsdiag.ppt>.

2.1 Points aberrants : détection univariée

Boîte à moustache et détection des points atypiques

L'outil le plus simple pour se faire une idée de la distribution d'une variable continue est la boîte à moustaches (Figure 2.2), dite *box-plot*¹. Elle offre une vue synthétique sur plusieurs indicateurs importants : le premier quartile (Q_1), la médiane (Me) et le troisième quartile (Q_3). On peut aussi jauger visuellement l'intervalle inter-quartile qui mesure la dispersion ($IQ = Q_3 - Q_1$).

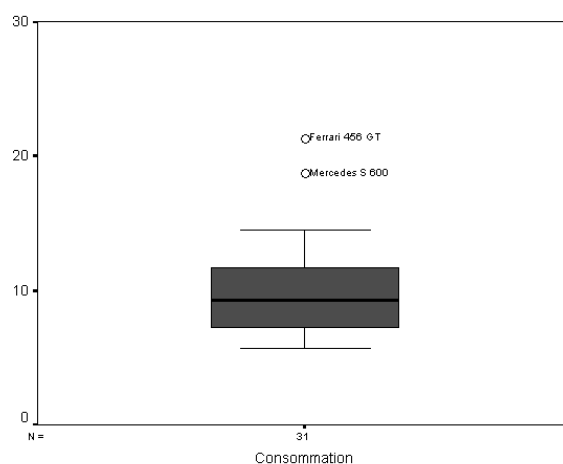


Fig. 2.2. Boxplot de la variable endogène "consommation (y)", 2 observations se démarquent

On pense à tort que les extrémités de la boîte correspondent aux valeurs minimales et maximales. En réalité il s'agit des valeurs minimales et maximales non atypiques. Les seuils désignant les valeurs atypiques sont définies par les règles suivantes² :

$$LIF = Q_1 - 1.5 \times IQ$$

$$UIF = Q_3 + 1.5 \times IQ$$

où LIF signifie "lower inner fence" et UIF "upper inner fence".

Les points situés au delà de ces limites sont souvent jugées *atypiques*. Il convient de se pencher attentivement sur les observations correspondantes.

1. http://en.wikipedia.org/wiki/Box_plot

2. <http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm>

Remarque 11 (Règle des 3-sigma). Une autre règle empirique est largement répandue dans la communauté statistique, il s'agit de la règle des 3-sigma. Elle fixe les bornes basses et hautes à 3 fois l'écart-type autour de la moyenne. Si l'on considère que la distribution est normale, 99.7% des observations sont situées dans cet intervalle. La principale faiblesse de cette approche est l'hypothèse de normalité sous-jacente qui en réduit la portée.

Les "outer fence"

Il est possible de durcir les conditions ci-dessus en élargissant les bornes des valeurs. On parle alors de *outer fence*. Elles sont définies de la manière suivante :

$$LOF = Q1 - 3 \times IQ$$

$$UOF = Q3 + 3 \times IQ$$

Pour distinguer les points détectés selon la règle *inner* ou *outer*, on parle de "points moyennement atypiques" (mild outlier) et "points extrêmement atypiques" (extreme outlier).

Application sur les données CONSO

Il est possible de produire une boîte à moustache pour chaque variable du fichier de données. Nous disposons ainsi très rapidement d'informations sur l'étalement de la distribution, de la présence de points qui s'écartent fortement des autres. Pour la variable endogène (Figure 2.2), nous détectons immédiatement 2 observations suspectes qui consomment largement plus que les autres véhicules : la Ferrari 456 GT et la Mercedes S 600.

Une autre manière de procéder est d'utiliser simplement le tableur EXCEL (Figure 2.3) :

1. de produire le 1er et le 3ème quartile ;
2. d'en déduire l'intervalle inter-quartile ;
3. de calculer les bornes *LIF* et *UIF* ;
4. et de s'appuyer sur la mise en forme conditionnelle pour distinguer les points "suspects" pour chaque variable.

Il semble que 3 véhicules soient assez différents du reste de l'échantillon, sur la quasi-totalité des variables. Nous produisons dans un tableau récapitulatif les associations "observation-variable" suspects (Tableau 2.1).

i	Modèle	Prix	Cylindrée	Puissance	Poids	Consommation
1	Daihatsu Cuore	11600	846	32	650	5.7
2	Suzuki Swift 1.0 GLS	12490	993	39	790	5.8
3	Fiat Panda Mambo L	10450	899	29	730	6.1
4	VW Polo 1.4 60	17140	1390	44	955	6.5
5	Opel Corsa 1.2i Eco	14825	1195	33	895	6.8
6	Subaru Vivio 4WD	13730	658	32	740	6.8
7	Toyota Corolla	19490	1331	55	1010	7.1
8	Ferrari 456 GT	285000	5474	325	1690	21.3
9	Mercedes S 600	183900	5987	300	2250	18.7
10	Maserati Ghibli GT	92500	2789	209	1485	14.5
11	Opel Astra 1.6i 16V	25000	1597	74	1080	7.4
12	Peugeot 306 XS 108	22350	1761	74	1100	9
13	Renault Safrane 2.2 V	36600	2165	101	1500	11.7
14	Seat Ibiza 2.0 GTI	22500	1983	85	1075	9.5
15	VW Golt 2.0 GTI	31580	1984	85	1155	9.5
16	Citroen ZX Volcane	28750	1998	89	1140	8.8
17	Fiat Tempira 1.6 Liberty	22600	1580	65	1080	9.3
18	Fort Escort 1.4i PT	20300	1390	54	1110	8.6
19	Honda Civic Joker 1.4	19900	1396	66	1140	7.7
20	Volvo 850 2.5	39800	2435	106	1370	10.8
21	Ford Fiesta 1.2 Zetec	19740	1242	55	940	6.6
22	Hyundai Sonata 3000	38990	2972	107	1400	11.7
23	Lancia K 3.0 LS	50800	2958	150	1550	11.9
24	Mazda Hachback V	36200	2497	122	1330	10.8
25	Mitsubishi Galant	31990	1998	66	1300	7.6
26	Opel Omega 2.5i V6	47700	2496	125	1670	11.3
27	Peugeot 806 2.0	36950	1998	89	1560	10.8
28	Nissan Primera 2.0	26950	1997	92	1240	9.2
29	Seat Alhambra 2.0	36400	1984	85	1635	11.6
30	Toyota Previa salon	50900	2438	97	1800	12.8
31	Volvo 960 Kombi aut	49300	2473	125	1570	12.7

Q1	19820	1390	55	1042.5	7.25
Q3	39395	2455.5	106.5	1525	11.65
IQ	19575	1065.5	51.5	482.5	4.4

LIF	-9542.5	-208.25	-22.25	318.75	0.65
UIF	68757.5	4053.75	183.75	2248.75	18.25

Fig. 2.3. Détection univariée des points atypiques pour chaque variable

Observations	Prix	Cylindrée	Puissance	Poids	Consommation
Ferrari 456 GT	*	*	*		*
Mercedes S 600	*	*	*	*	*
Maserati Ghibli GT	*		*		

Tableau 2.1. Points suspects fichier CONSO : détection univariée

2.2 Détection multivariée sur les exogènes : le levier

Le levier

La détection univariée donne déjà des informations intéressantes. Mais elle présente le défaut de ne pas tenir compte des interactions entre les variables. Dans cette section, nous étudions un outil capital pour l'étude des points atypiques et influents : le *levier*.

Son interprétation est relativement simple. Il indique, pour l'observation i , la distance avec le centre de gravité du nuage de points dans l'espace défini par les exogènes. La mesure a de particulier qu'elle tient compte de la forme du nuage de points, il s'agit de la *distance de Mahalanobis* (Tenenhaus, page 94). La prise en compte de la configuration des points dans l'espace de représentation permet de mieux juger de l'éloignement d'une observation par rapport aux autres (Figure 2.4).

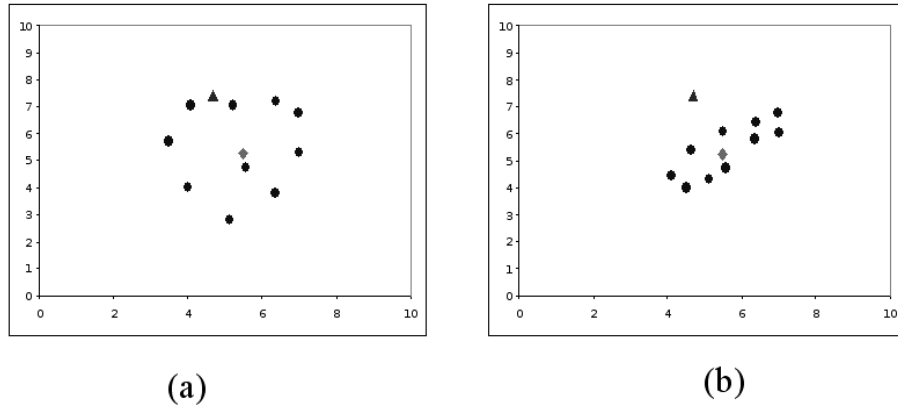


Fig. 2.4. Le point \triangle et le centre de gravité \diamond sont situés aux mêmes coordonnées dans les graphiques (a) et (b). Pourtant \triangle apparaît nettement atypique dans (b).

Le levier h_{ii} de l'observation i est lue sur la diagonale principale de la matrice H , dite *Hat Matrix*, définie de la manière suivante

$$H = X(X'X)^{-1}X' \quad (2.1)$$

La matrice H joue un rôle très important dans la régression, elle permet de passer des valeurs observées de Y vers les valeurs prédites \hat{Y} , elle permet aussi le passage de l'erreur théorique vers les résidus observés³. Les éléments h_{ij} de la matrice H présentent un certain nombre de propriétés. Concernant les éléments de la diagonale principale h_{ii} , on parle de *levier* car il détermine l'influence de l'observation i sur les estimateurs obtenus par les moindres carrés (Dodge, page 130). Même s'il n'utilise que les informations en provenance des exogènes X_j , le champ d'action du levier dépasse la détection multivariée des points aberrants. Nous le retrouverons dans la grande majorité des formules de détection des points atypiques et influents que nous présenterons dans la suite de ce chapitre.

Calcul des éléments diagonaux de la matrice H

La taille ($n \times n$) de la matrice H peut être considérable dès lors que la taille de l'échantillon augmente. Il est possible d'en calculer uniquement les éléments diagonaux en utilisant la formule

$$h_{ii} = h_i = x_i(X'X)^{-1}x_i'$$

où x_i représente la i -ème ligne de la matrice X .

Région critique

Nous disposons d'un indicateur. Il nous faut maintenant déterminer à partir de quelle valeur de h_i nous devons nous pencher attentivement sur une observation. Autrement dit, quelle est la valeur critique qui permet d'indiquer qu'un point est "suspect" ?

3. $\hat{\varepsilon} = [I - X(X'X)^{-1}X']\varepsilon$

Pour cela, penchons-nous sur quelques propriétés du levier. Par définition $0 \leq h_i \leq 1$, et surtout $\sum_{i=1}^n h_i = p + 1$, où $p + 1$ est le nombre de coefficients à estimer dans une régression avec constante. On considère que le levier d'une observation est anormalement élevé dès lors que :

$$R.C. : h_i > 2 \times \frac{p+1}{n} \quad (2.2)$$

Remarque 12 (Seuil de coupure et étude des points). La règle définie ci-dessus, aussi répandue soit-elle, est avant tout empirique. Dans la pratique, il est tout aussi pertinent de trier les observations selon la valeur de h_i de manière à mettre en évidence les cas extrêmes. Une étude approfondie de ces observations permet de statuer sur leur positionnement par rapport aux autres.

Application sur les données CONSO

Statistic	Leverage	RStandard	RStudent	DFFITS	Cook's D	COVRATIO
Lower Bound	-	-	-2.0000	-0.8032	-	0.5161
Upper Bound	0.3226	-	2.0000	0.8032	0.1538	1.4839
1	0.1397640	-0.0974596	-0.0955845	-0.0385280	0.0003086	1.4117430
2	0.0918145	-0.8678561	-0.8636034	-0.2745888	0.0152287	1.1565365
3	0.1130612	0.1536788	0.1507630	0.0538276	0.0006021	1.3655251
4	0.0808787	-0.8724912	-0.8683545	-0.2575893	0.0133972	1.1409514
5	0.1012776	0.1169945	0.1147527	0.0385218	0.0003085	1.3501986
6	0.1427420	0.6793855	0.6721855	0.2742896	0.0153710	1.2976569
7	0.0514944	-0.8353685	-0.8303655	-0.1934770	0.0075772	1.1195616
8	0.8685865	2.0573680	2.2048566	5.6684833	5.5953465	3.8078198
9	0.4842937	-2.3415866	-2.5847800	-2.5048213	1.0298092	0.7218781
10	0.6417805	0.3039078	0.2985368	0.3995918	0.0330941	3.3364913
11	0.0439766	-1.3895967	-1.4162163	-0.3037428	0.0177648	0.8652140
12	0.0486532	0.6807337	0.6735436	0.1523183	0.0047398	1.1688980
13	0.0773373	1.0379379	1.0395466	0.3009657	0.0180600	1.0671639
14	0.1049747	1.2672284	1.2828686	0.4393462	0.0376694	0.9883063
15	0.0475623	0.5792606	0.5717129	0.1277589	0.0033512	1.1970818
16	0.0623289	-0.2663410	-0.2615259	-0.0674271	0.0009431	1.2799220
17	0.0412885	1.3185245	1.3384416	0.2777603	0.0149743	0.8978415
18	0.0580546	0.5762327	0.5686857	0.1411814	0.0040930	1.2112412
19	0.0600194	-1.0810272	-1.0846914	-0.2740895	0.0149237	1.0284551
20	0.0579124	0.5049039	0.4975442	0.1233592	0.0031342	1.2293603
21	0.0620794	-1.1665739	-1.1750864	-0.3023155	0.0180151	0.9914220
22	0.2746014	2.0631542	2.2122695	1.3611345	0.3222697	0.6860567
23	0.1504667	-0.4128478	-0.4061641	-0.1709352	0.0060377	1.3858120
24	0.1232989	0.3548919	0.3488461	0.1308241	0.0035427	1.3544738
25	0.1135466	-2.0375178	-2.1795177	-0.7800446	0.1063533	0.5751144
26	0.1278381	-1.2491297	-1.2633678	-0.4836836	0.0457413	1.0237172
27	0.1520067	-0.1669817	-0.1638269	-0.0693618	0.0009996	1.4270668
28	0.0505854	-0.3334890	-0.3277145	-0.0756450	0.0011851	1.2543029
29	0.2258011	0.6487385	0.6413524	0.3463645	0.0245495	1.4483614
30	0.3154361	0.9039652	0.9006776	0.6113896	0.0753062	1.5149912
31	0.0865386	1.0479325	1.0499960	0.3231822	0.0208073	1.0734127

Fig. 2.5. Quelques indicateurs de points atypiques et influents dans TANAGRA. Données CONSO.

Nous appliquons les calculs ci-dessus sur les données CONSO. Nous avons utilisé le logiciel TANAGRA (Figure 2.5)⁴. La valeur de coupure est $2 \times \frac{4+1}{31} = 0.3226$, 3 points se démarquent immédiatement, les mêmes que pour la détection univariée : la Ferrari ($h_8 = 0.8686$), la Mercedes ($h_9 = 0.4843$) et la Maserati ($h_{10} = 0.6418$). Les raisons semblent évidentes : il s'agit de grosses cylindrées luxueuses, des limousines (Mercedes) ou des véhicules sportifs (Ferrari, Maserati).

Essayons d'approfondir notre analyse en triant cette fois-ci les observations de manière décroissante selon h_i . Les 3 observations ci-dessus arrivent bien évidemment en première place, mais nous constatons que d'autres observations présentaient un levier proche de la valeur seuil. Il s'agit de la Toyota Previa Salon, et dans une moindre mesure de la Hyundai Sonata 3000 (Figure 2.6). La première est un monospace (nous remarquons à proximité 2 autres monospaces, la Seat Alhambra et la Peugeot 806) qui se distingue par la conjonction d'un prix et d'un poids élevés ; la seconde est une voiture de luxe coréenne, les raisons de son éloignement par rapport aux autres véhicules tiennent, semble-t-il, en la conjonction peu courante d'un prix relativement moyen et d'une cylindrée élevée.

Modèle	const	Prix	Cylindrée	Puissance	Poids	Consomm	Prédiction	Résidus	0.3226
									Leverage
Ferrari 456 GT	1	285000	5474	325	1690	21.3	20.690	0.610	0.8686
Maserati Ghibli GT	1	92500	2789	209	1485	14.5	14.351	0.149	0.6418
Mercedes S 600	1	183900	5987	300	2250	18.7	20.074	-1.374	0.4843
Toyota Previa salon	1	50900	2438	97	1800	12.8	12.189	0.611	0.3154
Hyundai Sonata 3000	1	38990	2972	107	1400	11.7	10.264	1.436	0.2746
Seat Alhambra 2.0	1	36400	1984	85	1635	11.6	11.134	0.466	0.2258
Peugeot 806 2.0	1	36950	1998	89	1560	10.8	10.926	-0.126	0.1520
Lancia K 3.0 LS	1	50800	2958	150	1550	11.9	12.211	-0.311	0.1505
Subaru Vivio 4WD	1	13730	658	32	740	6.8	6.286	0.514	0.1427
Daihatsu Cuore	1	11600	846	32	650	5.7	5.774	-0.074	0.1398
Opel Omega 2.5i V6	1	47700	2496	125	1670	11.3	12.253	-0.953	0.1278
Mazda Hatchback V	1	36200	2497	122	1330	10.8	10.528	0.272	0.1233
Mitsubishi Galant	1	31990	1998	66	1300	7.6	9.168	-1.568	0.1135
Fiat Panda Mambo L	1	10450	899	29	730	6.1	5.982	0.118	0.1131
Seat Ibiza 2.0 GTI	1	22500	1983	85	1075	9.5	8.520	0.980	0.1050
Opel Corsa 1.2i Eco	1	14825	1195	33	895	6.8	6.709	0.091	0.1013
Suzuki Swift 1.0 GLS	1	12490	993	39	790	5.8	6.476	-0.676	0.0918
Volvo 960 Kombi aut	1	49300	2473	125	1570	12.7	11.881	0.819	0.0865
VW Polo 1.4 60	1	17140	1390	44	955	6.5	7.184	-0.684	0.0809
Renault Safrane 2.2. V	1	36600	2165	101	1500	11.7	10.885	0.815	0.0773
Citroen ZX Volcane	1	28750	1998	89	1140	8.8	9.011	-0.211	0.0623
Ford Fiesta 1.2 Zetec	1	19740	1242	55	940	6.6	7.523	-0.923	0.0621
Honda Civic Joker 1.4	1	19900	1396	66	1140	7.7	8.557	-0.857	0.0600
Fort Escort 1.4i PT	1	20300	1390	54	1110	8.6	8.143	0.457	0.0581
Volvo 850 2.5	1	39800	2435	106	1370	10.8	10.399	0.401	0.0579
Toyota Corolla	1	19490	1331	55	1010	7.1	7.765	-0.665	0.0515
Nissan Primera 2.0	1	26950	1997	92	1240	9.2	9.466	-0.266	0.0506
Peugeot 306 XS 108	1	22350	1761	74	1100	9	8.457	0.543	0.0487
VW Golf 2.0 GTI	1	31580	1984	85	1155	9.5	9.038	0.462	0.0476
Opel Astra 1.6i 16V	1	25000	1597	74	1080	7.4	8.510	-1.110	0.0440
Fiat Tempra 1.6 Liberty	1	22600	1580	65	1080	9.3	8.245	1.055	0.0413

Fig. 2.6. Trier les données CONSO selon la valeur du levier

4. Nous avons utilisé un logiciel spécialisé par commodité. L'enchaînement des calculs peut être facilement reproduit sur un tableur, il suffit d'utiliser à bon escient les fonctions matricielles.

2.3 Résidu standardisé

Résidu standardisé

Le résidu standardisé, appelé également *résidu studentisé interne* dans certains ouvrages, s'intéresse à l'importance du résidu observé $\hat{\varepsilon}_i = y_i - \hat{y}_i$. S'il est anormalement élevé, en valeur absolue, le point a été mal reconstitué par le modèle : il s'écarte ostensiblement de la relation modélisée entre les exogènes et l'endogène.

Si par hypothèse, la variance de l'erreur $\sigma_{\varepsilon_i}^2 = \sigma_\varepsilon^2$ est constante, il en va autrement du résidu $\sigma_{\hat{\varepsilon}_i}^2 = \sigma_\varepsilon^2(1 - h_i)$. Nous devons donc normaliser le résidu par son écart-type pour rendre les écarts comparables d'une observation à l'autre.

Lorsque nous travaillons sur un échantillon, nous ne disposons pas de la vraie valeur de σ_ε^2 , nous estimons la variance des résidus avec

$$\hat{\sigma}_{\hat{\varepsilon}_i}^2 = \hat{\sigma}_\varepsilon^2(1 - h_i) \quad (2.3)$$

où h_i est lue dans la *Hat Matrix* H , $\hat{\sigma}_\varepsilon^2 = \frac{\sum_i \hat{\varepsilon}_i^2}{n-p-1}$ est l'estimateur de la variance de l'erreur.

Le résidu standardisé est défini par le rapport

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{\hat{\varepsilon}_i}} = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_\varepsilon \sqrt{(1 - h_i)}} \quad (2.4)$$

Région critique

Pour décider du statut d'un point, il nous faut définir une valeur seuil au delà de laquelle le résidu standardisé est anormalement élevé (en valeur absolue).

Nous pouvons nous appuyer sur un appareillage statistique ici. En effet, par hypothèse $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon)$, nous en déduisons que $\hat{\varepsilon}_i \sim \mathcal{N}(0, \sigma_{\hat{\varepsilon}_i})$. On peut montrer facilement que $\hat{\sigma}_{\hat{\varepsilon}_i}^2$ suit une loi du χ^2 à $(n - p - 1)$ degrés de liberté.

De fait, le résidu standardisé, défini par le rapport (Equation 2.4) entre une loi normale et la racine carrée d'une loi du χ^2 normalisée), suit une loi de Student à $(n - p - 1)$ degrés de liberté

$$t_i \sim \mathcal{T}(n - p - 1) \quad (2.5)$$

Nous décidons qu'une observation est particulièrement mal reconstituée par le modèle (d'une certaine manière atypique) lorsque

$$R.C. : |t_i| > t_{1-\frac{\alpha}{2}}(n - p - 1)$$

où $t_{1-\frac{\alpha}{2}}(n - p - 1)$ est le fractile d'ordre $1 - \frac{\alpha}{2}$ de la loi de Student à $(n - p - 1)$ degrés de liberté.

Il s'agit bien d'un test bilatéral. Le résidu est suspect s'il est particulièrement élevé en valeur absolue.

Au final, un point apparaît comme aberrant avec un résidu standardisé élevé si :

- il est mal prédit c.-à-d. $\hat{\varepsilon}_i$ est élevé ;
- la régression est précise c.-à-d. $\hat{\sigma}_\varepsilon$ est faible ; en effet, si la régression est globalement précise, un point mal prédit apparaît comme d'autant plus suspect ;
- le point est éloigné des autres dans l'espace des exogènes ; en effet, plus h_i est élevé ($h_i \approx 1$), plus $(1 - h_i) \approx 0$, et le rapport est élevé.

Application sur les données CONSO

TANAGRA fournit automatiquement les résidus standardisés lors de l'analyse des points atypiques (Figure 2.5). Il faut comparer la valeur absolue de la colonne avec la valeur seuil $t_{0.95}(26) = 1.7056$ pour un risque à 10%.

Lorsque le nombre d'observations est élevé, il devient mal aisé d'inspecter le tableau des valeurs du résidu standardisé. Il est plus commode de revenir au graphique des résidus en mettant en abscisse l'endogène et en ordonnée le résidu standardisé. Nous traçons alors une ligne matérialisant les valeurs seuils $-t_{1-\frac{\alpha}{2}}$ et $+t_{1-\frac{\alpha}{2}}$ (Figure 2.7)⁵.

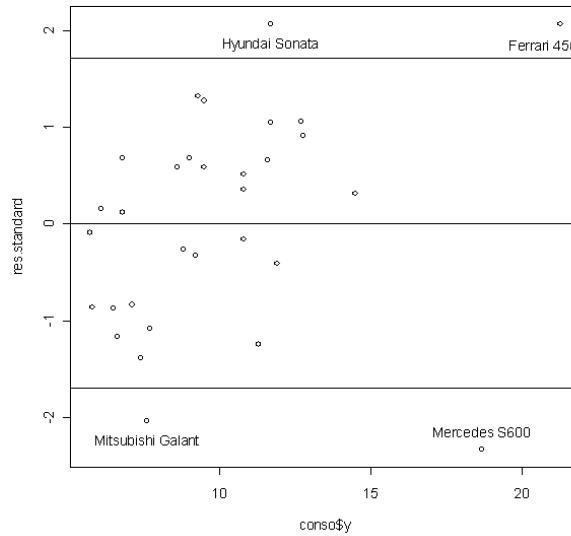


Fig. 2.7. Graphique des résidus standardisés vs. endogène - Données CONSO

Remarque 13 (Taille d'échantillon et risque α). Autre approche pragmatique, nous pouvons trier les données selon $|t_i|$. Les véhicules suspects sont très facilement mis en évidence (Figure 2.8). Cette technique est d'autant plus intéressante que le nombre de véhicules situés dans la région critique s'accroît mécaniquement à mesure que la taille n de l'échantillon augmente, laissant à croire un nombre élevé d'observations

5. Graphique réalisé avec le logiciel R, il est très facile de placer des étiquettes aux coordonnées choisies.

aberrantes. Il faudrait ajuster le risque α en accord avec la taille d'échantillon n . Mais il s'agit là d'une opération délicate. En utilisant un tri simple, nous pouvons considérer, par ordre d'importance, les points les moins bien reconnus par le modèle sans se poser la question d'un seuil critique convenable.

									0.3226	1.7056
Modèle	const	Prix	Cylindrée	Puissanc	Poids	Consomm	Prédiction	Résidus	Leverage	R.Standardisé
Mercedes S 600	1	183900	5967	300	2250	18.7	20.074	-1.374	0.4843	2.3416
Hyundai Sonata 3000	1	38990	2972	107	1400	11.7	10.264	1.436	0.2746	2.0632
Ferrari 456 GT	1	285000	5474	325	1690	21.3	20.690	0.610	0.8686	2.0574
Mitsubishi Galant	1	31990	1998	66	1300	7.6	9.168	-1.568	0.1135	2.0375
Opel Astra 1.6i 16V	1	25000	1597	74	1080	7.4	8.510	-1.110	0.0440	1.3896
Fiat Tempra 1.6 Liberty	1	22600	1580	65	1080	9.3	8.245	1.055	0.0413	1.3185
Seat Ibiza 2.0 GTI	1	22500	1983	85	1075	9.5	8.520	0.980	0.1050	1.2672
Opel Omega 2.5i V6	1	47700	2496	125	1670	11.3	12.253	-0.953	0.1278	1.2491
Ford Fiesta 1.2 Zetec	1	19740	1242	55	940	6.6	7.523	-0.923	0.0621	1.1666
Honda Civic Joker 1.4	1	19900	1396	66	1140	7.7	8.557	-0.857	0.0600	1.0810
Volvo 960 Kombi aut	1	49300	2473	125	1570	12.7	11.881	0.819	0.0865	1.0479
Renault Safrane 2.2. V	1	36600	2165	101	1500	11.7	10.885	0.815	0.0773	1.0379
Toyota Previa salon	1	50900	2438	97	1800	12.8	12.189	0.611	0.3154	0.9040
VW Polo 1.4 60	1	17140	1390	44	955	6.5	7.184	-0.684	0.0809	0.8725
Suzuki Swift 1.0 GLS	1	12490	993	39	790	5.8	6.476	-0.676	0.0918	0.8679
Toyota Corolla	1	19490	1331	55	1010	7.1	7.765	-0.665	0.0515	0.8354
Peugeot 306 XS 108	1	22350	1761	74	1100	9	8.457	0.543	0.0487	0.6807
Subaru Vivio 4WD	1	13730	658	32	740	6.8	6.286	0.514	0.1427	0.6794
Seat Alhambra 2.0	1	36400	1984	85	1635	11.6	11.134	0.466	0.2258	0.6487
VW Golf 2.0 GTI	1	31580	1984	85	1155	9.5	9.038	0.462	0.0476	0.5793
Fort Escort 1.4i PT	1	20300	1390	54	1110	8.6	8.143	0.457	0.0581	0.5762
Volvo 850 2.5	1	39800	2435	106	1370	10.8	10.399	0.401	0.0579	0.5049
Lancia K 3.0 LS	1	50800	2958	150	1550	11.9	12.211	-0.311	0.1505	0.4128
Mazda Hachtback V	1	36200	2497	122	1330	10.8	10.528	0.272	0.1233	0.3549
Nissan Primera 2.0	1	26950	1997	92	1240	9.2	9.466	-0.266	0.0506	0.3335
Maserati Ghibli GT	1	92500	2789	209	1485	14.5	14.351	0.149	0.6418	0.3039
Citroen ZX Volcane	1	28750	1998	89	1140	8.8	9.011	-0.211	0.0623	0.2663
Peugeot 806 2.0	1	36950	1998	89	1560	10.8	10.926	-0.126	0.1520	0.1670
Fiat Panda Mambo L	1	10450	899	29	730	6.1	5.982	0.118	0.1131	0.1537
Opel Corsa 1.2i Eco	1	14825	1195	33	895	6.8	6.709	0.091	0.1013	0.1170
Daihatsu Cuore	1	11600	846	32	650	5.7	5.774	-0.074	0.1398	0.0975

Fig. 2.8. Observations triées selon la valeur absolue du résidu standardisé

Les calculs aboutissent à des résultats contrastés, correspondant à des situations très différentes (Figure 2.8) :

- La Mercedes cumule un résidu fort (-1.374) et un levier élevé (0.4843). Ce type de véhicule appartient à une catégorie spécifique qui n'a rien en commun avec les voitures recensés dans ce fichier.
- La "Ferrari" est mal reconstituée parce qu'elle est avant tout très différente des autres $h = 0.8686$. Le résidu brut $\hat{\varepsilon} = 0.610$ n'est pas très élevé, on prédit correctement sa consommation au regard de ses caractéristiques. Mais le résidu rapporté à l'écart-type montre qu'il s'agit quand même d'un véhicule bien particulier.
- La Hyundai et la Mitsubishi Galant correspondent à une tout autre situation. Ces observations se fondent dans l'ensemble de la population, le levier est en deçà du seuil critique. En revanche ils n'obéissent pas à la relation mise en évidence entre les exogènes et l'endogène (Equation 0.1). La Hyundai consomme fortement par rapport à ses caractéristiques $\hat{\varepsilon} = y - \hat{y} = 11.7 - 10.264 = 1.436$; la Mitsubishi est en revanche particulièrement sobre (au regard de sa cylindrée) $\hat{\varepsilon} = 7.6 - 9.168 = -1.568$.

2.4 Résidu studentisé

Le résidu studentisé

Principe

Le résidu standardisé est un indicateur certes intéressant mais il présente un inconvénient fort : nous évaluons l'importance du résidu $\hat{\varepsilon}_i$ d'une observation qui a participé à la construction de la droite de régression. De fait, le point est juge et partie dans l'évaluation : on l'utilise pour construire le modèle, puis on regarde s'il a bien été modélisé. Si l'observation est fortement influente, au sens qu'elle "tire" exagérément les résultats de manière à présenter un résidu brut très faible $\hat{\varepsilon} \approx 0$, nous concluons à tort qu'elle est bien reconstituée et donc ne fausse en rien les résultats de la modélisation (Figure 2.9).

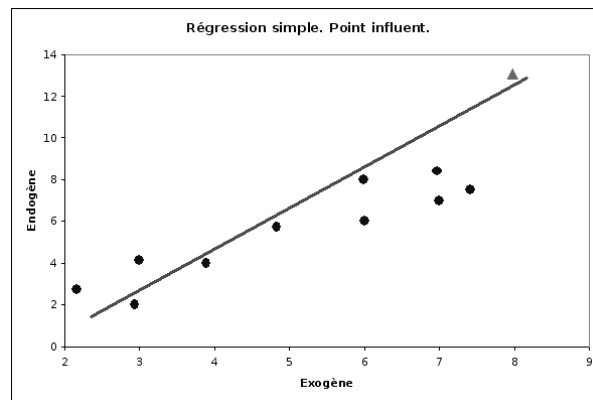


Fig. 2.9. Exemple de régression simple où l'observation \triangle est certes bien modélisée ($\hat{\varepsilon} \approx 0$) mais elle fausse totalement les calculs : on parle de point exagérément influent.

Il faudrait mettre en place une procédure qui permet de **confronter les résultats selon qu'une observation participe ou non aux calculs**. Parmi les pistes possible, nous nous penchons sur l'erreur de prédiction. Une mesure objective devrait ne pas faire participer le point i dans la construction du modèle utilisé pour prédire la valeur \hat{y}_i . Le résidu studentisé, on parle de *résidu studentisé externe ou RSTUDENT* dans certains ouvrages, s'appuie sur ce principe, il utilise la procédure suivante (Dodge, page 135) :

- Pour chaque observation i ,
- Nous la retirons de l'ensemble des données, et nous calculons les paramètres de la régression.
- Nous effectuons la prédiction sur l'observation i en donnée supplémentaire $\hat{y}_i(-i)$
- Nous obtenons aussi l'estimation de l'écart-type des erreurs $\hat{\sigma}_\varepsilon(-i)$, le levier $h_i(-i)$ obtenu avec la formule $h_i(-i) = x_i(X'_{-i}X_{-i})^{-1}x'_i$ où X_{-i} correspond à la matrice des X sans la ligne numéro i .
- A l'instar du résidu standardisé, nous formons le résidu studentisé à partir du rapport

$$t_i^* = \frac{y_i - \hat{y}_i(-i)}{\hat{\sigma}_\varepsilon(-i)\sqrt{(1 - h_i(-i))}} \quad (2.6)$$

Le principe de la donnée supplémentaire permet de mieux appréhender le rôle/le poids de l'observation i dans la régression. Si, exclue de la régression, elle reste bien prédite, elle est fondue dans la masse des points; en revanche, si son exclusion des calculs entraîne une très mauvaise prédiction, on peut penser qu'elle pèse fortement, peut-être à tort, sur les calculs (Figure 2.10).

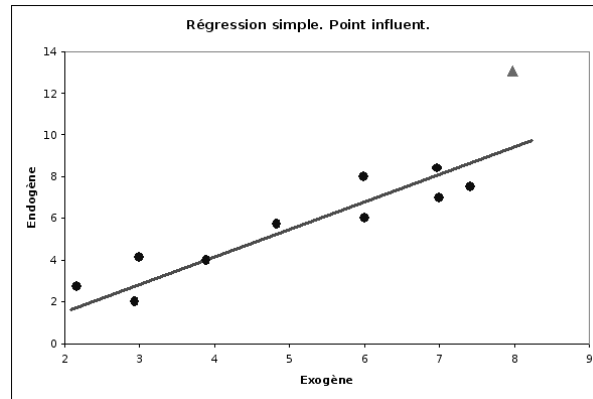


Fig. 2.10. Principe de la donnée supplémentaire : l'observation \triangle , exclue du calcul de la droite de régression, devient très mal prédite

Une autre interprétation

Il existe une autre manière de calculer le résidu studentisé. Elle ne facilite pas spécialement les calculs. En revanche, elle a le mérite de mettre en lumière la loi de distribution que nous pourrions utiliser par la suite pour définir la région critique du test.

Le principe est le suivant, nous effectuons n régressions avec toutes les observations. Pour la régression numéro i , nous introduisons une variable muette z définie de la manière suivante

$$\begin{aligned} z &= 1 \text{ pour l'observation numéro } i \\ &= 0 \text{ sinon} \end{aligned}$$

La régression numéro i s'écrit donc de la manière suivante :

$$y = a_0 + a_1x_1 + \dots + a_px_p + b \times z + \varepsilon \quad (2.7)$$

Le résidu studentisé correspond au t de Student du test de significativité du coefficient b . Nous savons que cette statistique suit une loi de Student $\mathcal{T}(n - p - 2)$ à $(n - p - 2)$ degrés de liberté. En effet, il y a bien $(p + 2)$ coefficients à estimer dans l'équation 2.7.

Calcul pratique

Si le concept sous-jacent semble relativement simple, il reste à produire les résultats. Quelle que soit l'approche adoptée, il faudrait effectuer n régressions. Si n est élevé, le calcul est très lourd, il peut se révéler rédhibitoire.

A ce stade intervient une propriété remarquable du résidu studentisé : **il est possible de le calculer pour chaque observation i sans avoir à procéder explicitement aux n régressions**. Nous utilisons pour cela d'une formule de transformation du résidu standardisé (Tenenhaus, page 95)⁶ :

$$t_i^* = t_i \sqrt{\frac{n - p - 2}{n - p - 1 - t_i^2}} \quad (2.8)$$

Le calcul supplémentaire demandé est négligeable.

Région critique

A partir de la formulation sous forme d'équation de régression (Équation 2.7), il est possible d'écrire rigoureusement le test d'hypothèses permettant de déterminer si une observation est atypique/influente ou non. On oppose :

$$H_0 : b = 0$$

$$H_1 : b \neq 0$$

Sous H_0 , la statistique $t_i^* \sim \mathcal{T}(n - p - 2)$, on en déduit la région critique du test :

$$R.C. : |t_i^*| > t_{1-\frac{\alpha}{2}}(n - p - 2)$$

où $t_{1-\frac{\alpha}{2}}(n - p - 2)$ est le fractile d'ordre $1 - \frac{\alpha}{2}$ de la loi de Student à $(n - p - 2)$ degrés de liberté.

Il s'agit bien d'un test bilatéral. Le résidu est suspect s'il est particulièrement élevé en valeur absolue.

Comparaisons multiples et contrôle du risque - I

En multipliant les tests, nous évaluons n observations, nous augmentons le risque de signaler à tort des points atypiques. Certains auteurs préconisent de rendre la détection plus exigeante en introduisant la correction de Bonferroni pour les comparaisons multiples : on divise le risque α par l'effectif n . Pour chaque observation à tester, nous comparons le résidu studentisé avec le fractile d'ordre $1 - \frac{\alpha}{2n}$. Dans l'exemple CONSO, le vrai risque à utiliser serait $1 - \frac{0.1}{2 \times 31} = 0.9984$ et le seuil critique $t_{0.9984}(25) = 3.539$. On constate que sur les données CONSO (Figure 2.11), aucune observation n'est atypique avec cette procédure.

6. La formule proposée dans Dodge semble erronée (page 135)

Comparaisons multiples et contrôle du risque – II

Si l'on comprend le principe de la correction du risque, multiplier les tests augmente les chances de désigner à tort un point aberrant, il faut donc être plus exigeant, la rectification ci-dessus est purement empirique. Pour dépasser ces problèmes, d'autres auteurs proposent tout simplement de comparer directement le résidu studentisé avec une valeur ad hoc, inspirée néanmoins des seuils fournis par la loi de Student, la valeur la plus utilisée est 2 en référence à un test à 5%. Pour ma part, je pense que le plus simple encore est de trier les observations selon $|t_i^*|$, cela nous donne plus de latitude pour juger de l'ampleur des écarts.

Application sur les données CONSO

Nous complétons le tableau EXCEL en ajoutant la colonne des résidus studentisés. La valeur seuil à 10% est 1.7081. Nous trions les données selon la valeur absolue de cette colonne. Nous constatons que ce sont les mêmes points que précédemment (cf. le résidu standardisé) qui se démarquent ((Mercedes S600, Hyundai Sonata, Ferrari 456 GT et Mitsubishi Galant, figure 2.11).

Modèle	const	Prix	Cylindrée	Puissance	Poids	Consomm	Prédiction	Résidus	0.3226		1.7081	
									Leverage	R.Standard	RSTUDENT	
Mercedes S 600	1	183900	5987	300	2250	18.7	20.074	-1.374	0.4843	2.3416	2.5848	
Hyundai Sonata 3000	1	38990	2972	107	1400	11.7	10.264	1.436	0.2746	2.0632	2.2123	
Ferrari 456 GT	1	285000	5474	325	1690	21.3	20.690	0.610	0.8686	2.0574	2.2049	
Mitsubishi Galant	1	31990	1998	66	1300	7.6	9.168	-1.568	0.1135	2.0375	2.1795	
Opel Astra 1.6i 16V	1	25000	1597	74	1080	7.4	8.510	-1.110	0.0440	1.3896	1.4162	
Fiat Tempra 1.6 Liberty	1	22600	1580	65	1080	9.3	8.245	1.055	0.0413	1.3185	1.3384	
Seat Ibiza 2.0 GTI	1	22500	1983	85	1075	9.5	8.520	0.980	0.1050	1.2672	1.2829	
Opel Omega 2.5i V6	1	47700	2496	125	1670	11.3	12.253	-0.953	0.1278	1.2491	1.2634	
Ford Fiesta 1.2 Zetec	1	19740	1242	55	940	6.6	7.523	-0.923	0.0621	1.1666	1.1751	
Honda Civic Joker 1.4	1	19900	1396	66	1140	7.7	8.557	-0.857	0.0600	1.0810	1.0847	
Volvo 960 Kombi aut	1	49300	2473	125	1570	12.7	11.881	0.819	0.0865	1.0479	1.0500	
Renault Safrane 2.2. V	1	36600	2165	101	1500	11.7	10.885	0.815	0.0773	1.0379	1.0395	
Toyota Previa salon	1	50900	2438	97	1800	12.8	12.189	0.611	0.3154	0.9040	0.9007	
VW Polo 1.4 60	1	17140	1390	44	955	6.5	7.184	-0.684	0.0809	0.8725	0.8684	
Suzuki Swift 1.0 GLS	1	12490	993	39	790	5.8	6.476	-0.676	0.0918	0.8679	0.8636	
Toyota Corolla	1	19490	1331	55	1010	7.1	7.765	-0.665	0.0515	0.8354	0.8304	
Peugeot 306 XS 108	1	22350	1761	74	1100	9	8.457	0.543	0.0487	0.6807	0.6735	
Subaru Vivio 4WD	1	13730	658	32	740	6.8	6.286	0.514	0.1427	0.6794	0.6722	
Seat Alhambra 2.0	1	36400	1984	85	1635	11.6	11.134	0.466	0.2258	0.6487	0.6414	
VW Golt 2.0 GTI	1	31580	1984	85	1155	9.5	9.038	0.462	0.0476	0.5793	0.5717	
Fort Escort 1.4i PT	1	20300	1390	54	1110	8.6	8.143	0.457	0.0581	0.5762	0.5687	
Volvo 850 2.5	1	39800	2435	106	1370	10.8	10.399	0.401	0.0579	0.5049	0.4975	
Lancia K 3.0 LS	1	50800	2958	150	1550	11.9	12.211	-0.311	0.1505	0.4128	0.4062	
Mazda Hachtback V	1	36200	2497	122	1330	10.8	10.528	0.272	0.1233	0.3549	0.3488	
Nissan Primera 2.0	1	26950	1997	92	1240	9.2	9.466	-0.266	0.0506	0.3335	0.3277	
Maserati Ghibli GT	1	92500	2789	209	1485	14.5	14.351	0.149	0.6418	0.3039	0.2985	
Citroen ZX Volcane	1	28750	1998	89	1140	8.8	9.011	-0.211	0.0623	0.2663	0.2615	
Peugeot 806 2.0	1	36950	1998	89	1560	10.8	10.926	-0.126	0.1520	0.1670	0.1638	
Fiat Panda Mambo L	1	10450	899	29	730	6.1	5.982	0.118	0.1131	0.1537	0.1508	
Opel Corsa 1.2i Eco	1	14825	1195	33	895	6.8	6.709	0.091	0.1013	0.1170	0.1148	
Daihatsu Cuore	1	11600	846	32	650	5.7	5.774	-0.074	0.1398	0.0975	0.0956	

Fig. 2.11. Observations triées selon la valeur absolue du résidu studentisé

Dans notre exemple, les deux indicateurs t_i et t_i^* concordent. Ce n'est pas toujours le cas en pratique. Il faut alors privilégier le résidu studentisé pour les raisons évoquées ci-dessus : le fait de considérer l'observation numéro i comme un point supplémentaire permet de mieux appréhender son influence sur la régression.

2.5 Autres indicateurs usuels

Dans cette section, nous énumérons d'autres indicateurs de points atypiques/influents couramment rencontrés dans les logiciels. Nous simplifions la présentation en mettant l'accent sur 3 aspects : le principe, la formule et la règle de détection. Les résultats relatifs au fichier de données CONSO ont été produites à l'aide du logiciel TANAGRA (Figure 2.5).

2.5.1 DFFITS

Le DFFITS s'appuie sur le même principe que le RSTUDENT, mais il compare cette fois-ci la prédiction en resubstitution \hat{y}_i et la prédiction en donnée supplémentaire $\hat{y}_i(-i)$. Dans le premier cas, l'observation a participé à la construction du modèle de prédiction, dans le second, non. Nous pouvons ainsi mesurer l'influence du point sur la régression. Dans notre exemple fictif (Figures 2.9 et 2.10), la différence serait très marquée, confirmant le rôle mystificateur de l'individu Δ .

Le DFFITS est normalisée de la manière suivante

$$DFFITS_i = \frac{\hat{y}_i - \hat{y}_i(-i)}{\hat{\sigma}_\varepsilon(-i)\sqrt{h_i}} \quad (2.9)$$

Nous considérons qu'une observation est influente lorsque

$$R.C. : |DFFITS_i| > 2\sqrt{\frac{p+1}{n}}$$

mais le plus simple toujours est de trier les observations selon $|DFFITS_i|$ pour mettre en évidence les points suspects.

Sur le fichier CONSO, le seuil critique est $2\sqrt{\frac{4+1}{31}} = 0.8032$. Nous constatons que la Ferrari (tout particulièrement), la Mercedes et la Hyundai se démarquent toujours. La Mitsubishi en revanche ne dépasse pas le seuil (0.7800) mais en est suffisamment proche pour qu'on ne remette pas en cause l'analyse proposée dans la section sur le résidu studentisé. On voit là tout l'intérêt de ne pas prendre pour argent comptant les valeurs seuils (Figure 2.12).

Calcul pratique du DFFITS

Il n'est heureusement pas nécessaire d'effectuer les n régressions pour calculer les $DFFITS_i$, on peut l'obtenir à partir du résidu studentisé

$$DFFITS_i = t_i^* \sqrt{\frac{h_i}{1-h_i}} \quad (2.10)$$

2.5.2 Distance de COOK

La distance de COOK généralise le DFFITS dans le sens où, au lieu de mesurer l'effet de la suppression de l'observation i sur la prédiction de y_i , il mesure son effet sur la prédiction des n valeurs de l'endogène.

Observation	Leverage	RStandard	RStudent	DFFITS	0.8032		
					DFFITS	Cook's D	COVRATIO
8 Ferrari 456 GT	0.8686	2.0574	2.2049	5.6685	5.6685	5.5953	3.8078
9 Mercedes S 600	0.4843	-2.3416	-2.5848	-2.5048	2.5048	1.0298	0.7219
22 Hyundai Sonata 3000	0.2746	2.0632	2.2123	1.3611	1.3611	0.3223	0.6861
25 Mitsubishi Galant	0.1135	-2.0375	-2.1795	-0.7800	0.7800	0.1064	0.5751
30 Toyota Previa salon	0.3154	0.9040	0.9007	0.6114	0.6114	0.0753	1.5150
26 Opel Omega 2.5i V6	0.1278	-1.2491	-1.2634	-0.4837	0.4837	0.0457	1.0237
14 Seat Ibiza 2.0 GTI	0.1050	1.2672	1.2829	0.4393	0.4393	0.0377	0.9883
10 Maserati Ghibli GT	0.6418	0.3039	0.2985	0.3996	0.3996	0.0331	3.3365
29 Seat Alhambra 2.0	0.2258	0.6487	0.6414	0.3464	0.3464	0.0245	1.4484
31 Volvo 960 Kombi aut	0.0865	1.0479	1.0500	0.3232	0.3232	0.0208	1.0734
11 Opel Astra 1.6i 16V	0.0440	-1.3896	-1.4162	-0.3037	0.3037	0.0178	0.8652
21 Ford Fiesta 1.2 Zetec	0.0621	-1.1666	-1.1751	-0.3023	0.3023	0.0180	0.9914
13 Renault Safrane 2.2 V	0.0773	1.0379	1.0395	0.3010	0.3010	0.0181	1.0672
17 Fiat Tempra 1.6 Liberty	0.0413	1.3185	1.3384	0.2778	0.2778	0.0150	0.8978
2 Suzuki Swift 1.0 GLS	0.0918	-0.8679	-0.8636	-0.2746	0.2746	0.0152	1.1565
6 Subaru Vivio 4WD	0.1427	0.6794	0.6722	0.2743	0.2743	0.0154	1.2977
19 Honda Civic Joker 1.4	0.0600	-1.0810	-1.0847	-0.2741	0.2741	0.0149	1.0285
4 VW Polo 1.4 60	0.0809	-0.8725	-0.8684	-0.2576	0.2576	0.0134	1.1410
7 Toyota Corolla	0.0515	-0.8354	-0.8304	-0.1935	0.1935	0.0076	1.1196
23 Lancia K 3.0 LS	0.1505	-0.4128	-0.4062	-0.1709	0.1709	0.0060	1.3858
12 Peugeot 306 XS 108	0.0487	0.6807	0.6735	0.1523	0.1523	0.0047	1.1689
18 Ford Escort 1.4i PT	0.0581	0.5762	0.5687	0.1412	0.1412	0.0041	1.2112
24 Mazda Hachback V	0.1233	0.3549	0.3488	0.1308	0.1308	0.0035	1.3545
15 VW Golf 2.0 GTI	0.0476	0.5793	0.5717	0.1278	0.1278	0.0034	1.1971
20 Volvo 850 2.5	0.0579	0.5049	0.4975	0.1234	0.1234	0.0031	1.2294
28 Nissan Primera 2.0	0.0506	-0.3335	-0.3277	-0.0756	0.0756	0.0012	1.2543
27 Peugeot 806 2.0	0.1520	-0.1670	-0.1638	-0.0694	0.0694	0.0010	1.4271
16 Citroen ZX Volcane	0.0623	-0.2663	-0.2615	-0.0674	0.0674	0.0009	1.2799
3 Fiat Panda Mambo L	0.1131	0.1537	0.1508	0.0538	0.0538	0.0006	1.3655
1 Daihatsu Cuore	0.1398	-0.0975	-0.0956	-0.0385	0.0385	0.0003	1.4117
5 Opel Corsa 1.2i Eco	0.1013	0.1170	0.1148	0.0385	0.0385	0.0003	1.3502

Fig. 2.12. Observations triées selon la valeur absolue du $DFFITS$

La première formulation de la distance de Cook D_i est la suivante :

$$D_i = \frac{\sum_{j=1}^n [\hat{y}_i - \hat{y}_i(-i)]^2}{\hat{\sigma}_\varepsilon^2(p+1)} \quad (2.11)$$

Ainsi, pour évaluer l'influence du point i sur la régression, nous la supprimons du calcul des coefficients, et nous comparons les prédictions avec le modèle complet (construit avec tous les points) et le modèle à évaluer (construit sans le point i). Si la différence est élevée, le point joue un rôle important dans l'estimation des coefficients.

Il nous faut définir la valeur seuil à partir de laquelle nous pouvons dire que l'influence est exagérée. La règle la plus simple est :

$$R.C. : D_i > 1 \quad (2.12)$$

Mais elle est jugée un peu trop permissive, laissant échapper à tort des points douteux, on lui préfère parfois la disposition plus exigeante suivante (Confais, page 309) :

$$R.C. : D_i > \frac{4}{n-p-1} \quad (2.13)$$

La distance de Cook a été calculée pour chaque observation du fichier CONSO. Les individus ont été triés selon D_i décroissants. La Ferrari, encore une fois très fortement, et la Mercedes se démarquent selon la première règle de détection (Équation 2.12). Si nous passons à la seconde règle $D_i > \frac{4}{n-p-1} = 0.1538$ (Équation 2.13), la Hyundai se révèle également suspecte (Figure 2.13).

Calcul pratique de la distance de Cook

De nouveau, il n'est pas question d'effectuer les n régressions en supprimant tour à tour chaque observation. Nous pouvons grandement simplifier les calculs en dérivant la distance de Cook à partir des résidus standardisés

0.1538						
Observation	Leverage	RStandard	RStudent	DFFITS	Cook's D	COVRATIO
8 Ferrari 456 GT	0.8686	2.0574	2.2049	5.6685	5.5953	3.8078
9 Mercedes S 600	0.4843	-2.3416	-2.5848	-2.5048	1.0298	0.7219
22 Hyundai Sonata 3000	0.2746	2.0632	2.2123	1.3611	0.3223	0.6861
25 Mitsubishi Galant	0.1135	-2.0375	-2.1795	-0.7800	0.1064	0.5751
30 Toyota Previa salon	0.3154	0.9040	0.9007	0.6114	0.0753	1.5150
26 Opel Omega 2.5i V6	0.1278	-1.2491	-1.2634	-0.4837	0.0457	1.0237
14 Seat Ibiza 2.0 GTI	0.1050	1.2672	1.2829	0.4393	0.0377	0.9883
10 Maserati Ghibli GT	0.6418	0.3039	0.2985	0.3996	0.0331	3.3365
29 Seat Alhambra 2.0	0.2258	0.6487	0.6414	0.3464	0.0245	1.4484
31 Volvo 960 Kombi aut	0.0865	1.0479	1.0500	0.3232	0.0208	1.0734
13 Renault Safrane 2.2. V	0.0773	1.0379	1.0395	0.3010	0.0181	1.0672
21 Ford Fiesta 1.2 Zetec	0.0621	-1.1666	-1.1751	-0.3023	0.0180	0.9914
11 Opel Astra 1.6i 16V	0.0440	-1.3896	-1.4162	-0.3037	0.0178	0.8652
6 Subaru Vivio 4WD	0.1427	0.6794	0.6722	0.2743	0.0154	1.2977
2 Suzuki Swift 1.0 GLS	0.0918	-0.8679	-0.8636	-0.2746	0.0152	1.1565
17 Fiat Tempra 1.6 Liberty	0.0413	1.3185	1.3384	0.2778	0.0150	0.8978
19 Honda Civic Joker 1.4	0.0600	-1.0810	-1.0847	-0.2741	0.0149	1.0285
4 VW Polo 1.4 60	0.0809	-0.8725	-0.8684	-0.2576	0.0134	1.1410
7 Toyota Corolla	0.0515	-0.8354	-0.8304	-0.1935	0.0076	1.1196
23 Lancia K 3.0 LS	0.1505	-0.4128	-0.4062	-0.1709	0.0060	1.3858
12 Peugeot 306 XS 108	0.0487	0.6807	0.6735	0.1523	0.0047	1.1689
18 Ford Escort 1.4i PT	0.0581	0.5762	0.5687	0.1412	0.0041	1.2112
24 Mazda Hachtback V	0.1233	0.3549	0.3488	0.1308	0.0035	1.3545
15 VW Golt 2.0 GTI	0.0476	0.5793	0.5717	0.1278	0.0034	1.1971
20 Volvo 850 2.5	0.0579	0.5049	0.4975	0.1234	0.0031	1.2294
28 Nissan Primera 2.0	0.0506	-0.3335	-0.3277	-0.0756	0.0012	1.2543
27 Peugeot 806 2.0	0.1520	-0.1670	-0.1638	-0.0694	0.0010	1.4271
16 Citroen ZX Volcane	0.0623	-0.2663	-0.2615	-0.0674	0.0009	1.2799
3 Fiat Panda Mambo L	0.1131	0.1537	0.1508	0.0538	0.0006	1.3655
1 Daihatsu Cuore	0.1398	-0.0975	-0.0956	-0.0385	0.0003	1.4117
5 Opel Corsa 1.2i Eco	0.1013	0.1170	0.1148	0.0385	0.0003	1.3502

Fig. 2.13. Observations triées selon la distance de Cook D_i

$$D_i = \frac{t_i^2}{(p+1)} \frac{h_i}{(1-h_i)} \quad (2.14)$$

Distance de Cook entre les coefficients estimés

Nous avons définis la distance de Cook comme un écart entre les prédictions. Il est également possible de la définir comme une distance entre les coefficients estimés, avec ou sans l'observation i à analyser. Dans ce cas, la distance de Cook s'écrit

$$D_i = \frac{(\hat{a} - \hat{a}(-i))'(X'X)^{-1}(\hat{a} - \hat{a}(-i))}{\hat{\sigma}_e^2(p+1)} \quad (2.15)$$

où \hat{a} est le vecteur des $(p+1)$ coefficients estimés $(\hat{a}_0, \hat{a}_1, \dots, \hat{a}_p)'$ avec les n observations; $\hat{a}(-i)$ le même vecteur estimé sans l'observation i .

La distance de Cook s'interprète, dans ce cas, comme l'amplitude de l'écart entre les coefficients estimés de la régression, avec et sans le point i . Il va sans dire que la valeur calculée D_i est exactement la même que celle obtenue avec la première définition (Équation 2.11).

De ce point de vue, la distance de Cook peut se lire comme la statistique du test de comparaison de deux vecteurs de coefficients. Sauf que qu'il ne peut s'agir d'un véritable test puisque les échantillons ne sont pas (pas du tout) indépendants. Néanmoins, si l'on poursuit l'idée, la distance de Cook suivrait une loi de Fisher à $(p+1, n-p-1)$ degrés de liberté. On s'appuie sur la p -value du test pour détecter les points atypiques : on considère qu'un point est suspect dès lors que la p -value calculée est inférieure à 50%⁷. On peut aussi imaginer une procédure plus souple et simplement trier les observations selon la p -value de la distance de Cook. Dans le cas du fichier CONSO, on constate que la Ferrari et la Mercedes se démarquent fortement par rapport aux autres véhicules (Figure 2.14).

7. <http://www-stat.stanford.edu/~jtaylo/courses/stats203/notes/diagnostics.pdf>

Observation	Leverage	RStandard	RStudent	DFBETS	Cook's D	p-value(Cook)
8 Ferrari 456 GT	0.8686	2.0574	2.2049	5.6685	5.5953	0.0013
9 Mercedes S 600	0.4843	-2.3416	-2.5848	-2.5048	1.0298	0.4209
22 Hyundai Sonata 3000	0.2746	2.0632	2.2123	1.3611	0.3223	0.8950
25 Mitsubishi Galant	0.1135	-2.0375	-2.1795	-0.7800	0.1064	0.9899
30 Toyota Previa salon	0.3154	0.9040	0.9007	0.6114	0.0753	0.9955
26 Opel Omega 2.5i V6	0.1278	-1.2491	-1.2634	-0.4837	0.0457	0.9986
14 Seat Ibiza 2.0 GTI	0.1050	1.2672	1.2829	0.4393	0.0377	0.9991
10 Maserati Ghibli GT	0.6418	0.3039	0.2985	0.3996	0.0331	0.9994
29 Seat Alhambra 2.0	0.2258	0.6487	0.6414	0.3464	0.0245	0.9997
31 Volvo 960 Kombi aut	0.0865	1.0479	1.0500	0.3232	0.0208	0.9998
13 Renault Safrane 2.2. V	0.0773	1.0379	1.0395	0.3010	0.0181	0.9999
21 Ford Fiesta 1.2 Zetec	0.0621	-1.1666	-1.1751	-0.3023	0.0180	0.9999
11 Opel Astra 1.6i 16V	0.0440	-1.3896	-1.4162	-0.3037	0.0178	0.9999
6 Subaru Vivio 4WD	0.1427	0.6794	0.6722	0.2743	0.0154	0.9999
2 Suzuki Swift 1.0 GLS	0.0918	-0.8679	-0.8636	-0.2746	0.0152	0.9999
17 Fiat Tempira 1.6 Liberty	0.0413	1.3185	1.3384	0.2778	0.0150	0.9999
19 Honda Civic Joker 1.4	0.0600	-1.0810	-1.0847	-0.2741	0.0149	0.9999
4 VW Polo 1.4 60	0.0809	-0.8725	-0.8684	-0.2576	0.0134	0.9999
7 Toyota Corolla	0.0515	-0.8354	-0.8304	-0.1935	0.0076	1.0000
23 Lancia K 3.0 LS	0.1505	-0.4128	-0.4062	-0.1709	0.0060	1.0000
12 Peugeot 306 XS 108	0.0487	0.6807	0.6735	0.1523	0.0047	1.0000
18 Ford Escort 1.4i PT	0.0581	0.5762	0.5687	0.1412	0.0041	1.0000
24 Mazda Hachtback V	0.1233	0.3549	0.3488	0.1308	0.0035	1.0000
15 VW Golt 2.0 GTI	0.0476	0.5793	0.5717	0.1278	0.0034	1.0000
20 Volvo 850 2.5	0.0579	0.5049	0.4975	0.1234	0.0031	1.0000
28 Nissan Primera 2.0	0.0506	-0.3335	-0.3277	-0.0756	0.0012	1.0000
27 Peugeot 806 2.0	0.1520	-0.1670	-0.1638	-0.0694	0.0010	1.0000
16 Citroen ZX Volcane	0.0623	-0.2663	-0.2615	-0.0674	0.0009	1.0000
3 Fiat Panda Mambo L	0.1131	0.1537	0.1508	0.0538	0.0006	1.0000
1 Daihatsu Cuore	0.1398	-0.0975	-0.0956	-0.0385	0.0003	1.0000
5 Opel Corsa 1.2i Eco	0.1013	0.1170	0.1148	0.0385	0.0003	1.0000

Fig. 2.14. Observations triées selon la p-value de la distance de Cook D_i

2.5.3 DFBETAS

La distance de Cook évalue globalement les disparités entre les coefficients de la régression utilisant ou pas l'observation numéro i . Si l'écart est important, on peut vouloir approfondir l'analyse en essayant d'identifier la variable qui est à l'origine de l'écart : c'est le rôle des DFBETAS.

Pour chaque observation i et pour chaque coefficient a_j , $j = 0, \dots, p$, nous calculons la quantité

$$DFBETAS_{j,i} = \frac{\hat{a}_j - \hat{a}_j(-i)}{\hat{\sigma}_\varepsilon(-i) \sqrt{(X'X)_j^{-1}}} \quad (2.16)$$

où \hat{a}_j est l'estimation du coefficient de la variable X_j (\hat{a}_0 pour la constante); $\hat{a}_j(-i)$ l'estimation du même coefficient lorsqu'on a omis l'observation i ; $\hat{\sigma}_\varepsilon(-i)$ l'estimation de l'écart-type de l'erreur de régression sans l'observation i ; $(X'X)_j^{-1}$ est lue sur la diagonale principale de la matrice $(X'X)^{-1}$.

On considère que l'observation i pèse indûment sur la variable X_j lorsque

$$R.C. : |DFBETAS_{j,i}| > 1 \quad (2.17)$$

Lorsque les observations sont nombreuses, on préférera la règle plus exigeante :

$$R.C. : |DFBETAS_{j,i}| > \frac{2}{\sqrt{n}} \quad (2.18)$$

Bien entendu, il est toujours possible de trier les observations selon les DFBETAS, mais cela peut être rapidement fastidieux lorsque le nombre de variables est élevé.

Appliqué sur les données CONSO, les DFBETAS nous permettent de mieux situer l'action des observations mis en avant par la distance de Cook. On compare les valeurs calculées avec le seuil $\frac{2}{\sqrt{31}} = 0.3592$.

On constate que la Ferrari et la Mercedes pèsent sur quasiment toutes les variables dès lors qu'on les retire ou qu'on les rajoute dans les effectifs pour la régression. La Hyundai, qui semble moins peser globalement (cf. D_i), a aussi une action sur l'ensemble des coefficients mis à part la constante. Enfin, la Maserati, la Mitsubishi et la Toyota Previa agissent de manière anecdotique sur quelques coefficients (Figure 2.15).

Modèle	DFBETAS				
	Intercept	Prix	Cylindrée	Puissance	Poids
Daihatsu Cuore	-0.0361	-0.0033	-0.0017	0.0000	0.0210
Suzuki Swift 1.0 GLS	-0.2353	-0.0343	0.0130	0.0014	0.1084
Fiat Panda Mambo L	0.0455	0.0118	0.0047	-0.0102	-0.0222
VW Polo 1.4 60	-0.1418	-0.0606	-0.1082	0.1393	0.0754
Opel Corsa 1.2i Eco	0.0210	0.0151	0.0121	-0.0226	-0.0075
Subaru Vivio 4WD	0.1934	0.0978	-0.1274	0.0328	-0.0162
Toyota Corolla	-0.1104	-0.0439	0.0311	0.0172	0.0086
Ferrari 456 GT	1.0398	3.4167	-0.5185	-0.8376	-0.3261
Mercedes S 600	0.8261	0.4977	-1.3736	0.3672	0.4475
Maserati Ghibli GT	0.0431	-0.1451	-0.2710	0.3734	0.0049
Opel Astra 1.6i 16V	-0.1770	0.0542	0.0519	-0.0883	0.0682
Peugeot 306 XS 108	0.0808	-0.0582	0.0515	0.0068	-0.0714
Renault Safrane 2.2 V	-0.1474	0.0098	-0.1119	0.0256	0.2056
Seat Ibiza 2.0 GTI	0.2318	-0.2902	0.2307	0.0817	-0.3221
VW Golf 2.0 GTI	0.0592	-0.0444	0.0578	-0.0064	-0.0616
Citroen ZX Volcane	-0.0334	0.0392	-0.0264	-0.0143	0.0403
Fiat Tempra 1.6 Liberty	0.1436	0.0067	0.0275	-0.0373	-0.0485
Fort Escort 1.4i PT	0.0295	0.0637	-0.0294	-0.0455	0.0471
Honda Civic Joker 1.4	-0.0568	-0.0362	0.1620	-0.0719	-0.0954
Volvo 850 2.5	-0.0050	-0.0552	0.0623	-0.0101	-0.0249
Ford Fiesta 1.2 Zetec	-0.2189	-0.0407	0.0701	-0.0304	0.0597
Hyundai Sonata 3000	-0.0042	-0.5261	1.2382	-0.5678	-0.6045
Lancia K 3.0 LS	0.0198	0.1351	-0.0227	-0.0938	0.0387
Mazda Hachback V	0.0222	-0.1092	0.0333	0.0674	-0.0615
Mitsubishi Galant	0.1202	-0.3202	-0.3484	0.6384	-0.1940
Opel Omega 2.5i V6	0.2891	0.0214	0.2247	-0.1193	-0.3439
Peugeot 806 2.0	0.0387	-0.0284	0.0312	0.0124	-0.0613
Nissan Primera 2.0	-0.0171	0.0451	-0.0072	-0.0284	0.0189
Seat Alhambra 2.0	-0.2082	0.1634	-0.1469	-0.0892	0.3176
Toyota Previa salon	-0.4118	0.3243	-0.1109	-0.2977	0.5301
Volvo 960 Kombi aut	-0.1496	-0.0511	-0.1392	0.1143	0.1801

Fig. 2.15. $DFBETAS_{j,i}$ pour le fichier CONSO

Calcul pratique

Encore une fois, il est hors de question d'effectuer n régressions, on s'en sort en utilisant la formule suivante

$$DFBETAS_{j,i} = t_i^* \left[\frac{[(X'X)^{-1}X']_{j,i}}{\sqrt{(X'X)^{-1}_{jj}(1-h_i)}} \right] \quad (2.19)$$

2.5.4 COVRATIO

A la différence de la distance de Cook, au lieu de mesurer la disparité entre les estimations des coefficients, avec ou sans l'intervention de l'observation i , le COVRATIO mesure les disparités entre les précisions des estimateurs c.-à-d. la variance des estimateurs.

A cet effet, il nous faut proposer une mesure de la variance globale des estimateurs, dite *variance généralisée*, elle est égale à

$$\text{var}(\hat{a}) = \hat{\sigma}_\varepsilon^2 \det(X'X)^{-1}$$

où $\det(X'X)^{-1}$ est le déterminant de la matrice $(X'X)^{-1}$.

On formule alors le $COVRATIO_i$ de l'observation i de la manière suivante :

$$COVRATIO_i = \frac{var(\hat{a}(-i))}{var(\hat{a})} \quad (2.20)$$

A première vue :

- Si $COVRATIO_i > 1$, la présence de l'observation i améliore la précision au sens où elle réduit la variance des estimateurs ;
- A l'inverse, si $COVRATIO_i < 1$ indique que la présence de l'observation i dégrade la variance.

Remarque 14. Attention, une diminution de la variance ($COVRATIO > 1$) n'est pas forcément un signe du rôle bénéfique de l'observation i . Une réduction excessive de la variance peut vouloir dire que l'observation pèse exagérément par rapport aux autres observations. Il faut manipuler avec beaucoup de précautions cet indicateur.

A partir de quel moment doit-on s'inquiéter de l'influence d'une observation ? La règle de détection la plus répandue est

$$R.C. : COVRATIO_i < 1 - \frac{3(p+1)}{n} \text{ ou } COVRATIO_i > 1 + \frac{3(p+1)}{n} \quad (2.21)$$

que l'on peut simplifier :

$$R.C. : |COVRATIO_i - 1| > \frac{3(p+1)}{n} \quad (2.22)$$

Le $COVRATIO$ a été calculé pour chaque observation du fichier CONSO. Le tableau est trié selon $|COVRATIO_i - 1|$ décroissant (Figure 2.16). Nous portons notre attention sur la première partie du tableau. Nous retrouvons la Ferrari, la Maserati et la Toyota Previa réapparaissent (cf. levier). Nous notons aussi qu'ils sont suivis d'autres monospaces (Seat Alhambra et Peugeot 806, même s'ils ne sont pas significatifs).

Calcul pratique

Il est possible d'obtenir le $COVRATIO$ à partir du résidu studentisé et du levier

$$COVRATIO_i = \frac{1}{\left[\frac{n-p-2}{n-p-1} + \frac{(t_i^*)^2}{n-p-1} \right]^{(p+1)} (1 - h_i)} \quad (2.23)$$

2.6 Bilan et traitement des données atypiques

Lecture des indicateurs

Trop d'information tue l'information a-t-on coutume de dire. C'est tout à fait vrai dans le cas de ce chapitre. La profusion d'outils peut rapidement donner le tournis. Confais (2006) propose un tableau récapitulatif, on ne peut plus salubre (pages 312 et 313). On discerne le type de lecture que l'on peut faire de chaque indicateur et les conclusions que l'on pourraient en tirer (Figure 2.17).

Observation	Leverage	RStandard	RStudent	< 0.5161 ; > 1.4839 COVRATIO	0.4839 COVRATIO-1
8 Ferrari 456 GT	0.8686	2.0574	2.2049	3.8078	2.808
10 Maserati Ghibli GT	0.6418	0.3039	0.2985	3.3365	2.336
30 Toyota Previa salon	0.3154	0.9040	0.9007	1.5150	0.515
29 Seat Alhambra 2.0	0.2258	0.6487	0.6414	1.4484	0.448
27 Peugeot 806 2.0	0.1520	-0.1670	-0.1638	1.4271	0.427
25 Mitsubishi Galant	0.1135	-2.0375	-2.1795	0.5751	0.425
1 Daihatsu Cuore	0.1398	-0.0975	-0.0956	1.4117	0.412
23 Lancia K 3.0 LS	0.1505	-0.4128	-0.4062	1.3858	0.386
3 Fiat Panda Mambo L	0.1131	0.1537	0.1508	1.3655	0.366
24 Mazda Hachtback V	0.1233	0.3549	0.3488	1.3545	0.354
5 Opel Corsa 1.2i Eco	0.1013	0.1170	0.1148	1.3502	0.350
22 Hyundai Sonata 3000	0.2746	2.0632	2.2123	0.6861	0.314
6 Subaru Vivio 4WD	0.1427	0.6794	0.6722	1.2977	0.298
16 Citroen ZX Volcane	0.0623	-0.2663	-0.2615	1.2799	0.280
9 Mercedes S 600	0.4843	-2.3416	-2.5848	0.7219	0.278
28 Nissan Primera 2.0	0.0506	-0.3335	-0.3277	1.2543	0.254
20 Volvo 850 2.5	0.0579	0.5049	0.4975	1.2294	0.229
18 Fort Escort 1.4i PT	0.0581	0.5762	0.5687	1.2112	0.211
15 VW Golf 2.0 GTI	0.0476	0.5793	0.5717	1.1971	0.197
12 Peugeot 306 XS 108	0.0487	0.6807	0.6735	1.1689	0.169
2 Suzuki Swift 1.0 GLS	0.0918	-0.8679	-0.8636	1.1565	0.157
4 VW Polo 1.4 60	0.0809	-0.8725	-0.8684	1.1410	0.141
11 Opel Astra 1.6i 16V	0.0440	-1.3896	-1.4162	0.8652	0.135
7 Toyota Corolla	0.0515	-0.8354	-0.8304	1.1196	0.120
17 Fiat Tempra 1.6 Liberty	0.0413	1.3185	1.3384	0.8978	0.102
31 Volvo 960 Kombi aut	0.0865	1.0479	1.0500	1.0734	0.073
13 Renault Safrane 2.2 V	0.0773	1.0379	1.0395	1.0672	0.067
19 Honda Civic Joker 1.4	0.0600	-1.0810	-1.0847	1.0285	0.028
26 Opel Omega 2.5i V6	0.1278	-1.2491	-1.2634	1.0237	0.024
14 Seat Ibiza 2.0 GTI	0.1050	1.2672	1.2829	0.9883	0.012
21 Ford Fiesta 1.2 Zetec	0.0621	-1.1666	-1.1751	0.9914	0.009

Fig. 2.16. Observations triées selon le $COVRATIO_i$ *Traitement des observations atypiques*

Reste alors la question délicate du traitement des données atypique, que peut-on faire des observations qui, manifestement, jouent un rôle particulier dans la régression ?

Tous les auteurs s'accordent à dire que la suppression automatique des observations atypiques n'est pas "la" solution. Il faut comprendre pourquoi l'observation se démarque autant et proposer des solutions appropriées :

- Premier réflexe : vérifier les données, y a-t-il des erreurs de saisie ou des erreurs de transcription ? Dans ce cas, il suffit de corriger les valeurs recensées.
- Si la distribution est très asymétrique (ex. salaires), il est plus indiqué de tenter de symétriser la distribution avec une transformation de variables adéquate (ex. log) avant de procéder à nouveau à l'analyse.
- Si l'on manipule des données longitudinales, on introduit une variable muette pour neutraliser l'effet de l'observation atypique (ex. guerre, famine).
- Il apparaît que les observations incriminées ne correspondent pas à la population étudiée (ex. des martiens se sont immiscés dans une enquête). Dans ce cas, et dans ce cas seulement, la suppression est réellement justifiée.

Dans notre exemple CONSO, il apparaît clairement que la Ferrari, voiture sportive d'exception, et la Mercedes, une limousine ultra-luxueuse, n'appartiennent pas au même monde que les autres véhicules de l'analyse. Ils se situent de plus à des niveaux de prix qui les situent définitivement hors de portée. Il paraît donc licite de les supprimer de nos données.

	Std Err Residual	Student Residual	Rstudent	Hat Diag H
signifiant	estimateur de l'erreur-type du résidu i	résidus studentisés internes, appelés standardized residual dans SAS-Insight	résidus studentisés externes, appelés studentized residual dans SAS- Insight	levier de l'obs. i
objet	permet de calculer l'intervalle de confiance autour du résidu i	test de significativité du résidu i	à comparer avec Student Residual écart-type calculé en retirant l'obs. i	mesure l'influence de l'obs.i à cause des valeurs xi
valeurs critiques		2	2	$\frac{2(p+1)}{n}$
Règle de décision		$ Student\ residual > 2$ alors le résidu i est significativement $\neq 0$	$ RStudent > 2$ alors l'observation i nécessite une investigation !	$h_i > \frac{2(p+1)}{n}$ nécessite une investigation
Option de PROC REG	R	R	Influence	Influence
	Cook's D	Df betas	Cov Ratio	Dffits
signifiant	distance de Cook	DFBETAS relatif à chaque coefficient β_j	Ratio de MSE sans et avec l'observation i	statistique DFFITS
objet	mesure le changement en retirant l'obs. i, sur les estimations de l'ensemble des coefficients	mesure normalisée de l'effet de l'obs. i sur l'estimation, pour chaque coefficient β_j	mesure l'effet de l'obs. i sur la précision	mesure normalisée du changement dans la valeur prédite, avec et sans l'obs. i
valeurs critiques	1 ou $\frac{4}{(n-p-1)}$	$\frac{2}{\sqrt{n}}$	$\frac{3(p+1)}{n}$	$2\sqrt{\frac{(p+1)}{n}}$
Règle de décision	CookD > 1 alors l'observation i est influente globalement	$ Dfbetas > \frac{2}{\sqrt{n}}$ indique une influence de l'obs. i sur l'estimation de β_j	$ Covratio - 1 > \frac{3(p+1)}{n}$ nécessite une investigation	$ Dffits > 2\sqrt{\frac{(p+1)}{n}}$ indique une influence de l'obs. i sur \hat{Y}_i
Option de PROC REG	R, Influence	Influence	R	R

Fig. 2.17. Tableau récapitulatif - Détection des observations atypiques (Confais et Le Guen, Modulad, 35, 2006)

Remarque 15 (Techniques graphiques vs. techniques numériques). A ce sujet, prenons toujours de la hauteur par rapport aux techniques numériques, on peut se demander si finalement cet attirail était bien nécessaire dans la mesure où, dès les graphiques des résidus, la Ferrari et la Mercedes étaient systématiquement à l'écart des autres. Elles auront surtout servi à confirmer et préciser le rôle perturbateur de ces 2 observations.

Nous effectuons la régression sur les 29 observations restantes. En étudiant de nouveau les points atypiques, nous constaterons que la Mitsubishi est particulièrement mal modélisée, ce n'est pas étonnant

car elle présente une consommation anormalement basse au regard de ses caractéristiques, sa cylindrée notamment. Nous mettrons également de côté la Maserati qui est un véhicule sportif turbo-compressé à hautes performances.

Remarque 16 (Quand la suppression des observations atypiques devient abusive?). Nous voyons bien là les limites de l'approche consistant à éliminer les observations considérées atypiques. En continuant ainsi, nous finirons par vider le fichier : aucun risque de voir des disparités entre les individus si nous n'avons plus qu'une seule observation.

Global results

Endogenous attribute	Consommation
Examples	27
R ²	0.929520
Adjusted-R ²	0.916706
Sigma error	0.651169
F-Test (4,22)	72.5365 (0.000000)

Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	123.0278	4	30.7570	72.5365	0.0000
Residual	9.3285	22	0.4240		
Total	132.3563	26			

Coefficients

Attribute	Coef.	std	t(22)	p-value
Intercept	1.838006	0.793367	2.316716	0.030220
Prix	0.000034	0.000045	0.752738	0.459587
Cylindrée	0.001208	0.000722	1.672661	0.108557
Puissance	-0.003742	0.015030	-0.248956	0.805704
Poids	0.003728	0.001300	2.868568	0.008926

Fig. 2.18. Résultats de la régression CONSO sans les observations atypiques

Dorénavant, nous utiliserons le fichier des 27 observations, expurgé des 4 véhicules énumérées ci-dessus, pour illustrer les autres thèmes abordés dans ce support (Figure 2.18). Nous obtenons des résultats bien différents avec des graphiques des résidus autrement plus sympathiques (Figure 2.19). La variable prix a disparu des paramètres significatifs. On s'étonne en revanche que ni puissance ni cylindrée ne soient pertinents pour expliquer la consommation. Peut-être faut-il y voir là l'effet de la colinéarité? Nous approfondirons cette question dans le chapitre suivant.

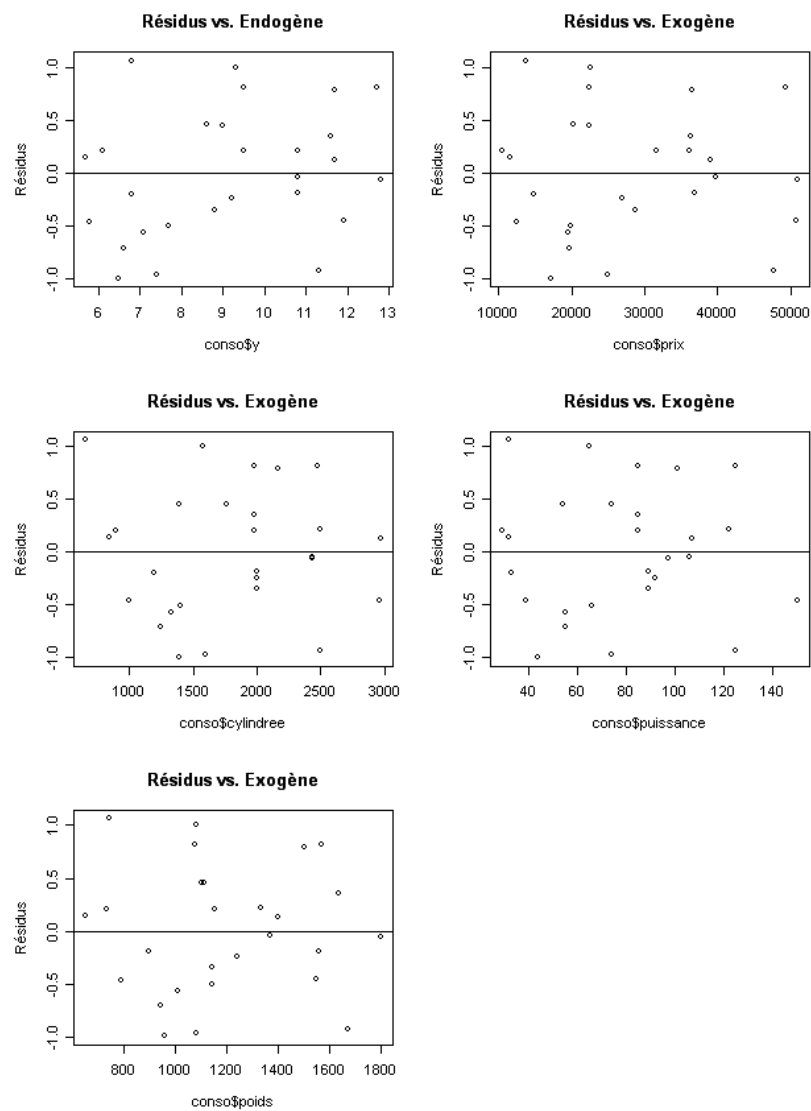


Fig. 2.19. Graphiques des résidus, fichier CONSO après suppression des 4 points atypiques

Colinéarité et sélection de variables

L'un des objectifs de la régression est d'essayer de décrire le processus de causalité entre les exogènes et l'endogène. Pour cela, nous étudions le signe et la valeur des coefficients. L'idée est de circonscrire au possible le rôle de telle ou telle variable dans l'explication des valeurs prises par Y . S'il est établi qu'une variable n'est d'aucune utilité, il est conseillé de l'éliminer, elle perturbe la lecture des résultats.

Les problèmes surgissent lorsqu'il va falloir définir une stratégie de sélection de variables. Peut-on simplement éliminer le bloc de variables qui ne sont pas significatifs au sens du test de Student ? Ce serait négliger l'effet conjoint des variables. Doit-on les éliminer une à une, comment doit-on organiser la suppression ? Est-ce que la suppression séquentielle est la meilleure stratégie, ne peut-on pas envisager une procédure où l'on sélectionne petit à petit les variables intéressantes ou lieu d'éliminer celles qui ne sont pas pertinentes ? etc.

Les procédures de sélection de variables que nous présentons dans ce chapitre répondent à ces questions. Souvent certaines variables exogènes sont redondantes, elles emmènent le même type d'information : c'est le problème de la colinéarité, elles se gênent mutuellement dans la régression.

Dans ce chapitre, nous décrivons quelques techniques simples de détection de la colinéarité. Puis, nous présentons une solution simple pour y remédier par le truchement de la sélection de variables.

3.1 Détection de la colinéarité

3.1.1 Conséquences de la colinéarité

On parle de colinéarité entre 2 variables exogènes lorsque la corrélation linéaire entre ces variables est élevée (ex. $r > 0.8$ a-t-on l'habitude d'indiquer¹ mais ce n'est pas une règle absolue). On peut généraliser cette première définition en définissant la colinéarité comme la corrélation entre une des exogènes avec une combinaison linéaire des autres exogènes.

1. Borcard, D., *Régression Multiple - Corrélation multiple et partielle*, 2001-2007 ; http://bio110.biol.umontreal.ca/BI02042/Regr_mult.pdf

Plusieurs problèmes peuvent surgir² :

- les valeurs/signes des coefficients sont contradictoires, elles ne concordent pas avec les connaissances du domaine ;
- les variances des estimateurs sont exagérées ;
- au point que les coefficients ne paraissent pas significatives (au sens du t de Student du test de nullité des coefficients), poussant le statisticien à les supprimer indûment ;
- les résultats sont très instables, l'adjonction ou la suppression de quelques observations modifie du tout au tout les valeurs et signes des coefficients.

Il y a un vrai risque de passer à côté d'une variable exogène importante tout simplement parce qu'elle est redondante avec une autre. La colinéarité entre variables exogènes rend illusoire la lecture des résultats sur la base des valeurs et de la significativité des coefficients. Il est indiqué de la détecter et de la traiter avant toute interprétation approfondie.

3.1.2 Illustration de l'effet nocif de la colinéarité

Essayons d'illustrer le mécanisme de la colinéarité.

- Si la colinéarité est parfaite, $\text{rang}(X'X) < p + 1 \rightarrow (X'X)^{-1}$ n'existe pas. Le calcul est impossible.
- Si la colinéarité est forte, $\det(X'X) \approx 0$, l'inverse³ $(X'X)^{-1} = \frac{1}{\det(X'X)} \text{com}A'$ contient des valeurs très élevées. Il en est de même pour la matrice de variance covariance des coefficients estimés $\hat{\Omega}_{\hat{a}} = \hat{\sigma}_{\varepsilon}^2 (X'X)^{-1}$. Dès lors, le t de Student $t_{\hat{a}_j} = \frac{\hat{a}_j}{\hat{\sigma}_{\hat{a}_j}}$ pour tester la significativité des coefficients présente mécaniquement de très faibles valeurs. La variable paraît non significative, elle est éliminée par le statisticien.

3.1.3 Quelques techniques de détection

Test de Klein

Il ne s'agit pas d'un test à proprement parler mais plutôt d'un indicateur simple pour détecter rapidement les situations à problèmes (Bourbonnais, pages 100 et 101). Le test de Klein repose sur le principe suivant

1. Nous calculons normalement la régression linéaire multiple $y = a_0 + a_1x_1 + \dots + a_px_p + \varepsilon$, nous recueillons le coefficient de détermination R^2 .
2. Nous calculons les corrélations croisées entre les variables exogènes X_{j1} et $X_{j2} : r_{j1,j2}$ avec $j_1 \neq j_2$.
3. Il y a présomption de colinéarité s'il existe au moins un couple de variables X_{ja}, X_{jb} tel que $R^2 < r_{ja,jb}^2$.

Dans la pratique, une simple proximité entre les valeurs R^2 et $r_{ja,jb}^2$ doit nous alerter.

2. Foucart, T., *Colinéarité et Régression linéaire*, in Mathématiques et Sciences Humaines, Numéro 173, pp. 5-25, 2006 ; <http://www.ehess.fr/revue-msh/pdf/N173R963.pdf>

3. Voir la méthode des cofacteurs, http://fr.wikipedia.org/wiki/Matrice_inversible

Application sur les données CONSO

Dans la régression sur 27 points, rappelons que le coefficient de détermination est $R^2 = 0.9295$ (Figure 2.18). Nous avons calculé les corrélations croisées entre les exogènes, puis leur carré (Figure 3.1). Nous constatons deux situations qui peuvent poser problème : la corrélation entre la puissance et la cylindrée ($r^2 = 0.91$) ; celle entre le poids et le prix ($r^2 = 0.90$)⁴.

Cela peut expliquer notamment pourquoi les variables puissance et cylindrée ne paraissent pas pertinentes pour expliquer la consommation. Ce qui est un non sens si on s'intéresse un tant soit peu aux véhicules automobiles.

Matrice des corrélations croisées				
	prix	cylindree	puissance	poids
prix	1	0.92	0.93	0.95
cylindree	0.92	1	0.96	0.86
puissance	0.93	0.96	1	0.85
poids	0.95	0.86	0.85	1

Matrice des corrélations croisées au carré				
	prix	cylindree	puissance	poids
prix	1	0.84	0.86	0.90
cylindree	0.84	1	0.91	0.74
puissance	0.86	0.91	1	0.73
poids	0.90	0.74	0.73	1

Fig. 3.1. Corrélation croisées et leur carrés. Données CONSO

Test de multicolinéarité - Facteur d'inflation de la variance (VIF)

Le test de Klein ne "détecte" que la colinéarité bivariée. Pour évaluer la multicolinéarité, il faudrait effectuer la régression de chaque exogène X_j avec les $(p - 1)$ autres exogènes, puis étudier le coefficient de détermination R_j^2 associé.

On appelle *facteur d'inflation de la variance (VIF)* la quantité (Saporta, page 422) :

$$v_j = \frac{1}{1 - R_j^2} \quad (3.1)$$

On parle de *facteur d'inflation* car nous avons la relation suivante

$$V(\hat{a}_j) = \frac{\sigma_\varepsilon^2}{n} v_j$$

L'écart-type de l'estimation est multiplié par un facteur $\sqrt{v_j}$.

Plus v_j sera élevé, plus la variance $V(\hat{a}_j)$ de l'estimation sera forte. L'estimation \hat{a}_j sera donc très instable, il aura moins de chances d'être significatif dans le test de nullité du coefficient dans la régression.

A partir de quelle valeur de v_j doit-on s'inquiéter ? Si les variables étaient 2 à 2 indépendantes, $v_j = 1$ et $V(\hat{a}_j) = \frac{\sigma_\varepsilon^2}{n}$. Nous pourrions obtenir les coefficients de la régression multiple à partir de p régressions

4. Les voitures sont vendues au poids maintenant ?

simples. Une règle usuelle de détection de la colinéarité est de prendre un seuil où l'on multiplierait d'un facteur de 2 l'écart-type de l'estimation. On décide qu'il y a un problème de colinéarité lorsque

$$v_j \geq 4$$

Certains utilisent une règle moins contraignante et préfèrent⁵ les seuils 5 ou même 10 c.-à-d. la multicollinéarité n'est signalée que si elle est vraiment élevée. A vrai dire, l'essentiel est d'identifier les variables qui posent problème dans la régression.

Tolérance. La quantité $1 - R_j^2$, appelée *tolérance*, est également fournie par les logiciels statistiques. Plus elle est faible, plus la variable X_j souffre de colinéarité. En dérivant la règle de détection du VIF, on s'inquiéterait dès que la tolérance est inférieure à 0.25.

Calcul pratique du VIF. Calculer p régressions croisées, chaque variable X_j contre les $(p - 1)$ autres pour obtenir les R_j^2 et donc v_j , serait vite fastidieux. Nous pouvons profiter des calculs existants pour produire le VIF. En effet, si C est la matrice des corrélations entre les exogènes, de taille $(p \times p)$, la quantité v_j peut être lue à la coordonnée j de la diagonale principale de la matrice inversée C^{-1} .

Nous en reparlerons plus loin (section 3.6), il est même possible de produire les résultats des régressions croisées à partir des valeurs de la matrice C^{-1} .

Application sur les données CONSO

Nous inversons la matrice de corrélation, nous lisons sur la diagonale principale les VIF. Même avec la règle de détection la plus permissive ($v_j \geq 10$), nous constatons que toutes les variables posent problème (Figure 3.2). Il y a réellement une très forte colinéarité des exogènes dans ce fichier. La variable *prix* en particulier est fortement liée avec les autres variables. Ce qui n'est étonnant finalement. Le prix est un indicateur du niveau de gamme des voitures. On s'attend à ce qu'il soit, un tant soit peu, en relation avec des critères objectifs tels que la puissance ou la cylindrée.

inverse matrice de corrélation				
	PRIX	CYLINDREE	PUISSANC	POIDS
PRIX	19.79	-1.45	-7.51	-11.09
CYLINDREE	-1.45	12.87	-9.80	-1.36
PUISSANC	-7.51	-9.80	14.89	2.86
POIDS	-11.09	-1.36	2.86	10.23

Fig. 3.2. Inverse de la matrice des corrélations - Sur la diagonale principale le VIF

Autres tests statistiques de multicollinéarité

Il existe des tests statistiques plus rigoureux basés sur la matrice des corrélations C : soit à partir du déterminant de la matrice, le test de Farrar et Glauber par exemple (Bournonnais, page 101) ; soit

5. Voir <http://www2.chass.ncsu.edu/garson/PA765/regress.htm>, section **Multicollinearity**, pour une description détaillée des critères et des seuils critiques.

à partir de ses valeurs propres (ex. l'indice de multicollinéarité - <http://www.ehess.fr/revue-msh/pdf/N173R963.pdf> ; voir aussi Saporta, section 17.3.2.2, page 422, sur les relations entre le VIF et les valeurs propres de la matrice C). Ils s'appuient tous sur une démarche similaire, l'hypothèse nulle est l'orthogonalité des variables exogènes, on évalue dans quelle mesure on s'écarte de cette hypothèse.

Sans remettre en doute la pertinence de ces tests, force est de constater que les approches simples suffisent souvent pour apprécier au mieux les multiples situations.

Cohérence des signes

Il existe une autre approche très simple pour détecter la colinéarité, comparer les signes des coefficients de la régression avec le signe des corrélations simples entre les exogènes et l'endogène. La procédure est la suivante :

1. Nous calculons normalement la régression linéaire multiple $y = a_0 + a_1x_1 + \dots + a_px_p + \varepsilon$, nous recueillons les signes des coefficients estimés \hat{a}_j .
2. Nous calculons les corrélations croisées entre chaque variable exogène X_j et l'endogène : r_{y,x_j} .
3. Il y a présomption de colinéarité s'il existe des situations où $\text{signe}(\hat{a}_j) \neq \text{signe}(r_{y,x_j})$. En effet, cela indique que les autres variables perturbent la relation entre Y et X_j .

Application aux données CONSO

Nous calculons les corrélations simples entre chaque exogène et l'endogène. Nous comparons les résultats avec les coefficients de la régression (Figure 3.3). Il y a un conflit pour la variable puissance que nous soupçonnons justement d'être écartée à tort.

	a_j	r_{y,x}
prix	0.000034	0.942597
cylindree	0.001208	0.908790
puissance	-0.003742	0.888304
poids	0.003728	0.944740

Fig. 3.3. Comparaison des corrélations individuelles et des coefficients. Données CONSO

3.2 Traitement de la colinéarité - Sélection de variables

Il existe plusieurs pistes pour traiter la colinéarité. On note principalement la régression ridge qui est une technique de régularisation visant à rendre l'inversion de $(X'X)$ plus stable ; la régression sur les axes principaux de l'analyse en composantes principales, qui sont des variables synthétiques deux à deux linéairement indépendantes produites à partir des exogènes initiales ; la régression PLS (Partial Least Squares) qui impose une contrainte dans la recherche des solutions ; etc.

Dans ce chapitre, nous traiterons plus particulièrement de la sélection de variables. L'objectif est de trouver un sous-ensemble de q variables exogènes ($q \leq p$) qui soient, autant que possible, *pertinentes* et *non-redondantes* pour expliquer l'endogène Y . Deux problèmes se posent alors :

1. quelle est la bonne valeur de q ?
2. comment choisir ces q variables ?

Outre le traitement de la colinéarité, la sélection de variables répond à une autre motivation : la préférence à la simplicité. A pouvoir explicatif sensiblement équivalent, on choisit les modèles parcimonieux pour plusieurs raisons : le modèle est plus lisible, il est plus facile à interpréter ; le nombre de variables à collecter est plus faible ; le modèle est plus robuste, c'est le principe du Rasoir d'Occam.

3.2.1 Sélection par optimisation

Cette approche consiste à produire toutes les combinaisons possibles de variables exogènes, puis de choisir la régression qui maximise un critère de qualité. Le premier écueil est le nombre de cas à évaluer, il est égal à $2^p - 1$, ce qui peut se révéler prohibitif lorsque p est élevé. Il faut donc choisir une stratégie de recherche non-exhaustive mais qui a de bonnes chances de trouver la solution optimale. Il existe un grand nombre de techniques d'exploration dans la littérature (ex. approches gloutonnes, approches best first search, algorithmes génétiques, etc.). Elles se distinguent par leur complexité et leur aptitude à trouver la solution maximisant le critère.

Mais quel critère justement ? C'est ce que nous allons étudier maintenant.

Critère du R^2

Le R^2 semble de prime abord évident. Il exprime la part de la variance expliquée par le modèle. C'est le premier critère que l'on regarde dans une régression. On essaie de trouver la combinaison de variables qui maximise le R^2 .

En réalité, il ne convient pas. En effet, le R^2 augmente de manière mécanique avec le nombre de variables : plus on ajoute de variables, meilleur il est, même si ces variables ne sont absolument pas pertinentes. A la limite, on connaît d'office la solution optimale : c'est le modèle comportant les p variables candidates.

Dans un processus de sélection de modèle, le R^2 conviendrait uniquement pour comparer des solutions comportant le même nombre de variables.

Critère du R^2 corrigé

Le R^2 corrigé, noté \bar{R}^2 , tient compte des degrés de liberté, donc du nombre de variables introduits dans le modèle. Il rend comparable des régressions comportant un nombre d'exogènes différent. Pour bien

comprendre la différence, rappelons la formule du R^2

$$R^2 = 1 - \frac{SCR}{SCT} \quad (3.2)$$

où $SCR = \sum_i (y_i - \hat{y}_i)^2$ est la somme des carrés résiduels, $SCT = \sum_i (y_i - \bar{y})^2$ est la somme des carrés totaux, ceux de l'endogène.

Le \bar{R}^2 introduit une correction par les degrés de liberté, il s'écrit

$$\bar{R}^2 = 1 - \frac{CMR}{CMT} = 1 - \frac{SCR/(n - q - 1)}{SCT/(n - 1)} \quad (3.3)$$

où CMR sont les carrés moyens résiduels, CMT les carrés moyens totaux, q est le nombre de variables dans le modèle évalué.

Il est possible d'exprimer le \bar{R}^2 à partir du R^2

$$\bar{R}^2 = 1 - \frac{n - 1}{n - q - 1} (1 - R^2) \quad (3.4)$$

On voit bien le mécanisme qui se met en place. Deux effets antagonistes s'opposent lorsque l'on ajoute une variable supplémentaire dans le modèle : \bar{R}^2 augmente parce que R^2 s'améliore, \bar{R}^2 diminue parce que le nombre d'exogènes q prend une valeur plus élevée. Tant que la précision du modèle quantifiée par R^2 prend le pas sur la complexité du modèle quantifiée par q , nous pouvons ajouter de nouvelles variables.

Si le principe est sain, on se rend compte dans la pratique que ce critère est trop permissif. L'effet contraignant de q n'est pas assez fort dans la formulation du \bar{R}^2 (Équation 3.4). Le critère favorise les solutions comportant un grand nombre de variables. Il faut trouver des formulations plus restrictives.

Critères AIC et BIC

Ces critères s'appuient sur la même idée : mettre en balance la précision du modèle quantifié par le R^2 (ou le SCR , c'est la même chose puisque SCT est constant quel que soit le modèle à évaluer) avec la complexité du modèle quantifiée par le nombre de variables qu'il comporte.

Avec le critère Akaike (AIC), nous cherchons la régression qui *minimise* la quantité suivante :

$$AIC = n \ln \frac{SCR}{n} + 2(q + 1) \quad (3.5)$$

Avec le critère BIC de Schwartz, nous cherchons à optimiser

$$BIC = n \ln \frac{SCR}{n} + \ln(n)(q + 1) \quad (3.6)$$

Dès que $n > e^2 \approx 7$, on constate que le critère BIC pénalise plus fortement les modèles complexes. Il favorise les solutions comportant peu de variables.

Remarque 17 (Complexité et colinéarité entre les exogènes). Notons que ces techniques de sélection ne tiennent pas compte explicitement de la redondance entre les variables. Cela est fait de manière implicite

avec la pénalisation de la complexité : deux explicatives corrélées n'améliorent guère le SCR mais sont pénalisées parce que la complexité augmente, elles ne peuvent pas être simultanément présentes dans le modèle.

Critère du PRESS

Maximiser le coefficient de détermination R^2 n'est pas approprié. Rappelons que

$$R^2 = 1 - \frac{SCR}{SCT}$$

où SCT , la somme des carrés totaux est constante quelle que soit la régression considérée; SCR est définie de la manière suivante :

$$SCR = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

Lorsque l'on rajoute de nouvelles variables dans le modèle, même non pertinentes, SCR diminue mécaniquement (au pire il reste constant), et par conséquent R^2 augmente. Cela provient du fait que l'on confronte la vraie valeur y_i avec la prédiction \hat{y}_i alors que l'observation i a participé à l'élaboration du modèle. A l'extrême, si on se contente que créer autant de dummy variable qu'il y a d'observations, nous sommes assurés d'obtenir un $R^2 = 1$ puisque nous réalisons une interpolation.

Pour avoir une estimation honnête des performances en prédiction, il ne faudrait pas que l'observation i participe à la construction du modèle lorsqu'on veut prédire sa valeur de l'endogène. Elle intervient ainsi comme une observation supplémentaire⁶. On déduit alors un indicateur similaire au SCR que l'on appelle PRESS (Predicted Residual Sum of Squares)⁷ :

$$PRESS = \sum_{i=1}^n (y_i - \hat{y}_i(-i))^2 \quad (3.7)$$

où $\hat{y}_i(-i)$ est la prédiction de la valeur de l'endogène pour l'observation i utilisée en donnée supplémentaire dans la régression numéro i .

Calcul pratique du PRESS

Tout comme lors du calcul de certains indicateurs lors de la détection des points atypiques, nous ne saurions effectuer réellement n régressions, surtout lorsque les effectifs sont élevés. Encore une fois la matrice H nous sauve la mise, il est possible de calculer le PRESS à partir de la seule régression sur l'ensemble des observations en utilisant la relation suivante

$$y_i - \hat{y}_i(-i) = \frac{y_i - \hat{y}_i}{1 - h_i} \quad (3.8)$$

6. Cela n'est pas sans rappeler la distinction que nous faisons entre les résidus standardisés et studentisés dans la détection des points atypiques.

7. http://www.ltrr.arizona.edu/~dmeko/notes_12.pdf

Procédure de sélection basée sur le PRESS

A la différence du R^2 , nous disposons d'un critère *honnête* d'évaluation des performances en prédiction. Il est possible dès lors de définir une stratégie de sélection de variables uniquement basé sur ce critère de performances, sans tenir compte explicitement de la complexité du modèle. En effet, dans la pratique, on se rend compte que si l'on rajoute des variables non-pertinentes, sans pouvoir explicatif, le R^2 peut s'améliorer (fallacieusement), le PRESS lui en revanche se dégrade, indiquant par là l'inutilité de la variable.

Remarque 18 (Wrapper). Notons pour l'anecdote que ce type de stratégie de sélection de variables dans le domaine de l'apprentissage automatique (grosso modo, il s'agit de problèmes de prédiction où la variable à prédire est qualitative) est connu sous le terme générique *wrapper*. Sauf, qu'à ma connaissance, les procédures construisent explicitement les n modèles de prédiction (moins si on décide d'exclure non pas une seule mais k observations à chaque phase de construction de modèle)⁸.

Application : calcul du PRESS sur les données CONSO

Calculons le PRESS à partir des coefficients de la régression estimées sur les 27 observations (Figure 2.18). Nous procédons par étapes (Figure 3.4) :

MODÈLE	PRIX	YLINDREE	VISSANC	POIDS	MATION	Y-chapeau	e	h	e/(1-h)	PRESS
Daihatsu Cuore	11600.00	846.00	32.00	650.00	5.70	5.56	0.14	0.22	0.18	0.03
Suzuki Swift 1.0 GL	12490.00	993.00	39.00	790.00	5.80	6.26	-0.46	0.11	-0.52	0.27
Fiat Panda Mambo L	10450.00	899.00	29.00	730.00	6.10	5.89	0.21	0.14	0.24	0.06
VW Polo 1.4 60	17140.00	1390.00	44.00	955.00	6.50	7.49	-0.99	0.13	-1.15	1.31
Opel Corsa 1.2i Eco	14825.00	1195.00	33.00	895.00	6.80	7.00	-0.20	0.17	-0.24	0.06
Subaru Vivio 4WD	13730.00	658.00	32.00	740.00	6.80	5.74	1.06	0.29	1.49	2.21
Toyota Corolla	19490.00	1331.00	55.00	1010.00	7.10	7.67	-0.57	0.06	-0.60	0.36
Opel Astra 1.6i 16V	25000.00	1597.00	74.00	1080.00	7.40	8.36	-0.96	0.06	-1.03	1.05
Peugeot 306 XS 108	22350.00	1761.00	74.00	1100.00	9.00	8.55	0.45	0.09	0.50	0.25
Renault Safrane 2.2	36600.00	2165.00	101.00	1500.00	11.70	10.91	0.79	0.12	0.89	0.80
Seat Ibiza 2.0 GTI	22500.00	1983.00	85.00	1075.00	9.50	8.69	0.81	0.19	1.00	1.01
VW Golf 2.0 GTI	31580.00	1984.00	85.00	1155.00	9.50	9.29	0.21	0.10	0.23	0.05
Citroen ZX Volcane	28750.00	1998.00	89.00	1140.00	8.80	9.14	-0.34	0.07	-0.37	0.14
Fiat Tempra 1.6 Lib	22600.00	1580.00	65.00	1080.00	9.30	8.30	1.00	0.05	1.05	1.11
Fort Escort 1.4i PT	20300.00	1390.00	54.00	1110.00	8.60	8.14	0.46	0.09	0.51	0.26
Honda Civic Joker 1	19900.00	1396.00	66.00	1140.00	7.70	8.20	-0.50	0.20	-0.63	0.40
Volvo 850 2.5	39800.00	2435.00	106.00	1370.00	10.80	10.84	-0.04	0.12	-0.05	0.00
Ford Fiesta 1.2 Zet	19740.00	1242.00	55.00	940.00	6.60	7.31	-0.71	0.09	-0.77	0.60
Hyundai Sonata 3000	38990.00	2972.00	107.00	1400.00	11.70	11.57	0.13	0.58	0.31	0.09
Lancia K 3.0 LS	50800.00	2958.00	150.00	1550.00	11.90	12.35	-0.45	0.33	-0.68	0.46
Mazda Hachback V	36200.00	2497.00	122.00	1330.00	10.80	10.58	0.22	0.21	0.27	0.07
Opel Omega 2.5i V6	47700.00	2496.00	125.00	1670.00	11.30	12.23	-0.93	0.18	-1.14	1.30
Peugeot 806 2.0	36950.00	1998.00	89.00	1560.00	10.80	10.99	-0.19	0.17	-0.23	0.05
Nissan Primera 2.0	26950.00	1997.00	92.00	1240.00	9.20	9.44	-0.24	0.16	-0.29	0.08
Seat Alhambra 2.0	36400.00	1984.00	85.00	1635.00	11.60	11.25	0.35	0.30	0.51	0.26
Toyota Previa salon	50900.00	2438.00	97.00	1800.00	12.80	12.86	-0.06	0.50	-0.12	0.01
Volvo 960 Kombi aut	49300.00	2473.00	125.00	1570.00	12.70	11.88	0.82	0.27	1.12	1.25
PRESS										13.54
SCR										9.33

Fig. 3.4. Calcul du PRESS sur les données CONSO - Régression à 4 explicatives

8. Kohavi, R., John, G., *Wrappers for Feature Subset Selection*, in Artificial Intelligence, (97)1-2, P. 273-324, 1997 – <http://citeseer.ist.psu.edu/cache/papers/cs/124/http://zSzzSzrobotics.stanford.edu/zSzzronnykzSzwappers.pdf/kohavi97wrappers.pdf>

1. Nous utilisons les coefficients de la régression pour calculer la prédiction en resubstitution \hat{y}_i ;
2. Nous formons alors l'erreur de prédiction $\hat{\varepsilon}_i = y_i - \hat{y}_i$;
3. Nous calculons les éléments diagonaux de la *Hat Matrix*, qui sont ni plus ni moins que les leviers (leverage) $h_i = [X(X'X)^{-1}X']_{ii}$;
4. Nous formons l'erreur de prédiction *en donnée supplémentaire* $y_i - \hat{y}_i(-i) = \frac{\hat{\varepsilon}_i}{1-h_i}$;
5. Nous en déduisons le $PRESS = \sum_{i=1}^n [y_i - \hat{y}_i(-i)]^2 = 13.54$.

Notons pour rappel que $SCR = 9.33$ (Figure 2.18), nous avons systématiquement la relation $SCR \leq PRESS$. Plus l'écart entre ces deux indicateurs est élevé, plus nous suspectons un **sur-apprentissage** c.-à-d. le modèle "colle" trop aux données, il intègre des spécificités du fichier et ne restitue plus la vraie relation qui existe dans la population.

Calcul du PRESS sur les données CONSO - Modèle simplifié

A titre de comparaison, nous avons calculé le PRESS du modèle n'utilisant que CYLINDRÉE et POIDS comme explicatives. A priori le modèle est de moins bonne qualité puisque le $R^2 = 0.92768$ et la $SCR = 9.57211$ sont moins avantageux que ceux de la régression à 4 variables (PRIX, CYLINDRÉE, PUISSANCE, POIDS) avec respectivement $R^2 = 0.92952$ et $SCR = 9.3285$. Et pourtant, le modèle simplifié s'avère **plus performant en prédiction** avec $PRESS = 11.694$ (Figure 3.5), contre $PRESS = 13.54$ précédemment.

Cela montre combien la recherche d'un modèle parcimonieux est tout sauf une élucubration esthétique. Elle permet d'améliorer (souvent) l'efficacité des modèles lors du déploiement dans la population. Les prédictions sont plus précises.

Sélection de variables sur les données CONSO - Critère AIC

Nous allons essayer de trouver le modèle optimal qui minimise le critère AIC. Nous adoptons une démarche *backward*. Elle consiste, à partir du modèle complet comportant toutes les variables, à éliminer unes à unes les variables qui permettent de diminuer l'AIC, et de continuer ainsi tant que la suppression d'une variable améliore le critère.

Voici le détail de la procédure :

1. calculer l'AIC pour le modèle comportant l'ensemble courant de variables ;
2. évaluer l'AIC consécutive à la suppression de chaque variable du modèle, choisir la suppression entraînant la plus forte diminution et vérifier qu'elle propose une amélioration du critère par rapport à la situation précédente ;
3. si NON, arrêt de l'algorithme ; si OUI, retour en (1).

MODÈLE	CYLINDREE	POIDS	CONSO	Y^	e	h	e/(1-h)	PRESS
Daihatsu Cuore	846.00	650.00	5.70	5.43	0.27	0.152	0.319	0.102
Suzuki Swift 1.0 GL	993.00	790.00	5.80	6.25	-0.45	0.105	-0.506	0.256
Fiat Panda Mambo L	899.00	730.00	6.10	5.86	0.24	0.124	0.275	0.076
VW Polo 1.4 60	1390.00	955.00	6.50	7.52	-1.02	0.059	-1.080	1.167
Opel Corsa 1.2i Eco	1195.00	895.00	6.80	6.99	-0.19	0.075	-0.206	0.042
Subaru Vivio 4WD	658.00	740.00	6.80	5.59	1.21	0.164	1.449	2.100
Toyota Corolla	1331.00	1010.00	7.10	7.69	-0.59	0.059	-0.624	0.389
Opel Astra 1.6i 16V	1597.00	1080.00	7.40	8.35	-0.95	0.042	-0.993	0.986
Peugeot 306 XS 108	1761.00	1100.00	9.00	8.66	0.34	0.046	0.360	0.130
Renault Safrane 2.2	2165.00	1500.00	11.70	10.99	0.71	0.084	0.778	0.605
Seat Ibiza 2.0 GTI	1983.00	1075.00	9.50	8.83	0.67	0.098	0.737	0.544
VW Golf 2.0 GTI	1984.00	1155.00	9.50	9.20	0.30	0.060	0.323	0.104
Citroen ZX Volcane	1998.00	1140.00	8.80	9.15	-0.35	0.069	-0.373	0.139
Fiat Tempra 1.6 Lib	1580.00	1080.00	9.30	8.33	0.97	0.042	1.014	1.028
Fort Escort 1.4i PT	1390.00	1110.00	8.60	8.21	0.39	0.066	0.412	0.170
Honda Civic Joker 1	1396.00	1140.00	7.70	8.36	-0.66	0.074	-0.711	0.505
Volvo 850 2.5	2435.00	1370.00	10.80	10.76	0.04	0.088	0.048	0.002
Ford Fiesta 1.2 Zet	1242.00	940.00	6.60	7.26	-0.66	0.067	-0.702	0.493
Hyundai Sonata 3000	2972.00	1400.00	11.70	11.60	0.10	0.296	0.149	0.022
Lancia K 3.0 LS	2958.00	1550.00	11.90	12.25	-0.35	0.193	-0.437	0.191
Mazda Hachtback V	2497.00	1330.00	10.80	10.66	0.14	0.122	0.163	0.026
Opel Omega 2.5i V6	2496.00	1670.00	11.30	12.19	-0.89	0.132	-1.023	1.046
Peugeot 806 2.0	1998.00	1560.00	10.80	11.04	-0.24	0.162	-0.285	0.081
Nissan Primera 2.0	1997.00	1240.00	9.20	9.60	-0.40	0.043	-0.414	0.171
Seat Alhambra 2.0	1984.00	1635.00	11.60	11.36	0.24	0.240	0.318	0.101
Toyota Previa salon	2438.00	1800.00	12.80	12.70	0.10	0.245	0.136	0.019
Volvo 960 Kombi aut	2473.00	1570.00	12.70	11.71	0.99	0.092	1.094	1.198
								PRESS 11.694
DTROITEREG								
POIDS CYLINDREE CONST								
a^	0.00450	0.00131	1.39228					
	0.00078	0.00038	0.49688					
R^2	0.92768	0.63154	#N/A					
	153.92749	24	#N/A					
	122.78419	9.57211	#N/A					

Fig. 3.5. Calcul du PRESS sur les données CONSO - Régression à 2 explicatives (CYLINDRÉE, POIDS)

Appliqué sur le fichier CONSO de 27 observations, nous obtenons la séquence de calculs⁹ :

Étape	Modèle courant (cte = constante)	AIC	Suppression d'une variable (AIC)
1	y = prix + cylindrée + puissance + poids + cte	-18.69	puissance → -20.6188 prix → -20.0081 cylindrée → -17.4625 poids → -12.1155
2	y = prix + cylindrée + poids + cte	-20.6188	prix → -21.9986 cylindrée → -17.6000 poids → -13.3381
3	y = cylindrée + poids + cte	-21.9986	cylindrée → -13.3049 poids → -0.2785

Au départ, étape 1, avec toutes les variables, $AIC = -18.69 = 27 \ln \frac{9.328}{27} + 2(4 + 1)$. La suppression de la variable *puissance* entraîne la plus grande diminution du critère, il passe alors à -20.6188, etc. A l'étape 3, on constate qu'aucune suppression de variable n'améliore le modèle courant.

9. Nous avons utilisé la fonction **stepAIC** du package MASS du logiciel R

Le modèle optimal au sens du critère AIC est

$$y = 1.392276 + 0.01311 \times \text{cylindree} + 0.004505 \times \text{poids}$$

Remarque 19 (Recherche forward). Si nous avons adopté une recherche *forward* c.-à-d. partir du modèle composé de la seule constante, ajouter au fur et à mesure une variable de manière à diminuer au possible le critère AIC, nous aurions obtenu le même ensemble final de variables exogènes.

3.2.2 Techniques basées sur le F partiel de Fisher

Les techniques présentées dans cette section s'appuient sur le F partiel de Fisher. Grosso modo, on ajoute une variable si le carré du t de Student (qui suit une loi de Fisher) indique que le coefficient associé est significativement différent de 0 ; on supprime une variable si son coefficient n'est pas significatif (Tenenhaus, pages 100 à 108).

Sélection par avant - Forward Selection

Comme son nom l'indique, il s'agit d'une technique incrémentale qui consiste à repérer à chaque étape la variable proposant un t de Student le plus élevé en valeur absolue (ou dont le carré est le plus élevé), de l'ajouter dans le pool courant si le coefficient est significatif, et de continuer ainsi tant que les ajouts sont possibles.

On commence par p régressions simples. Si une variable a été ajoutée, on poursuit avec $p-1$ régressions à 2 variables, etc. L'ajout d'une variable dépend de la significativité du coefficient de la variable choisie, il dépend donc du risque α défini par l'utilisateur. Si on souhaite obtenir peu de variables, on fixe un risque faible.

Il faut être prudent par rapport à ce risque. En effet, la variable à tester est celle qui maximise le $F = t^2$. Nous sommes en situation de comparaisons multiples. La loi sous l'hypothèse nulle est modifiée. On n'est pas sûr de prendre réellement un risque α d'accepter à tort une variable. Pour éviter cet aspect trompeur, certains logiciels proposent de fixer directement une valeur seuil de F pour accepter ou rejeter la meilleure variable à chaque étape. Cela peut paraître arbitraire, d'autant que les valeurs par défaut correspondent peu ou prou à des niveaux de risques usuels (ex. Dans STATISTICA, le seuil de 3.84 proposé est à peu près le fractile de la loi de Fisher à 5%). Mais au moins, le statisticien évitera de faire référence explicitement à un niveau de risque erroné.

D'autres logiciels tels que SPSS offrent les deux possibilités à l'utilisateur : il peut fixer un risque critique ou directement un seuil critique. L'essentiel étant de bien comprendre ce que l'on est en train de manipuler.

Enfin, le principal reproche que l'on peut adresser à cette approche est qu'une variable choisie à une étape n'est plus jamais remise en cause par la suite.

Application sur les données CONSO

Nous avons appliqué ce processus de sélection aux données CONSO avec 27 observations. Nous avons choisi un risque de 5%, avec bien entendu toutes les réserves d'usages ci-dessus. Le processus de sélection est résumé dans le tableau 3.1.

Étape	Modèle courant (cte = constante)	R^2	$t_{\hat{a}_j}^2 = F$ (p-value)
1	$y = \text{cte}$	-	<p>poids $\rightarrow 207.63$ (0.0000)</p> <p>prix $\rightarrow 199.19$ (0.0000)</p> <p>cylindrée $\rightarrow 118.60$ (0.0000)</p> <p>puissance $\rightarrow 93.53$ (0.0000)</p>
2	$y = \text{poids} + \text{cte}$	0.8925	<p>cylindrée $\rightarrow 11.66$ (0.0023)</p> <p>puissance $\rightarrow 7.42$ (0.0118)</p> <p>prix $\rightarrow 6.32$ (0.0190)</p>
2	$y = \text{poids} + \text{cylindrée} + \text{cte}$	0.9277	<p>prix $\rightarrow 0.53$ (0.4721)</p> <p>puissance $\rightarrow 0.01$ (0.9288)</p>

Tableau 3.1. Sélection forward basé sur le t^2 - Données CONSO

Parmi les 4 régressions simples, c'est la variable *poids* qui présente un $t^2 = F = 207.63$ le plus élevé, elle est très significative, en tous les cas avec un p-value largement en-deçà du niveau que l'on s'est fixé (5%). La variable *poids* est donc intégrée. À l'étape 2, nous essayons de voir quelle est la variable qu'on pourrait lui adjoindre. Nous effectuons 3 régressions à 2 variables (*poids* et une autre) : *cylindrée* se révèle être la plus intéressante, avec un $F = 11.66$, elle est significative à 5%. Elle est intégrée. À l'étape 3, nous avons 2 régressions à 3 variables (*poids*, *cylindrée* et une autre) à tester. Nous constatons que la variable la plus intéressante, *prix* avec un $F = 0.53$, n'est plus significative (pvalue > 5%). On s'en tient donc au modèle à 2 variables : *poids* et *cylindrée*.

Dans le fichier CONSO, l'optimisation du AIC et la sélection forward basé sur le F donnent des résultats identiques. Ce n'est pas toujours vrai dans la pratique.

Élimination en arrière - Backward Selection

Cette procédure fonctionne à l'inverse de la précédente. Elle commence avec la régression comportant toutes les exogènes, regarde quelle est la variable la moins pertinente au sens du t de Student (le carré du t de Student le plus faible), élimine la variable si elle n'est pas significative au risque α . Elle recommence avec les variables restantes. Le processus est interrompu lorsqu'il n'est plus possible de supprimer une variable.

Si l'on met de côté les réserves d'usages par rapport au vrai sens à donner au risque des tests successifs, on fixe généralement un risque α plus élevé pour la suppression : la possibilité de retenir une variable est

favorisée par rapport à celle d'en ajouter. Notamment parce que la colinéarité peut masquer le rôle de certaines d'entre elles¹⁰. La valeur $\alpha = 10\%$ est proposée par défaut dans la logiciel SPSS par exemple. La plupart des logiciels procèdent ainsi.

Application sur les données CONSO

Nous appliquons la procédure au fichier CONSO, voici le détail des calculs :

Étape	Modèle courant (cte = constante)	R^2	Évaluation $t^2 = F$ (pvalue)
1	$y = \text{prix} + \text{cylindrée} + \text{puissance} + \text{poids} + \text{cte}$	0.9295	<p>puissance $\rightarrow 0.0620$ (0.8057)</p> <p>prix $\rightarrow 0.5666$ (0.4596)</p> <p>cylindrée $\rightarrow 2.7978$ (0.1086)</p> <p>poids $\rightarrow 8.2287$ (0.0089)</p>
2	$y = \text{prix} + \text{cylindrée} + \text{poids} + \text{cte}$	0.9293	<p>prix $\rightarrow 0.5344$ (0.4721)</p> <p>cylindrée $\rightarrow 4.6779$ (0.0412)</p> <p>poids $\rightarrow 9.4345$ (0.0054)</p>
3	$y = \text{cylindrée} + \text{poids} + \text{cte}$	0.9277	<p>cylindrée $\rightarrow 11.6631$ (0.0023)</p> <p>poids $\rightarrow 33.7761$ (0.0000)</p>

Le modèle complet à 4 variables propose un $R^2 = 0.9295$. La variable la moins intéressante est *puissance* avec un $t^2 = 0.0620$, elle n'est pas significative à 10% (pvalue = 0.8057). Nous pouvons la retirer. Le modèle suivant, à 3 exogènes, propose un $R^2 = 0.9293$. La variable la moins pertinente est *prix* qui n'est pas non plus significative, elle est également éliminée. La régression à 2 exogènes, *cylindrée* et *poids*, possède des variables qui sont toutes significatives à 10% : c'est notre modèle définitif avec un $R^2 = 0.9277$.

On note que le R^2 diminue mécaniquement à mesure que nous supprimons des variables. Mais la dégradation est minime au regard du gain en simplicité obtenu en réduisant le nombre de variables du modèle.

Procédure stepwise - Stepwise regression

Cette procédure est un *mix* des approches *forward* et *backward*. A la première étape, on commence par construire le meilleur modèle à 1 exogène. Par la suite, à chaque étape, on regarde si l'ajout d'une variable ne provoque pas le retrait d'une autre. Cela est possible lorsqu'une variable exogène expulse une autre variable qui lui est corrélée, et qui semblait pourtant plus significative dans les étapes précédentes.

10. Merci à Matthieu Buisine pour m'avoir indiqué les incohérences de la version précédente de ce document. Avec un seuil plus élevé, on a tendance à plus retenir les variables et non l'inverse. Merci Matthieu. C'est avec ce type de commentaires qu'on peut faire avancer les choses.

Généralement, on fixe un risque plus exigeant pour la sélection (ex. 5%, on ne fait entrer la meilleure variable que si elle est significative à 5%) que pour la suppression (ex. 10%, on supprime la variable la moins pertinente si elle est non significative à 10%).

Application sur les données CONSO

Appliqué sur les données CONSO avec le logiciel SPSS, cette technique nous renvoie le modèle à 2 variables

$$y = 1.392276 + 0.01311 \times \text{cylindree} + 0.004505 \times \text{poids}$$

3.3 Régression stagewise

La régression *stagewise* est une procédure *forward* qui consiste à ajouter, au fur et à mesure, une variable qui explique au mieux la fraction de Y non-expliquée par les variables déjà sélectionnées (Bourbonnais, page 105 ; Dodge¹¹, page 161 à 164).

On peut résumer l'approche de la manière suivante :

1. On sélectionne la variable X_a qui est la plus corrélée, en valeur absolue, avec Y . On la sélectionne si la corrélation est significativement différent de 0 au risque α . Nous utilisons un test de Student à $(n - 2)$ degrés de liberté

$$t_a = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}.$$

Comme il s'agit de tester un coefficient qui a fait l'objet d'une optimisation préalable, le vrai risque du test n'est pas α . Mais dans la pratique, il ne faut pas attacher trop d'importance à un calcul prétendument pointu du vrai risque qui, de toute manière, dépend de la préférence à la simplicité de l'utilisateur : on diminue α si on veut moins de variables dans le modèle, on l'augmente si on en veut plus. C'est plus en ce sens qu'il faut lire la valeur de α .

2. On veut choisir la variable X_b qui est la plus corrélée avec la fraction de Y non-expliquée par X_a . Pour ce faire, on calcule le résidu de la régression

$$e_1 = y - (\hat{a}_0 + \hat{a}_1 x_a)$$

La variable X_b est celle qui est la plus corrélée avec e_1 . On l'intègre dans le modèle si la corrélation est significativement différent de 0 au risque α . Attention, les degrés de liberté sont modifiés $(n - 3)$, il en est de même pour la statistique du test¹². On utilise

11. La description donnée par Dodge est un peu différente. La méthode Stagewise est utilisée pour sélectionner les variables, et les coefficients de la régression finale sont déduits des calculs intermédiaires. Il distingue donc les paramètres fournis par stagewise des paramètres estimés à l'aide de la MCO.

12. Lorsque les effectifs sont élevés, cette correction a peu d'effet

$$t_b = \frac{r}{\sqrt{\frac{1-r^2}{n-3}}}.$$

3. Si la variable X_b est intégrée, nous cherchons la variable suivante X_c qui explique au mieux la fraction de Y non-expliquée conjointement par X_a et X_b . Le plus simple toujours est de prendre le résidu

$$e_2 = y - (\hat{b}_0 + \hat{b}_1 x_a + \hat{b}_2 x_b)$$

de choisir la variable qui lui le plus corrélé, et de tester la significativité du coefficient de corrélation avec un t_c de Student à $(n - 4)$ degrés de liberté

$$t_c = \frac{r}{\sqrt{\frac{1-r^2}{n-4}}}.$$

4. on continue ainsi jusqu'à ce qu'aucun ajout de variable ne soit possible.
5. Au final, le plus simple est de re-estimer la droite de régression avec les variables sélectionnées.

Application sur les données CONSO

Nous appliquons la régression stagewise sur les données CONSO. Nous détaillons les calculs :

1. Nous calculons les corrélations brutes entre Y et les exogènes r_{Y,X_j} . Nous obtenons le tableau suivant :

X_j	r
poids	0.9447
prix	0.9426
cylindrée	0.9088
puissance	0.8883

La variable la plus corrélée avec l'endogène est *poids* : $r = 0.9447$

2. Vérifions si la corrélation est significativement différente de 0. Pour ce faire, nous formons la statistique de Student $t = \frac{0.9447}{\sqrt{\frac{1-0.9447^2}{27-2}}} = 14.4094$ et calculons la p-value associée $p\text{-value} = 0.0000$. La corrélation est significativement supérieure à zéro en valeur absolue, elle est acceptée.
3. Pour choisir la variable suivante, nous procédons en deux temps : (a) nous calculons les coefficients de la régression $y = 1.0353 + 0.0068 \times \text{poids}$; (b) nous calculons le résidu $e_1 = y - (1.0353 + 0.0068 \times \text{poids})$.
4. Nous calculons les corrélations r_{e_1, X_j} pour déterminer la variable la plus corrélée avec e_1

X_j	r
cylindrée	0.2908
puissance	0.2544
prix	0.1471
poids	0.0000

Bien évidemment, la corrélation $r_{e_1, \text{poids}} = 0$ puisque e_1 est la fraction de Y qui n'est pas expliquée par *poids*.

5. La variable la plus intéressante est *cylindrée*, nous formons le t de Student $t = \frac{0.2908}{\sqrt{\frac{1-0.2908^2}{27-3}}} = 1.4891$, avec une p -value égale à 0.1495.
6. Au risque de 5%, la variable *cylindrée* n'est significativement corrélée avec e_1 . Le processus de sélection de variables est stoppée.

Au final, le "meilleur" modèle d'explication de la consommation selon la procédure stagewise intègre uniquement la variable *poids* :

$$y = 1.0353 + 0.0068 \times \text{poids}$$

3.4 Coefficient de corrélation partielle et sélection de variables

3.4.1 Coefficient de corrélation brute

Le coefficient de corrélation¹³ quantifie le degré de liaison **linéaire** entre deux variables continues Y et X . Elle est définie par

$$\rho_{y,x} = \frac{\text{cov}(y,x)}{\sigma_y \cdot \sigma_x} \quad (3.9)$$

C'est une mesure symétrique. Par définition $-1 \leq \rho \leq +1$, $\rho > 0$ (resp. $\rho < 0$) si la liaison est positive (resp. négative). Lorsque les variables sont indépendantes, $\rho = 0$, l'inverse n'est pas vrai.

Le coefficient de corrélation empirique est l'estimation de ρ sur un fichier de n observations :

$$r_{y,x} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \cdot \sqrt{\sum_i (y_i - \bar{y})^2}} \quad (3.10)$$

On parle de corrélation brute parce que l'on mesure directement la liaison entre Y et X sans qu'aucune autre variable n'intervienne. Nous l'opposerons à la corrélation partielle exposée plus bas.

Pour vérifier que la corrélation entre deux variables est significativement différent de zéro, nous posons le test d'hypothèses

$$H_0 : \rho = 0$$

$$H_1 : \rho \neq 0$$

La statistique du test s'écrit

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

La région critique du test au risque α , rejet de H_0 , est définie par

$$R.C. : |t| > t_{1-\frac{\alpha}{2}}(n-2)$$

où $t_{1-\frac{\alpha}{2}}(n-2)$ est le fractile d'ordre $1 - \frac{\alpha}{2}$ de la loi de Student à $(n-2)$ degrés de liberté.

13. <http://en.wikipedia.org/wiki/Correlation>

Quelques exemples sur les données CONSO

Prenons quelques variables du fichier CONSO et calculons le coefficient de corrélation linéaire (Tableau 3.2).

variable 1	variable 2	r	t	p-value
y	puissance	0.8883	9.6711	0.0000
y	cylindrée	0.9088	10.8901	0.0000
puissance	cylindrée	0.9559	16.2700	0.0000

Tableau 3.2. Corrélation entre quelques variables du fichier CONSO

Nous constatons que toutes ces corrélations sont élevées et très significativement différentes de zéro.

3.4.2 Coefficient de corrélation partielle

Mesurer la corrélation partielle

Corrélation n'est pas causalité a-t-on coutume de dire : ce n'est pas parce que 2 variables varient de manière concomitante, dans le même sens ou en sens opposé, qu'il faut y voir forcément une relation de cause à effet.

Parfois, la corrélation peut être totalement fortuite, il s'agit simplement d'un artefact statistique auquel on ne peut donner aucune interprétation valable. Parfois aussi, et c'est le cas qui nous intéresse ici, elle est due à une tierce variable qui joue le rôle d'intermédiaire entre les 2 variables étudiées.

Exemple 2. Ventes de lunettes de soleil et ventes de glaces : aucune des deux n'a un effet sur l'autre, il s'agit plutôt de la température qui les fait varier dans le même sens.

Exemple 3. La corrélation entre la taille des personnes et la longueur de leurs cheveux est négative. Avant d'y voir un quelconque phénomène de compensation, on se rend compte qu'il y a 2 populations dans le fichier : les hommes et les femmes (Figure 3.6). En général, les hommes sont plus grands et ont les cheveux plus courts. La variable "sexe" est la variable intermédiaire qui fait apparaître une relation factice entre la taille et la longueur des cheveux.

L'idée de la *corrélation partielle* justement est de mesurer le degré de liaison entre 2 variables en neutralisant (en contrôlant) les effets d'une troisième variable. Il peut y avoir plusieurs types d'effets (Figure 3.7 ; le texte en ligne qui accompagne ce schéma est très instructif - <http://www2.chass.ncsu.edu/garson/pa765/partialr.htm>).

Pour calculer la corrélation partielle, nous utilisons les corrélations brutes

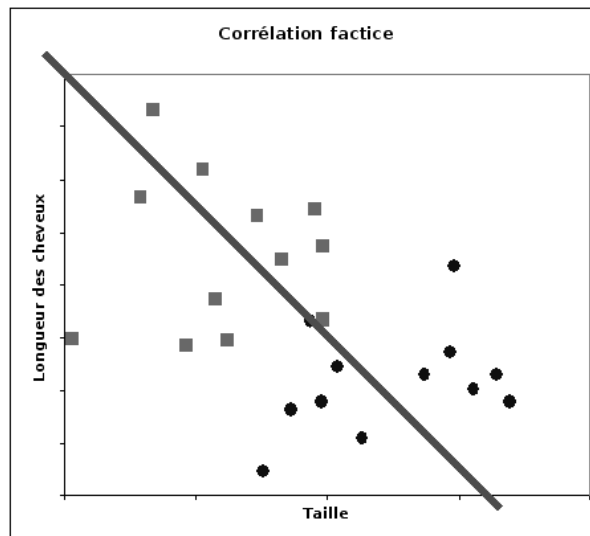


Fig. 3.6. La corrélation est la conséquence de la présence de 2 populations distinctes dans le fichier

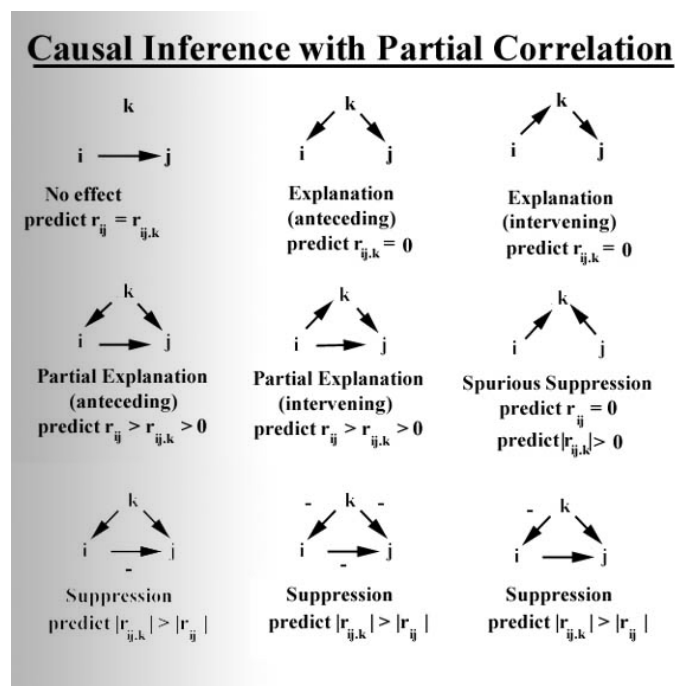


Fig. 3.7. Différentes interactions dans la mesure de la corrélation partielle

$$r_{y,x/z} = \frac{r_{y,x} - r_{y,z}r_{x,z}}{\sqrt{1 - r_{y,z}^2} \cdot \sqrt{1 - r_{x,z}^2}} \quad (3.11)$$

L'idée sous-jacente est simple : on retranche de la liaison brute mesurée entre y et x , l'effet induit par z .

Tester la corrélation partielle

Pour vérifier la significativité d'une corrélation partielle, nous adoptons la même démarche que pour la corrélation brute. Les hypothèses à tester sont :

$$H_0 : \rho_{y,x/z} = 0$$

$$H_1 : \rho_{y,x/z} \neq 0$$

La statistique du test s'écrit :

$$t = \frac{r_{y,x/z}}{\sqrt{\frac{1-r_{y,x/z}^2}{n-3}}}$$

Et la région critique du test est définie par :

$$R.C. : |t| > t_{1-\frac{\alpha}{2}}(n-3)$$

où $t_{1-\frac{\alpha}{2}}(n-3)$ est le fractile d'ordre $1 - \frac{\alpha}{2}$ de la loi de Student à $(n-3)$ degrés de liberté. Il faut faire attention au degré de liberté, il y a bien 3 paramètres estimés dans la statistique étudiée.

Exemple sur les données CONSO

Nous voulons mesurer les relations entre la consommation et la puissance, en contrôlant l'effet de la cylindrée (la taille du moteur). Nous appliquons directement la formule ci-dessus (Équation 3.11) en utilisant les corrélations brutes calculées précédemment (Tableau 3.2) :

$$r_{y,\text{puissance}/\text{cylindrée}} = \frac{0.8883 - 0.9088 \cdot 0.9559}{\sqrt{1 - 0.9088^2} \cdot \sqrt{1 - 0.9559^2}} = 0.1600$$

Pour tester la nullité du coefficient, nous formons la statistique

$$t = \frac{0.1600}{\sqrt{\frac{1-0.1600^2}{27-3}}} = 0.7940$$

Le t calculé est 0.7940, avec une p-value de 0.4350.

Au risque de 5% (et bien au-delà), on ne constate pas de liaison significative entre *consommation* (y) et *puissance*, une fois retranchée l'explication apportée par la *cylindrée*.

Autre lecture : à cylindrée égale, la consommation ne varie pas avec la puissance.

3.4.3 Calcul de la corrélation partielle d'ordre supérieur à 1

Nous savons maintenant calculer la corrélation partielle d'ordre 1. Comment faire pour calculer les corrélations partielles d'ordre supérieur ? c.-à-d. mesurer la liaison entre y et x en contrôlant l'effet induit par d'autres (z_1, z_2, \dots) variables.

Il existe une formule de passage qui permet de généraliser la première expression (Équation 3.11). Mais elle devient difficile à manipuler à mesure que le nombre de variables z_j augmente, d'autant plus qu'elle impose de calculer de proche en proche toutes les corrélations croisées. Il est plus aisé d'utiliser une autre formulation de la corrélation partielle.

Pour calculer la corrélation partielle $r_{y,x/z_1,z_2}$, nous procédons par étapes :

1. nous enlevons de y toute l'information acheminée par z_1 et z_2 en calculant le résidu de la régression

$$e_1 = y - (\hat{a}_0 + \hat{a}_1 z_1 + \hat{a}_2 z_2)$$

2. nous procédons de même pour la variable x

$$e_2 = x - (\hat{b}_0 + \hat{b}_1 z_1 + \hat{b}_2 z_2)$$

3. la corrélation partielle peut être obtenue par la corrélation brute entre les 2 résidus

$$r_{y,x/z_1,z_2} = r_{e_1,e_2}$$

4. et nous pouvons tester la nullité du coefficient en formant la statistique

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-4}}}$$

5. qui suit une loi de Student à $(n-4)$ degrés de liberté.

De manière générale, lorsque nous avons k variables de contrôle z_j , pour tester :

$$H_0 : \rho_{y,x/z_1,\dots,z_k} = 0$$

$$H_1 : \rho_{y,x/z_1,\dots,z_k} \neq 0$$

Nous calculons la corrélation r entre les résidus

$$e_1 = y - (\hat{a}_0 + \hat{a}_1 z_1 + \hat{a}_k z_k)$$

$$e_2 = x - (\hat{b}_0 + \hat{b}_1 z_1 + \hat{b}_k z_k)$$

Et la statistique du test s'écrit

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-k-2}}}$$

Elle suit une loi de Student à $(n-k-2)$ degrés de liberté.

Exemple sur les données CONSO

Nous voulons calculer et tester la corrélation partielle $r_{y,puissance/cylindree,poids}$. Nous procédons selon les étapes ci-dessus :

1. former le résidu $e_1 = y - (1.3923 + 0.0045 \cdot poids + 0.0013 \cdot cylindree)$;
2. idem, former $e_2 = puissance - (-15.8347 + 0.0117 \cdot poids + 0.0444 \cdot cylindree)$
3. calculer alors la corrélation $r = r_{e_1,e_2} = 0.0188$;
4. la statistique du test $t = \frac{0.0188}{\sqrt{\frac{1-0.0188^2}{27-2-2}}} = 0.0903$;
5. et la p-value = 0.9288.

En conclusion, la liaison entre la *consommation* (y) et la *puissance* est nulle (au risque de 5%) dès lors que l'on retranche l'effet induit par les variables *poids* et *cylindrée*.

Il est intéressant d'ailleurs de récapituler le lien entre la consommation (y) et la puissance à mesure que l'on fait intervenir d'autres variables (Tableau 3.3).

Corrélation	r	t	p-value
$r_{y,puissance}$	0.8883	9.6711	0.0000
$r_{y,puissance/cylindree}$	0.1600	0.7940	0.4350
$r_{y,puissance/cylindree,poids}$	0.0188	0.0903	0.9288

Tableau 3.3. Corrélations partielles entre CONSO (Y) et PUISSANCE

3.4.4 Procédure de sélection fondée sur la corrélation partielle

La notion de corrélation partielle s'accorde bien avec la sélection de variables de type *forward* : on veut mesurer la relation d'une variable candidate avec l'endogène sachant les valeurs prises par les variables déjà choisies ; ou encore, on veut mesurer l'information additionnelle apportée par une variable supplémentaire dans l'explication des valeurs prises par l'endogène.

L'enchaînement des opérations serait :

1. détecter la variable exogène X_a la plus corrélée (**en valeur absolue**) avec l'endogène, la sélectionner si la liaison est significative ;
2. détecter la seconde variable X_b exogène qui maximise la corrélation partielle $r_{y,X_b/X_a}$, on l'introduit dans le modèle si elle est significativement différente de zéro ;
3. à l'étape q , il s'agit de calculer la corrélation partielle d'ordre $q - 1$ pour sélectionner ou pas la q -ème variable.

La règle d'arrêt est simplement une corrélation partielle non-significative de la meilleure variable à une étape donnée.

Exemple sur les données CONSO

Forward Selection Process

partial corr. F (p-value)	Step 1	Step 2	Step 3
d.f.	25	24	23
$r(Y, X_j^*/X_{j1}, X_{j2}, \dots)$	Poids : 0.9447	Cylindrée : 0.5719	-
R^2	0.8925	0.9277	-
Prix	0.9426 199.19 (0.0000)	0.4567 6.32 (0.0190)	0.1507 0.53 (0.4721)
Cylindrée	0.9088 118.60 (0.0000)	0.5719 11.66 (0.0023)	-
Puissance	0.8883 93.53 (0.0000)	0.4859 7.42 (0.0118)	0.0188 0.01 (0.9288)
Poids	0.9447 207.63 (0.0000)	-	-

Fig. 3.8. Sélection de variables fondée sur la corrélation partielle - Données CONSO

Appliquée sur les données CONSO, le modèle choisi comporte les exogènes *poids* et *cylindrée* (Figure 3.8). Détaillons ces résultats :

1. A la première étape, la variable la plus corrélée avec l'endogène est *poids* avec $r = 0.9447$ et $t^2 = F = 207.63$. La liaison est très significative $p - value < 0.0001$. Elle est donc intégrée dans le modèle dont le coefficient de détermination serait $R^2 = 0.8925$.
2. La variable la plus corrélée avec l'endogène, conditionnellement à *poids*, est *cylindrée* avec $r_{y,cylindree/poids} = 0.5719$ et $t^2 = F = 11.66$. La liaison est significative, $p - value = 0.0023$. Nous sélectionnons donc cette seconde variable, le coefficient de détermination du modèle $y = a_0 + a_1poids + a_2cylindree$ est $R^2 = 0.9277$.
3. La variable la plus corrélée avec l'endogène, conditionnellement à *poids* et *cylindrée*, est *prix* avec $r = 0.1507$ et $t^2 = F = 0.53$. La liaison n'est plus significative à 5% puisque la $p - value = 0.4721$. Nous stoppons la procédure de sélection.
4. Au final, le modèle définitif comprend les variables *poids* et *cylindrée*.

3.4.5 Équivalence avec la sélection fondée sur le t de Student de la régression

Les valeurs des $t^2 = F$ manipulées dans le processus de sélection basé sur la corrélation partielle (Figure 3.8) ne sont pas sans rappeler celles de la régression forward basée sur le F -partiel (Tableau 3.1). Ce n'est absolument pas fortuit.

En effet, dans un modèle à q variables explicatives, il y a une relation directe entre la corrélation partielle d'ordre $(q - 1)$, $r_{y,x_q/x_1,\dots,x_{q-1}}$, et le t de Student du test de nullité du q -ème coefficient $t_{\hat{a}_q}$ dans une régression à q exogènes (Bourbonnais, page 93) :

$$r_{y,x_q/x_1,\dots,x_{q-1}}^2 = \frac{t_{\hat{a}_q}^2}{t_{\hat{a}_q}^2 + (n - q - 1)} \quad (3.12)$$

Ainsi, tester la nullité du coefficient de X_q dans la régression à q variables équivaut à tester la nullité du coefficient de corrélation partielle d'ordre $(q-1)$. Il est tout à fait normal que l'on retrouve exactement les mêmes tests, avec les mêmes degrés de liberté, à chaque étape du processus de sélection.

De même, nous comprenons mieux maintenant pourquoi nous faisons référence à un *F-partiel* dans le processus de sélection forward basé sur le *t* de Student des coefficients de régression (Section 3.2.2).

3.5 Les régressions partielles

3.5.1 Principe des régression partielles

La régression partielle permet d'évaluer graphiquement l'apport d'une variable additionnelle dans une régression où $(p-1)$ variables explicatives sont déjà présentes. La procédure permet de produire un "nuage de points", le **graphique des régressions partielles**, directement lié à la notion de corrélation partielle. Il permet également d'identifier les observations atypiques et/ou influentes de la régression.

Pour fixer les idées, mettons que l'on souhaite évaluer l'influence de X_p dans la régression

$$Y = a_0 + a_1X_1 + \dots + a_{p-1}X_{p-1} + a_pX_p + \varepsilon$$

Après estimation des paramètres \hat{a}_j , nous pouvons produire les résidus $\hat{\varepsilon}$ de cette régression.

Le graphique de la régression partielle pour la variable X_p est construit de la manière suivante¹⁴ :

1. Nous réalisons la régression de Y sur les $(p-1)$ explicatives

$$Y = b_0 + b_1X_1 + \dots + b_{p-1}X_{p-1} + \varepsilon_Y$$

Avec les coefficients estimés, nous calculons les résidus de la régression $\hat{\varepsilon}_Y$.

2. Nous expliquons maintenant X_p à l'aide toujours des $(p-1)$ explicatives

$$X_p = c_0 + c_1X_1 + \dots + c_{p-1}X_{p-1} + \varepsilon_{X_p}$$

Nous en déduisons les résidus $\hat{\varepsilon}_{X_p}$.

3. Le graphique de la régression partielle pour X_p est le nuage de points $(\hat{\varepsilon}_{X_p}, \hat{\varepsilon}_Y)$ c.-à-d. avec $\hat{\varepsilon}_{X_p}$ en abscisse et $\hat{\varepsilon}_Y$ en ordonnée.
4. Le coefficient de corrélation linéaire calculé sur les résidus $(\hat{\varepsilon}_{X_p}, \hat{\varepsilon}_Y)$ nous fournit le coefficient de corrélation partielle entre Y et X_p . Cette approche est très pratique pour calculer les corrélations partielles d'ordre supérieur à 1 (section 3.4.3).

14. <http://www.itl.nist.gov/div898/software/dataplot/refman1/auxillar/partregr.htm>; et, ouh là il y a du copier-coller dans l'air, http://en.wikipedia.org/wiki/Partial_regression_plot

5. A partir de ce nuage de points, nous pouvons calculer la régression

$$\hat{\varepsilon}_Y = d \times \hat{\varepsilon}_{X_j} + e + \varepsilon_r$$

Et en tirer les résidus $\hat{\varepsilon}_r$.

Le graphique des résidus $\hat{\varepsilon}_r$ cumule des propriétés très intéressantes¹⁵ :

1. Nous constatons que $\hat{e} = 0$, c'est tout à fait normal puisque les variables intervenant dans la régression sont centrées c.-à-d. $\bar{\hat{\varepsilon}}_Y = \bar{\hat{\varepsilon}}_{X_p} = 0$.
2. Nous constatons surtout que $\hat{d} = \hat{a}_p$. Nous retrouvons le coefficient du modèle complet incluant X_p .
3. Le résidu $\hat{\varepsilon}_r$ est identique au résidu du modèle complet \hat{e} c.-à-d. $\hat{\varepsilon}_{i,r} = \hat{\varepsilon}_i$, $\forall i = 1, \dots, n$.
4. Les observations influentes sont facilement identifiables dans ce graphique.
5. Nous pouvons également détecter rapidement les autres situations pathologiques (ex. hétéroscédasticité, groupes d'observations atypiques, non linéarité...).

Dans le cadre de l'évaluation influence de la variable X_p dans la régression, nous pouvons lire le graphique sous l'angle suivant (Cornillon et Matzner-Lober¹⁶, page 96) :

1. Si le nuage de points ne présente pas de "forme particulière", s'il est globalement horizontal (de pente nulle), la variable X_p n'apporte pas d'information supplémentaire pertinente par rapport aux variables déjà présentes.
2. Si les points forment une droite de pente non nulle, X_p influe linéairement dans la régression.
3. S'ils forment une courbe non linéaire, il est judicieux de remplacer X_p par une de ses transformées (ex. en appliquant le logarithme, en passant au carré, etc.).

Cette dernière idée n'est pas sans rappeler la notion de "résidus partiels" développée par ailleurs (section 6.2.2). Mais, à la différence de celle-ci, qui est un outil dédié à la détection de la forme de liaison la plus appropriée entre X_p et Y en présence des $(p - 1)$ autres explicatives, le nuage de points des régressions partielles, notamment parce que les valeurs de X_p n'apparaissent pas explicitement dans le graphique, ne donne pas d'indications sur la fonction à utiliser pour transformer X_p et linéariser la relation. **Pour le traitement de la non-linéarité, il est préférable de passer par les résidus partiels.**

3.5.2 Traitement des données CONSO

Nous souhaitons évaluer la contribution de puissance (X_p) dans l'explication de la consommation (Y), sachant que les variables cylindrée et poids ont déjà été sélectionnées. Nous avons une série de régression à construire (Figure 3.9) :

15. http://en.wikipedia.org/wiki/Partial_regression_plot

16. Cornillon, P-A., Matzner-Lober, E., *Régression - Théorie et applications.*, Springer, 2007.

MODÈLE	PUISSANCE	CYLINDREE	POIDS	CONSO	RES.Y	RES.X	RES.R	RES
Daihatsu Cuore	32	846	650	5.7	0.271	2.665	0.267	0.267
Suzuki Swift 1.0 GL	39	993	790	5.8	-0.453	1.500	-0.455	-0.455
Fiat Panda Mambo L	29	899	730	6.1	0.241	-3.624	0.245	0.245
VW Polo 1.4 60	44	1390	955	6.5	-1.017	-13.062	-1.001	-1.001
Opel Corsa 1.2i Eco	33	1195	895	6.8	-0.191	-14.699	-0.173	-0.173
Subaru Vivio 4WD	32	658	740	6.8	1.212	9.966	1.200	1.200
Toyota Corolla	55	1331	1010	7.1	-0.587	-0.083	-0.587	-0.587
Opel Astra 1.6i 16V	74	1597	1080	7.4	-0.951	6.283	-0.959	-0.959
Peugeot 306 XS 108	74	1761	1100	9.0	0.344	-1.236	0.345	0.345
Renault Safrane 2.2	101	2165	1500	11.7	0.712	3.147	0.708	0.708
Seat Ibiza 2.0 GTI	85	1983	1075	9.5	0.665	0.194	0.665	0.665
VW Golf 2.0 GTI	85	1984	1155	9.5	0.304	-0.785	0.305	0.305
Citroen ZX Volcane	89	1998	1140	8.8	-0.347	2.769	-0.350	-0.350
Fiat Tempra 1.6 Lib	65	1580	1080	9.3	0.971	-1.962	0.974	0.974
Fort Escort 1.4i PT	54	1390	1110	8.6	0.385	-4.872	0.391	0.391
Honda Civic Joker 1	66	1396	1140	7.7	-0.658	6.512	-0.666	-0.666
Volvo 850 2.5	106	2435	1370	10.8	0.044	-2.330	0.047	0.047
Ford Fiesta 1.2 Zet	55	1242	940	6.6	-0.655	4.687	-0.661	-0.661
Hyundai Sonata 3000	107	2972	1400	11.7	0.105	-25.535	0.136	0.136
Lancia K 3.0 LS	150	2958	1550	11.9	-0.353	16.336	-0.372	-0.372
Mazda Hachtback V	122	2497	1330	10.8	0.143	11.383	0.129	0.129
Opel Omega 2.5i V6	125	2496	1670	11.3	-0.887	10.459	-0.900	-0.900
Peugeot 806 2.0	89	1998	1560	10.8	-0.239	-2.134	-0.236	-0.236
Nissan Primera 2.0	92	1997	1240	9.2	-0.396	4.646	-0.402	-0.402
Seat Alhambra 2.0	85	1984	1635	11.6	0.241	-6.388	0.249	0.249
Toyota Previa salon	97	2438	1800	12.8	0.103	-16.482	0.123	0.123
Volvo 960 Kombi aut	125	2473	1570	12.7	0.993	12.648	0.978	0.978

1 - CONSO = f (cylindrée, poids)

poids	cylindrée	constante
0.00450	0.00131	1.39228

2 - PUISSANCE = f (cylindrée, poids)

poids	cylindrée	constante
0.01167	0.04442	-15.83469

3 - RES.Y = f (RES.X)

RES.X	CONST
0.00121	0.00000

4 - CONS = f (cylindrée, poids, puissance)

poids	cylindrée	puissance	constante
0.00449	0.00126	0.00121	1.41143

5 - Corrélation (RES.Y, RES.X)

r	0.01884
r ²	0.00035

Fig. 3.9. Régression partielle pour la variable *puissance*- Données CONSO

1. Nous régressons la consommation sur cylindrée et poids, nous obtenons

$$conso = 0.00450 \times poids + 0.00131 \times cylindree + 1.39228$$

Nous en déduisons la colonne des résidus RES.Y ($\hat{\epsilon}_{conso}$).

2. Nous construisons la régression pour puissance

$$puissance = 0.01167 \times poids + 0.04442 \times cylindree - 15.83469$$

Nous construisons également la colonne des résidus RES.X ($\hat{\epsilon}_{puissance}$).

Nous pouvons former le graphique de la régression partielle pour la variable *puissance* (Figure 3.10). Sans s'avancer outre mesure, le nuage de points ne présente pas une pente particulière. L'explication additionnelle de la puissance sur la consommation par rapport à la cylindrée et le poids n'est pas décisive. Notons cependant une certaine dissymétrie de la distribution sur l'axe des abscisses ($\hat{\epsilon}_{puissance}$). Elle est en grande partie due à un point atypique, la *Hyundai Sonata 3000* qui est singulièrement peu puissante (107 ch) au regard de sa cylindrée (2972 cm³). C'est aussi le cas, dans une moindre mesure cependant, de la *Toyota Previa Salon*.

3. Nous calculons la régression sur les résidus

$$\hat{\epsilon}_{conso} = 0.00121 \times \hat{\epsilon}_{puissance} + 0.0000$$

Effectivement, la constante de la régression est nulle. Quant à la pente $\hat{d} = 0.00121$, conformément au graphique, elle est très proche de 0, confirmant l'idée d'une faible influence additionnelle de puissance dans la régression. Le coefficient de détermination est $R^2 = 0.00035$.

Nous formons les résidus $\hat{\epsilon}_r$ (RES.R).

4. Voyons maintenant ce qu'il en est de la régression incluant toutes les explicatives, nous avons

$$conso = 0.00449 \times poids + 0.00126 \times cylindree + 0.00121 \times puissance + 1.41143$$

- a) Premier résultat qui saute aux yeux, nous constatons bien l'égalité entre les coefficients $\hat{a}_p = \hat{d} = 0.00121$.
 - b) Autre résultat important, en calculant les résidus $\hat{\varepsilon}$ (RES) de cette régression, nous retrouvons exactement les valeurs de $\hat{\varepsilon}_r$ (RES.R).
5. Enfin, dernière information importante, en calculant la corrélation entre $\hat{\varepsilon}_Y$ et $\hat{\varepsilon}_{puissance}$, nous retrouvons effectivement sur la corrélation partielle obtenues par ailleurs (Tableau 3.3), soit

$$r_{\hat{\varepsilon}_{conso}, \hat{\varepsilon}_{puissance}} = r_{conso, puissance / cylindree, poids} = 0.01884$$

Bien évidemment, en passant cette corrélation au carré, nous retrouvons le coefficient de détermination de la régression de $\hat{\varepsilon}_{conso}$ sur $\hat{\varepsilon}_{puissance}$: $r^2 = (0.01884)^2 = 0.00035$.

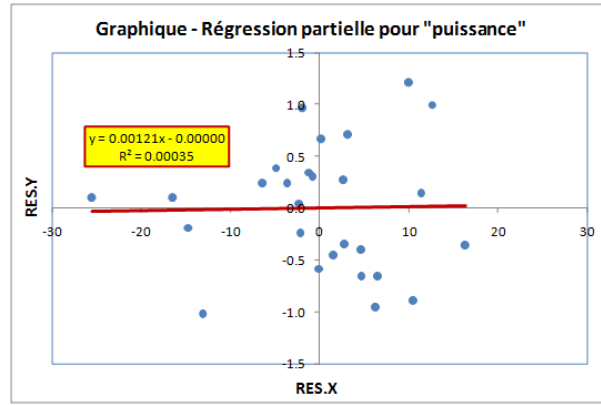


Fig. 3.10. Graphique de la régression partielle pour la variable *puissance*- Données CONSO

3.6 Régressions croisées

3.6.1 Principe des régressions croisées

Nous avons mis en avant le critère VIF (variance inflation factor - section 3.1.3) pour mesurer la multicolinéarité entre les exogènes c.-à-d. la liaison de chaque X_j avec une combinaison linéaire des $(p-1)$ autres explicatives. Dans son principe, le critère v_j est basé sur le coefficient de détermination R_j^2 de la régression de X_j sur les autres. On parle de **régressions croisées**. Dans les faits, nous obtenons directement la valeur de v_j sur la diagonale de l'inverse de la matrice des corrélations C^{-1} .

Dans cette section, nous approfondissons cette idée. Nous montrons qu'il est possible de reconstituer les résultats des régressions croisées à partir de la matrice C^{-1} .

Notons v_{kj} les éléments de la matrice C^{-1} , qui est de dimension $(p \times p)$; $v_{jj} = v_j$ est le VIF de la variable X_j lue sur la diagonale de C^{-1} . Nous nous intéressons à la régression

$$X_l = a_0 + a_1 X_1 + \cdots + a_{l-1} X_{l-1} + a_{l+1} X_{l+1} + \cdots + a_p X_p + \varepsilon_l$$

Coefficient de détermination. Comme le VIF v_l peut être exprimé en fonction du coefficient de détermination R_l^2 de cette régression (équation 3.1), l'inverse est aussi vrai, soit :

$$R_l^2 = 1 - \frac{1}{v_l} \quad (3.13)$$

Test de significativité globale. Il y a $(p - 1)$ explicatives dans la régression, les degrés de liberté doivent être ajustés en conséquence. Pour tester la significativité globale de la régression, nous formons la statistique F_l

$$F_l = \frac{R_l^2 / (p - 1)}{(1 - R_l^2) / (n - (p - 1) - 1)} = \frac{R_l^2 / (p - 1)}{(1 - R_l^2) / (n - p)} \quad (3.14)$$

Sous H_0 , tous les coefficients de la régression sont nuls (hormis la constante), F_l suit une loi de Fisher à $(p - 1, n - p)$ degrés de liberté.

Estimation de la variance de l'erreur $\hat{\sigma}_{\varepsilon_l}^2$. La variance de l'erreur de régression, autre indicateur de qualité de l'ajustement, est aussi déduite du VIF. Elle est corrigée par la variance de la variable :

$$\hat{\sigma}_{\varepsilon_l}^2 = \frac{(n - 1) \frac{s_{x_l}^2}{v_l}}{n - (p - 1) - 1} = \frac{(n - 1) \frac{s_{x_l}^2}{v_l}}{n - p} \quad (3.15)$$

où $s_{x_l}^2$ est la variance estimée de la variable X_l

$$s_{x_l}^2 = \frac{1}{n - 1} \sum_{i=1}^n (x_{il} - \bar{x}_l)^2$$

Coefficients standardisés de la régression. Dans un premier temps, nous produisons les coefficients standardisés de la régression. Contrairement aux coefficients usuels, ils permettent la comparaison de l'impact des variables indépendantes sur la variable dépendante en les ramenant sur une échelle commune. Leur obtention est immédiate à partir de la matrice C^{-1}

$$\hat{\beta}_{kl} = - \frac{v_{kl}}{v_l} \quad (3.16)$$

Coefficients de la régression. Les coefficients de la régression sont alors obtenus en les dénormalisant des écart-types des variables, soit

$$\hat{a}_{kl} = \hat{\beta}_{kl} \times \frac{s_{x_l}}{s_{x_k}}, k \neq 0 \quad (3.17)$$

Pour la constante \hat{a}_0 , nous avons besoin des moyennes

$$\hat{a}_{0l} = \bar{x}_l - \sum_{k \neq l} \hat{a}_{kl} \times \bar{x}_k \quad (3.18)$$

Corrélations partielles. Dernier résultats intéressant, il est possible de produire les corrélations partielles entre les variables, prises deux à deux, à partir de la matrice C^{-1} . Pour mesurer la liaison entre les variables X_k et X_j en contrôlant l'influence des autres, nous calculons

$$r_{x_k, x_j / X - \{x_k, x_j\}} = - \frac{v_{kj}}{\sqrt{v_k \times v_j}} \quad (3.19)$$

L'objectif est de mesurer la "véritable" relation entre les variables, en dehors de toute influence. Un décalage éventuel (forte réduction en valeur absolue) entre les valeurs de r_{x_k, x_j} et $r_{x_k, x_j / X - \{x_k, x_j\}}$ est révélateur du caractère artificiel de la relation mesurée à l'aide de la corrélation brute que l'on peut lire dans la matrice C .

3.6.2 Régressions croisées sur les données CONSO

Reprenons notre exemple CONSO pour fixer les idées, nous avons la matrice des corrélations croisées C (Figure 3.1 - l'ordre des variables est *PRIX*, *CYLINDREE*, *PUISSANCE*, *POIDS*)

$$C = \begin{pmatrix} 1 & 0.92 & 0.93 & 0.95 \\ 0.92 & 1 & 0.96 & 0.86 \\ 0.93 & 0.96 & 1 & 0.85 \\ 0.95 & 0.86 & 0.85 & 1 \end{pmatrix}$$

Et son inverse C^{-1} (Figure 3.2)

$$C^{-1} = \begin{pmatrix} 19.79 & -1.45 & -7.51 & -11.09 \\ -1.45 & 12.87 & -9.80 & -1.36 \\ -7.51 & -9.80 & 14.89 & 2.86 \\ 2.86 & -1.36 & 2.86 & 10.23 \end{pmatrix}$$

Nous souhaitons retrouver les caractéristiques de la régression¹⁷

$$PRIX = a_0 + a_2 \times CYLINDREE + a_3 \times PUISSANCE + a_4 \times POIDS$$

Nous connaissons le VIF de la variable *PRIX*, $v_1 = 19.79$ lue dans la matrice C^{-1} . Nous en déduisons le coefficient de détermination de la régression

17. Nous avons décalé sciemment les indices pour respecter la position des variables parmi l'ensemble des explicatives.

$$R_1^2 = 1 - \frac{1}{v_j} = 1 - \frac{1}{19.79} = 0.9495$$

Le coefficient de détermination est très élevé, la variable PRIX est fortement corrélée avec une combinaison linéaire des autres variables. On peut vérifier la significativité globale de la régression en utilisant le test F, avec

$$F_1 = \frac{R_1^2/(p-1)}{(1-R_1^2)/(n-p)} = \frac{0.9495/(4-1)}{(1-0.9495)/(27-4)} = 144.0716$$

Sachant que la variance de PRIX est égal à $s_{prix}^2 = 158812189.1$, nous calculons l'écart-type de l'erreur de la régression

$$\hat{\sigma}_{\varepsilon_1} = \sqrt{\frac{(n-1) \frac{s_{x_1}^2}{v_1}}{n-p}} = \sqrt{\frac{(27-1) \frac{158812189.1}{19.79}}{27-4}} = 3011.7605$$

Pour obtenir les coefficients standardisés de la régression de PRIX, nous nous intéressons à la 1^{ère} colonne de la matrice C^{-1} , nous calculons

$$\begin{aligned}\hat{\beta}_{21} &= -\frac{v_{21}}{v_{11}} = -\frac{-1.45}{19.79} = 0.0734 \\ \hat{\beta}_{31} &= -\frac{v_{31}}{v_{11}} = -\frac{-7.51}{19.79} = 0.3796 \\ \hat{\beta}_{41} &= -\frac{v_{41}}{v_{11}} = -\frac{-11.09}{19.79} = 0.5601\end{aligned}$$

Nous les "dé-standardisons" en utilisant les écarts-type estimés $s_1 = 12602.07$, $s_2 = 634.42$, $s_3 = 32.79$ et $s_4 = 314.21$; soit :

$$\begin{aligned}\hat{a}_{21} &= \hat{\beta}_{21} \times \frac{s_1}{s_2} = 0.0734 \times \frac{12602.07}{634.42} = 1.4572 \\ \hat{a}_{31} &= \hat{\beta}_{31} \times \frac{s_1}{s_3} = 0.3796 \times \frac{12602.07}{32.79} = 145.9061 \\ \hat{a}_{41} &= \hat{\beta}_{41} \times \frac{s_1}{s_4} = 0.5601 \times \frac{12602.07}{314.21} = 22.4638\end{aligned}$$

A l'aide des moyennes des variables \bar{x}_j , nous pouvons produire la constante de la régression

$$\begin{aligned}\hat{a}_{01} &= \bar{x}_1 - \sum_{k \neq 1} \hat{a}_{k1} \times \bar{x}_k \\ &= 28260.56 - (1.4572 \times 1802.07 + 145.9061 \times 78.15 + 22.4638 \times 1193.15) \\ &= -12570.3173\end{aligned}$$

Toutes ces opérations sont résumés dans une feuille Excel (Figure 3.11). Bien évidemment, avec la même démarche nous pouvons produire les régressions des autres exogènes versus les $(p-1)$ autres variables.

Concernant les corrélations partielles, le logiciel LAZSTATS¹⁸ nous les fournit directement¹⁹. Si, au regard de la matrice C , les liaisons brutes sont fortes quelles que soient les variables, nous constatons

18. <http://statpages.org/miller/openstat/LazStatsPage.htm>

19. La régression croisée est également programmée dans la version 1.4.40 (et ultérieures) de TANAGRA - <http://tutoriels-data-mining.blogspot.com/2011/07/tanagra-version-1440.html>.

	C				
	PRIX	CYLINDREE	PUISSANC	POIDS	
PRIX	1.00	0.92	0.93	0.95	PRIX vs. autres R^2 0.9495 F 144.0716 p -value 4.74913E-15 $\sigma^2(\epsilon)$ 3011.7605 betas cylindree 0.0734 puissance 0.3796 poids 0.5601 coefs cylindree 1.4572 puissance 145.9061 poids 22.4638 const -12570.3173
CYLINDREE	0.92	1.00	0.96	0.86	
PUISSANC	0.93	0.96	1.00	0.85	
POIDS	0.95	0.86	0.85	1.00	
VIF	C ⁻¹				
	PRIX	CYLINDREE	PUISSANC	POIDS	
	19.79	-1.45	-7.51	-11.09	
	-1.45	12.87	-9.80	-1.36	
	-7.51	-9.80	14.89	2.86	
	-11.09	-1.36	2.86	10.23	
variance	158812189.10	402489.38	1075.05	98725.28	
ecart-type	12602.07	634.42	32.79	314.21	
moyenne	28260.56	1802.07	78.15	1193.15	

Fig. 3.11. Régressions croisées - PRIX = f(CYLINDREE, PUISSANCE, POIDS)

par exemple que la relation entre prix et cylindrée ($r_{prix,cylindree} = 0.92$ - lue dans la matrice C) est en réalité influencée par puissance et poids ($r_{prix,cylindree/puissance,poids} = 0.091$) (Figure 3.12). En effet, à partir de la matrice C^{-1} ,

$$r_{prix,cylindree/puissance,poids} = -\frac{v_{12}}{\sqrt{v_1 \times v_2}} = -\frac{-1.45}{\sqrt{19.79 \times 12.87}} = 0.091$$

En revanche, la liaison entre puissance et cylindrée (0.96) reste forte même après avoir retranché l'influence de prix et poids (0.708).

Partial Correlations with 27 cases.				
Variables	Prix	Cylindree	Puissance	Poids
Prix	-1.000	0.091	0.438	0.779
Cylindree	0.091	-1.000	0.708	0.118
Puissance	0.438	0.708	-1.000	-0.232
Poids	0.779	0.118	-0.232	-1.000

Fig. 3.12. Régressions croisées - Corrélations partielles

Vérification avec la régression explicite. A titre de vérification, nous avons calculé explicitement sur les données le modèle $PRIX = f(CYLINDREE, PUISSANCE, POIDS)$ à l'aide du logiciel TANAGRA (Figure 3.13). Nous constatons que les résultats concordent en tous points (R^2 , F , $\hat{\sigma}_\epsilon$, \hat{a}_j) avec les valeurs issues du post-traitement de la matrice C^{-1} (Figure 3.11).

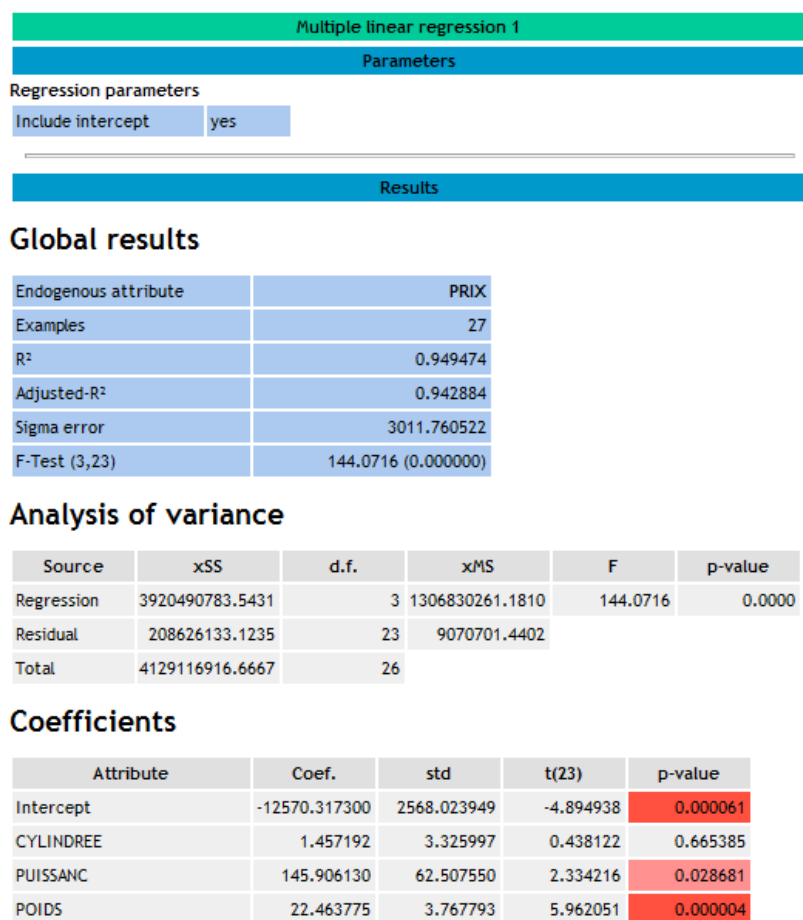


Fig. 3.13. Régressions croisées - Régression explicite : Prix vs. les autres variables

3.7 Conclusion

La colinéarité devient un problème dès lors que l'on veut lire et interpréter les résultats de la régression. La sélection de variables compte parmi les solutions possibles. Néanmoins, il ne faut surtout pas prendre au pied de la lettre les sous-ensembles de variables fournis par les algorithmes de sélection. Étudier de près les résultats intermédiaires en compagnie d'un expert du domaine (ex. un médecin, un économiste, etc.) est indispensable pour bien appréhender les interdépendances en jeu et repérer les aléas qui peuvent altérer les résultats.

Régression sur des exogènes qualitatives

La régression telle que nous l'étudions met en relation des variables exclusivement continues. Si on veut introduire des variables qualitatives nominales, la stratégie consistant à procéder au simple recodage des variables incriminées, le codage 0/1 dit *codage disjonctif complet* est certainement la plus connue. Mais il faut vérifier la validité des hypothèses probabilistes et structurelles liées à la technique des MCO. Il faut également savoir interpréter les résultats.

Si c'est l'endogène qui est qualitative, on parle de *régression logistique*. Les hypothèses liées aux erreurs de la MCO ne sont plus respectées. Nous entrons dans un cadre qui dépasse notre propos, nous ne l'aborderons pas dans ce chapitre. Pour les lecteurs intéressés par le sujet, je conseille la lecture du document accessible en ligne "Pratique de la Régression Logistique - Régression Logistique Binaire et Polytomique" (<http://eric.univ-lyon2.fr/~ricco/cours/ouvrages.html>).

Si ce sont les exogènes qui sont qualitatives, nous pouvons procéder au codage, mais encore faut-il :

1. définir le type de codage à utiliser ;
2. donner un sens aux résultats et comprendre les coefficients fournis par la régression.

Le cas des exogènes qualitatives nous fait mettre un pied dans le vaste domaine de la comparaison de populations. La technique paramétrique privilégiée dans ce cadre est l'*analyse de variance (ANOVA)*. Nous présentons très brièvement un cas particulier de cette technique¹.

4.1 Analyse de variance à 1 facteur et transposition à la régression

L'analyse de variance (ANOVA) à un facteur est une généralisation de la comparaison de moyennes dans K populations. Pour fixer les idées, travaillons sur un jeu de données.

1. La présentation adoptée ici s'appuie en grande partie sur l'excellent document en ligne de D. Mouchiroud, <http://spiral.univ-lyon1.fr/mathsv/cours/pdf/stat/Chapitre9.pdf>. Le chapitre 9 fait partie d'un document plus général "Probabilité et Statistique", <http://spiral.univ-lyon1.fr/mathsv/>

4.1.1 Un exemple introductif

Le fichier LOYER (Figure 4.1) décrit le montant du loyer au m^2 de 15 habitations situées dans différentes zones de la ville. On distingue 3 types de lieu d'habitation : banlieue, campagne et centre.

Loyer (Euro au m^2)	Lieu Habitation
6.9	banlieue
6.3	banlieue
6.7	banlieue
6.5	banlieue
7.2	banlieue
5.6	campagne
4.9	campagne
5.3	campagne
7	centre
7.5	centre
8	centre
7.2	centre
8.4	centre
7.4	centre
8.3	centre

Fig. 4.1. Loyer au m^2 selon le lieu d'habitation - Fichier LOYER

On veut répondre à la question : le loyer au m^2 est-il significativement différent d'une zone à l'autre ?

4.1.2 ANOVA à 1 facteur

Test d'hypothèses

Le problème que nous décrivons est une comparaison de moyennes de K populations. On peut décrire le test d'hypothèses de la manière suivante

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_K = \mu$$

$$H_1 : \text{une des moyennes au moins diffère des autres}$$

où μ_k est la moyenne de la variable d'intérêt Y pour la population k .

Une manière simple de visualiser les différences est d'afficher les boîtes à moustaches de la variable Y selon le groupe d'appartenance (Figure 4.2). Plus les boxplot seront décalés, plus forte sera la différenciation. Autre information très importante que nous communique ce graphique, nous pouvons nous faire une idée de la dispersion des valeurs dans chaque groupe. Nous verrons plus loin la portée de cette information.

Remarque 20 (Facteurs fixes et facteurs aléatoires). On parle de *facteurs fixes* lorsque tous les groupes sont représentés dans le fichier de données, de *facteurs aléatoires* lorsque seulement un échantillon des groupes sont présents. Dans le cas de l'ANOVA à 1 facteur, cette distinction n'a aucune conséquence sur les calculs.

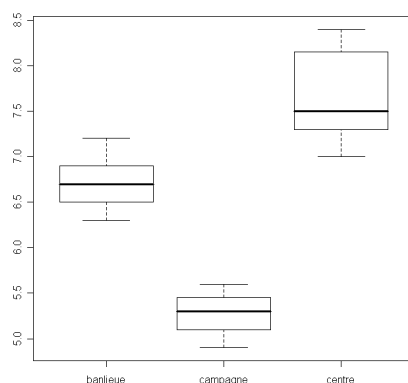


Fig. 4.2. Boîtes à moustaches des loyers selon le lieu d'habitation - Fichier LOYER

Statistique du test

L'équation de décomposition de la variance permet de construire la statistique du test. Elle s'écrit

$$SCT = SCE + SCR$$

$$\sum_{k=1}^K \sum_{i=1}^{n_k} (y_{i,k} - \bar{y})^2 = \sum_{k=1}^K n_k (\bar{y}_k - \bar{y})^2 + \sum_{k=1}^K \sum_{i=1}^{n_k} (y_{i,k} - \bar{y}_k)^2$$

où $y_{i,k}$ représente la valeur de Y pour l'individu i du groupe k ; \bar{y} est la moyenne globale de Y ; \bar{y}_k est la moyenne conditionnelle c.-à-d. la moyenne de Y dans le groupe k .

Cette décomposition se lit comme suit :

- SCT est la somme des carrés des écarts totaux, elle indique la variabilité totale de Y ;
- SCE est la somme des carrés des écarts inter-groupes, expliqués par l'appartenance aux groupes ;
- SCR est la somme des carrés des écarts intra-groupes, résiduels à l'intérieur des groupes.

La somme SCT est constante. Par conséquent, une valeur de SCE élevée indique que l'appartenance aux groupes détermine fortement la valeur de la variable d'intérêt. A l'extrême, lorsque $SCE = SCT$, connaître le groupe d'appartenance d'un individu permet de connaître à coup sûr la valeur de Y associée.

Nous construisons le tableau d'analyse de variance à partir de ces informations

Sources de variation	Degrés de liberté (ddl)	Somme des carrés (SC)	Carrés moyens (CM)
Expliqués (inter)	$K - 1$	SCE	$CME = \frac{SCE}{K-1}$
Résiduels (intra)	$n - K$	SCR	$CMR = \frac{SCR}{n-K}$
Totaux	$n - 1$	SCT	-

Sous H_0 , la statistique $F = \frac{CME}{CMR}$ suit une loi de Fisher à $(K - 1, n - K)$ degrés de liberté.

La région critique du test s'écrit

$$R.C. : F > F_{1-\alpha}(K-1, n-K)$$

où $F_{1-\alpha}(K-1, n-K)$ est le quantile d'ordre $(1-\alpha)$ de la loi de Fisher.

Conditions d'applications

L'ANOVA à 1 facteur est un test paramétrique. Elle est assortie d'un certain nombre de conditions pour être réellement opérationnelle : les observations doivent être indépendantes, notamment les K échantillons comparés doivent être indépendants ; la variable d'intérêt doit suivre une loi normale ; la variance de Y dans les groupes doit être homogène (homoscédasticité).

Notons 2 points importants : l'ANOVA à 1 facteur est assez robuste ; ces conditions, et c'est ce qui nous intéresse ici, ne sont pas sans rappeler certaines hypothèses de la régression linéaire multiple. Nous y reviendrons plus loin.

Application aux données LOYER

Lieu Habitation	Moyenne	n	n x Ecart moyenne ²
banlieue	6.7200	5	0.1280
campagne	5.2667	3	7.8085
centre	7.6857	7	4.5442
Globale	6.8800	15	

Tableau ANOVA			
Source	ddl	SC	CM
SCE	2	12.48076	6.24038
SCR	12	2.54324	0.21194
SCT	14	15.02400	-

F	29.44458
p-value	0.00002

Loyer (Euro au m ²)	Lieu Habitation
6.9	banlieue
6.3	banlieue
6.7	banlieue
6.5	banlieue
7.2	banlieue
5.6	campagne
4.9	campagne
5.3	campagne
7	centre
7.5	centre
8	centre
7.2	centre
8.4	centre
7.4	centre
8.3	centre

Fig. 4.3. Tableau de calcul de l'ANOVA à 1 facteur - Données LOYER

Nous appliquons ces calculs sur les données LOYER (Figure 4.3), voici les étapes :

1. Recenser les effectifs n_k et les moyennes \bar{y}_k conditionnelles ;
2. Calculer la moyenne globale $\bar{y} = 6.88$;
3. Former $SCT = 15.02400$ et $SCE = 5(6.72 - 6.88)^2 + 3(5.27 - 6.88)^2 + 7(7.69 - 6.88)^2 = 12.48076$;
4. En déduire $SCR = 15.024 - 12.48076 = 2.54324$;
5. Calculer la statistique du test $F = \frac{12.48076/2}{2.54324/12} = 29.44458$;
6. Obtenir enfin la p-value à l'aide de la loi de Fisher à (2, 12) degrés de liberté, $p.value = 0.00002$.

Au risque de 5%, l'hypothèse d'égalité des moyennes est rejetée. Le lieu d'habitation a une influence sur le montant du loyer.

Remarque 21 (Analyse des contrastes). On complète généralement l'ANOVA avec l'analyse des contrastes. Elle vise à déterminer quelle est la moyenne qui diffère le plus des autres, ou encore quelles sont les couples (triplets, etc.) de moyennes qui s'opposent le plus. Nous garderons à l'esprit cette idée car elle nous aidera à mieux comprendre les résultats de la régression appliquée aux exogènes qualitatives.

Analogie avec la régression

Quel est le rapport avec la régression? On comprend mieux l'objet de ce chapitre si l'on reformule le test de comparaison de moyennes. Les valeurs prises par la variable d'intérêt peut s'écrire sous la forme suivante :

$$y_{i,k} = \mu + \alpha_k + \varepsilon_{i,k}$$

où α_k est l'effet du facteur k , $\varepsilon_{i,k} \sim \mathcal{N}(0, \sigma)$.

Il s'agit, ni plus ni moins, d'une droite de régression que l'on peut résoudre avec la MCO. Il suffit de coder convenablement la variable exogène qualitative. L'hypothèse nulle de l'ANOVA devient

$$H_0 : \alpha_1 = \alpha_2 = \dots = \alpha_K = 0$$

qui s'apparente au test de significativité globale d'une régression linéaire multiple.

Il nous faut donc définir une transformation appropriée de la variable exogène qualitative pour que la régression puisse résoudre un problème d'ANOVA. Le codage est d'autant plus important qu'il conditionne l'interprétation des coefficients de l'équation de régression. C'est ce que nous allons voir maintenant.

4.2 Inadéquation du codage disjonctif complet

4.2.1 Codage disjonctif complet

La méthode la plus simple/connue pour transformer une variable qualitative X à K modalités en une variable numérique est le *codage disjonctif complet*. A chaque modalité k de X , on associe une **variable indicatrice**² Z_k telle que

$$Z_{i,k} = \begin{cases} 1 & \text{si } X_i = k \\ 0 & \text{sinon} \end{cases}$$

Sur l'exemple LOYER, nous aurions 3 indicatrices Z_k définies à partir des correspondances suivantes

Lieu	$Z_{banlieue}$	$Z_{campagne}$	$Z_{centre-ville}$
Banlieue	1	0	0
Campagne	0	1	0
Centre	0	0	1

2. *Dummy variable* en anglais.

Loyer	Habitation	banlieue	campagne	centre
6.9	banlieue	1	0	0
6.3	banlieue	1	0	0
6.7	banlieue	1	0	0
6.5	banlieue	1	0	0
7.2	banlieue	1	0	0
5.6	campagne	0	1	0
4.9	campagne	0	1	0
5.3	campagne	0	1	0
7	centre	0	0	1
7.5	centre	0	0	1
8	centre	0	0	1
7.2	centre	0	0	1
8.4	centre	0	0	1
7.4	centre	0	0	1
8.3	centre	0	0	1

Fig. 4.4. Codage disjonctif complet de la variable *habitation*

Nous disposons d'un nouveau tableau de données (Figure 4.4), et nous écrivons naturellement la régression comme suit

$$\text{loyer} = a_0 + a_1 Z_{\text{banlieue}} + a_2 Z_{\text{campagne}} + a_3 Z_{\text{centre-ville}} + \varepsilon$$

Pourtant, effectuer cette régression provoquerait immédiatement une erreur en raison d'un problème flagrant de colinéarité. En effet, pour tout individu i

$$Z_{i,\text{banlieue}} + Z_{i,\text{campagne}} + Z_{i,\text{centre-ville}} = 1$$

La matrice $(Z'Z)$ n'est pas inversible car la première colonne de Z est composée de la valeur 1, et la somme des 3 colonnes suivantes est aussi égale à 1.

4.2.2 Régression sans constante et lecture des coefficients

Pour éviter cet écueil, une solution serait de définir une régression sans constante. L'équation devient

$$\text{loyer} = a_1 Z_{\text{banlieue}} + a_2 Z_{\text{campagne}} + a_3 Z_{\text{centre-ville}} + \varepsilon$$

Loyer	Habitation	banlieue	campagne	centre
6.9	banlieue	1	0	0
6.3	banlieue	1	0	0
6.7	banlieue	1	0	0
6.5	banlieue	1	0	0
7.2	banlieue	1	0	0
5.6	campagne	0	1	0
4.9	campagne	0	1	0
5.3	campagne	0	1	0
7	centre	0	0	1
7.5	centre	0	0	1
8	centre	0	0	1
7.2	centre	0	0	1
8.4	centre	0	0	1
7.4	centre	0	0	1
8.3	centre	0	0	1

	centre	campagne	banlieue
coef.	7.68571	5.26667	6.72000
std.dev	0.17400	0.26579	0.20588
R ²	0.99649	0.46037	#N/A
	1136.34152	12	#N/A
	722.49676	2.54324	#N/A

Fig. 4.5. Régression sans constante - Données LOYER

Nous lançons les MCO pour obtenir les coefficients (Figure 4.5).

Lecture de coefficients

Penchons nous sur les coefficients. Nous ne sommes pas sans noter une certaine similitude avec les valeurs des moyennes conditionnelles présentées dans le tableau de l'ANOVA à 1 facteur (Figure 4.3). Nous observons que $\hat{a}_1 = \bar{y}_{banlieue}$, $\hat{a}_2 = \bar{y}_{campagne}$ et $\hat{a}_3 = \bar{y}_{centre}$.

Remarque 22 (Moyenne conditionnelle). Pour rappel, nous pouvons définir la moyenne conditionnelle \bar{y}_k de la manière suivante, selon qu'on utilise ou non la variable recodée

$$\begin{aligned}\bar{y}_k &= \frac{1}{n_k} \sum_{i: z_{i,k}=1} y_i \\ &= \frac{1}{n_k} \sum_{i: x_i=k} y_i\end{aligned}$$

Dans la régression sans constante mettant en jeu des exogènes codées 0/1 à partir d'une variable qualitative, les coefficients s'interprètent comme des moyennes conditionnelles de la variable endogène.

Décomposition de la variance

La décomposition de la variance en revanche est incorrecte. Si la $SCR = 2.54324$ est calculée correctement par DROITEREG, la $SCE = 722.49676$ est fausse (cf. celle fournie par l'ANOVA à un facteur, figure 4.3, où $SCE = 12.48076$). Tout simplement parce que dans la régression sans constante, l'équation d'analyse de variance décomposant la variabilité totale en variabilité expliquée et résiduelle n'est plus licite. **Nous ne pouvons donc pas déduire des résultats de la régression (sans constante) la solution du test de comparaison des K moyennes conditionnelles.**

4.2.3 Vers des solutions plus générales

Malgré son intérêt, cette technique n'est pas généralisable, notamment parce qu'il n'est pas possible d'introduire plusieurs (≥ 2) variables qualitatives recodées dans la régression. Nous devons nous tourner vers d'autres solutions qui peuvent s'appliquer dans un cadre plus large.

Pour contourner le problème de la colinéarité, une solution simple serait tout simplement d'omettre la dernière modalité dans le codage. Pour une variable qualitative à K modalités, nous produisons ainsi $(K-1)$ variables indicatrices. Reste à savoir comment introduire dans ces nouvelles variables l'information sur la dernière modalité. Ce point est loin d'être anodin, il définit le mode de lecture des coefficients de la régression lorsqu'on introduit les variables exogènes recodées dans l'analyse.

4.3 Codage "Cornered effect" de l'exogène qualitative

4.3.1 Principe

On part du constat que la dernière modalité K peut être déduite des autres dans le codage disjonctif complet

$$Z_{i,K} = 1 - (Z_{i,1} + Z_{i,2} + \dots + Z_{i,K-1})$$

On omet tout simplement la variable Z_K dans la régression. On sait que

$$X_i = K \Leftrightarrow Z_{i,1} = Z_{i,2} = \dots = Z_{i,K-1} = 0$$

Lorsque X prend la modalité K , toutes les indicatrices Z_1, \dots, Z_{K-1} prennent la valeur zéro. Ainsi, nous avons le tableau de correspondance suivant pour les données LOYER avec 2 indicatrices Z_k :

Lieu	$Z_{banlieue}$	$Z_{campagne}$
Banlieue	1	0
Campagne	0	1
Centre	0	0

L'équation de régression estimée à l'aide des MCO pour les données LOYER en omettant la variable Z_{centre} devient

$$loyer = \hat{\alpha}_0 + \hat{\alpha}_1 Z_{banlieue} + \hat{\alpha}_2 Z_{campagne} \quad (4.1)$$

Reste à interpréter les coefficients de la régression.

Remarque 23 (Choix de la modalité de référence). Le choix de la modalité de référence K est très important. Il faut qu'elle soit bien circonscrite pour que les interprétations aient une certaine consistance. Par exemple, prendre la modalité "autres" comme référence n'est pas une bonne idée parce qu'on ne sait pas très bien souvent ce qu'elle recouvre. De fait, la différenciation avec les autres groupes est mal définie. Prenons le cas des symptômes grippaux, il y a 4 cas possibles : pas de symptômes, toux grasse, toux sèche, autres symptômes. Pour l'interprétation, il semble judicieux de prendre comme référence l'absence de symptômes. En ce qui concerne les effectifs, il est souhaitable que le groupe de référence comporte suffisamment d'observations pour que l'estimation de la moyenne conditionnelle associée soit précise.

4.3.2 Lecture des résultats

Voyons quelques cas particuliers pour mieux appréhender la situation :

- Si l'habitation i^* est en *centre-ville*, nous savons que $Z_{i^*,banlieue} = Z_{i^*,campagne} = 0$. Par conséquent $\hat{y}_{i^*} = \hat{\alpha}_0$, le loyer prédit est $\hat{\alpha}_0$.
- Si l'habitation i^* est en *banlieue*, nous savons que $Z_{i^*,banlieue} = 1$, les autres indicatrices sont égales à 0. Nous en déduisons la valeur prédite du loyer $\hat{y}_{i^*} = \hat{\alpha}_0 + \hat{\alpha}_1$.

En généralisant, nous observons les relations suivantes :

- $\hat{a}_0 = \bar{y}_{centre}$
- $\hat{a}_1 = \bar{y}_{banlieue} - \bar{y}_{centre}$
- $\hat{a}_2 = \bar{y}_{campagne} - \bar{y}_{centre}$

Cela nous emmène à tirer plusieurs conclusions :

1. Les coefficients de la régression s'interprètent comme une moyenne conditionnelle de l'endogène (la constante) ou comme des écarts à cette moyenne (les autres coefficients).
2. On parle de *cornered effect* car la constante représente la moyenne conditionnelle de l'endogène pour les observations portant la modalité exclue. Elle nous sert de moyenne de référence. L'appellation *dummy coding* est également utilisée pour qualifier ce type de codage.
3. Le test de significativité globale de la régression correspond **exactement** à une ANOVA à 1 facteur c.-à-d. tester l'égalité globale des moyennes conditionnelles.
4. Pour le cas particulier de ($K = 2$) groupes, nous avons une régression simple avec seul indicatrice. Le test de significativité globale est équivalent au test de significativité de la pente ([18], section 3.6.1), il correspond à un test de comparaison de moyennes de 2 sous-populations ([18], section 13.3).
5. Nous pouvons même aller plus loin, les tests de significativité des coefficients a_k ($k \geq 1$) s'apparentent à un test de comparaison de la moyenne conditionnelle μ_k avec la moyenne de référence μ_K . "S'apparente" car, d'une part, l'estimation de l'écart-type n'est pas la même, la statistique réduite n'est donc pas exactement la même; d'autre part, il y a des disparités entre les degrés de liberté. Nous y reviendrons en détail ci-dessous.
6. De même, la comparaison des coefficients des indicatrices s'apparente à une comparaison de moyennes entre 2 groupes quelconques.

4.3.3 Application aux données LOYER

Loyer	Habitation	banlieue	campagne
6.9	banlieue	1	0
6.3	banlieue	1	0
6.7	banlieue	1	0
6.5	banlieue	1	0
7.2	banlieue	1	0
5.6	campagne	0	1
4.9	campagne	0	1
5.3	campagne	0	1
7	centre	0	0
7.5	centre	0	0
8	centre	0	0
7.2	centre	0	0
8.4	centre	0	0
7.4	centre	0	0
8.3	centre	0	0

	campagne	banlieue	constante
coef.	-2.4190	-0.9657	7.6857
	0.3177	0.2696	0.1740
	0.8307	0.4604	#N/A
	29.4446	12	#N/A
	12.48076	2.54324	#N/A

Moyenne conditionnelles		
campagne	banlieue	centre
5.27	6.72	7.69

Test significativité globale	
F	29.44458
ddl1	2
ddl2	12
p-value	0.00002

Fig. 4.6. Régression avec données codées "cornered effect" - Données LOYER

Nous effectuons la régression sur notre fichier de données codé selon la technique "cornered effect" (Figure 4.6). Il y a bien $p = 2$ variables exogènes. Nous obtenons les résultats de l'équation de régression (Equation 4.1), nous en déduisons les moyennes conditionnelles :

- $\hat{a}_0 = \bar{y}_{centre} = 7.69$;
- $\hat{a}_1 = -0.97 \Rightarrow \bar{y}_{banlieue} = 7.69 + (-0.97) = 6.72$;
- $\hat{a}_2 = -2.42 \Rightarrow \bar{y}_{campagne} = 7.69 + (-2.42) = 5.27$

Pour tester la significativité globale de la régression, nous exploitons les sorties du tableur EXCEL :

Indicateur	Valeur
<i>SCE</i>	12.48076
<i>SCR</i>	2.54324
<i>ddl1</i> = p	2
<i>ddl2</i> = $n - p - 1$	12
<i>F</i>	$\frac{12.48076/2}{2.54324/12} = 29.44458$
p-value	0.00002

Ces résultats - la décomposition de la variance ($SCT = SCE + SCR$) et les degrés de liberté - correspondent exactement à ceux de l'ANOVA à 1 facteur (Figure 4.3). Les deux approches sont équivalentes.

4.4 Comparaisons entres groupes

4.4.1 Comparaisons avec le groupe de référence

Principe du test

Les coefficients des indicatrices se lisent comme des écarts à la moyenne de référence (la moyenne de Y pour le groupe de référence). De fait, le test

$$\begin{cases} H_0 : \mu_j = \mu_K \\ H_1 : \mu_j \neq \mu_K \end{cases}$$

Peut s'écrire sous la forme d'un test de significativité des paramètres de la régression

$$\begin{cases} H_0 : a_j = 0 \\ H_1 : a_j \neq 0 \end{cases}$$

La statistique de test s'écrit

$$t_{\hat{a}_j} = \frac{\hat{a}_j}{\hat{\sigma}_{\hat{a}_j}} \quad (4.2)$$

Elle suit une loi de Student à $ddl = (n - p - 1)$ degrés de liberté sous H_0 . N'oublions pas que dans notre configuration, $p = K - 1$, ainsi $ddl = n - K$. Le test est bilatéral.

Application aux données CONSO

Nous souhaitons comparer les moyennes des loyers en banlieue et en centre-ville. Nous disposons de toutes les informations nécessaires via DROITEREG (Figure 4.6) :

$$\begin{aligned}\hat{a}_{banlieue} &= -0.9657 \\ \hat{\sigma}_{\hat{a}_{banlieue}} &= 0.2696 \\ t_{\hat{a}_{banlieue}} &= \frac{-0.97}{0.27} = -3.5825 \\ ddl &= 15 - 3 = 12 \\ p - value &= 0.0038\end{aligned}$$

Au risque $\alpha = 5\%$, nous rejetons l'hypothèse nulle. Le loyer moyen en banlieue est différent de celui du centre-ville.

Équivalence avec le test usuel de comparaison de moyennes

Nous avons vu plus haut que le test de significativité globale de la régression était complètement équivalente à une ANOVA à un facteur. Est-ce que ce résultat est transposable à la comparaison d'un groupe avec la référence ?

Le test de comparaison de moyennes, avec une hypothèse d'égalité des variances dans les groupes, est basé sur l'indicateur

$$D = \bar{y}_j - \bar{y}_K$$

Nous constatons immédiatement que $D = \hat{a}_j$. La différence se joue sur l'estimation de la variance. En effet,

$$\hat{\sigma}_D^2 = s^2 \left(\frac{1}{n_j} + \frac{1}{n_K} \right)$$

Avec

$$s^2 = \frac{(n_j - 1)s_j^2 + (n_K - 1)s_K^2}{n_j + n_K - 2}$$

s_k^2 étant l'estimateur sans biais de la variance pour le groupe k .

Sous H_0 , la statistique $t_D = \frac{t}{\hat{\sigma}_D}$ suit une loi de Student à $(n_j + n_K - 2)$ degrés de liberté.

Si l'estimation de l'écart est la même ($\hat{a}_j = D$), il n'y a aucune raison en revanche que les estimations des variances coïncident. Les degrés de liberté sont différents. Numériquement, les régions critiques ne seront pas identiques.

Notons un élément très important, *les autres groupes n'interviennent pas dans cette écriture de la comparaison directe*. Alors que dans la régression, ils pèsent dans le calcul de la variance de la statistique de test et dans la définition des degrés de liberté.

Loyer	Habitation
6.9	banlieue
6.3	banlieue
6.7	banlieue
6.5	banlieue
7.2	banlieue
7	centre
7.5	centre
8	centre
7.2	centre
8.4	centre
7.4	centre
8.3	centre

Données			
Habitation	Nombre de Loyer	Var de Loyer	Moyenne de Loyer
banlieue	5	0.1220	6.7200
centre	7	0.3014	7.6857
Total général	12	0.4561	7.2833

D	-0.9657
---	---------

s ²	0.2297
sigma ² _D	0.0787

t_D	-3.4415
ddl	10

p-value	0.0063
---------	--------

Fig. 4.7. Comparaison des moyennes - Loyers en banlieue et en centre-ville

Application aux données LOYER

Nous comparons directement les moyennes des loyers pour le centre-ville et la banlieue, à l'exclusion des autres groupes (campagne). Les calculs sont résumés dans une feuille Excel (Figure 4.7) :

1. Avec l'outil "Tableaux croisés dynamiques", nous obtenons

Habitation	n_k	s_k^2	\bar{y}_k
banlieue	5	0.1220	6.7200
centre	7	0.3014	7.6857

2. Nous en déduisons $D = 6.7200 - 7.6857 = -0.9657$, ce qui correspond exactement au coefficient de l'indicatrice "banlieue" obtenue dans la régression.
3. Concernant la variance de D , nous calculons successivement

$$s^2 = \frac{(5-1) \times 0.1220 + (7-1) \times 0.3014}{5+7-2} = 0.2297$$

$$\hat{\sigma}_D^2 = 0.2297 \times \left(\frac{1}{5} + \frac{1}{7} \right) = 0.0787$$

4. Nous formons le rapport

$$t_D = \frac{D}{\hat{\sigma}_D} = \frac{-0.9657}{\sqrt{0.0787}} = \frac{-0.9657}{0.2806} = -3.4415$$

5. Avec un loi $\mathcal{T}(5+7-2) \equiv \mathcal{T}(10)$, nous avons une probabilité critique de 0.0063.
6. Au risque 5%, tout comme avec la régression, nous rejetons l'hypothèse d'égalité des moyennes.

Les conclusions sont identiques, certes. Mais que cela ne masque pas les disparités au niveau de la variance de la statistique de test et des degrés de liberté.

Une autre estimation de la variance commune dans la comparaison de moyennes

La divergence entre les deux procédures tient essentiellement à une estimation différente de la variance commune de Y dans les groupes. Dans cette section, on montre que l'on peut retrouver les résultats de la régression avec la comparaison de moyennes en utilisant la variance intra-classes de l'ANOVA, prenant en compte tous les groupes.

Reprenons la procédure de comparaison de deux moyennes. La statistique $D = \bar{y}_j - \bar{y}_K$ ne change pas, elle est en relation directe avec les moyennes à confronter.

En revanche, nous utilisons une autre estimation de la variance commune, la variance intra-classes vue dans l'ANOVA (section 4.1), c.-à-d.

$$\tilde{s}^2 = \frac{SCR}{n - K} = \frac{\sum_{k=1}^K (n_k - 1) \times s_k^2}{n - K}$$

SCR est la somme des carrés des écarts à la moyenne intra-groupes. Elle correspond également à la somme des carrés résiduels de la régression. Les degrés de liberté deviennent $(n - K)$ dans ce contexte.

La variance de la statistique de test est obtenu avec

$$\tilde{\sigma}_D^2 = \tilde{s}^2 \left(\frac{1}{n_j} + \frac{1}{n_K} \right)$$

Et au final, nous avons

$$\tilde{t}_D = \frac{D}{\tilde{\sigma}_D}$$

Qui, sous H_0 suit une loi de Student à $(n - K)$ degrés de liberté.

Application sur les données LOYER

Comparons de nouveau la moyenne des loyers en banlieue et en centre-ville. Nous avons déjà $D = -0.9657$. Pour la variance intra-classes, nous reprenons les résultats de l'ANOVA (Figure 4.3),

$$\tilde{s}^2 = \frac{SCR}{n - K} = \frac{2.54324}{12} = 0.2119$$

Puis, nous calculons

$$\tilde{\sigma}_D^2 = \tilde{s}^2 \left(\frac{1}{n_j} + \frac{1}{n_K} \right) = 0.2119 \left(\frac{1}{5} + \frac{1}{7} \right) = 0.0727$$

Enfin,

$$\tilde{t}_D = \frac{D}{\tilde{\sigma}_D} = \frac{-0.9657}{\sqrt{0.0727}} = \frac{-0.9657}{0.2696} = -3.5825$$

Exactement la même valeur que la statistique $t_{\hat{a}_{banlieue}}$, le degrés de liberté étant également identiques.

Le test issu de la régression et le test de comparaison directe des moyennes sont donc équivalents si, et seulement si, nous prenons la variance empirique intra-classes intégrant tous les groupes pour estimer la variance σ_Y^2 de Y dans la seconde procédure³.

4.4.2 Comparaisons entre deux groupes quelconques

Construction du test

Toujours à partir des résultats de la régression, nous pouvons élaborer le test de comparaison de moyennes de Y pour deux groupes l et j quelconques. Il s'écrit

$$\begin{cases} H_0 : \mu_l = \mu_j \\ H_1 : \mu_l \neq \mu_j \end{cases}$$

Comment transposer cela à la régression ? Nous savons que

$$\begin{aligned} a_l &= \mu_l - \mu_K \\ a_j &= \mu_j - \mu_K \end{aligned}$$

On montre très facilement que le test de comparaison de moyennes est équivalent au test de comparaison de coefficients

$$\begin{cases} H_0 : a_l = a_j \\ H_1 : a_l \neq a_j \end{cases}$$

Pour mettre en application ce test, nous formons la statistique E , avec

$$E = \hat{a}_l - \hat{a}_j \quad (4.3)$$

Jusque là, c'est plutôt facile. La vraie gageure est de calculer correctement la variance de E . Elle est définie comme suit

$$V(E) = \sigma_E^2 = V(\hat{a}_l) + V(\hat{a}_j) - 2 \times COV(\hat{a}_l, \hat{a}_j) \quad (4.4)$$

Nous introduisons une nouvelle notion : la covariance entre les coefficients estimés. En effet, puisque les variables (les indicatrices) ne sont pas indépendantes, la covariance entre les coefficients n'est pas nulle. Elle est lue dans la matrice de variance covariance des coefficients qui est estimée avec

$$\hat{\Omega}_{\hat{a}} = \hat{\sigma}_\varepsilon^2 (Z'Z)^{-1}$$

Où $\hat{\sigma}_\varepsilon^2$ est l'estimation de la variance de l'erreur ; Z est la matrice des indicatrices avec, à la première colonne, le vecteur des valeurs 1 pour matérialiser la constante ([18], section 9.6.2). Rappelons que l'on retrouve sur la diagonale principale de la matrice $\hat{\Omega}_{\hat{a}}$ les variances estimées des coefficients.

3. J'adore ce type de configuration. Un même et seul problème traité avec deux prismes a priori très différents - comparaison de moyennes d'un côté, la régression de l'autre - et, au final, nous obtenons un résultat numérique identique. C'est bon ça. Comme quoi, il doit y avoir une certaine forme de vérité derrière toutes ces techniques.

Sous H_0 ,

$$t_E = \frac{E}{\hat{\sigma}_E}$$

suit une loi de Student à $(n - K)$ degrés de liberté. Le test est bilatéral. La région critique correspond aux valeurs extrêmes de t_E .

Remarque 24 (Test de contraintes linéaire sur les coefficients). La comparaison des coefficients de la régression peut s'inscrire dans un cadre plus général, les tests de contraintes linéaires. Nous décrivons en détail l'approche dans notre premier support ([18], section 11.3). Ainsi, nous pouvons comparer plusieurs (≥ 2) moyennes, ou encore tester des formes linéaires plus générales (ex. $\mu_1 = 2 \times \mu_2 + 7 \times \mu_3$, etc.). Notre champ d'investigation est considérablement élargi.

Application aux données CONSO

Nous souhaitons comparer la moyenne des loyers en banlieue et en campagne (Figure 4.8) :

Z				
Loyer	Habitation	constante	banlieue	campagne
6.9	banlieue	1	1	0
6.3	banlieue	1	1	0
6.7	banlieue	1	1	0
6.5	banlieue	1	1	0
7.2	banlieue	1	1	0
5.6	campagne	1	0	1
4.9	campagne	1	0	1
5.3	campagne	1	0	1
7	centre	1	0	0
7.5	centre	1	0	0
8	centre	1	0	0
7.2	centre	1	0	0
8.4	centre	1	0	0
7.4	centre	1	0	0
8.3	centre	1	0	0

DROITEREG			
	campagne	banlieue	constante
coef.	-2.4190	-0.9657	7.6857
sigma^(a^)	0.31768	0.26956	0.17400
	0.8307	0.4604	#N/A
	29.4446	12	#N/A
	12.4808	2.5432	#N/A
var^(a^)	0.10092	0.07266	0.03028

Moyenne conditionnelles		
campagne	banlieue	centre
5.26667	6.72000	7.68571

(Z'Z)		
15.0	5.0	3.0
5.0	5.0	0.0
3.0	0.0	3.0

(Z'Z)^(-1)		
0.14286	-0.14286	-0.14286
-0.14286	0.34286	0.14286
-0.14286	0.14286	0.47619

Omega^(a^)			
	constante	banlieue	campagne
constante	0.03028	-0.03028	-0.03028
banlieue	-0.03028	0.07266	0.03028
campagne	-0.03028	0.03028	0.10092

E	1.45333
sigma^2_E	0.11303
sigma_E	0.33620
t_E	4.32278
ddl	12
p-value	0.00099

Fig. 4.8. Comparaison des moyennes via la régression - Loyers en banlieue et à la campagne

- A partir des coefficients estimés via la fonction DROITEREG, nous pouvons calculer E ,

$$E = \hat{a}_{banlieue} - \hat{a}_{campagne} = -0.9657 - (-2.4190) = 1.45333$$

- Plusieurs étapes sont nécessaires pour aboutir à l'estimation de la variance de E . D'abord, il faut disposer de la matrice Z , composée des indicatrices et de la colonne de 1 (en fond jaune dans la figure 4.8).
- Nous calculons ensuite la matrice $(Z'Z)$ que nous inversons. Nous avons $(Z'Z)^{-1}$.

- Enfin, nous la pré-multipions par l'estimation de la variance de l'erreur fournie par DROITE-REG $\hat{\sigma}_\varepsilon^2 = (0.4604)^2 = 0.21194$ pour obtenir l'estimation de la matrice de variance covariance des coefficients estimés

$$\hat{\Omega}_{\hat{a}} = (0.4604)^2 \times \begin{pmatrix} 0.14286 & -0.14286 & -0.14286 \\ -0.14286 & 0.34286 & 0.14286 \\ -0.14286 & 0.14286 & 0.47619 \end{pmatrix} = \begin{pmatrix} 0.03028 & -0.03028 & -0.03028 \\ -0.03028 & 0.07266 & 0.03028 \\ -0.03028 & 0.03028 & 0.10092 \end{pmatrix}$$

- Nous piochons dans cette matrice les valeurs nécessaires à l'estimation de la variance de E ,

$$\begin{aligned} \hat{\sigma}_E^2 &= \hat{V}(\hat{a}_{campagne}) + \hat{V}(\hat{a}_{banlieue}) - 2 \times \widehat{COV}(\hat{a}_{campagne}, \hat{a}_{banlieue}) \\ &= 0.10092 + 0.07266 - 2 \times 0.03028 \\ &= 0.11303 \end{aligned}$$

- Il nous reste à former

$$t_E = \frac{1.45333}{\sqrt{0.11303}} = \frac{1.45333}{0.33620} = 4.32278$$

- Avec la loi $\mathcal{T}(12)$, nous avons la probabilité critique $p.value = 0.00099$.
- Au risque 5%, nous rejetons l'hypothèse nulle d'égalité des loyers moyens en banlieue et à la campagne.

Équivalence avec la comparaison de moyennes

Curieux comme nous sommes, voyons ce qu'il en est si nous passons par la comparaison directe des moyennes, toujours en utilisant la variance intra-classe $\tilde{s}^2 = 0.2119$ de l'ANOVA comme estimation de la variance de Y .

Nous calculons l'écart entre les moyennes

$$D = \bar{y}_{banlieue} - \bar{y}_{campagne} = 6.72000 - 5.26667 = 1.45333$$

Nous obtenons sa variance avec

$$\tilde{\sigma}_D^2 = \tilde{s}^2 \left(\frac{1}{n_{banlieue}} + \frac{1}{n_{campagne}} \right) = 0.2119 \left(\frac{1}{5} + \frac{1}{3} \right) = 0.11303$$

Reste à former le rapport,

$$\tilde{t}_D = \frac{1.45333}{\sqrt{0.11303}} = 4.32278 = t_E$$

Encore une fois, nous constatons que le test basé sur les résultats de la régression et celui basé sur une comparaison directe des moyennes, pour peu que l'on utilise l'estimation \tilde{s}^2 de la variance de Y , sont totalement équivalents.

4.5 Régression avec plusieurs explicatives qualitatives

Approfondissons l'analyse en ajoutant d'autres variables qualitatives explicatives. Plusieurs questions se posent alors : comme lire les paramètres estimés, en effet les variables ne sont pas indépendantes, nous avons des coefficients partiels maintenant ([18], section 13.1.2) ; comment tester l'influence d'une explicative qualitative, représentée par plusieurs variables indicatrices, dans la régression ; comment prendre en compte l'effet conjoint - l'interaction - des explicatives sur l'endogène.

Nous utilisons un nouveau fichier pour illustrer notre propos. Il s'agit de la base *Auto Pollution Filter Noise* du serveur DASL⁴. On cherche à expliquer le niveau de pollution sonore de véhicules (NOISE, en décibels) à partir de la taille (SIZE, 3 valeurs possibles) et du type de silencieux (TYPE, 2 valeurs). Nous avons choisi d'ignorer la variable SIDE présente dans la base, elle indiquait le côté du véhicule à partir duquel la mesure a été effectuée.

NOISE	SIZE	TYPE
810	S1	T1
820	S1	T1
820	S1	T1
840	S2	T1
840	S2	T1
845	S2	T1
785	S3	T1
790	S3	T1
785	S3	T1
835	S1	T1
835	S1	T1
835	S1	T1
845	S2	T1
855	S2	T1
850	S2	T1
760	S3	T1
760	S3	T1
770	S3	T1
820	S1	T2
820	S1	T2
820	S1	T2
820	S2	T2
820	S2	T2
825	S2	T2
775	S3	T2
775	S3	T2
775	S3	T2
825	S1	T2
825	S1	T2
825	S1	T2
815	S2	T2
825	S2	T2
825	S2	T2
770	S3	T2
760	S3	T2
765	S3	T2

Moyenne de NOISE Étiquettes	T1	T2	Total général
Étiquettes de lig			
S1	825.8333	822.5000	824.1667
S2	845.8333	821.6667	833.7500
S3	775.0000	770.0000	772.5000
Total général	815.5556	804.7222	810.1389

Fig. 4.9. Données NOISE - Valeurs et moyennes conditionnelles

4. <http://lib.stat.cmu.edu/DASL/Datafiles/airpullutionfiltersdat.html>

Première étape pour fixer les idées, nous calculons les moyennes de la variable d'intérêt conditionnellement aux valeurs de SIZE et TYPE (Figure 4.9). Dans ce qui suit, nous noterons μ_{kl} les espérances conditionnelles et \bar{y}_{kl} les moyennes empiriques ; $\mu_{k.}$ (resp. $\mu_{.l}$) est la moyenne de Y conditionnellement aux seules valeurs de la première explicative (resp. la seconde explicative) ; $\mu_{..} = \mu$ est la moyenne globale de Y . Dans notre exemple, nous lisons :

Moyennes	T1	T2	TYPE
S1	$\bar{y}_{11} = 825.8333$	$\bar{y}_{12} = 822.5000$	$\bar{y}_{1.} = 824.1667$
S2	$\bar{y}_{21} = 845.8333$	$\bar{y}_{22} = 821.6667$	$\bar{y}_{2.} = 833.7500$
S3	$\bar{y}_{31} = 775.0000$	$\bar{y}_{32} = 770.0000$	$\bar{y}_{3.} = 772.5000$
SIZE	$\bar{y}_{.1} = 815.5556$	$\bar{y}_{.2} = 804.7222$	$\bar{y}_{..} = \bar{y} = 810.1389$

La moyenne du bruit chez les voitures (SIZE = S1), quel que soit le type de silencieux utilisé, est $\bar{y}_{1.} = 824.1667$; elle est de $\bar{y}_{21} = 845.8333$ chez les véhicules (SIZE = S2) et (TYPE = T1) ; etc.

Manifestement, il y a des différences entre les moyennes conditionnelles. Notre objectif consiste à évaluer jusqu'à quel point et selon que processus ces écarts sont significatifs. Les connaisseurs auront reconnu un problème d'analyse de variance (ANOVA) à 2 facteurs.

Nous avons donc un double objectif en réalisant les régressions sur indicatrices :

1. Voir dans quelle mesure la régression peut répondre à la problématique de l'analyse de variance c.-à-d. évaluer l'impact des exogènes sur la variable d'intérêt Y , en faisant la part entre les explicatives. Mieux même, est-ce qu'il est possible de retrouver les résultats numériques de l'ANOVA ?
2. Montrer de quelle manière et à quelles conditions nous pouvons retrouver le tableau des moyennes conditionnelles ci-dessus à partir des coefficients de la régression.

4.5.1 Régression sur les indicatrices

A l'aide du logiciel R, nous avons mené une analyse de variance sans prise en compte de l'interaction entre les deux explicatives (Figure 4.10). Ce faisant, nous émettons l'hypothèse que l'influence de TYPE (resp. SIZE) sur le bruit des véhicules (NOISE) ne dépend pas de (est la même quelle que soit) la valeur prise par SIZE (resp. TYPE).

Nous constatons que les deux variables impactent significativement sur le bruit au risque 5%.

Ces résultats nous serviront de référence dans cette section.

Effet global des explicatives

Nous créons les indicatrices adéquates pour les variables SIZE et TYPE. Dans les deux cas, nous prenons la première modalité comme référence. Nous avons donc 3 nouvelles colonnes : $S1$, $S2$ et $T2$. Nous réalisons la régression sur ces indicatrices

```

R Console
> #anova sans interaction
> print(summary(aov(NOISE ~ SIZE + TYPE, data = noise.data)))
              Df Sum Sq Mean Sq F value    Pr(>F)
SIZE           2 26051.4 13025.7 150.659 < 2.2e-16 ***
TYPE           1  1056.2   1056.2  12.217  0.001411 **
Residuals     32  2766.7     86.5
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Fig. 4.10. Données NOISE - ANOVA sans interaction entre SIZE et TYPE

$$NOISE = a_1 \times T2 + a_2 \times S3 + a_3 \times S2 + a_0$$

En introduisant les indicatrices de cette manière, nous considérons que les variables n'interagissent pas dans la définition de NOISE.

NOISE	SIZE	TYPE	S2	S3	T2
810	S1	T1	0	0	0
820	S1	T1	0	0	0
820	S1	T1	0	0	0
840	S2	T1	1	0	0
840	S2	T1	1	0	0
845	S2	T1	1	0	0
785	S3	T1	0	1	0
790	S3	T1	0	1	0
785	S3	T1	0	1	0
835	S1	T1	0	0	0
835	S1	T1	0	0	0
835	S1	T1	0	0	0
845	S2	T1	1	0	0
855	S2	T1	1	0	0
850	S2	T1	1	0	0
760	S3	T1	0	1	0
760	S3	T1	0	1	0
770	S3	T1	0	1	0
820	S1	T2	0	0	1
820	S1	T2	0	0	1
820	S1	T2	0	0	1
820	S2	T2	1	0	1
820	S2	T2	1	0	1
825	S2	T2	1	0	1
775	S3	T2	0	1	1
775	S3	T2	0	1	1
775	S3	T2	0	1	1
825	S1	T2	0	0	1
825	S1	T2	0	0	1
825	S1	T2	0	0	1
815	S2	T2	1	0	1
825	S2	T2	1	0	1
825	S2	T2	1	0	1
770	S3	T2	0	1	1
760	S3	T2	0	1	1
765	S3	T2	0	1	1

DROITEREG				
	T2	S3	S2	constante
a^	-10.83333	-51.66667	9.58333	829.58333
sigma^(a^)	3.09943	3.79601	3.79601	3.09943
R^2	0.90739	9.29830	#N/A	#N/A
F	104.51138	32	#N/A	#N/A
SCE	27107.63889	2766.66667	#N/A	#N/A

SCR	
F	104.51138
ddl1	3
ddl2	32
p-value	1.29189E-16

Fig. 4.11. Données NOISE - Régression sur indicatrices, sans prise en compte des interactions

DROITEREG nous fournit tous les éléments nécessaires à l'analyse (Figure 4.11) :

- La régression est de bonne qualité avec $R^2 = 0.90739$. Elle est globalement significative avec un $F = 104.51138$ et une probabilité critique très faible < 0.00001 .
- La variabilité résiduelle, non expliquée par la régression ($SCR = 2766.66667$) est identique à celle de l'ANOVA sans interaction (Figure 4.10), avec les mêmes degrés de liberté $ddl = 36 - 3 - 1 = 32$.
- Par conséquent, la partie expliquée est cohérente également $SCE = 27107.63889$, à la différence que la fonction AOV de R décompose directement la portion dévolue à SIZE ($SCE_{size} = 26051.4$) et à TYPE ($SCE_{type} = 1056.2$).

Première conclusion, notre appréhension du problème est tout à fait conforme avec une analyse de variance à 2 facteurs sans prise en compte des interactions. À vrai dire, ce n'est pas trop une surprise. En lisant la documentation de R, on se rend compte qu'AOV procède par régressions dans les sous-groupes "*AOV Fit an analysis of variance model by a call to lm for each stratum*"; $lm()$ étant la fonction dévolue à la régression.

Décomposition de l'effet des variables

Deuxième étape, il faut discerner ce qui revient à SIZE et TYPE, comme le fait l'analyse de variance (Figure 4.10).

Cas des variables binaires

Si la variable n'est représentée que par une indicatrice, on peut passer directement par le test de significativité du coefficient associé. C'est le cas justement de la variable TYPE. Via la fonction DROITEREG (Figure 4.11), nous avons $\hat{a}_1 = -10.83333$ et $\hat{\sigma}_{\hat{a}_1} = 3.09943$. Nous formons la statistique de test

$$t_{\hat{a}_1} = \frac{-10.83333}{3.09943} = -3.49526$$

Avec un loi de Student à 32 degrés de liberté, nous avons une probabilité critique de 0.001411.

Où est l'analogie avec l'analyse de variance sans interaction ?

On sait que qu'il y a une relation directe entre la loi de Student et la loi de Fisher, $\mathcal{T}(ddl) \equiv \mathcal{F}(1, ddl)$. Dans notre exemple, on constate aisément que

$$(t_{\hat{a}_1})^2 = (-3.49526)^2 = 12.217$$

Correspond exactement au carré moyen associé à la variable TYPE dans l'ANOVA sans interaction (Figure 4.10). En conclusion : au risque 5%, à taille égale des véhicules, on considère que le type de silencieux influe sur leur niveau sonore.

La régression nous fournit une information supplémentaire, le silencieux de type T2 permet de réduire le niveau sonore puisque que $\hat{a}_1 = -10.83333$ est de signe négatif.

Cas des variables à plus de 2 modalités

Quand la variable est représentée par ($q \geq 2$) indicatrices, il faut tester la significativité simultanée des coefficients associés. Dans le cas de la variable SIZE, il s'agit de tester

$$H_0 : a_2 = a_3 = 0$$

$$H_1 : \text{un des deux au moins est non nul}$$

La manière la plus simple de procéder est de réaliser deux régressions : la première avec l'ensemble des p indicatrices, nous obtenons un coefficient de détermination R_1^2 ; la seconde sans les q indicatrices incriminées, nous avons R_0^2 . La statistique de test s'écrit alors ([18], section 10.4)

$$F = \frac{(R_1^2 - R_0^2)/q}{(1 - R_1^2)/(n - p - 1)} \quad (4.5)$$

Sous H_0 , elle suit une loi de Fisher à $(q, n - p - 1)$ degrés de liberté.

Pour les données NOISE, nous disposons déjà de $R_1^2 = 0.90739$ via la régression sur toutes les indicatrices (Figure 4.11). Reste à réaliser la régression sans les indicatrices de SIZE (Figure 4.12), nous avons $R_0^2 = 0.03536$. Nous formons la statistique destinée à évaluer la significativité de SIZE :

$$F_{size} = \frac{(R_1^2 - R_0^2)/q}{(1 - R_1^2)/(n - p - 1)} = \frac{(0.90739 - 0.03536)/2}{(1 - 0.90739)/(36 - 3 - 1)} = 150.659$$

DROITEREG					DROITEREG		
	T2	S3	S2	constante		T2	constante
a^	-10.83333	-51.66667	9.58333	829.58333	a^	-10.83333	815.55556
sigma^(a^)	3.09943	3.79601	3.79601	3.09943	sigma^(a^)	9.70447	6.86209
R^2	0.90739	9.29830	#N/A	#N/A	R^2	0.03536	29.11340
F	104.51138	32	#N/A	#N/A	F	1.24618	34
SCE	27107.63889	2766.66667	#N/A	#N/A	SCE	1056.25000	28818.05556
	SCR					SCR	

Test significativité de SIZE	
F_size	150.659
ddl1	2
ddl2	32
p-value	5.20804E-17

Fig. 4.12. Données NOISE - Tester la significativité de SIZE (2 indicatrices)

Au risque 5%, en contrôlant l'effet de TYPE, on conclut que la taille (SIZE) des véhicules influe sur le niveau sonore puisque la p-value est (largement) inférieure au seuil. Nous retrouvons exactement les valeurs (F, degrés de liberté, p-value) fournies par l'ANOVA sans interaction (Figure 4.10).

TYPE seul n'est pas significatif, TYPE en contrôlant SIZE le devient ???

Notons un résultat important qui va nous éclairer lors de la lecture des coefficients que nous aborderons plus bas. La régression où TYPE intervient seul (à travers l'indicatrice T2, figure 4.12) n'est pas signi-

ficative ($F = 1.24618$, la p-value est 0.2721). En nous référant au tableau des moyennes conditionnelles (Figure 4.9), l'écart de 10.83333 ($\bar{y}_{.1} - \bar{y}_{.2} = 815.5556 - 804.7222$) n'est pas concluant.

Pourtant dans la régression incluant les deux variables (Figure 4.11), TYPE devient significative à travers son coefficient ($t_{\hat{a}_1} = -3.49526$, avec une p-value = 0.001411).

D'où vient cette contradiction?

L'analyse est modifiée par la variable SIZE. En effet, en compilant les écarts entre les deux colonnes *pour chaque ligne* du tableau des moyennes conditionnelles (Figure 4.9) (c.-à-d. $\bar{y}_{11} - \bar{y}_{12} = 825.8333 - 822.5000 = 3.3333$, $\bar{y}_{21} - \bar{y}_{22} = 845.8333 - 821.6667 = 24.1667$ et $\bar{y}_{31} - \bar{y}_{32} = 775.0000 - 770.0000 = 5.0000$), on aboutit à un indicateur d'écart "moyen" qui devient significatif. **L'hypothèse sous-jacente est que TYPE (resp. SIZE) pèse de la même manière sur NOISE, quelle que soit la valeur de SIZE (resp. TYPE).** Ce qui n'est pas tout à fait vrai si l'on se réfère au tableau des moyennes conditionnelles. Nous constatons que cette hypothèse simplificatrice n'est pas sans répercussions sur la lecture des coefficients.

Interprétation des coefficients

La constante $\hat{a}_0 = 829.58333$ devrait correspondre à la moyenne du groupe de référence, à savoir (SIZE = S1) et (TYPE = T1), les modalités dont les indicatrices ont été exclues. Or $\bar{y}_{11} = 825.8333$ dans le tableau des moyennes (Figure 4.9). Cette distorsion est la conséquence de l'hypothèse selon laquelle l'impact de l'une des explicatives ne dépend pas de la valeur prise par l'autre.

Pour nous en convaincre, nous avons construit le tableau des moyennes conditionnelles reconstituées (\tilde{y}_{kl}) à partir des résultats de la régression (Figure 4.11). *Les calculs sont facilités par le fait que nous avons des effectifs totalement équilibrés c.-à-d. $n_{kl} = 6, \forall k, l$.*

Pour rappel, $\hat{a}_0 = 829.58$ est la constante, $\hat{a}_1 = -10.83$ le coefficient de T2, $\hat{a}_2 = -51.67$ celui de S3 et $\hat{a}_3 = 9.58$ celui de S2.

Moyennes	T1	T2	Total
S1	$\tilde{y}_{11} = \hat{a}_0 = 829.58$	$\tilde{y}_{12} = \hat{a}_0 + \hat{a}_1 = 818.75$	$\tilde{y}_{1.} = \frac{829.58 + 818.75}{2} = 824.17$
S2	$\tilde{y}_{21} = \hat{a}_0 + \hat{a}_3 = 839.17$	$\tilde{y}_{22} = \hat{a}_0 + \hat{a}_1 + \hat{a}_3 = 828.33$	$\tilde{y}_{2.} = \frac{839.17 + 828.33}{2} = 833.75$
S3	$\tilde{y}_{31} = \hat{a}_0 + \hat{a}_2 = 777.92$	$\tilde{y}_{32} = \hat{a}_0 + \hat{a}_2 + \hat{a}_1 = 767.08$	$\tilde{y}_{3.} = \frac{777.92 + 767.08}{2} = 772.50$
Total	$\tilde{y}_{.1} = \frac{829.58 + 839.17 + 777.92}{3} = 815.56$	$\tilde{y}_{.2} = \frac{818.75 + 828.33 + 767.08}{3} = 804.72$	$\tilde{y}_{..} = \bar{y} = 810.14$

Nous récapitulons les résultats en confrontant les moyennes calculées et les moyennes reconstituées (Figure 4.13) :

- Une première information très importante saute aux yeux : les moyennes marginales sont parfaitement reconstituées, tant pour SIZE ($\bar{y}_{k.} = \tilde{y}_{k.}, \forall k$) que pour TYPE ($\bar{y}_{.l} = \tilde{y}_{.l}, \forall l$).
- Il en est de même en ce que concerne la moyenne globale $\bar{y}_{..} = \tilde{y}_{..} = 810.14$
- Les divergences apparaissent lorsque nous calculons les moyennes conditionnelles.

Tableau des Moyennes calculées				Tableau des Moyennes reconstituées			
	T1	T2	Total		T1	T2	Total
S1	825.83	822.50	824.17	Ecart	S1	829.58	818.75
S2	845.83	821.67	833.75	3.33	S2	839.17	828.33
S3	775.00	770.00	772.50	24.17	S3	777.92	767.08
Total	815.56	804.72	810.14	5.00	Total	815.56	804.72
SCE				SCE			
16002.78 10852.78				13025.69 13025.69			

Fig. 4.13. Données NOISE - Moyennes reconstituées, régression sans interaction

- Preuve que nous ne tenons pas compte des interactions dans la régressions, nous constatons que les écarts sont constants entre les deux colonnes $T1$ et $T2$ (colonne écarts) quelle que soit la valeur de $SIZE$ (S1, S2 ou S3) c.-à-d. $(\tilde{y}_{k1} - \tilde{y}_{k2}) = 10.83, \forall k$.
- Pour $SIZE$, la démonstration est un peu plus difficile. Il faut calculer la sommes des carrés des écarts (variabilité expliquée) de $SIZE$ selon les valeurs de $TYPE$, nous avons $SCE_l = \sum_k 6 \times (\tilde{y}_{kl} - \tilde{y}_{.l})^2 = 13025.69, \forall l$.
- Ces deux résultats sont en contradiction avec ceux obtenus via le tableau des moyennes calculées directement à partir des données, moyennes qui tiennent compte des interactions entre $SIZE$ et $TYPE$. Les écarts ne sont pas constants d'une ligne à l'autre, les SCE ne sont pas les mêmes d'une colonne à l'autre.

4.5.2 Prise en compte des interactions

De nouveau avec R, nous avons réalisé une ANOVA en prenant en compte les interactions entre $SIZE$ et $TYPE$ cette fois-ci (Figure 4.14). Maintenant, **nous considérons que l'effet de $TYPE$ (resp. $SIZE$) sur le bruit peut dépendre de la valeur prise par $SIZE$ (resp. $TYPE$).**

Voyons de quelle manière nous pouvons retrouver ces résultats à l'aide de la régression.

```

R Console
> #anova avec interaction
> print(summary(aov(NOISE ~ SIZE * TYPE, data = noise.data)))
              Df Sum Sq Mean Sq  F value    Pr(>F)
SIZE           2 26051.4 13025.7 199.1189 < 2.2e-16 ***
TYPE           1  1056.2   1056.2  16.1465 0.0003631 ***
SIZE:TYPE       2    804.2    402.1   6.1465 0.0057915 **
Residuals     30   1962.5     65.4
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Fig. 4.14. Données NOISE - ANOVA avec interaction entre $SIZE$ et $TYPE$

4.5.3 Ajout de nouvelles indicatrices

Prendre en compte l'interaction signifie que l'impact de $TYPE$ (resp. $SIZE$) dépend de la valeur prise par $SIZE$ (resp. $TYPE$). Pour ce faire, nous ajoutons de nouvelles variables explicatives dans la régression,

elles sont formées à partir du produit des indicatrices. Concrètement, nous travaillons maintenant sur le modèle :

$$NOISE = b_1 \times S3 * T2 + b_2 \times S2 * T2 + b_3 \times T2 + b_4 \times S3 + b_5 \times S2 + b_0$$

La variable $S3 * T2$ (resp. $S2 * T2$) est aussi une indicatrice. Elle prend la valeur 1 si et seulement si les indicatrices $S3$ et $T2$ (resp. $S2$ et $T2$) prennent simultanément la valeur 1. Elle est égale à zéro dans tous les autres cas.

Voyons deux exemples pour bien situer le rôle des coefficients. Par rapport à la référence ($S1, T1$) avec un niveau de bruit moyen b_0 :

- un véhicule ($S3, T1$) présente un niveau de bruit ($b_0 + b_4$) ;
- un véhicule ($S3, T2$) présente un niveau de bruit ($b_0 + b_4 + b_3 + b_1$).

Nous verrons plus loin que notre modèle étant *saturé*, c.-à-d. tous les effets possibles sont pris en compte dans la régression, il sera possible de reconstituer à l'identique le tableau des moyennes conditionnelles à partir des coefficients du modèle.

NOISE	SIZE	TYPE	S2	S3	T2	S2*T2	S3*T2
810	S1	T1	0	0	0	0	0
820	S1	T1	0	0	0	0	0
820	S1	T1	0	0	0	0	0
840	S2	T1	1	0	0	0	0
840	S2	T1	1	0	0	0	0
845	S2	T1	1	0	0	0	0
785	S3	T1	0	1	0	0	0
790	S3	T1	0	1	0	0	0
785	S3	T1	0	1	0	0	0
835	S1	T1	0	0	0	0	0
835	S1	T1	0	0	0	0	0
835	S1	T1	0	0	0	0	0
845	S2	T1	1	0	0	0	0
855	S2	T1	1	0	0	0	0
850	S2	T1	1	0	0	0	0
760	S3	T1	0	1	0	0	0
760	S3	T1	0	1	0	0	0
770	S3	T1	0	1	0	0	0
820	S1	T2	0	0	1	0	0
820	S1	T2	0	0	1	0	0
820	S2	T2	1	0	1	1	0
820	S2	T2	1	0	1	1	0
825	S2	T2	1	0	1	1	0
775	S3	T2	0	1	1	0	1
775	S3	T2	0	1	1	0	1
775	S3	T2	0	1	1	0	1
825	S1	T2	0	0	1	0	0
825	S1	T2	0	0	1	0	0
825	S1	T2	0	0	1	0	0
815	S2	T2	1	0	1	1	0
825	S2	T2	1	0	1	1	0
825	S2	T2	1	0	1	1	0
770	S3	T2	0	1	1	0	1
760	S3	T2	0	1	1	0	1
765	S3	T2	0	1	1	0	1

DROITEREG						
	S3*T2	S2*T2	T2	S3	S2	constante
a^	-1.6667	-20.8333	-3.3333	-50.8333	20.0000	825.8333
sigma^2(a^)	6.6039	6.6039	4.6696	4.6696	4.6696	3.3019
R^2	0.93431	8.0881	#N/A	#N/A	#N/A	#N/A
F	85.3355	30	#N/A	#N/A	#N/A	#N/A
SCE	27911.8056	1962.5000	#N/A	#N/A	#N/A	#N/A
SCR						
F	85.3355					
ddl1	5					
ddl2	30					
p-value	8.1995E-17					

Fig. 4.15. Données NOISE - Régression sur indicatrices, avec prise en compte des interactions

Pour l'heure, lançons la fonction DROITEREG pour obtenir les estimations (Figure 4.15) :

- La régression est de très bonne qualité avec un $R^2 = 0.93431$.

- Elle est globalement significative à 5% avec une p-value < 0.00001 .
- Par rapport à l'ANOVA avec interaction (Figure 4.14), nous constatons que la variabilité résiduelle, non expliquée par le modèle, est la même : $SCR = 1962.50$, avec les mêmes degrés de liberté $ddl = 30$. Notre spécification de la régression semble donc convenir.

4.5.4 Tester la significativité de l'interaction

L'étape suivante consiste à vérifier la pertinence de l'introduction du terme d'interaction dans notre analyse (toujours à 5%). L'ANOVA l'affirme avec $F_{size:type} = 6.1465$ et une p-value de 0.0057915 (Figure 4.14). Voyons si la régression fournit les mêmes résultats.

DROITEREG - H1						
	S3*T2	S2*T2	T2	S3	S2	constante
a^	-1.6667	-20.8333	-3.3333	-50.8333	20.0000	825.8333
sigma^(a^)	6.6039	6.6039	4.6696	4.6696	4.6696	3.3019
R^2_1	0.93431	8.0881	#N/A	#N/A	#N/A	#N/A
F	85.3355	30	#N/A	#N/A	#N/A	#N/A
SCE	27911.8056	1962.5000	#N/A	#N/A	#N/A	#N/A

Test interaction	
R^2_1	0.93431
R^2_0	0.90739

q	2
n	36
p	5

DROITEREG - H0			
	T2	S3	S2
a^	-10.8333	-51.6667	9.5833
sigma^(a^)	3.0994	3.7960	3.7960
R^2_0	0.90739	9.2983	#N/A
F	104.5114	32	#N/A
SCE	27107.6389	2766.6667	#N/A

F	6.1465
ddl1	2
ddl2	30
p-value	0.0057915

Fig. 4.16. Données NOISE - Tester l'interaction entre SIZE et TYPE

Pour tester la pertinence de l'interaction, nous devons tester la significativité simultanée des coefficients de $(S3 * T2)$ et $(S2 * T2)$ (Figure 4.16). Nous appliquons une démarche analogue à celle présentée précédemment pour tester la nullité des coefficients de plusieurs indicatrices (≥ 2) associées à une variable qualitative (page 103) :

- La régression sur toutes les variables, y compris les ($q = 2$) termes d'interaction $(S3 * T2, S2 * T2)$, présente un coefficient de détermination $R_1^2 = 0.93431$.
- La régression sans les termes d'interaction propose un $R_0^2 = 0.90739$.
- R_1^2 est forcément supérieur à R_0^2 puisque que nous avons des variables additionnelles, mais l'est-il significativement ? Pour le savoir, nous utilisons la statistique

$$F_{size:type} = \frac{(R_1^2 - R_0^2)/q}{(1 - R_1^2)/(n - p - 1)} = \frac{(0.93431 - 0.90739)/2}{(1 - 0.93431)/(36 - 5 - 1)} = 6.1465$$

- Avec une distribution de Fisher $\mathcal{F}(2, 30)$, nous obtenons une p-value de 0.0057915. Les termes d'interaction sont justifiés dans la régression.
- Et par la même occasion, nous constatons que nous avons reproduit à l'identique les résultats de l'ANOVA avec interaction (Figure 4.14). Nous sommes contents.

4.5.5 Interprétation des coefficients

Dernière étape de notre exploration, voyons de quelle manière nous pouvons obtenir les "vraies" moyennes conditionnelles à partir de la régression. Cela est possible maintenant parce que nous prenons en compte tous les effets des exogènes sur l'endogène. On dit que le *modèle est saturé*.

Pour éviter les renvois répétés vers d'autres pages, récapitulons les coefficients estimés de la régression :

Variable	S3*T2	S2*T2	T2	S3	S2	Constante
Coefficient	\hat{b}_1	\hat{b}_2	\hat{b}_3	\hat{b}_4	\hat{b}_5	\hat{b}_0
-	-1.67	-20.83	-3.33	-50.83	20.00	825.83

Première vérification immédiate, la constante $\hat{b}_0 = 825.83$ correspond bien à la moyenne conditionnelle de Y pour la combinaison des modalités exclues (S1, T1). C'est plutôt rassurant. Reste à reconstituer les autres moyennes conditionnelles à partir de l'association de ces coefficients. Notons la logique d'obtention des moyennes à partir des \hat{b}_j .

Enfin, nous gardons toujours à l'esprit que les effectifs sont totalement équilibrés, $n_{kl} = 6, \forall k, l$; cela explique les formules simplifiées utilisées pour le calcul des moyennes marginales.

\bar{y}_{kl}	T1	T2	TYPE
S1	$\bar{y}_{11} = \hat{b}_0 = 825.83$	$\bar{y}_{12} = \hat{b}_0 + \hat{b}_3 = 822.50$	$\bar{y}_{1.} = \frac{825.83+822.50}{2} = 824.17$
S2	$\bar{y}_{21} = \hat{b}_0 + \hat{b}_5 = 845.83$	$\bar{y}_{22} = \hat{b}_0 + \hat{b}_5 + \hat{b}_3 + \hat{b}_2 = 821.67$	$\bar{y}_{2.} = \frac{845.83+821.67}{2} = 833.75$
S3	$\bar{y}_{31} = \hat{b}_0 + \hat{b}_4 = 775.00$	$\bar{y}_{32} = \hat{b}_0 + \hat{b}_4 + \hat{b}_3 + \hat{b}_1 = 770.00$	$\bar{y}_{3.} = \frac{775.00+770.00}{2} = 772.50$
SIZE	$\bar{y}_{.1} = \frac{825.83+845.83+775.00}{3} = 815.56$	$\bar{y}_{.2} = \frac{822.50+821.67+770.00}{3} = 804.72$	$\bar{y}_{..} = \bar{y} = 810.14$

Le tableau des moyennes conditionnelles est parfaitement reconstitué!

4.6 Régression avec un mix d'explicatives qualitatives et quantitatives

Nous avons détaillé l'interprétation des coefficients dans le cas d'un mélange d'explicatives qualitatives et quantitatives dans notre support consacré à la régression ([18], section 13.3). Nous y essayons d'expliquer le niveau de salaire à partir du genre (homme vs. femme) et du nombre d'années d'études.

Nous suivrons la même démarche dans cette section. Nous la compléterons avec l'étude de l'interaction entre ces variables. Nous utiliserons cependant un autre exemple pour varier les plaisirs. Avec les données CARBURANT, nous essayons d'expliquer la consommation des véhicules⁵ (CONSO, en litres par 100 km) en fonction du type de carburant [CARBURANT : gazole (0) ou essence (1)] et la cylindrée (en cm^3). Le fichier représente un échantillon de $n = 32$ observations extrait de la base **Automobile Data Set**⁶ accessible sur le serveur *UCI Machine Learning Repository*.

5. C'est vrai qu'il adore les voitures le monsieur, j'ai du être garagiste dans une autre vie.

6. <http://archive.ics.uci.edu/ml/datasets/Automobile>, les unités ont été converties.

4.6.1 Interprétation des coefficients

Régression avec la variable CARBURANT seule

Dans un premier temps, nous tentons d'expliquer la consommation uniquement à l'aide du carburant utilisé. Le modèle s'écrit

$$CONSO = a_1 \times CARBURANT + a_0$$

CARBURANT étant binaire, nous savons que a_0 est la moyenne de la consommation des véhicules fonctionnant au gazole; a_1 représente alors l'écart entre cette moyenne et celle des véhicules essence.

conso.ville	carburant	carburant	cylindrée
6.50	gazole	0	1590
6.36	gazole	0	1590
7.84	gazole	0	1803
9.05	gazole	0	2376
9.41	gazole	0	2491
8.40	gazole	0	2491
9.41	gazole	0	2491
8.40	gazole	0	2491
8.40	gazole	0	2491
10.69	gazole	0	2999
10.69	gazole	0	2999
10.69	gazole	0	2999
10.90	gazole	0	2999
9.80	essence	1	1606
9.80	essence	1	1606
9.80	essence	1	1606
9.80	essence	1	1606
9.80	essence	1	1770
10.23	essence	1	1770
10.23	essence	1	1803
10.23	essence	1	1803
12.38	essence	1	1983
12.38	essence	1	1983
12.38	essence	1	2294
12.38	essence	1	2311
12.38	essence	1	2311
12.38	essence	1	2556
12.38	essence	1	2556
12.38	essence	1	2556
12.38	essence	1	2556
12.38	essence	1	2556
14.50	essence	1	2966

a^	DROITEREG - 1			F	22.428
	carburant	constante			
	2.49316	8.98000			
	0.52644	0.40565			
	0.42779	1.46260			
sigma^(a^)	22.42823	30	ddl1	1	
R^2	47.97849	64.17601	ddl2	30	
t	4.73585		p-value	4.90725E-05	
p-value	4.90725E-05				

a^	DROITEREG - 2			F	187.962
	cylindrée	carburant	constante		
	0.00296	3.47220	1.74761		
	0.00021	0.20152	0.52854		
	0.92838	0.52629	#N/A		
sigma^(a^)	187.96236	29	ddl1	2	
R^2	104.12219	8.03231	ddl2	29	
t	14.23736	17.22964	p-value	2.49945E-17	
p-value	1.28036E-14	8.9472E-17			

Fig. 4.17. Données CARBURANT - Régressions sans prise en compte de l'interaction

Voyons ce que nous fournit DROITEREG (Figure 4.17) :

- Le modèle n'explique que $R^2 = 42.8\%$ de la variance de l'endogène.
- Mais il est globalement significatif à 5% avec un $F = 22.428$ et une p-value de 4.90725×10^{-5} .
- La pente de la droite, qui représente l'écart entre les moyennes conditionnelles de consommation des véhicules essence et diesel, est significative également ($t_{\hat{a}_1} = 4.73585$, avec une p-value de

4.90725×10^{-5}). Ce résultats n'est pas étonnant. Tester le modèle dans sa globalité et tester la pente sont équivalents dans la régression simple.

- Nous pouvons donc dire que les véhicules à essence consomment significativement plus que les diesels. L'écart est estimé à $\hat{a}_1 = 2.49316$ litres au 100 km.
- Pour obtenir les consommations moyennes conditionnelles, nous formons :

$$\begin{aligned}\bar{y}_{gazole} &= \hat{a}_0 = 8.98 \\ \bar{y}_{essence} &= \hat{a}_0 + \hat{a}_1 = 8.98 + 2.49316 = 11.47316\end{aligned}$$

Régression avec CARBURANT et CYLINDRÉE

Nous souhaitons approfondir l'analyse en introduisant la variable CYLINDRÉE. Le modèle s'écrit maintenant :

$$CONSO = b_1 \times CYLINDREE + b_2 \times CARBURANT + b_0$$

L'estimation à l'aide de DROITREG fournit de nouveaux résultats (Figure 4.17) :

- La régression est de meilleure qualité avec un $R^2 = 0.92838$, elle est globalement significative bien évidemment ($F = 187.962$, p-value = 2.49945×10^{-7}).
- Les deux variables CYLINDRÉE et CARBURANT sont largement significatives à 5%.
- La constante $\hat{b}_0 = 1.74761$ n'est pas très intéressante. Elle correspondrait à la consommation moyenne des véhicules de cylindrée nulle fonctionnant au gazole.
- En revanche, le coefficient de CARBURANT, $\hat{b}_2 = 3.47220$, revêt une signification très intéressante. *A cylindrée égale*, les essences consomment 3.47220 litres au 100km de plus que les gazoles. **L'idée est que cet écart reste constant quelle que soit la cylindrée des véhicules.**
- De même le coefficient de CARBURANT $\hat{b}_1 = 0.00296$ propose une lecture très instructive : *à type de carburant égal*, l'augmentation de la cylindrée d' 1 cm^3 entraîne une augmentation de la consommation de 0.00296 litres au 100 km. **On considère ainsi que la variation est identique chez les essences et les gazoles.**

Ces deux hypothèses sous-jacentes à la régression sans interactions introduisent une lecture intéressante des coefficients. Il permettent d'analyser l'impact de chaque explicative en contrôlant l'effet des autres variables. Mais elles en limitent la portée également. Nous n'explorons pas la configuration selon laquelle l'effet de l'une dépend des valeurs prises par l'autre. C'est précisément l'objet de l'introduction des termes d'interaction.

4.6.2 Prise en compte des interactions

Pour prendre en compte l'interaction, nous introduisons une nouvelle variable formée par le produit de l'indicatrice et de l'explicative quantitative. Si la variable qualitative comporte plus de deux modalités, il faudra former le produit de chaque indicatrice avec la variable quantitative. Dans notre exemple, le modèle devient :

$$CONSO = c_1 \times CARB \times CYL + c_2 \times CYLINDREE + c_3 \times CARBURANT + c_0$$

conso.ville	carburant	carburant	cylindrée	carb*cyl
6.50	gazole	0	1590	0
6.36	gazole	0	1590	0
7.84	gazole	0	1803	0
9.05	gazole	0	2376	0
9.41	gazole	0	2491	0
8.40	gazole	0	2491	0
9.41	gazole	0	2491	0
8.40	gazole	0	2491	0
8.40	gazole	0	2491	0
10.69	gazole	0	2999	0
10.69	gazole	0	2999	0
10.69	gazole	0	2999	0
10.90	gazole	0	2999	0
9.80	essence	1	1606	1606
9.80	essence	1	1606	1606
9.80	essence	1	1606	1606
9.80	essence	1	1606	1606
9.80	essence	1	1770	1770
10.23	essence	1	1770	1770
10.23	essence	1	1803	1803
10.23	essence	1	1803	1803
12.38	essence	1	1983	1983
12.38	essence	1	1983	1983
12.38	essence	1	2294	2294
12.38	essence	1	2311	2311
12.38	essence	1	2311	2311
12.38	essence	1	2556	2556
12.38	essence	1	2556	2556
12.38	essence	1	2556	2556
12.38	essence	1	2556	2556
12.38	essence	1	2556	2556
14.50	essence	1	2966	2966

DROITEREG				
	carb*cyl	cylindrée	carburant	constante
a^	0.000162	0.002872	3.10093	1.95224
sigma^2(a^)	0.00042	0.00030	0.98541	0.75502
R^2	0.92876	0.53419	#N/A	#N/A
	121.67774	28	#N/A	#N/A
	104.16452	7.98998	#N/A	#N/A
t	0.38516	9.49256	3.14683	
p-value	0.70303	0.00000	0.00389	

Fig. 4.18. Données CARBURANT - Régressions avec prise en compte de l'interaction

Les valeurs prises par les termes d'interactions sont très particulières (Figure 4.18). Elle sont égales à la variable quantitative lorsqu'elle correspond à l'indicatrice concernée, 0 dans tous les autres cas. C'est comme si les données ont été morcelées et, pour chaque morceau, nous analysons de manière (plus ou moins) séparée l'impact de l'exogène quantitative. Nous approfondirons cette idée lorsque nous ferons le parallèle entre la régression avec interaction et la comparaison de régressions séparées.

Pour l'heure, voyons les résultats de notre régression (Figure 4.18) :

- Le coefficient de détermination est très légèrement amélioré ($R^2 = 0.92876$).
- Attention, le coefficient de CARBURANT $\hat{c}_3 = 3.10093$ correspond au décalage de consommation lorsque les véhicules sont de cylindrée nulle. Dans le cas présent, sa lecture n'est pas très intéressante.
- Parce qu'il y a interaction, le décalage du niveau de consommation selon le carburant dépend de la valeur de la cylindrée. Par exemple, pour les véhicules de 1500 cm^3 , l'écart est de $(0.000162 \times 0 \times 1500 + 0.002872 \times 1500 + 3.10093 \times 0 + 1.95224) - (0.000162 \times 1 \times 1500 + 0.002872 \times 1500 + 3.10093 \times 1 + 1.95224) = 3.34452$; il devient 3.42572 si la cylindrée passe à 2000 cm^3 .

- Le coefficient de CYLINDREE ($\hat{c}_2 = 0.002872$) correspond à l'augmentation de la consommation consécutive à un accroissement de 1 cm^3 de la cylindrée du moteur *pour les véhicules gazole*.
- Si nous souhaitons obtenir la même information pour les véhicules essence, il faut former $\hat{c}_2 + \hat{c}_1 = 0.002872 + 0.000162 = 0.003034$.
- Pour savoir si ce différentiel de comportement entre les essences et les gazoles est bien réel, il faut alors tester la significativité de c_1 . Dans notre exemple, on se rend compte qu'il ne l'est pas avec $t_{\hat{c}_1} = 0.38516$ et une p-value de 0.70303. Les données ne contredisent pas l'hypothèse ($H_0 : c_1 = 0$), on peut considérer que le surcroît de consommation consécutif à une augmentation de cylindrée est le même chez les gazoles et les essences.

Nous pouvons nous contenter de la régression sans interaction dans l'explication de la consommation à partir du type de carburant et de la cylindrée.

Remarque 25 (Explicative qualitative à plus de 2 modalités). Dans le cas où l'explicative qualitative est exprimée par plusieurs indicatrices, il faudrait tester la nullité simultanée des coefficients associés à tous les termes d'interactions.

4.6.3 Lien avec la comparaison de régressions

La régression avec un mix d'exogènes qualitatives et quantitatives a de fortes connexions avec la comparaison de régressions ([18], chapitre 8) et l'analyse des ruptures de structures (chapitre 5). Le rapprochement est facilité par le fait que nous n'avons que deux exogènes dans notre exemple illustratif, l'une qualitative et l'autre quantitative. Nous pouvons représenter graphiquement les deux régressions (Figure 4.19) :

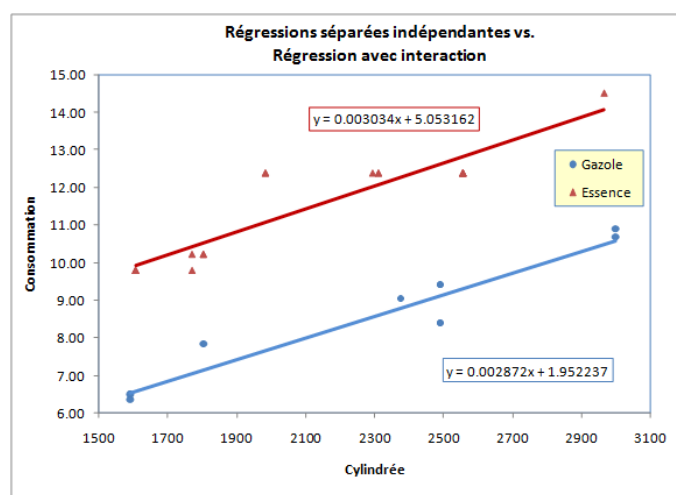


Fig. 4.19. Données CARBURANT - Régressions séparées indépendantes - Prise en compte de l'interaction

- Nous avons autant de régressions que de modalités de l'explicative qualitative. Ici, nous avons 2 modèles, l'un pour les véhicules fonctionnant au **gazole** avec $a_{gazole} \times CYLINDREE + b_{gazole}$,

$$M_{gazole} : CONSO = 0.002872 \times CYLINDREE + 1.952237$$

L'autre pour ceux roulant à l'**essence**, $a_{essence} \times CYLINDREE + b_{essence}$

$$M_{essence} : CONSO = 0.003034 \times CYLINDREE + 5.053162$$

- Le deux régressions ont quasiment la même pente, c'est le signe de l'absence d'interactions. Dans le cas contraire, l'écart entre les régressions ne serait pas constant, les droites pourraient même se croiser.
- Par rapport aux coefficients \hat{c}_j de la régression avec prise en compte des interactions (Figure 4.18), voyons les relations :
 1. La constante de M_{gazole} coïncide avec la constante de la régression, soit $\hat{b}_{gazole} = \hat{c}_0 = 1.952237$. Son interprétation est purement théorique, elle correspondrait à la consommation des véhicules gazole de cylindrée nulle.
 2. La pente de M_{gazole} est identique au coefficient de CYLINDREE $\hat{a}_{gazole} = \hat{c}_2 = 0.002872$. Le mécanisme est relativement simple, lorsque CARBURANT = GAZOLE, CARBURANT vaut 0, le terme d'interaction $CARB * CYL$ également. De fait, le coefficient c_2 revient à mesurer l'impact de la cylindrée uniquement chez les véhicules roulant au gazole.
 3. Passons à la régression chez les véhicules à essence $M_{essence}$. La constante mesure la consommation pour les véhicules de cylindrée nulle, le décalage entre les deux constantes correspond au coefficient de CARBURANT dans la régression avec interaction : $\hat{b}_{essence} - \hat{b}_{gazole} = 5.053162 - 1.952237 = 3.10093 = \hat{c}_3$.
 4. Concernant la pente, nous l'avons déjà mis en exergue précédemment, nous constatons que $\hat{a}_{essence} = \hat{c}_2 + \hat{c}_1 = 0.002872 + 0.000162 = 0.003034$.
 5. Le terme d'interaction permet de situer la concomitance entre les pentes. Si le coefficient associé est nul, l'écart entre les droites serait stable. De fait, **dans la régression sans terme d'interaction, nous les obligeons explicitement à être parallèles**. Les régressions séparées seraient contraintes par cette exigence (Figure 4.20 ; les coefficients sont à comparer avec ceux de la régression sans interaction $CONSO = 0.00296 \times CYLINDREE + 3.47220 \times CARBURANT + 1.74761$, figure 4.17).

Bien évidemment, la lecture est moins facile lorsque l'explicative qualitative possède plusieurs (> 2) modalités ou lorsque nous avons plusieurs explicatives quantitatives. Mais fondamentalement, les mécanismes sous-jacents sont identiques.

4.7 Sélection de variables en présence d'exogènes qualitatives

L'introduction d'exogènes qualitatives représentées par plusieurs indicatrices pose une question clé dans la sélection de variables : doit-on traiter ces indicatrices en bloc ou individuellement ?

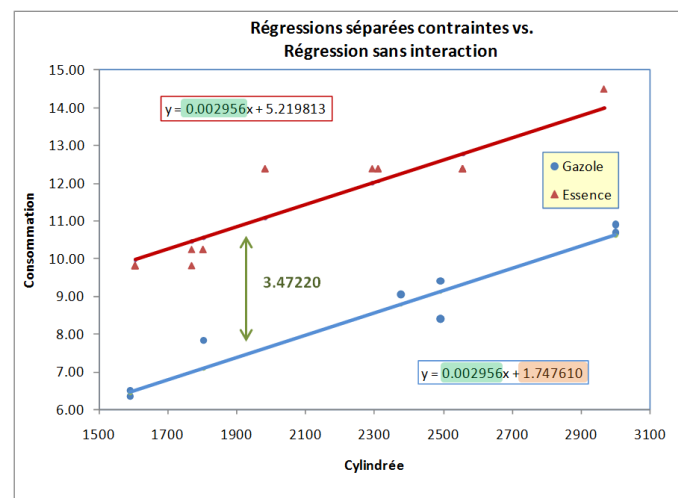


Fig. 4.20. Régressions séparées avec contrainte d'égalité des pentes - Non prise en compte de l'interaction

4.7.1 Traitement groupé des indicatrices

Risque	FUMER	SEXE	IMC	F ANCIEN	F OCCAS	F REGULIER
0.1	JAMAIS	1	29.38	0	0	0
0.8	JAMAIS	1	21.93	0	0	0
0.8	ANCIEN	1	20.30	1	0	0
1.5	JAMAIS	1	30.47	0	0	0
1.7	JAMAIS	1	21.25	0	0	0
1.9	ANCIEN	0	28.86	1	0	0
2	OCCASIONNEL	0	30.48	0	1	0
2	JAMAIS	0	23.57	0	0	0
2.1	JAMAIS	1	29.87	0	0	0
2.2	REGULIER	1	28.26	0	0	1
2.2	REGULIER	1	28.26	0	0	1
2.3	JAMAIS	0	26.02	0	0	0
2.6	JAMAIS	1	24.42	0	0	0
2.6	ANCIEN	0	29.24	1	0	0
2.7	REGULIER	0	26.87	0	0	1
2.7	JAMAIS	1	32.71	0	0	0
2.7	ANCIEN	1	22.80	1	0	0
2.9	REGULIER	1	23.67	0	0	1
2.9	ANCIEN	0	23.87	1	0	0
2.9	ANCIEN	1	29.67	1	0	0
3	ANCIEN	0	36.22	1	0	0
3.3	REGULIER	1	24.77	0	0	1
3.3	ANCIEN	0	30.05	1	0	0
3.5	REGULIER	1	25.52	0	0	1
3.5	REGULIER	0	24.54	0	0	1
3.6	JAMAIS	0	23.66	0	0	0
3.6	JAMAIS	1	25.45	0	0	0
3.8	REGULIER	0	21.68	0	0	1
3.9	OCCASIONNEL	1	28.28	0	1	0
4	JAMAIS	1	25.85	0	0	0
4	ANCIEN	0	38.23	1	0	0
4.4	JAMAIS	1	40.55	0	0	0
4.7	JAMAIS	1	25.99	0	0	0
4.9	REGULIER	1	31.87	0	0	1
4.9	REGULIER	0	26.48	0	0	1

Fig. 4.21. Données CARDIO - Expliquer le risque de maladie cardio-vasculaire

La première approche consiste à traiter en bloc les indicatrices relatives à une exogène tout simplement parce que nous sommes dans un schéma de sélection de variables. On cherche à identifier celles qui sont pertinentes pour expliquer l'endogène. Dissocier les indicatrices d'une exogène qualitative ne paraît pas judicieux car l'interprétation des coefficients qui en découlerait devient hasardeux (*apparemment*, nous reviendrons sur cet aspect dans la section suivante).

Prenons un schéma backward pour fixer les idées (section 3.2.2). Pour rappel, il s'agit d'une procédure de retrait pas-à-pas des variables jusqu'à ce qu'elles soient toutes significatives à un risque α qui constitue le paramètre de l'algorithme. À chaque étape, on retire la variable la moins significative c.-à-d. dont la p-value est la plus élevée, et supérieure à α bien évidemment, puis on relance la régression.

Pour les variables quantitatives, la décision est basée sur le t de Student du test de significativité ([18], section 10.3). Pour les variables qualitatives, on se base sur le F de Fisher de nullité simultanée des coefficients associées aux indicatrices ([18], section 10.4).

Nous utilisons le fichier "CARDIO" pour illustrer la procédure (Figure 4.21, $n = 35$ observations). On souhaite expliquer le risque de maladie cardio-vasculaire. Il s'agit d'une note comprise entre 0 et 5 attribuée par des experts à la suite de la lecture du dossier médical de patients, plus la note est élevée, plus le risque de maladie est élevé. Les variables explicatives candidates sont : le sexe, codée 0 (femme) / 1 (homme); l'indice de masse corporelle (IMC, voir http://fr.wikipedia.org/wiki/Indice_de_masse_corporelle); le comportement par rapport au tabac (FUMER), avec 4 valeurs possibles ("jamais", "ancien", "occasionnel", "régulier"). Cette dernière a été recodée en 3 indicatrices, la modalité "jamais" servant de référence.

Nous réalisons tous les tests à $\alpha = 10\%$ dans tout ce qui suit.

La régression avec la totalité des variables fournit une régression qui n'est pas globalement significative ($F = 1.4502$, p-value = 0.236412) et un $R^2 = 0.200021$. Du côté des explicatives, l'IMC et l'indicatrice (FUMER = REGULIER) sont significatives (Figure 4.22). Ce dernier résultat n'a pas trop de sens pour nous puisque nous voulons traiter la variable FUMER comme un tout. Nous calculons donc le F-partiel et la p-value associée pour chaque exogène. Pour les variables individuelles (quantitatives ou binaires), il s'agit tout simplement du carré du t de Student. Pour FUMER, il s'agit de confronter le coefficient de détermination de la régression comprenant toutes les variables avec celui de la régression avec les seules variables IMC et SEXE (Figure 4.23, $R^2 = 0.075983$).

Nous formons le tableau de F-partiels.

Variable	F	ddl_1	ddl_2	p-value
SEXE	$F = (-0.702823)^2 = 0.493960$	1	29	0.48777
IMC	$F = (1.810742)^2 = 3.278787$	1	29	0.080553
FUMER	$F = \frac{(0.200021 - 0.075983)/3}{(1 - 0.200021)/29} = 1.498832$	3	29	0.235665

Ce sont les résultats que l'on obtiendrait avec la PROC GLM de SAS par exemple (Figure 4.24).

Global results	
Endogenous attribute	Risque
Examples	35
R ²	0.200021
Adjusted-R ²	0.062094
Sigma error	1.109763
F-Test (5,29)	1.4502 (0.236412)

Analysis of variance					
Source	xSS	d.f.	xMS	F	p-value
Regression	8.9301	5	1.7860	1.4502	0.2364
Residual	35.7156	29	1.2316		
Total	44.6457	34			

Coefficients				
Attribute	Coef.	std	t(29)	p-value
Intercept	0.723624	1.235823	0.585540	0.562714
SEXE	-0.289566	0.412005	-0.702823	0.487770
IMC	0.076497	0.042246	1.810742	0.080553
F_ANCIEN	-0.152774	0.513437	-0.297553	0.768166
F_OCCAS	0.123681	0.851808	0.145198	0.885559
F_REGULIER	0.836509	0.467957	1.787576	0.084300

Fig. 4.22. Données CARDIO - Régression avec SEXE, IMC, et toutes les indicatrices de FUMER

Global results	
Endogenous attribute	Risque
Examples	35
R ²	0.075983
Adjusted-R ²	0.018232
Sigma error	1.135416
F-Test (2,32)	1.3157 (0.282410)

Analysis of variance					
Source	xSS	d.f.	xMS	F	p-value
Regression	3.3923	2	1.6962	1.3157	0.2824
Residual	41.2534	32	1.2892		
Total	44.6457	34			

Coefficients				
Attribute	Coef.	std	t(32)	p-value
Intercept	1.355763	1.210760	1.119762	0.271150
SEXE	-0.259158	0.392674	-0.659985	0.513986
IMC	0.060342	0.042101	1.433264	0.161481

Fig. 4.23. Données CARDIO - Régression avec SEXE et IMC

La variable la moins intéressante est SEXE, elle n'est pas significative à 10%. Nous la retirons. Nous relançons la régression avec IMC et toutes les indicatrices de FUMER (Figure 4.25). Nous obtenons $R^2 = 0.186395$. De nouveau, il nous faut calculer les F-partiels. Pour cela nous avons besoin de la régression avec IMC seule (Figure 4.26, $R^2 = 0.063405$).

Sortie - (Sans titre) Le Système SAS 03:41 Saturday, June 18, 2011

The GLM Procedure

Table: Risque Risque

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Model	5	8.92971128	1.78594226	1.45	0.2364
Error	29	35.71600301	1.23158631		
Corrected Total	34	44.64571429			

	R-carré	Coef de Var	Racine MSE	Risque Moyenne
	0.200013	38.84190	1.109769	2.857143

Source	DDL	Type I SS	Moyenne quadratique	Valeur F	Pr > F
SEXE	1	0.74404762	0.74404762	0.60	0.4433
IMC	1	2.64754048	2.64754048	2.15	0.1534
FUMER	3	5.53812318	1.84604106	1.50	0.2356

Source	DDL	Type III SS	Moyenne quadratique	Valeur F	Pr > F
SEXE	1	0.60866232	0.60866232	0.49	0.4877
IMC	1	4.03768028	4.03768028	3.28	0.0806
FUMER	3	5.53812318	1.84604106	1.50	0.2356

Fig. 4.24. Données CARDIO - PROC GLM de SAS avec SEXE, IMC, et FUMER

Results

Global results

Endogenous attribute	Risque
Examples	35
R ²	0.186395
Adjusted-R ²	0.077915
Sigma error	1.100363
F-Test (4,30)	1.7182 (0.171997)

Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	8.3217	4	2.0804	1.7182	0.1720
Residual	36.3240	30	1.2108		
Total	44.6457	34			

Coefficients

Attribute	Coef.	std	t(30)	p-value
Intercept	0.483239	1.177492	0.410397	0.684431
IMC	0.076970	0.041883	1.837727	0.076025
F_ANCIEN	-0.022528	0.474770	-0.047450	0.962469
F_OCCAS	0.205394	0.836689	0.245485	0.807753
F_REGULIER	0.890773	0.457635	1.946472	0.061016

Fig. 4.25. Données CARDIO - Régression avec IMC et toutes les indicatrices de FUMER

Variable	F	ddl ₁	ddl ₂	p-value
IMC	$F = (1.837727)^2 = 3.377241$	1	29	0.076025
FUMER	$F = \frac{(0.186395 - 0.063405)/3}{(1 - 0.186395)/30} = 1.511667$	3	29	0.231622

Results

Global results

Endogenous attribute	Risque
Examples	35
R ²	0.063405
Adjusted-R ²	0.035024
Sigma error	1.125664
F-Test (1,33)	2.2340 (0.144503)

Analysis of variance

Source	xSS	d.f.	xMS	F	p-value
Regression	2.8308	1	2.8308	2.2340	0.1445
Residual	41.8149	33	1.2671		
Total	44.6457	34			

Coefficients

Attribute	Coef.	std	t(33)	p-value
Intercept	1.148130	1.159131	0.990509	0.329133
IMC	0.062240	0.041642	1.494667	0.144503

Fig. 4.26. Données CARDIO - Régression avec IMC

La variable la moins intéressante est FUMER, que nous retirons puisque la p-value est plus grande que 10%. Reste donc que la variable IMC qui est éliminée aussi finalement puisque, seule, elle n'est pas significative au risque $\alpha = 10\%$ (Figure 4.26).

Finalement, il n'est pas possible d'expliquer le risque cardio-vasculaire avec les variables initialement disponibles.

4.7.2 Traitement individuel des indicatrices

Dans ce section, nous renouvelons le processus de sélection *backward*. Mais en ignorant sciemment que les indicatrices F_ANCIEN , F_OCCAS et $F_REGULIER$ sont relatives à la même variable FUMER. Cela peut paraître étrange parce qu'on omet (perd ?) de l'information ainsi. Mais en analysant attentivement les sorties du logiciel, on se rend compte que l'on obtient des résultats qui ne sont pas inintéressants.

Nous avons utilisé la procédure BACKWARD ELIMINATION REG de TANAGRA, toujours avec un seuil $\alpha = 10\%$. Détaillons les informations fournies par le logiciel (Figure 4.27) :

- Nous avons un $R^2 = 0.184447$ et, surtout, le R^2 ajusté qui tient compte des degrés de liberté, atteint ici sa valeur la plus élevée avec $\bar{R}^2 = 0.133475$. Nous avons là le modèle le plus avantageux - compte tenu du nombre d'explicatives utilisées - parmi toutes les tentatives effectuées jusqu'à présent.
- Le modèle est maintenant globalement significatif à 10% avec $F = 3.6186$ et une p-value de 0.0038303.
- Penchons-nous sur le processus de sélection *Backward Elimination Process*. Initialement le coefficient de détermination ajusté de la régression avec la totalité des explicatives est de $\bar{R}^2 = 0.062$.

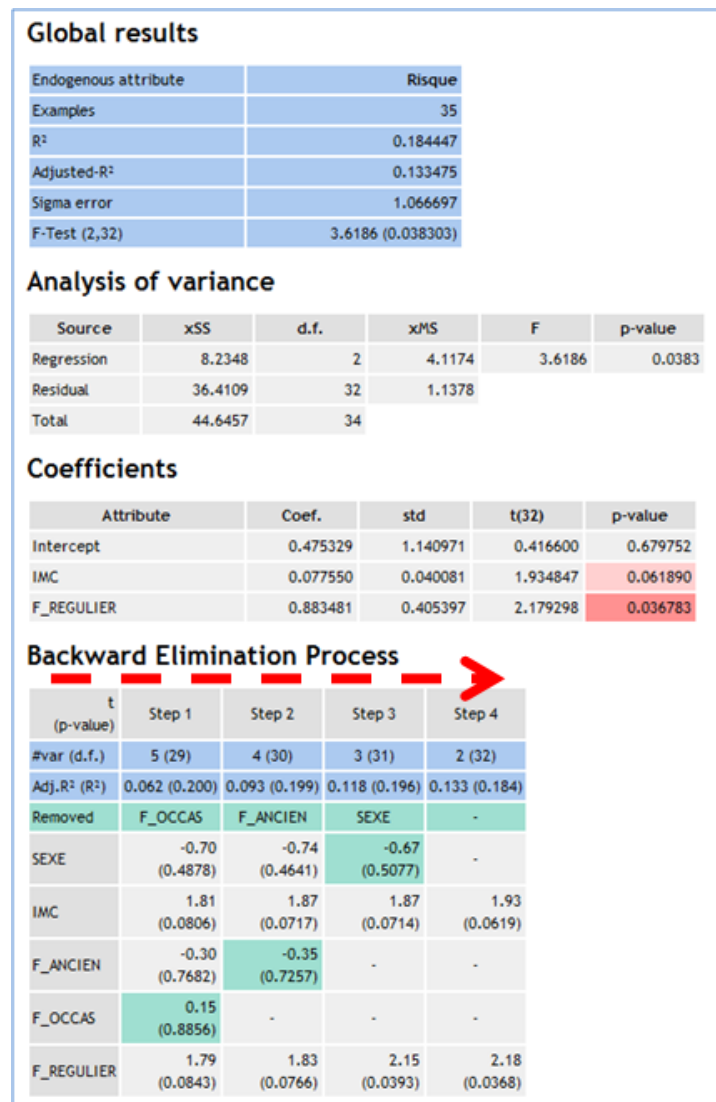


Fig. 4.27. Données CARDIO - Régression *backward*, indicatrices traitées individuellement

1. La première variable éliminée a été l'indicatrice F_OCCAS , avec une p-value de 0.8856. Le R^2 ajusté de la régression qui en résulte est de $\bar{R}^2 = 0.093$.
 2. La seconde est F_ANCIEN , \bar{R}^2 passe à 0.118.
 3. La troisième est $SEXE$, \bar{R}^2 passe à 0.133.
 4. Les deux variables restantes, IMC (p-value = 0.0619) et $F_REGULIER$ (p-value = 0.0368) sont significatives à 10%. Le processus est stoppé.
- Les caractéristiques de la régression avec les deux variables retenues sont affichées dans la partie haute de la fenêtre (Figure 4.27).

Comment expliquer ce résultat ? En s'accordant la possibilité de traiter individuellement les indicatrices, on s'offre une liberté supplémentaire dans le processus exploratoire. La contrainte de traitement

en bloc des indicatrices de variables qualitatives ne pèse plus. Nous avons accès à des combinaisons plus fines des variables explicatives. Clairement, la solution proposée ici est introuvable si nous traitons les indicatrices de FUMER en bloc. Par ailleurs, je me suis rendu compte qu'à la sortie, nous obtenons souvent des modèles plus performants en prédiction (si on se base sur le critère du PRESS par exemple, section 3.2.1).

Comment interpréter les coefficients ? Si les performances sont bonnes, qu'en est-il de l'interprétation ? Est-ce que les résultats ont un sens ? Le noeud du problème est ici. La lecture du coefficient de IMC, explicative quantitative, n'est pas modifiée. Quand l'IMC augmente d'une unité, le risque est augmenté de 0.077550. Concernant le comportement tabagique, les modalités de référence deviennent maintenant ("jamais", "ancien" et "occasionnel"). **Les indicatrices exclues désignent les modalités de référence.** C'est lorsque l'on est un fumeur régulier qu'il y a un surcroît de risque de maladie. Ici, nous lisons : fumer régulièrement, par rapport aux 3 autres types comportements (qui sont mis dans le même panier), entraîne un risque de maladie accru de 0.883481.

Quelques commentaires. Le traitement individuel des indicatrices d'une explicative catégorielle n'est pas très usuel. Les praticiens, essentiellement pour des questions d'interprétations, préfèrent le traitement en bloc. Pourtant, nous le constatons sur notre exemple, en se donnant la possibilité de traiter individuellement les indicatrices, nous avons accès à des solutions (des modèles) plus performantes. La seule contrainte est qu'il nous faut reconsidérer les modalités de références pour les variables catégorielles dont les indicatrices auraient été dissociées. L'interprétation est modifiée. Elle est peut être aussi un peu plus subtile.

4.8 Autres stratégies de codage d'une exogène qualitative nominale

4.8.1 Le codage "centered effect" d'une exogène qualitative nominale

Principe

Nous comprenons que le type de codage définit l'interprétation des coefficients. Nous proposons dans cette section une autre approche. Certes nous créons toujours $(K - 1)$ variables en excluant la K -ème modalité, mais nous attribuons des valeurs différentes. Pour la variable Z_j correspondant à la modalité j de X ($j = 1, \dots, K - 1$) :

$$Z_{i,j} = \begin{cases} 1 & \text{si } X_i = j \\ -1 & \text{si } X_i = K \\ 0 & \text{sinon} \end{cases}$$

La modalité K (*centre-ville*) sert toujours de référence. Mais, cette fois-ci, nous signalons explicitement sa présence pour l'individu i en attribuant la valeur -1 à toutes les variables recodées Z_j . Voici le tableau de correspondance

Lieu	$Z_{banlieue}$	$Z_{campagne}$
Banlieue	1	0
Campagne	0	1
Centre	-1	-1

Nous estimons les coefficients de la régression :

$$loyer = b_0 + b_1 Z_{banlieue} + b_2 Z_{campagne} + \varepsilon \quad (4.6)$$

Comment lire ces coefficients ?

Lecture des résultats

Voyons à nouveau quelques cas particuliers :

- Si l'habitation i^* est en *centre-ville*, nous savons que $Z_{i^*,banlieue} = Z_{i^*,campagne} = -1$. Par conséquent, le loyer prédit est $\hat{y}_{i^*} = \hat{b}_0 - (\hat{b}_1 + \hat{b}_2)$.
- Si l'habitation i^* est en *banlieue*, nous savons que $Z_{i^*,banlieue} = 1$, les autres indicatrices sont égales à 0. Nous en déduisons la valeur prédite du loyer $\hat{y}_{i^*} = \hat{b}_0 + \hat{b}_1$.

En généralisant, nous observons les relations suivantes :

- $\bar{y}_{banlieue} = \hat{b}_0 + \hat{b}_1 \Rightarrow \hat{b}_1 = \bar{y}_{banlieue} - \hat{b}_0$;
- $\bar{y}_{campagne} = \hat{b}_0 + \hat{b}_2 \Rightarrow \hat{b}_2 = \bar{y}_{campagne} - \hat{b}_0$
- $\bar{y}_{centre} = \hat{b}_0 - (\hat{b}_1 + \hat{b}_2)$

Cela nous emmène à tirer plusieurs conclusions :

- La constante de la régression s'interprète maintenant comme une valeur centrale, moyenne non-pondérée des moyennes conditionnelles

$$\hat{b}_0 = \frac{1}{3}(\bar{y}_{banlieue} + \bar{y}_{campagne} + \bar{y}_{centre})$$

D'où l'appellation "centered effect". On parle également de "deviation coding" dans la littérature.

- De manière générale, cette valeur centrale ne coïncide pas avec la moyenne globale de l'endogène $\hat{b}_0 \neq \bar{y}$. Ce sera le cas uniquement si les effectifs dans les groupes étaient équilibrés c.-à-d.

$$\hat{b}_0 = \bar{y} \text{ si et seulement si } n_j = \frac{n}{K}$$

- Les autres coefficients se lisent comme la différence entre la moyenne conditionnelle et cette valeur centrale. Pour le cas de la banlieue, $\hat{b}_1 = \bar{y}_{banlieue} - \hat{b}_0$
- Le test de significativité globale de la régression (tous les coefficients exceptés la constante sont-ils tous égaux à zéro ?) correspond toujours au test d'égalité des moyennes conditionnelles. Nous devrions retrouver les résultats de l'ANOVA à 1 facteur.

Application aux données LOYER

Loyer	Habitation	banlieue	campagne
6.9	banlieue	1	0
6.3	banlieue	1	0
6.7	banlieue	1	0
6.5	banlieue	1	0
7.2	banlieue	1	0
5.6	campagne	0	1
4.9	campagne	0	1
5.3	campagne	0	1
7	centre	-1	-1
7.5	centre	-1	-1
8	centre	-1	-1
7.2	centre	-1	-1
8.4	centre	-1	-1
7.4	centre	-1	-1
8.3	centre	-1	-1

	campagne	banlieue	constante
coef.	-1.29	0.16	6.56
	0.20	0.17	0.13
	0.83	0.46	#N/A
	29.44	12	#N/A
	12.48	2.54	#N/A

Moyenne conditionnelles		
campagne	banlieue	centre
5.27	6.72	7.69

Test significativité globale	
F	29.44
ddl1	2
ddl2	12
p-value	0.0000

Fig. 4.28. Régression avec données codées "centered effect" - Données LOYER

Nous effectuons la régression sur les données LOYER recodées (Figure 4.28). Nous obtenons les coefficients \hat{b} (Équation 4.6) et nous en déduisons les moyennes conditionnelles :

- $\hat{b}_2 = -1.29 \Rightarrow \bar{y}_{campagne} = \hat{b}_2 + \hat{b}_0 = -1.29 + 6.56 = 5.27$;
- $\hat{b}_1 = 0.16 \Rightarrow \bar{y}_{banlieue} = \hat{b}_1 + \hat{b}_0 = 0.16 + 6.56 = 6.72$;
- $\bar{y}_{centre} = \hat{b}_0 - (\hat{b}_1 + \hat{b}_2) = 6.56 - (0.16 + (-1.29)) = 7.69$.

Le test de significativité globale de la régression nous fournit un $F = 29.44$ à $(2, 12)$ degrés de liberté, la $p - value < 0.0001$, ce qui est conforme avec les résultats de l'ANOVA à 1 facteur (Figure 4.3). Les tests sont totalement équivalents. **Le codage n'influe pas sur la qualité de la modélisation. Il pèse en revanche sur la lecture des coefficients.** Ce constat est confirmé dans la 3-ème solution proposée dans la section suivante.

Le codage "simple coding"

La stratégie *simple coding* aboutit à un résultat mixte entre les solutions *cornered effect* et *centered effect*. Nous créons toujours $(K - 1)$ variables qui ne sont plus vraiment des indicatrices :

$$Z_{i,j} = \begin{cases} \frac{K-1}{K} & \text{si } X_i = j \\ \frac{-1}{K} & \text{si } X_i \neq j \end{cases}$$

Nous avons ainsi les correspondances suivantes pour les données LOYER

Lieu	$Z_{banlieue}$	$Z_{campagne}$
Banlieue	2/3	-1/3
Campagne	-1/3	2/3
Centre	-1/3	-1/3

Lorsque nous lançons la régression, les coefficients obtenus (Figure 4.29) mélangent ceux de la solution "cornered effect" (Figure 4.6, les paramètres des indicatrices) avec ceux de "centered effect" (Figure 4.28, la constante).

Loyer	Habitation	banlieue	campagne
6.9	banlieue	0.67	-0.33
6.3	banlieue	0.67	-0.33
6.7	banlieue	0.67	-0.33
6.5	banlieue	0.67	-0.33
7.2	banlieue	0.67	-0.33
5.6	campagne	-0.33	0.67
4.9	campagne	-0.33	0.67
5.3	campagne	-0.33	0.67
7	centre	-0.33	-0.33
7.5	centre	-0.33	-0.33
8	centre	-0.33	-0.33
7.2	centre	-0.33	-0.33
8.4	centre	-0.33	-0.33
7.4	centre	-0.33	-0.33
8.3	centre	-0.33	-0.33

	campagne	banlieue	constante
coef.	-2.42	-0.97	6.56
	0.32	0.27	0.13
	0.83	0.46	#N/A
	29.44	12	#N/A
	12.48	2.54	#N/A

Moyenne conditionnelles		
campagne	banlieue	centre
5.27	6.72	7.69

Test significativité globale	
F	29.44
ddl1	2
ddl2	12
p-value	0.0000

Fig. 4.29. Régression avec données codées "simple coding" - Données LOYER

Nous pouvons établir les relations entre les coefficients et les moyennes conditionnelles :

$$\begin{aligned}
 -\hat{b}_0 &= 6.56 = \frac{1}{3}(\bar{y}_{\text{banlieue}} + \bar{y}_{\text{campagne}} + \bar{y}_{\text{centre}}); \\
 -\hat{b}_1 &= -0.97 = \bar{y}_{\text{banlieue}} - \bar{y}_{\text{centre}}; \\
 -\hat{b}_2 &= -2.42 = \bar{y}_{\text{campagne}} - \bar{y}_{\text{centre}}.
 \end{aligned}$$

Nous avons 3 équations à 3 inconnues, nous pouvons facilement déduire les moyennes conditionnelles.

Commentaire sur le codage "simple coding". Cette solution est référencée dans la littérature, j'en parle uniquement pour cette raison. Personnellement, je ne vois pas très bien ce qu'elle apporte de plus par rapport aux codages *cornered effect* et *centered effect*.

4.8.2 Le codage "contrast effect" d'une exogène qualitative

L'utilisation des contrastes est une alternative à l'utilisation des indicatrices. L'objectif est de comparer les moyennes de la variables dépendante par groupes de modalités. **Les blocs sont construits de manière à mettre en évidence les oppositions les plus intéressantes.**

Reprenons l'exemple du fichier LOYER (Figure 4.1, page 84). Nous souhaitons mener une analyse en deux temps : (1) tout d'abord, vérifier que la moyenne des loyers en centre-ville est différent de la moyenne des loyers à la campagne et en banlieue ; (2) puis, dans ce second temps, effectuer la comparaison à l'intérieur de ce second groupe c.-à-d. comparer les loyers en banlieue et à la campagne.

Nous devons utiliser un codage de type "contrastes" pour réaliser cette analyse. Il repose sur une série de spécifications bien précises ([5], pages 71 à 75) : (a) si l'explicative nominale possède K modalités, nous créerons $(K - 1)$ variables contrastes ; (b) les codes de chaque contraste doit être définis de manière à créer une opposition entre groupes, la somme des codes du premier bloc doit être égal à 1, la somme

pour le second égale à -1 (ou *vice versa*), la somme totale doit être nulle; (c) les codes des variables contrastes doivent être deux à deux orthogonaux c.-à-d. le produit scalaire doit être nul.

Vite un exemple pour bien comprendre le mécanisme. La variable "lieu d'habitation" est composée de 3 modalités, nous créons donc 2 variables contrastes $C1$ et $C2$. Nous adoptons le codage suivant pour réaliser l'analyse en deux temps spécifiée ci-dessus c.-à-d. opposer centre-ville au couple (campagne, banlieue), puis opposer campagne et banlieue.

Lieu	C1	C2
Banlieue	0.5	1
Campagne	0.5	-1
Centre	-1	0

Notons que :

- La somme des codes tant pour $C1$ [$0.5 + 0.5 + (-1) = 0$] que pour $C2$ [$1 + (-1) + 0 = 0$] sont nuls.
 - Pour chaque contraste, la somme des codes positifs est égal à 1, celle des codes négatifs -1 .
 - Les signes sont opposés : ("banlieue", "campagne") d'une part, ("centre") d'autre part pour $C1$.
 - Concernant $C2$, "centre" n'entrant plus en ligne de compte, son code est égal à 0. "Campagne" et "banlieue" sont de signes opposés.
 - Enfin, les deux contrastes sont bien orthogonaux puisque [$0.5 \times 1 + 0.5 \times (-1) + (-1) \times 0$] = 0.
- Ces vérifications faites, nous pouvons construire notre tableau de données et lancer la régression

$$LOYER = a_2 \times C2 + a_1 \times C1 + a_0$$

Pour faciliter la lecture, nous avons reporté dans la feuille Excel les moyennes conditionnelles et les résultats de l'ANOVA (section 4.1). Voyons voir tout cela (Figure 4.30) :

- Première conclusion importante, encore une fois, la qualité globale de l'ajustement n'est pas affectée par le type de codage. La variabilité résiduelle est exactement la même $SCR = 2.54324$ pour la régression et l'analyse de variance. **Le type de codage ne modifie pas le pouvoir explicatif du modèle. En revanche, il met en lumière des aspects différents des informations que recèlent les données.** L'intérêt est de pouvoir en tirer des interprétations en rapport avec les objectifs de notre étude.
- La constante $\hat{a}_0 = 6.55746$ correspond à la moyenne non pondérée des moyennes conditionnelles

$$\hat{a}_0 = \frac{\bar{y}_{banlieue} + \bar{y}_{campagne} + \bar{y}_{centre}}{3} = \frac{6.72000 + 5.26667 + 7.68571}{3} = 6.55746$$

Ce résultat rejoint celui du codage "centered effect" (Figure 4.28).

Les coefficients \hat{a}_j nous permettent d'obtenir les écarts entre les moyennes (non pondérées des moyennes) des modalités dans les groupes que l'on oppose. Si k_1 (resp. k_2) est le nombre de modalités dans le premier (resp. second) groupe, nous avons :

$$e_j = \hat{a}_j \times \frac{k_1 + k_2}{k_1 \times k_2} \quad (4.7)$$

Loyer (Euro au m²)	Lieu Habitation	C1	C2
6.9	banlieue	0.5	1
6.3	banlieue	0.5	1
6.7	banlieue	0.5	1
6.5	banlieue	0.5	1
7.2	banlieue	0.5	1
5.6	campagne	0.5	-1
4.9	campagne	0.5	-1
5.3	campagne	0.5	-1
7	centre	-1	0
7.5	centre	-1	0
8	centre	-1	0
7.2	centre	-1	0
8.4	centre	-1	0
7.4	centre	-1	0
8.3	centre	-1	0

Moyennes conditionnelles		
Lieu Habitation	Moyenne	n
banlieue	6.72000	5
campagne	5.26667	3
centre	7.68571	7
Globale	6.88000	15

Tableau d'analyse de variance (ANOVA)			
Source	ddl	SC	CM
SCE	2	12.48076	6.24038
SCR	12	2.54324	0.21194
SCT	14	15.02400	-

F	29.44458
p-value	2.35293E-05

DROITEREG		
C2	C1	constante
0.72667	-1.12825	6.55746
0.16810	0.16129	0.12619
0.83072	0.46037	#N/A
29.44458	12	#N/A
12.48076	2.54324	#N/A

t	4.32278	-6.99505
p-value	0.00099124	1.4443E-05

moy.non.pondérée	6.55746
------------------	---------

e1	-1.69238
ecart.moy1	-1.69238

e2	1.45333
ecart.moy2	1.45333

Fig. 4.30. Régression avec données codées "contrast effect" - Données LOYER

- Pour le premier contraste, nous obtenons $\hat{a}_1 = -1.12825$. Nous avons $k_1 = 2$ (banlieue et campagne) dans le 1^{er} groupe, et $k_2 = 1$ (centre) dans le 2nd. Nous calculons

$$e_1 = -1.12825 \times \frac{2+1}{2 \times 1} = -1.69238$$

Qui correspond à l'écart entre les moyennes (non pondérée des moyennes conditionnelles) dans les groupes c.-à-d.

$$\frac{\bar{y}_{banlieue} + \bar{y}_{campagne}}{2} - \bar{y}_{centre} = \frac{6.72000 + 5.26667}{2} - 7.68571 = -1.69238 = e_1$$

- Pour le second contraste opposant "banlieue" et "campagne", la modalité "centre" étant mise de côté, nous $k_1 = 1$ et $k_2 = 2$, nous en déduisons

$$e_2 = \hat{a}_2 \times \frac{1+1}{1 \times 1} = 0.72667 \times 2 = 1.45333$$

Il correspond à l'écart

$$\bar{y}_{banlieue} - \bar{y}_{campagne} = 6.72000 - 5.26667 = 1.45333 = e_2$$

Dans les deux cas, les écarts sont significatifs à 5% selon la régression puisque nous avons :

$$t_{\hat{a}_1} = -6.99505 \rightarrow p\text{-value} = 1.4443 \times 10^{-5}$$

$$t_{\hat{a}_2} = 4.32278 \rightarrow p\text{-value} = 0.00099124$$

Conclusion. Certes, l'outil n'est pas très limpide au premier abord. Il faut proposer un codage qui répond à des spécifications assez restrictives. L'affaire devient compliquée lorsque le nombre de modalités est élevé. Mais une fois que nous avons mis en place le bon schéma de codage, les avantages sont appréciables. Nous pouvons décomposer l'analyse en une cascade d'oppositions entre groupes. Nous obtenons une estimation des écarts, et nous pouvons tester de surcroît s'ils sont significatifs. Tout cela à la lecture des résultats d'une seule régression. L'effort initial est largement récompensé.

Remarque 26 (Lorsque les effectifs sont équilibrés.). Lorsque les effectifs sont équilibrés c.-à-d. nous avons les mêmes effectifs dans chaque groupe, une pratique quand même bien répandue en statistique, nous opposons bien les moyennes conditionnelles. Ainsi, la procédure n'est pas sans rappeler les schémas de comparaisons multiples que l'on initie à la suite d'une ANOVA détectant des différences globalement significatives entre les moyennes conditionnelles.

4.9 Codage d'une exogène qualitative ordinale

On parle de variable qualitative ordinale lorsque (1) la variable prend un nombre fini de modalités (de valeurs); (2) il y a une relation d'ordre entre ces modalités. L'exemple le plus souvent cité est la satisfaction. On peut imaginer 3 valeurs possibles : mécontent, satisfait, très satisfait. Manifestement, le passage d'un niveau à l'autre implique une amélioration. Mais nous n'avons pas d'indications sur son amplitude. Le codage numérique simple (1, 2, 3) peut nous induire en erreur justement parce qu'il introduit une valorisation de l'amplitude de l'écart qui - peut-être, on ne le sait pas en réalité - n'a pas lieu d'être. Nous reviendrons en détail sur ce type de codage plus loin (section 4.11.2).

Lorsque l'exogène est qualitative ordinale, l'utilisation d'indicateurs telle que décrite dans les sections précédentes remplit son office. Mais nous perdons le caractère ordonné des modalités. Une information importante est omise. La modélisation n'en tient pas compte. L'interprétation en pâtit.

4.9.1 Un exemple introductif

Nous utilisons des données artificielles dans cette section. Nous cherchons à expliquer Y à partir de X . Nous disposons de $n = 30$ observations.

Manifestement, la liaison est non linéaire (Figure 4.31). Plutôt que de chercher la forme de la liaison la plus appropriée, nous préférons découper le domaine de X en 3 intervalles. Cette stratégie est très pratique pour traiter les problèmes de non-linéarité. Le premier intervalle I_1 est défini sur $(X < 10)$, le I_2 second sur $(10 \leq X < 20)$ et le troisième I_3 sur $(X \geq 20)$ ⁷ (Figure 4.32).

La variable qualitative ordinale Z à $K = 3$ modalités (z_1, z_2, z_3) est déduite de ces intervalles, soit

7. Ce qui correspond *grossa modo* à la technique des intervalles de largeur égales. Elle a pour mérite de ne pas modifier la distribution des données; elle a pour inconvénient d'être très sensible aux points atypiques, certains intervalles peuvent être vides. Nos données étant très simples, elle donne entièrement satisfaction.

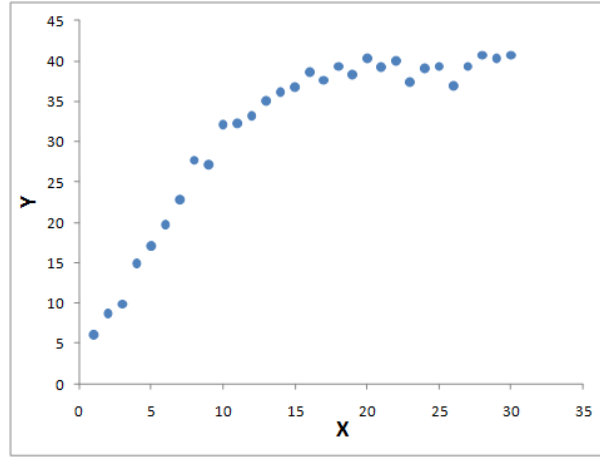


Fig. 4.31. Codage ordinal - Nuage de points

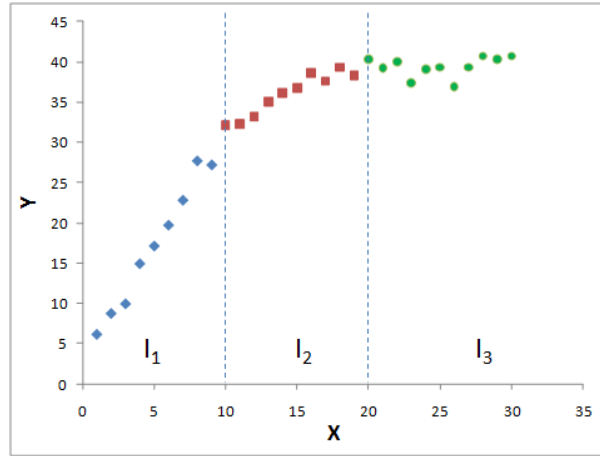


Fig. 4.32. Codage ordinal - Découpage en 3 intervalles

$$Z_i = \begin{cases} z_1, & \text{si } x_i < 10 \\ z_2, & \text{si } (10 \leq x_i < 20) \\ z_3, & \text{si } x_i \geq 20 \end{cases}$$

La variable qualitative ordinale Z s'ajoute au tableau de données. Nous remarquons que la représentation des valeurs de Y en fonction des modalités de Z montre qu'une discrétisation induit toujours une perte d'information (Figure 4.33). Nous espérons qu'elle se fera au profit d'une meilleure appréhension de la relation entre Y et X ⁸.

Dans la suite de cette section, nous verrons (1) comment coder numériquement Z pour pouvoir analyser la relation entre Y et Z via une régression ; (2) comment par la suite interpréter les coefficients estimés selon la stratégie de codage choisie.

8. La discrétisation, c'est pas automatique... On perd en variance ce qu'on espère gagner en biais dans la modélisation. Le tout est de délimiter jusqu'à quel point.

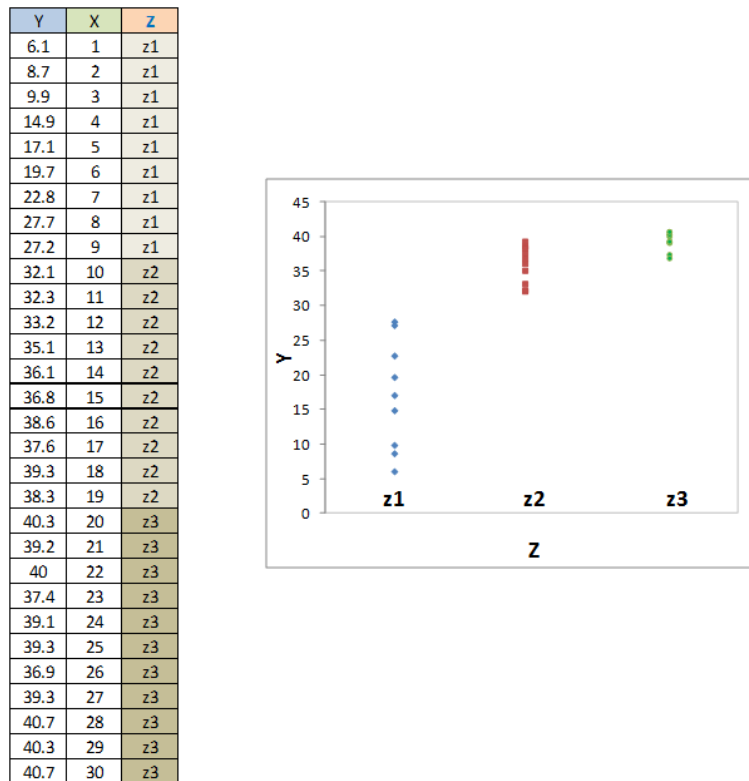


Fig. 4.33. Tableau de données avec la variable ordinaire Z - Valeurs de Y en fonction des modalités de Z

4.9.2 (In)adéquation du codage pour variable qualitative nominale

La première modalité z_1 servant de référence, nous créons deux indicatrices en se référant au codage "cornered effect" étudiée lors du traitement des variables exogènes qualitatives nominales :

$$C2_i = \begin{cases} 1, & \text{si } Z_i = z_1 \\ 0, & \text{sinon} \end{cases}$$

Et

$$C3_i = \begin{cases} 1, & \text{si } Z_i = z_2 \\ 0, & \text{sinon} \end{cases}$$

La correspondance peut s'écrire sous forme de tableau

Z	C_2	C_3
z_1	0	0
z_2	1	0
z_3	0	1

Le fichier de données est transformé. Nous pouvons lancer la régression

$$Y = a_0 + a_2 \times C2 + a_3 \times C3 + \varepsilon$$

Y	Z	C2	C3
6.1	z1	0	0
8.7	z1	0	0
9.9	z1	0	0
14.9	z1	0	0
17.1	z1	0	0
19.7	z1	0	0
22.8	z1	0	0
27.7	z1	0	0
27.2	z1	0	0
32.1	z2	1	0
32.3	z2	1	0
33.2	z2	1	0
35.1	z2	1	0
36.1	z2	1	0
36.8	z2	1	0
38.6	z2	1	0
37.6	z2	1	0
39.3	z2	1	0
38.3	z2	1	0
40.3	z3	0	1
39.2	z3	0	1
40	z3	0	1
37.4	z3	0	1
39.1	z3	0	1
39.3	z3	0	1
36.9	z3	0	1
39.3	z3	0	1
40.7	z3	0	1
40.3	z3	0	1
40.7	z3	0	1

DROITEREG - Y = f (C2, C3)			
	C3	C2	constante
a^	22.26	18.82	17.12
R²	2.085	2.132	1.547
F	0.825	4.640	#N/A
	63.656	27	#N/A
	2740.878	581.276	#N/A

t	10.674	8.827
p-value	3.44339E-11	1.9233E-09

Fig. 4.34. Codage disjonctif - Tableau de données et régression sur les indicatrices

Auscultons les résultats (Figure 4.34) :

- Le coefficient de détermination est $R^2 = 0.825$.
- La régression est globalement pertinente avec $F = 63.656$, à comparer avec un $F_{0.95}(2, 27) = 3.35$ pour un test à 5%.
- La constante $\hat{a}_0 = 17.12$ correspond à la moyenne de Y pour la modalité z_1 , la référence. Nous l'appellerons moyenne de référence $\bar{y}_1 = 17.12$. Ce résultat est tout à fait conforme au comportement des régressions sur variables indicatrices.
- Le second coefficient $\hat{a}_2 = 18.82$ indique le décalage entre la moyenne pour z_2 et la référence z_1 c.-à-d. $\bar{y}_2 = \bar{y}_1 + \hat{a}_2 = 17.12 + 18.82 = 35.94$. L'écart entre les moyennes conditionnelles est significatif à 5% d'après le t de Student du test de significativité du coefficient de la régression ($t_{\hat{a}_2} = 10.674$).
- Le troisième coefficient enfin matérialise l'écart entre la moyenne pour z_3 et la référence z_1 . Ainsi, $\bar{y}_3 = \bar{y}_1 + \hat{a}_3 = 17.12 + 22.26 = 39.38$. Ici aussi la différence est significative.
- Représentées graphiquement, les relations entre les moyennes conditionnelles et les coefficients de la régression prennent tout leur sens (Figure 4.35).

Tout cela est cohérent. Il reste pourtant une information importante qui apparaît clairement dans le graphique, et que la régression n'a pas mis en évidence. La différence entre les moyennes des 2ème et 3ème

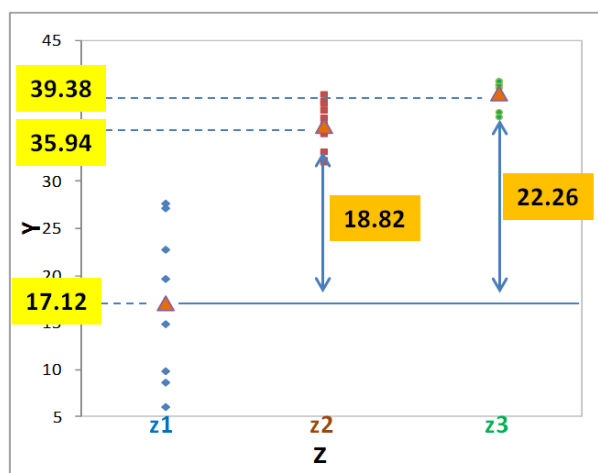


Fig. 4.35. Codage disjonctif - Moyennes conditionnelles et coefficients de la régression

modalités est visiblement faible, voire non significative (*à confirmer par les calculs...*). Or dans le cadre du traitement des variables ordinales, ce n'est pas tant l'écart par rapport à la première modalité qui nous intéresse, mais plutôt l'étude du passage d'un palier (modalité) à un autre (le suivant). Comment coder la variable explicative ordinale pour que la régression fournisse des coefficients propices à ce type d'analyse ?

4.9.3 Utilisation du codage cumulatif

Le codage cumulatif est une solution simple à ce problème. Il s'agit toujours d'utiliser des indicatrices, mais en intégrant l'information de passage aux paliers supérieurs. Pour notre exemple, nous créerons $D2$ et $D3$ telles que :

$$D2_i = \begin{cases} 1, & \text{si } (Z_i \geq z_2) \\ 0, & \text{sinon} \end{cases}$$

Et

$$D3_i = \begin{cases} 1, & \text{si } (Z_i \geq z_3) \\ 0, & \text{sinon} \end{cases}$$

La nouvelle correspondance entre les modalités et les indicatrices devient

Z	D ₂	D ₃
z ₁	0	0
z ₂	1	0
z ₃	1	1

Nous introduisons ainsi des contraintes supplémentaires dans la construction des indicatrices. Nous notons plusieurs particularités :

- A l'instar du codage usuel, si un individu appartient à la modalité de référence z_1 , toutes les indicatrices prennent la valeur 0 ($D2_i = D3_i = 0$).
- Si un individu appartient à la k -ème modalité ($k > 2$), l'indicatrice associée prend la valeur 1, et toutes les indicatrices de niveau inférieur également c.-à-d. $Dk_i = 1 \Rightarrow D2_i = \dots = D(k-1)_i = 1$.
- Seule la dernière modalité z_K est codée de manière identique dans les deux approches $CK_i = DK_i$.

Y	Z	D2	D3
6.1	z1	0	0
8.7	z1	0	0
9.9	z1	0	0
14.9	z1	0	0
17.1	z1	0	0
19.7	z1	0	0
22.8	z1	0	0
27.7	z1	0	0
27.2	z1	0	0
32.1	z2	1	0
32.3	z2	1	0
33.2	z2	1	0
35.1	z2	1	0
36.1	z2	1	0
36.8	z2	1	0
38.6	z2	1	0
37.6	z2	1	0
39.3	z2	1	0
38.3	z2	1	0
40.3	z3	1	1
39.2	z3	1	1
40	z3	1	1
37.4	z3	1	1
39.1	z3	1	1
39.3	z3	1	1
36.9	z3	1	1
39.3	z3	1	1
40.7	z3	1	1
40.3	z3	1	1
40.7	z3	1	1

DROITEREG - Y = f (D2, D3)			
	D3	D2	constante
a^	3.44	18.82	17.12
	2.027	2.132	1.547
R²	0.825	4.640	#N/A
F	63.656	27	#N/A
	2740.878	581.276	#N/A

t	1.698	8.827
p-value	0.1011	1.9233E-09

Fig. 4.36. Codage cumulatif - Tableau de données et régression sur les indicatrices

Quelles sont les conséquences de ce codage dit "cumulatif" sur la régression ? Nous nous efforçons de vérifier cela tout de suite. Nous implémentons la régression

$$Y = b_0 + b_2 \times D2 + b_3 \times D3 + \varepsilon$$

Analysons les résultats (Figure 4.36) :

- Par rapport à la régression précédente, la qualité globale n'est pas modifiée (R^2 , F - test). Ce constat est très important. **L'introduction du nouveau codage ne dégrade pas les qualités prédictives et explicatives du modèle.**
- La vraie nouveauté se situe au niveau des coefficients. La constante $\hat{b}_0 = 17.12$ correspond toujours à la moyenne $\bar{y}_1 = 17.12$ de la modalité de référence.

- Le coefficient de la première indicatrice $\hat{b}_2 = 18.82$ constitue toujours au décalage entre les moyennes conditionnelles $\bar{y}_2 = \bar{y}_1 + \hat{b}_2 = 17.12 + 18.82 = 35.94$.
- En revanche, le coefficient de l'indicatrice D_3 indique le décalage entre la moyenne conditionnelle de la 3ème modalité et la précédente ! Ici, $\bar{y}_3 = \bar{y}_2 + \hat{b}_3 = 35.94 + 3.44 = 39.38$. Il apparaît que cet écart n'est pas significatif à 5% puisque dans la régression $t_{\hat{b}_3} = 1.698$ avec une p-value de 0.1011.

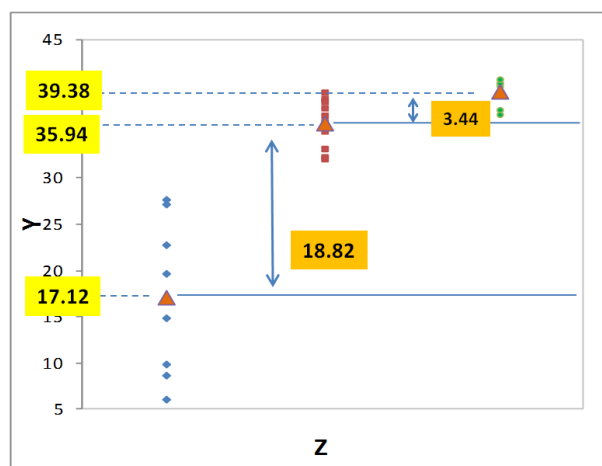


Fig. 4.37. Codage cumulatif - Moyennes conditionnelles et coefficients de la régression

La généralisation est immédiate : tous les coefficients des indicatrices correspondent au décalage des moyennes conditionnelles entre les modalités successives. Nous mettons bien en évidence l'effet du passage d'un pallier à un autre. La représentation des moyennes et des coefficients dans le nuage de points illustre parfaitement le phénomène (Figure 4.37).

4.9.4 Codage "backward difference"

Le codage "backward difference" est une variante du codage cumulatif. Il consiste toujours à comparer la moyenne conditionnelle d'un niveau par rapport à la précédente. Mais la constante est modifiée : elle indique la moyenne non pondérée des moyennes conditionnelles.

Le tableau de correspondance pour notre exemple est modifié. Nous n'avons plus vraiment des indicatrices à proprement parler :

Z	D_2	D_3
z_1	$-\frac{K-1}{K} = -\frac{2}{3}$	$-\frac{K-2}{K} = -\frac{1}{3}$
z_2	$\frac{1}{K} = \frac{1}{3}$	$-\frac{K-2}{K} = -\frac{1}{3}$
z_3	$\frac{1}{K} = \frac{1}{3}$	$\frac{2}{K} = \frac{2}{3}$

Y	Z	D2	D3
6.1	z1	-0.67	-0.33
8.7	z1	-0.67	-0.33
9.9	z1	-0.67	-0.33
14.9	z1	-0.67	-0.33
17.1	z1	-0.67	-0.33
19.7	z1	-0.67	-0.33
22.8	z1	-0.67	-0.33
27.7	z1	-0.67	-0.33
27.2	z1	-0.67	-0.33
32.1	z2	0.33	-0.33
32.3	z2	0.33	-0.33
33.2	z2	0.33	-0.33
35.1	z2	0.33	-0.33
36.1	z2	0.33	-0.33
36.8	z2	0.33	-0.33
38.6	z2	0.33	-0.33
37.6	z2	0.33	-0.33
39.3	z2	0.33	-0.33
38.3	z2	0.33	-0.33
40.3	z3	0.33	0.67
39.2	z3	0.33	0.67
40	z3	0.33	0.67
37.4	z3	0.33	0.67
39.1	z3	0.33	0.67
39.3	z3	0.33	0.67
36.9	z3	0.33	0.67
39.3	z3	0.33	0.67
40.7	z3	0.33	0.67
40.3	z3	0.33	0.67
40.7	z3	0.33	0.67

DROITEREG - Y = f (D2, D3)			
	D3	D2	constante
a^	3.44	18.82	30.81
	2.027	2.132	0.850
R²	0.825	4.640	#N/A
F	63.656	27	#N/A
	2740.878	581.276	#N/A

t	1.698	8.827
p-value	0.1011	1.9233E-09

Fig. 4.38. Codage "backward difference" - Tableau de données et régression

Le schéma semble différent par rapport au codage cumulatif. Mais, à bien y regarder, nous percevons l'effet de cascade lorsque nous passons d'un niveau à un autre. Les coefficients de la régression refléteront cette spécification.

Le jeu de données est modifié en conséquence et la régression est relancée (Figure 4.38). Les coefficients de D_2 et D_3 indiquent toujours l'écart entre chaque niveau successif. La constante en revanche indique la moyenne non-pondérée des moyennes conditionnelles c.-à-d.

$$\hat{b}_0 = 30.81 = \frac{1}{3}(17.12 + 35.94 + 39 + 38)$$

4.9.5 Codage "forward difference"

Référencé dans la littérature, le codage "forward difference" agit à rebroussement chemin du "backward" c.-à-d. il permet d'obtenir l'écart entre les moyennes conditionnelles lorsque nous analysons les niveaux dans le sens inverse.

Dans notre exemple, nous utiliserions ce tableau de conversion :

Z	D_2	D_3
z_1	$\frac{K-1}{K} = \frac{2}{3}$	$\frac{K-2}{K} = \frac{1}{3}$
z_2	$-\frac{1}{K} = -\frac{1}{3}$	$\frac{K-2}{K} = \frac{1}{3}$
z_3	$-\frac{1}{K} = -\frac{1}{3}$	$-\frac{2}{K} = -\frac{2}{3}$

A l'issue de la régression, les signes des coefficients de D_2 et D_3 sont inversés. La constante, elle, n'est pas modifiée, elle correspond toujours à la moyenne non pondérée des moyennes conditionnelles.

4.9.6 Codage "Helmert"

Le codage "Helmert" permet de mettre en évidence la différence entre la moyenne conditionnelle de la variable endogène pour un niveau par rapport à celle de l'ensemble des niveaux qui lui succèdent (Helmert coding) ou qui le précèdent (Reverse Helmert coding).

Intéressons-nous à cette seconde configuration pour nos données. Nous calculons les moyennes conditionnelles de Y pour chaque modalité de Z , individuellement puis cumulativement.

Z	Moyenne	Moyenne cumulée
z_1	17.12	17.12
z_2	35.94	26.53
z_3	39.38	30.81

Comment lire ces chiffres ?

- Pour la modalité z_1 , la moyenne de Y est de 17.12.
- Pour z_2 , elle est de 35.94.
- La moyenne cumulée des 2 premières modalités est égale à la moyenne non pondérée des deux moyennes c.-à-d. $26.53 = \frac{1}{2}(17.12 + 35.94)$.
- Pour la 3ème z_3 , la moyenne conditionnelle est de 39.38. Cumulativement sur les 3 modalités, nous aurions $30.81 = \frac{1}{3}(17.12 + 35.94 + 39.38)$.

Remarque 27 (Moyenne pondérée ou non-pondérée des moyennes conditionnelles). La notion de moyenne cumulée telle qu'elle est décrite ici peut paraître étrange. On aurait été tenté de penser que la moyenne de Y pour les 2 premières modalités doit être égale à la moyenne calculée sur l'ensemble des observations composant ces deux niveaux c.-à-d. en effectuant une moyenne pondérée des moyennes conditionnelles. Le résultat aurait été différent. Les deux approches ne coïncident que dans les cas des groupes équilibrés, lorsque les effectifs sont identiques pour chaque modalité. On rencontre souvent cette configuration dans les données expérimentales c.-à-d. lorsque les données sont issues d'une expérimentation. Les statisticiens procèdent ainsi notamment parce que les tests paramétriques de comparaison de populations s'avèrent nettement plus robustes dans ce cas.

Il y aurait donc 2 écarts à caractériser pour ces données : la différence entre la moyenne du 2nd niveau et celle du premier $18.82 = (35.94 - 17.12)$; la différence entre le 3ème niveau et les deux premiers

$12.85 = (39.38 - 26.53)$. Voyons comment obtenir cela avec la régression

$$Y = c_0 + c_2 \times H2 + c_3 \times H3 + \varepsilon$$

Voici le tableau de conversion

Z	H_2	H_3
z_1	$-\frac{1}{2}$	$-\frac{1}{3}$
z_2	$\frac{1}{2}$	$-\frac{1}{3}$
z_3	0	$\frac{2}{3}$

Y	Z	H2	H3
6.1	z1	-0.50	-0.33
8.7	z1	-0.50	-0.33
9.9	z1	-0.50	-0.33
14.9	z1	-0.50	-0.33
17.1	z1	-0.50	-0.33
19.7	z1	-0.50	-0.33
22.8	z1	-0.50	-0.33
27.7	z1	-0.50	-0.33
27.2	z1	-0.50	-0.33
32.1	z2	0.50	-0.33
32.3	z2	0.50	-0.33
33.2	z2	0.50	-0.33
35.1	z2	0.50	-0.33
36.1	z2	0.50	-0.33
36.8	z2	0.50	-0.33
38.6	z2	0.50	-0.33
37.6	z2	0.50	-0.33
39.3	z2	0.50	-0.33
38.3	z2	0.50	-0.33
40.3	z3	0.00	0.67
39.2	z3	0.00	0.67
40	z3	0.00	0.67
37.4	z3	0.00	0.67
39.1	z3	0.00	0.67
39.3	z3	0.00	0.67
36.9	z3	0.00	0.67
39.3	z3	0.00	0.67
40.7	z3	0.00	0.67
40.3	z3	0.00	0.67
40.7	z3	0.00	0.67

DROITEREG - Y = f (H2, H3)			
	H3	H2	constante
a^	12.85	18.82	30.81
	1.759	2.132	0.850
R^2	0.825	4.640	#N/A
F	63.656	27	#N/A
	2740.878	581.276	#N/A

t	7.306	8.827
p-value	0.0000	1.9233E-09

Fig. 4.39. Codage "Reverse Helmert" - Tableau de données et régression

Que faut-il en penser ?

- H_2 confronte le premier niveau au second. Les codes sont de même valeur mais de signe opposé. Le troisième n'entre pas en ligne de compte, son code est égal à 0.
- H_3 oppose les deux premiers niveaux (ils partagent la même valeur) au troisième, ce dernier est codé de manière à ce que la somme des codes soit nulle.
- Nous observons l'effet cumulatif du codage en passant de H_2 à H_3 .

Nous recodons notre jeu de données et nous réalisons la régression sous Excel (Figure 4.39). Les coefficients des variables correspondent aux écarts entre les moyennes successives telles que nous les avons définis ci-dessus, $\hat{c}_2 = 18.82$ et $\hat{c}_3 = 12.85$. La constante quant à elle est égale à la moyenne non pondérée des moyennes conditionnelles $\hat{c}_0 = 30.81 = \frac{1}{3}(17.12 + 35.94 + 39.38)$

4.10 Codage polynomial orthogonal d'une exogène qualitative ordinale

Nous l'avons vu précédemment, tenir compte du caractère ordinal de l'exogène Z est intéressant pour mettre en évidence les écarts entre les moyennes successives de l'endogène Y . Pour des motifs pédagogiques, nous nous étions arrangés pour que la relation entre Y et Z soit monotone dans notre jeu de données illustratif afin que les écarts successifs soient toujours positifs. En réalité, le codage n'introduit aucune hypothèse quant à la nature de la relation entre les variables.

Dans cette section, nous étudions le codage par polynômes orthogonaux. Sa particularité est d'introduire une hypothèse de tendance dans la relation liant Y et Z ordinal. Elle (la tendance) peut être *linéaire*, *quadratique*, etc. On peut - en théorie - prendre des polynômes de degrés élevés lorsque le nombre de modalités augmente. Bien évidemment, il faut que Z soit ordinal pour que ce type d'analyse soit possible, mais on recommande également que ses modalités soient également espacées⁹. Ce commentaire est loin d'être anodin. En effet, si les modalités sont régulièrement espacées, on peut être tenté de coder numériquement $(0, 1, 2, \dots)$ les modalités de Z et d'introduire telle quelle la nouvelle variable. Dans ce cas, nous faisons l'hypothèse que la relation entre Y et Z est linéaire. Le codage polynomial orthogonal nous permet de couvrir cette configuration, mais aussi d'explorer d'autres hypothèses de liaison.

Notre fichier exemple se prête bien à ce type d'analyse puisque les niveaux sont issus d'une discrétisation à intervalles de largeurs égales d'une variable quantitative.

4.10.1 Construction du codage

L'objectif est de construire un ensemble de vecteurs orthonormés permettant de traduire les différents types de liaison - une tendance que l'on peut traduire sous forme de polynôme - existant entre Y et Z . Ne disposant que de 3 modalités pour la variable Z , nous irons jusqu'à un polynôme de degré 2.

Tendance linéaire

Nous devons traduire les valeurs initiales de Z , mettons $(0, 1, 2)$ pour fixer les idées, en un ensemble de valeurs $V = (v_1, v_2, v_3)$ avec une relation linéaire

$$V = a \times Z + b$$

Énumérons les contraintes définissant les valeurs de V :

9. http://www.ats.ucla.edu/stat/r/library/contrast_coding.htm#ORTHOGONAL

1. $v_1 = a \times 0 + b = b$ pour la première valeur de Z .
2. $v_2 = a \times 1 + b = a + b$.
3. $v_3 = a \times 2 + b = 2a + b$.
4. $v_2 - v_1 = v_3 - v_2 = a$, tout à fait logiquement puisque la liaison est linéaire.
5. $v_1 + v_2 + v_3 = 0$, la somme des codes doit être nulle.
6. $\sqrt{v_1^2 + v_2^2 + v_3^2} = 1$, parce que le vecteur doit être normé.

Il vient de ces spécifications (je laisse la résolution aux férus de mathématiques, mais ce n'est pas bien compliqué en vérité) que le vecteur V s'écrit :

$$V = \left(-\frac{\sqrt{2}}{2}, 0, \frac{\sqrt{2}}{2}\right)$$

Par rapport aux valeurs de Z , V traduit l'idée d'une progression linéaire avec une pente positive.

Tendance quadratique

Nous devons mettre en évidence une relation quadratique cette fois-ci, on souhaite de plus que le nouveau vecteur $W = (w_1, w_2, w_3)$ soit orthogonal au précédent. W est relié à Z comme suit

$$W = \alpha \times Z^2 + \beta \times Z + \gamma$$

De nouveau, nous pouvons écrire :

1. $w_1 = \alpha \times 0^2 + \beta \times 0 + \gamma = \gamma$.
2. $w_2 = \alpha \times 1^2 + \beta \times 1 + \gamma = \alpha + \beta + \gamma$.
3. $w_3 = \alpha \times 2^2 + \beta \times 2 + \gamma = 4\alpha + 2\beta + \gamma$.
4. $w_1 + w_2 + w_3 = 0$.
5. $\sqrt{w_1^2 + w_2^2 + w_3^2} = 1$.
6. $v_1 \times w_1 + v_2 \times w_2 + v_3 \times w_3 = 0$ parce que le nouveau vecteur doit être orthogonal au précédent.

Nous obtenons cette fois-ci (de nouveau, je laisse résolution aux matheux) :

$$W = \left(\frac{1}{\sqrt{6}}, \frac{-2}{\sqrt{6}}, \frac{1}{\sqrt{6}}\right)$$

Par rapport aux valeurs de Z , W représente une parabole à concavité tournée vers le haut.

4.10.2 Régression sur les variables recodées

Nous utilisons le tableau de conversion suivant...

Z	V	W
z_1	$-\frac{\sqrt{2}}{2}$	$\frac{1}{\sqrt{6}}$
z_2	0	$\frac{-2}{\sqrt{6}}$
z_3	$\frac{\sqrt{2}}{2}$	$\frac{1}{\sqrt{6}}$

... pour élaborer les données utilisées dans la régression

$$Y = d_0 + d_1 \times V + d_2 \times W + \varepsilon$$

DROITEREG d'Excel nous fournit les coefficients du modèle (Figure 4.40).

Y	Z	V	W
6.1	0	-0.707	0.408
8.7	0	-0.707	0.408
9.9	0	-0.707	0.408
14.9	0	-0.707	0.408
17.1	0	-0.707	0.408
19.7	0	-0.707	0.408
22.8	0	-0.707	0.408
27.7	0	-0.707	0.408
27.2	0	-0.707	0.408
32.1	1	0.000	-0.816
32.3	1	0.000	-0.816
33.2	1	0.000	-0.816
35.1	1	0.000	-0.816
36.1	1	0.000	-0.816
36.8	1	0.000	-0.816
38.6	1	0.000	-0.816
37.6	1	0.000	-0.816
39.3	1	0.000	-0.816
38.3	1	0.000	-0.816
40.3	2	0.707	0.408
39.2	2	0.707	0.408
40	2	0.707	0.408
37.4	2	0.707	0.408
39.1	2	0.707	0.408
39.3	2	0.707	0.408
36.9	2	0.707	0.408
39.3	2	0.707	0.408
40.7	2	0.707	0.408
40.3	2	0.707	0.408
40.7	2	0.707	0.408

DROITEREG - Y = f (V, W)			
	W	V	constante
a^	-6.28	15.74	30.81
	1.470	1.475	0.850
R²	0.825	4.640	#N/A
F	63.656	27	#N/A
	2740.878	581.276	#N/A

t	-4.271	10.674
p-value	0.0002	3.4434E-11

Fig. 4.40. Codage polynomial orthogonal - Tableau de données et régression

Nous devons lire ces résultats en gardant à l'esprit la forme de la relation entre Y et Z (Figure 4.33) :

1. Avec $R^2 = 0.825$, la régression est de qualité équivalente aux précédents, le codage n'induit pas une perte d'information dans l'explication de la relation entre Y et Z .
2. La constante $\hat{d}_0 = 30.81$ correspond à la moyenne non pondérée des moyennes conditionnelles.
3. Le coefficient de V , $\hat{d}_1 = 15.74$ est très largement significatif. Il y a donc une très forte relation linéaire positive entre Y et Z .
4. Et elle se double d'une liaison quadratique puisque le coefficient de W , $\hat{d}_2 = -6.28$ est également significatif à 5%. Le coefficient estimé est négatif en revanche. Cela veut dire que la parabole est

inversée par rapport au codage défini, elle est à concavité tournée vers le bas. L'étude visuelle du nuage de points (Figure 4.33) confirme cette analyse.

La construction des codes des polynômes orthogonaux est le principe écueil de cette approche. Les calculs seront d'autant plus difficile que le nombre de modalités augmente. Fort heureusement, certains logiciels les fournissent automatiquement. La commande `contr.poly()` du logiciel *R* par exemple permet d'obtenir automatiquement les $(K - 1)$ vecteurs de codes pour le traitement de K modalités. C'est d'ailleurs ainsi que j'ai pu vérifier mes calculs dans la section 4.10.1.

4.11 Les erreurs à ne pas commettre

Comme nous pouvons le constater, le codage conditionne la lecture des résultats. Le véritable danger est d'utiliser une transformation qui occasionne une perte d'information, ou qui introduit une information supplémentaire qui n'existe pas dans les données. Dans cette section, nous nous penchons sur le codage numérique $\{1, 2, 3, \dots\}$ des variables qualitatives.

4.11.1 Codage numérique d'une variable discrète nominale

On parle de variable discrète nominale lorsque (1) la variable prend un nombre fini de modalités (de valeurs); (2) il n'y a pas de relation d'ordre entre les modalités. On peut appréhender ainsi la variable *habitation* du fichier LOYER, il n'y a pas de hiérarchie entre les zones de résidence : vivre à la campagne n'est pas mieux que vivre en ville, etc. Dans ce cas, le codage suivant est totalement inapproprié

$$Z_i = \begin{cases} 1 & \text{si } X_i = \text{"banlieue"} \\ 2 & \text{si } X_i = \text{"campagne"} \\ 3 & \text{si } X_i = \text{"centre"} \end{cases}$$

En effet, nous introduisons dans la variable recodée une relation d'ordre qui n'existe pas dans les données initiales, information que la régression va utiliser pour calculer les coefficients.

Dans ce cas, les différents codages décrits plus haut (*cornered effect*, *centered effect*, *contrastes*) sont plus adaptés, à charge au statisticien de choisir celui qui convient le mieux au problème traité.

4.11.2 Codage numérique d'une variable discrète ordinale

Une variable ordinale est une variable qualitative dont les modalités sont ordonnées (section 4.9).

Parfois, le caractère ordinal repose tout simplement sur un point de vue différent des mêmes données. Considérons la variable *habitation* comme un indicateur d'éloignement par rapport au centre-ville où seraient situés la majorité des lieux de travail. Dans ce cas, il y a bien une relation d'ordre dans les modalités prises par la variable et coder

$$Z_i = \begin{cases} 1 & \text{si } X_i = \text{"centre"} \\ 2 & \text{si } X_i = \text{"banlieue"} \\ 3 & \text{si } X_i = \text{"campagne"} \end{cases}$$

semble tout à fait licite.

Notons cependant que ce codage n'est pas totalement innocent, il introduit une information supplémentaire dont tiendra compte la régression dans le calcul des coefficients : l'amplitude de l'écart. Avec ce codage nous sommes en train de dire que

- l'écart entre "centre" et "banlieue" est de 1, il en est de même pour l'écart entre "banlieue" et "campagne" ;
- et de plus, nous affirmons également que l'écart entre "campagne" et "centre" est 2 fois plus élevé que l'écart entre "centre" et "banlieue".

De fait, nous faisons l'hypothèse d'une forme linéaire de liaison. En réalité, nous n'en savons rien. Peut-être est-ce vrai, peut être est-ce erroné. L'étude du codage polynomial ci-dessus (section 4.10) montre que l'on peut explorer des solutions plus sophistiquées dès lors que l'on émet des hypothèses sur le forme de la relation entre l'endogène et l'exogène. Quoi qu'il en soit, le pire serait de lancer les calculs sans être conscient de ce qu'on manipule.

4.12 Conclusion pour le traitement des exogènes qualitatives

Il y a 2 idées maîtresses à retenir de ce chapitre :

1. Il est possible d'effectuer une régression linéaire multiple avec des exogènes qualitatives, le tout est de produire une transformation appropriée des données ;
2. Le codage est primordial car il détermine les informations que nous extrayons des données initiales et, par conséquent, l'interprétation des coefficients fournis par la régression.

L'analyse devient particulièrement intéressante lorsque nous introduisons plusieurs exogènes qualitatives ou un mélange de variables qualitatives et quantitatives. La technique est riche et ses applications multiples. L'étude des interactions, entres autres, se révèle particulièrement passionnante (voir [6]).

Tester les changements structurels

Le test de changement structurel est défini naturellement pour les données longitudinales : l'idée est de vérifier qu'au fil du temps, la nature de la relation entre l'endogène et les exogènes n'a pas été modifiée. Statistiquement, il s'agit de contrôler que les coefficients de la régression sont les mêmes quelle que soit la sous-période étudiée.

Prenons un cas simple pour illustrer cela. On veut expliquer le niveau de production des entreprises d'un secteur en fonction du temps. En abscisse, nous avons l'année, en ordonnée la production. A une date donnée, nous observons que la relation est modifiée brutalement, parce qu'il y a eu, par exemple, une mutation technologique introduisant une hausse de la productivité (Figure 5.1). Il est évident dans ce cas qu'il n'est pas possible d'effectuer une seule régression pour toute la période, la pente de la droite de régression est modifiée.

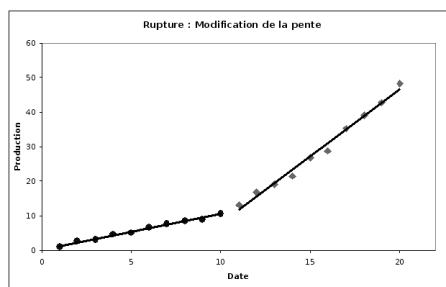


Fig. 5.1. Rupture de structure : modification de la pente à la date $t = 11$

Mettons maintenant qu'à la date $t = 11$ est survenue une catastrophe détruisant une partie de l'outil de travail. Dans ce cas, la production connaît un recul fort, puis évolue de la même manière que naguère. Dans ce cas, la pente de la régression reste identique, seule est modifiée l'origine (la constante) de la régression (Figure 5.2).

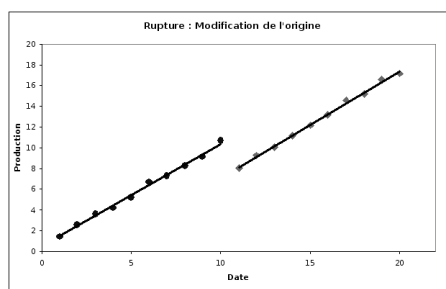


Fig. 5.2. Rupture de structure : modification de l'origine à la date $t = 11$

Extension aux données transversales

Chercher des points d'inflexion. La notion de rupture de structure est extensible aux données transversales. Il suffit d'imaginer la relation entre la puissance et la taille du moteur. À partir d'un certain stade, augmenter indéfiniment la cylindrée entraîne une amélioration infime de la puissance (Figure 5.3). La relation est peut-être non-linéaire. Le test de changement structurel permet de localiser le point d'inflexion de la courbe de régression si l'on triait les données selon l'exogène.

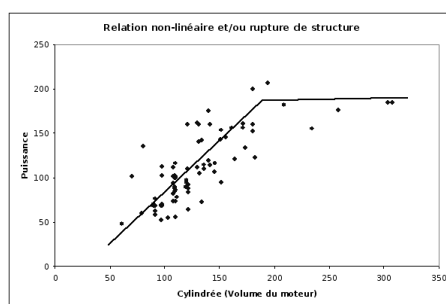


Fig. 5.3. Relation non-linéaire ou linéaire par morceaux ?

Travailler sur des populations différentes. Toujours dans le même domaine, on sait que la technologie des moteurs fonctionnant au gazole et à l'essence est quelque peu différente. Fractionner les données en 2 parties, selon le type de carburant, permet de mettre à jour l'existence de 2 populations avec des comportements, éventuellement, différents.

Bref, le test de changement structurel vise avant tout à constater statistiquement des modifications de comportement dans l'échantillon étudié. À charge au statisticien de caractériser au mieux ce qui permet de définir les sous-ensembles que l'on confronte (en utilisant des informations externes ou une variable supplémentaire disponible dans les données) et déceler la nature du changement survenu (modification des coefficients relatifs à quelles variables?).

Pour une étude approfondie de la détection et de la caractérisation des changements structurels dans la régression, je conseille la lecture attentive du chapitre 4 de l'ouvrage de Johnston (pages 111 à 145).

C'est une des rares références, en français, qui explicite avec autant de détails l'étude des ruptures de structure dans la régression.

5.1 Régression contrainte et régression non-contrainte - Test de Chow

5.1.1 Formulation et test statistique

Les tests de changements structurels reposent sur la confrontation d'une régression contrainte (a) avec une régression non-contrainte (b) (ou tout du moins, avec moins de contraintes)¹. L'objectif est de déterminer si, sur les deux sous-ensembles (sous-périodes) étudiées, certains coefficients de la régression sont les mêmes. On peut comparer plusieurs coefficients simultanément.

La démarche est la suivante :

- (a) On effectue la régression sur l'échantillon complet (n observations). C'est la régression "contrainte" dans le sens où les coefficients doivent être les mêmes quelle que soit la sous-population (sous-période) étudiée.

$$y_i = a_0 + a_1 x_{i,1} + \dots + a_p x_{i,p} + \varepsilon_i, i = 1, \dots, n \quad (5.1)$$

- (b) On effectue 2 régressions indépendantes sur les 2 sous-populations. Ce sont les régressions "non-contraintes" dans le sens où nous n'imposons pas que les coefficients soient les mêmes sur les 2 sous-populations (sous-périodes).

$$\begin{aligned} y_i &= a_{0,1} + a_{1,1} x_{i,1} + \dots + a_{p,1} x_{i,p} + \varepsilon_i, i = 1, \dots, n_1 \\ y_i &= a_{0,2} + a_{1,2} x_{i,1} + \dots + a_{p,2} x_{i,p} + \varepsilon_i, i = n_1 + 1, \dots, n \text{ (} n_2 \text{ obs.)} \end{aligned}$$

Il y a alors plusieurs manières d'appréhender le test de rupture de structure.

1. Est-ce que la régression contrainte est d'aussi bonne qualité que les 2 régressions non-contraintes ? Si oui, cela indiquerait qu'il n'y a pas à distinguer les régressions dans les 2 sous-populations : ce sont les mêmes. Pour cela, nous confrontons la somme des carrés des résidus (qui est un indicateur de qualité de la régression, plus elle faible, meilleure est l'approximation)

(a) Régression contrainte : SCR

(b) Régressions non-contraintes : SCR_1 et SCR_2

1. Sur l'idée de confronter 2 régressions, dont une serait une restriction de l'autre, voir l'excellent document de T. Duchesne, Chapitre 3, Section 3.6 "Le principe de somme de carrés résiduels additionnelle" ; <http://archimede.mat.ulaval.ca/pages/genest/regression/chap3.pdf>. La réflexion sur le mode de calcul des degrés de liberté est très instructive.

Par construction,

$$SCR \geq SCR_1 + SCR_2$$

Si SCR est "significativement" plus grand que $SCR_1 + SCR_2$, il y a bien une différence. Reste bien sûr à quantifier le "significativement".

2. On peut y répondre en appréhender le problème sous forme d'un test d'hypothèses. Nous opposons

$$H_0 : \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} a_{0,1} \\ a_{1,1} \\ \vdots \\ a_{p,1} \end{pmatrix} = \begin{pmatrix} a_{0,2} \\ a_{1,2} \\ \vdots \\ a_{p,2} \end{pmatrix}$$

H_1 : un des coefficients (au moins) diffère des autres

La statistique du test de Chow² s'appuie sur les sommes des carrés résiduels des régressions contraintes (SCR) et non-contraintes (SCR_1 et SCR_2). Elle s'écrit :

$$F = \frac{[SCR - (SCR_1 + SCR_2)] / ddl_n}{(SCR_1 + SCR_2) / ddl_d}$$

Plus que les valeurs génériques des degrés de liberté, voyons en détail le mécanisme de leur formation afin que nous puissions le reproduire dans d'autres configurations.

Pour ddl_d , qui est le plus facile à appréhender, nous avons la réunion de 2 régressions indépendantes :

$$\begin{aligned} ddl_d &= (n_1 - p - 1) + (n_2 - p - 1) \\ &= (n_1 + n_2) - 2p - 2 \\ &= n - 2p - 2 \\ &= n - 2(p + 1) \end{aligned}$$

Pour ddl_n , la situation est un peu plus complexe :

$$\begin{aligned} ddl_n &= (n - p - 1) - [(n_1 - p - 1) + (n_2 - p - 1)] \\ &= (n - p - 1) - (n - 2p - 2) \\ &= p + 1 \end{aligned}$$

A posteriori, ($ddl_n = p + 1$) semble évident. En effet, nous avons bien $(p + 1)$ contraintes linéaires dans l'hypothèse nulle de notre test de comparaison des coefficients.

2. Gregory C. Chow (1960). *Tests of Equality Between Sets of Coefficients in Two Linear Regressions*. in *Econometrica* 28(3) : 591-605.

Sous H_0 , la statistique F suit une loi de Fisher à $(p+1, n-2p-2)$ degrés de liberté. La région critique du test s'écrit

$$R.C. : F > F_{1-\alpha}(p+1, n-2p-2)$$

où $F_{1-\alpha}(p+1, n-2p-2)$ est le quantile d'ordre $(1-\alpha)$ de la loi de Fisher à $(p+1, n-2p-2)$ degrés de liberté.

5.1.2 Un exemple

Nous reprenons un exemple décrit dans Johnston (pages 134 à 138). Nous voulons effectuer une régression linéaire simple $Y = aX + b + \varepsilon$. Les données (fichier CHOW) peuvent être subdivisées en 2 sous-parties (sous-périodes) correspondant à une variable supplémentaire³ (Figure 5.4).

Obs	Période	Y	X
1	1	1	2
2	1	2	4
3	1	2	6
4	1	4	10
5	1	6	13
6	2	1	2
7	2	3	4
8	2	3	6
9	2	5	8
10	2	6	10
11	2	6	12
12	2	7	14
13	2	9	16
14	2	9	18
15	2	11	20

Fig. 5.4. Données pour le test de Chow (Johnston, page 134)

Pour réaliser le test global de Chow c.-à-d. la régression est-elle la même dans les 2 sous-parties du fichier ?, nous réalisons 3 régressions : (a) sur la totalité du fichier, (b) sur la première partie, (c) sur la seconde partie. Nous obtenons les résultats suivants (Figure 5.5) :

a : $Y = 0.52X - 0.07$ avec $SCR = 6.56$ et $ddl = 13$;

b : $Y = 0.44X - 0.06$ avec $SCR_1 = 0.69$ et $ddl_1 = 3$;

c : $Y = 0.51X + 0.40$ avec $SCR_2 = 2.47$ et $ddl_2 = 8$.

Calculons les degrés de liberté : $ddl_n = 13 - (3 + 8) = 2$ et $ddl_d = 3 + 8 = 11$. La statistique du test est donc égale à

$$F = \frac{[6.56 - (0.69 + 2.47)]/2}{(0.69 + 2.47)/11} = 5.91$$

3. C'est un peu abstrait j'en conviens. Mettons que l'on veut expliquer la consommation (Y) en fonction de la taille du moteur (X). Les données regroupent les véhicules fonctionnant au gazole et à l'essence.

Obs	Periode	Y	X
1	1	1	2
2	1	2	4
3	1	2	6
4	1	4	10
5	1	6	13
6	2	1	2
7	2	3	4
8	2	3	6
9	2	5	8
10	2	6	10
11	2	6	12
12	2	7	14
13	2	9	16
14	2	9	18
15	2	11	20

Régression globale		
	X	const
coef.	0.52	-0.07
	0.03	0.37
	0.95	0.71
	252.71	13
	127.44	6.56
	SCR	

Régression période 1		
	X	const
coef.	0.44	-0.06
	0.05	0.43
	0.96	0.48
	66.82	3
	15.31	0.69
	SCR1	

Régression période 2		
	X	const
coef.	0.51	0.40
	0.03	0.38
	0.97	0.56
	276.71	8
	85.53	2.47
	SCR2	

ddl n	2
ddl d	11
SCR-(SCR1+SCR2)	3.40
SCR1+SCR2	3.16
F	5.91
p-value	0.0181

Fig. 5.5. Test global de Chow

La p-value associée est 0.0181.

Au risque de 5%, ces deux sous-parties du fichier donnent bien lieu à 2 régressions différentes ⁴.

5.2 Détecter la nature de la rupture

Il y a 2 types de ruptures dans la régression :

1. une modification de niveau, la constante n'est pas la même dans les 2 sous-périodes ;
2. une modification de pente, la relation entre l'endogène et une ou plusieurs exogènes a été modifiée.

Nous verrons dans cette section quels tests mettre en place pour caractériser ces situations.

5.2.1 Tester la stabilité de la constante

Dans ce cas, les coefficients des exogènes sont communs aux deux sous populations, seule la constante est modifiée. Le test d'hypothèses s'écrit :

$$H_0 : a_{0,1} = a_{0,2} = a_0$$

$$H_1 : a_{0,1} \neq a_{0,2}$$

En pratique, nous construisons deux variables auxiliaires qui permettent de spécifier les 2 sous-parties du fichier :

4. Au risque de 1%, la conclusion aurait été différente. Mais au vu de la taille de l'échantillon, prendre un risque critique aussi bas nous conduirait quasi-systématiquement à accepter l'hypothèse nulle.

$$d_{i,1} = \begin{cases} 1, & i = 1, \dots, n_1 \\ 0, & i = n_1 + 1, \dots, n \end{cases}$$

$$d_{i,2} = \begin{cases} 0, & i = 1, \dots, n_1 \\ 1, & i = n_1 + 1, \dots, n \end{cases}$$

Et nous construisons la régression suivante (Equation 5.2), c'est la régression non-contrainte que nous opposons à l'équation initiale (Equation 5.1) où la constante est la même sur les deux périodes.

$$y_i = a_{0,1}d_{i,1} + a_{0,2}d_{i,2} + a_1x_{i,1} + \dots + a_px_{i,p} + \varepsilon_i \quad (5.2)$$

Attention, nous n'introduisons plus de constante dans cette régression car $d_{i,1} + d_{i,2} = 1$, le calcul ne serait pas possible.

Bien entendu, nous pourrions effectuer le test d'hypothèses ($H_0 : a_{0,1} = a_{0,2}$) directement sur l'équation 5.2 (Voir "Tests de comparaisons de coefficients et tests de combinaisons linéaires de coefficients"; Bourbonnais, page 69; Johnston, pages 95 à 101). Mais il est plus simple, et plus cohérent avec notre démarche dans ce chapitre, de procéder en opposant le modèle contraint et le(s) modèle(s) non contraint(s).

Obs	Periode	Y	X	D1	D2
1	1	1	2	1	0
2	1	2	4	1	0
3	1	2	6	1	0
4	1	4	10	1	0
5	1	6	13	1	0
6	2	1	2	0	1
7	2	3	4	0	1
8	2	3	6	0	1
9	2	5	8	0	1
10	2	6	10	0	1
11	2	6	12	0	1
12	2	7	14	0	1
13	2	9	16	0	1
14	2	9	18	0	1
15	2	11	20	0	1

	D2	D1	X
coef.	0.55	-0.47	0.50
	0.34	0.30	0.03
	0.97	0.54	#N/A
	149.57	12	#N/A
	130.51	3.49	#N/A

SCR	6.56
SCR3	3.49
SCR-SCR3	3.07

ddl_n	1
ddl_d	12

F	10.5409
p-value	0.0070

Fig. 5.6. Test de la constante de régression

Pour illustrer notre propos, nous reprenons notre exemple ci-dessus (Figure 5.4). Rappelons que la régression contrainte (Équation 5.1) a fourni (Figure 5.5) : $SCR = 6.56$ et $ddl = 13$. Nous réalisons maintenant la régression non-contrainte destinée à tester la stabilité de la constante (Équation 5.2), elle nous propose les résultats suivants (Figure 5.6) :

- $SCR_3 = 3.49$ et $ddl_3 = 12$;
- pour opposer les modèles contraints et non-contraints (resp. équations 5.1 et 5.2), nous calculons tout d'abord les degrés de liberté : $ddl_n = ddl - ddl_3 = 13 - 12 = 1$ et $ddl_d = ddl_3 = 12$;
- nous formons alors la statistique $F = \frac{(SCR - SCR_3)/ddl_n}{SCR_3/ddl_3} = \frac{3.07/1}{3.49/12} = 10.54$;
- avec un $p\text{-value} = 0.0070$.

Conclusion : la différence de structure détectée par le test global de Chow serait due, au moins en partie, à une différence entre les constantes des régressions construites dans chaque sous-échantillon. "En partie" car nous n'avons pas encore testé l'influence de la pente de régression, c'est l'objet de la section suivante.

5.2.2 Tester la stabilité du coefficient d'une des exogènes

Une formulation erronée

Il s'agit maintenant de tester si la rupture est imputable à une modification de la pente de la régression c.-à-d. un ou plusieurs coefficients associés à des exogènes ne sont pas les mêmes sur les deux périodes.

Nous traitons dans cette section, sans nuire à la généralité du discours, du test du coefficient associé à la variable x_1 de la régression.

Forts des schémas décrit précédemment, nous dérivons deux variables intermédiaires z_1 et z_2 à partir de la variable x_1 avec :

$$z_{i,1} = \begin{cases} x_{i,1}, & i = 1, \dots, n_1 \\ 0 & , i = n_1 + 1, \dots, n \end{cases}$$

$$z_{i,2} = \begin{cases} 0 & , i = 1, \dots, n_1 \\ x_{i,1}, & i = n_1 + 1, \dots, n \end{cases}$$

Nous pourrions alors être tenté de proposer comme formulation non-contrainte de la régression :

$$y_i = a_0 + a_{1,1}z_{i,1} + a_{1,2}z_{i,2} + \dots + a_px_{i,p} + \varepsilon_i \quad (5.3)$$

Que nous opposerions au modèle initial (Équation 5.1).

En fait, cette formulation du test est erronée, principalement pour 2 raisons :

1. Une modification de la pente entraîne *de facto* une modification de l'origine de la régression. Un exemple fictif, construit sur une régression simple illustre bien la situation (Figure 5.7).
2. En contraignant les deux équations, contraints et non-contraints, à avoir la même origine, nous faussons les résultats relatifs au test de la pente (Figure 5.8).

En conclusion, pour tester la stabilité des coefficients sur 2 sous-ensembles de données, il faut absolument relâcher, dans le modèle de référence, la contrainte de stabilité de la constante.

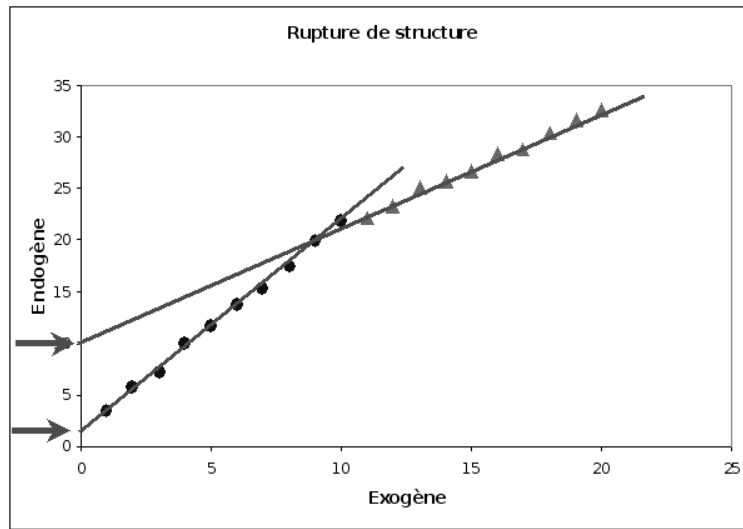


Fig. 5.7. Un changement de pente entraîne automatiquement une modification de l'origine

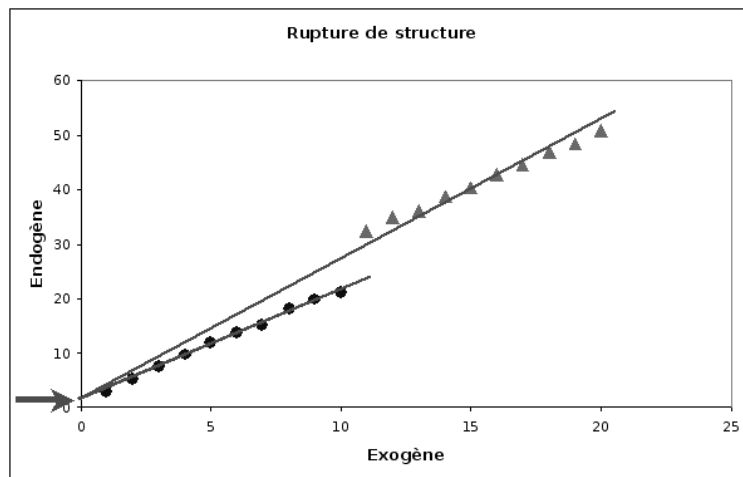


Fig. 5.8. En imposant la même origine aux deux régressions, on fausse l'appréciation des pentes

Tester la pente en relâchant la contrainte sur la constante

A la lumière de ces éléments, il apparaît que le modèle de référence, le modèle contraint, devient maintenant celui où les constantes sont possiblement différentes sur les 2 sous-parties du fichier (Équation 5.2). Et nous lui opposons le modèle :

$$y_i = a_{0,1}d_{i,1} + a_{0,2}d_{i,2} + a_{1,1}z_{i,1} + a_{1,2}z_{i,2} + \dots + a_px_{i,p} + \varepsilon_i \quad (5.4)$$

L'hypothèse nulle du test est naturellement $H_0 : a_{1,1} = a_{1,2}$.

Obs	Periode	Y	X	D1	D2	Z1	Z2
1	1	1	2	1	0	2	0
2	1	2	4	1	0	4	0
3	1	2	6	1	0	6	0
4	1	4	10	1	0	10	0
5	1	6	13	1	0	13	0
6	2	1	2	0	1	0	2
7	2	3	4	0	1	0	4
8	2	3	6	0	1	0	6
9	2	5	8	0	1	0	8
10	2	6	10	0	1	0	10
11	2	6	12	0	1	0	12
12	2	7	14	0	1	0	14
13	2	9	16	0	1	0	16
14	2	9	18	0	1	0	18
15	2	11	20	0	1	0	20

	Z2	Z1	D2	D1
coef.	0.51	0.44	0.40	-0.06
	0.03	0.06	0.37	0.48
	0.98	0.54	#N/A	#N/A
	113.86	11	#N/A	#N/A
	130.84	3.16	#N/A	#N/A

SCR_3	3.49
SCR_4	3.16
SCR_3-SCR_4	0.33

ddl_n	1
ddl_d	11

F	1.15
p-value	0.3068

Fig. 5.9. Test de la pente de régression

Reprenons notre fichier de données et mettons en place ces calculs. Pour notre modèle de référence (Équation 5.2), nous avons obtenu $SCR_3 = 3.49$ et $ddl_3 = 12$. Dans la nouvelle régression (Equation 5.4), nous avons (Figure 5.9) :

- $SCR_4 = 3.16$ et $ddl_4 = 11$;
- on calcule les degrés de libertés $ddl_n = ddl_3 - ddl_4 = 12 - 11 = 1$ et $ddl_d = ddl_4 = 11$;
- la statistique du test s'écrit alors $F = \frac{(SCR_3 - SCR_4)/ddl_n}{SCR_4/ddl_d} = \frac{(3.49 - 3.16)/1}{3.16/11} = 1.15$;
- avec une p-value = 0.3068.

Les différences détectées entre les régressions sur les 2 sous-parties du fichier ne sont pas imputables à une modification de la pente. En d'autres termes, la pente de la régression est la même dans les 2 sous-populations.

Moralité de tout ceci, concernant notre fichier de données : il y a bien une rupture de structure entre les 2 sous-populations, elle est essentiellement due à une modification de la constante. A vrai dire, un nuage de points nous aurait permis de très vite aboutir aux mêmes conclusions (Figure 5.10), à la différence que la démarche décrite dans cette section est applicable quelle que soit le nombre de variables exogènes.

5.3 Conclusion

L'étude des changements structurels peut être étendue à l'analyse de k sous-populations (ou sous-périodes). Il s'agit tout simple de définir correctement le modèle contraint, qui sert de référence, et le(s) modèle(s) non-contraint(s), qui servent à identifier la nature de la rupture. Seulement, les tests et la compréhension des résultats deviennent difficiles, voire périlleux, il faut procéder avec beaucoup de prudence.

Le véritable goulot d'étranglement de cette démarche est la détection *intuitive* du point de rupture. Encore pour les données longitudinales, quelques connaissances approfondies du domaine donnent des indications sur les événements (économiques, politiques, etc.) qui peuvent infléchir les relations entre les variables. En revanche, pour les données transversales, deviner le point d'inflexion sur une variable

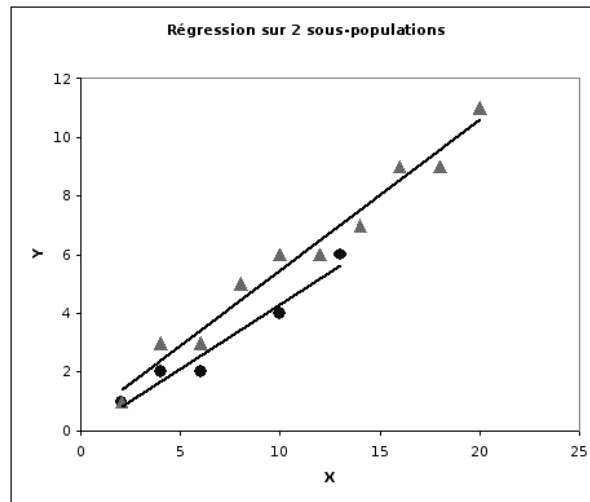


Fig. 5.10. Nuage de points (X,Y) et droites de régression pour les deux sous-populations de notre fichier exemple (Figure 5.4)

exogène, ou encore déterminer le facteur externe qui vient modifier la structure des dépendances, relève du saut dans l'inconnu. Très souvent, les graphiques, notamment des résidus, sont d'une aide précieuse pour *flairer* les ruptures dans les données.

Détection et traitement de la non linéarité

Nous avons abordé le sujet de la non-linéarité dans notre support consacré à la régression linéaire simple (et multiple) [18](chapitre 6). Nous nous étions surtout intéressés aux configurations où, sur la base des connaissances du domaine et de l'interprétation que l'on souhaitait donner aux résultats, nous choissions une forme particulière de la relation. Il était possible de retrouver une forme linéaire, dont les coefficients pouvaient être estimés avec la méthode des moindres carrés ordinaires, en appliquant les fonctions de transformations idoines.

Nous allons plus loin dans ce chapitre. Nous nous basons sur un processus guidé par les données et non plus par les connaissances du domaine. Il y a toujours une double étape : évaluer la compatibilité des données avec l'hypothèse d'une relation linéaire ; si elle est réfutée, trouver la transformation de variables la plus appropriée de manière à améliorer la qualité de l'ajustement. Bien évidemment, il y a une limite (très difficile à trouver) à ne pas dépasser dans la recherche du meilleur modèle. Il s'agit d'exploiter au mieux les informations véhiculées par les données, sans en ingérer les spécificités qui ne sont pas généralisables dans la population. En apprentissage supervisé, on parlerait du problème de sur-ajustement des modèles.

6.1 Non linéarité dans la régression simple

6.1.1 Linéarisation par transformation de variables

Dans le cadre de la régression simple, la détection graphique est une approche privilégiée. Elle permet de détecter l'éventuelle non-linéarité de la relation et, de plus, elle donne une idée sur la transformation à opérer pour obtenir un modèle satisfaisant.

Pour illustrer notre propos, nous reprenons un exemple tiré de l'ouvrage de Aïvazian (pages 148 et 149, données "Éprouvettes"). On étudie la résistance à la rupture des éprouvettes de ciment (Y , en kg/m^2) en fonction de la durée de maintien (X , en jours) (Figure 6.1). Manifestement, la relation est non-linéaire. L'ajustement à l'aide de la régression linéaire simple n'est pas très satisfaisante ($R^2 = 0.6199$).

X	Y
1	13
1	13.3
1	11.8
2	21.9
2	24.5
2	24.7
3	29.8
3	28
3	24.1
3	24.2
3	26.2
7	32.4
7	30.4
7	34.5
7	33.1
7	35.7
28	41.8
28	42.6
28	40.3
28	35.7
28	37.5

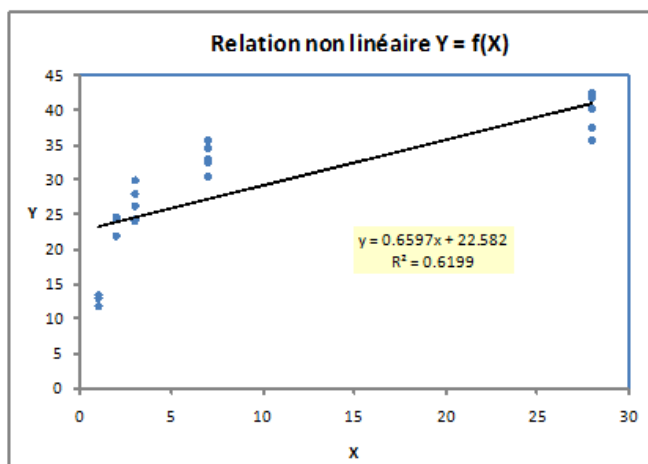
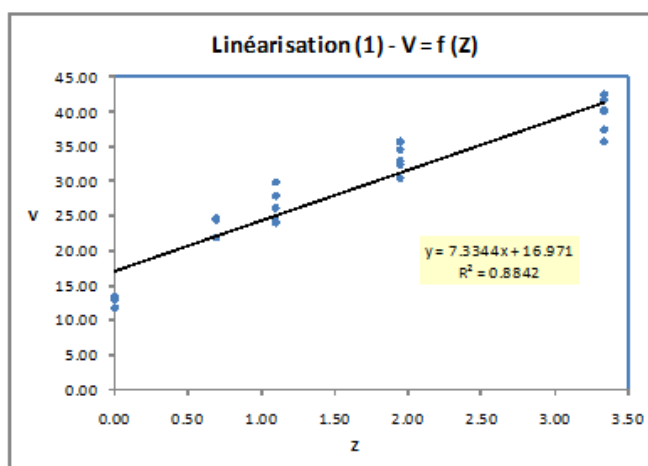


Fig. 6.1. Liaison linéaire - Données "Éprouvettes"

Il nous faut proposer les bonnes transformations de variables. L'affaire est loin d'être évidente. Dans notre exemple, on voit bien que la relation est logarithmique. Pourtant, en tentant la transformation $Z = \ln(X)$, $V = Y$ n'étant pas modifiée, la qualité de l'ajustement ($V = a \times Z + b$) laisse à désirer avec $R^2 = 0.8842$ (Figure 6.2). Nous constatons également un autre élément gênant : la dispersion de V selon les valeurs de Z n'est visiblement pas constante, nous sommes en situation d'hétéroscédasticité.

$Z = \ln(X)$	$V = Y$
0.00	13.00
0.00	13.30
0.00	11.80
0.69	21.90
0.69	24.50
0.69	24.70
1.10	29.80
1.10	28.00
1.10	24.10
1.10	24.20
1.10	26.20
1.95	32.40
1.95	30.40
1.95	34.50
1.95	33.10
1.95	35.70
3.33	41.80
3.33	42.60
3.33	40.30
3.33	35.70
3.33	37.50

Fig. 6.2. Données "Éprouvettes" - $V = f(Z)$ avec $V = Y$ et $Z = \ln(X)$.

La "bonne" solution nous est fournie par Aïvazian (pages 149 et 150), il propose les transformations

$$V = \log_{10}(Y)$$

$$Z = \frac{1}{X}$$

Bien malin aurait été celui qui y aurait pensé. Il faut à la fois de l'intuition, de bonnes connaissances du domaine, une certaine pratique, pour proposer rapidement les bonnes formules. Concernant notre exemple, elles sont tout à fait justifiées puisque la qualité de l'ajustement est grandement améliorée ($R^2 = 0.9612$) (Figure 6.3).

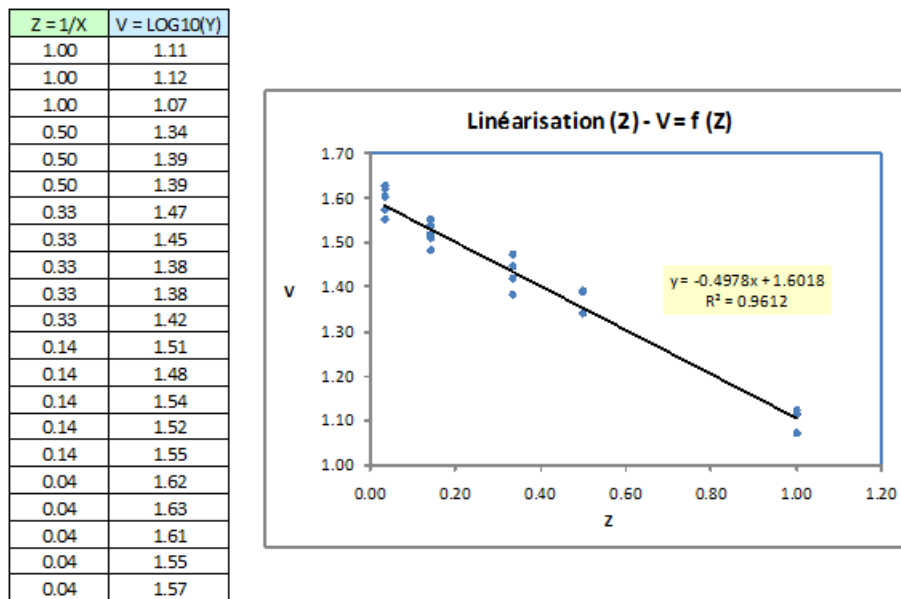


Fig. 6.3. Données "Éprouvettes" - $V = f(Z)$ avec $V = \log_{10}(Y)$ et $Z = 1/X$.

Bref, la recherche du "bon" modèle à travers les transformations de variables peut s'avérer très fructueux. Passer d'un R^2 de 0.6199 à 0.9612 est une avancée indéniable. Mais il faut être capable d'introduire les bonnes transformations de variables. La solution n'est pas toujours évidente.

6.1.2 Détecter numériquement la non-linéarité dans la régression simple

La détection graphique présente un double avantage : nous pouvons déceler une éventuelle non-linéarité, nous disposons de pistes sur la "bonne" forme de la relation. Elle n'est pas adaptée en revanche pour le traitement automatisé d'un grand nombre de variables. Dans ce contexte, rien ne remplace un critère numérique qui permet, au moins dans une première étape, d'isoler les cas à problèmes sur lesquels nous devrions nous pencher plus attentivement. Dans cette section, nous présentons une approche pour détecter numériquement - avec une procédure statistique, le résultat est probabiliste - les relations non linéaires. Pour ce faire, nous opposerons deux mesures d'associations des variables quantitatives :

le premier, le rapport de corrélation, ne fait aucune hypothèse sur la forme de la liaison ; le second, le coefficient de corrélation linéaire de Pearson, mesure la force d'une liaison linéaire. La divergence entre ces indicateurs permet de caractériser la nature non-linéaire de la relation.

Le rapport de corrélation

Le rapport de corrélation est une mesure "universelle" de la liaison entre deux variables quantitatives. "Universelle" car elle est valable quelle que soit la forme de la relation, même si cette dernière est non-monotone.

Le rapport de corrélation est une mesure asymétrique. Il résulte de la confrontation de deux espérances mathématiques ([17], section 3.6.) : la moyenne de la variable dépendante, et sa moyenne conditionnellement aux valeurs de X . Plus fort sera l'impact de X sur Y , plus élevé sera l'écart entre ces deux quantités.

Concrètement, sur un échantillon de taille n pour lequel nous disposons de K valeurs distinctes de X , et pour chaque valeur de X , nous disposons de n_k valeurs de Y , le rapport de corrélation empirique s'écrit :

$$\eta_{y/x}^2 = \frac{\sum_{k=1}^K n_k (\bar{y}_k - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (6.1)$$

On distingue au numérateur la variance inter-classes, la variabilité de Y expliquée par X ; et au dénominateur, la variance totale de Y . Les habitués de l'analyse de variance ne seront pas dépayés. On peut d'ailleurs utiliser le rapport de corrélation pour mesurer l'influence d'une variable indépendante qualitative nominale sur une variable dépendante quantitative.

Par définition, le rapport de corrélation varie entre 0 et 1 ($0 \leq \eta_{y/x}^2 \leq 1$). Il est nul si la liaison n'existe pas ; il est égal à 1 si X explique parfaitement les valeurs prises par Y . Il est possible de mettre en place un test de significativité ([17], section 3.6.2).

Remarque 28 (Cas des données individuelles, non-groupées). Dans le cas des données non-groupées c.-à-d. à chaque valeur de X , on dispose d'une seule valeur de Y , nous avons $K = n$ et $n_k = 1, \forall k$. Par construction, $\eta_{y/x}^2$ est toujours égal à 1, laissant croire une liaison parfaite. Dans ce cas, il est judicieux de procéder artificiellement à des regroupements en découpant en intervalles les valeurs de X . On peut, par exemple, utiliser la technique des intervalles d'amplitudes égales¹. Le choix du nombre de classes est crucial. Il faut qu'il soit suffisamment faible pour que l'effectif dans chaque classe permettent d'obtenir des moyennes \bar{y}_k qui aient un sens ; il faut qu'il soit suffisamment élevé pour que la forme de la relation entre Y et X ne soit pas occultée. Il dépend aussi du nombre d'observations disponibles.

Le coefficient de corrélation linéaire

Le coefficient de corrélation de Pearson mesure le degré de liaison *linéaire* entre deux variables Y et X ([17], chapitre 2). Le coefficient de corrélation empirique, calculé à partir d'un échantillon, est obtenu

1. <http://www.info.univ-angers.fr/~gh/wstat/discr>

de la manière suivante :

$$r_{yx} = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_i (y_i - \bar{y})^2 \times \sum_i (x_i - \bar{x})^2}} \quad (6.2)$$

On reconnaît au numérateur la covariance entre Y et X , elle est normalisée au dénominateur par le produit des écarts-type. Le coefficient est compris entre -1 et $+1$ ($-1 \leq r_{yx} \leq +1$), le signe tient compte du sens de la relation. Nous perdons cette information en passant au carré r_{yx}^2 ($r_{yx}^2 \geq 0$). Mais nous y gagnons en interprétation car l'indicateur correspond au coefficient de détermination de la régression² de Y sur X dans ce cas c.-à-d. il représente la part de variance de Y expliquée par X .

Tester la non linéarité par comparaison de $\eta_{y/x}^2$ et r_{yx}^2

$\eta_{y/x}^2$ et r_{yx}^2 se rejoignent en termes d'interprétation, à la différence que l'on introduit une contrainte de linéarité dans le second indicateur. De fait

$$\eta_{y/x}^2 \geq r_{yx}^2$$

Il y aurait égalité si et seulement si la relation est parfaitement linéaire. Nous exploitons l'amplitude de la différence entre ces indicateurs pour caractériser le caractère non linéaire de la liaison entre Y et X .

Le test de linéarité de la relation revient donc à tester la significativité de la différence entre ces deux indicateurs. Nous utilisons la statistique de test (Aïvazian, page 121 ; Veyseyre, page 368 ; Dagnelie³, page 483)

$$W^2 = \frac{(\eta_{y/x}^2 - r_{yx}^2)/(K - 2)}{(1 - \eta_{y/x}^2)/(n - K)} \quad (6.3)$$

Sous H_0 , W^2 suit une loi de Fisher à $(K - 2, n - K)$ degrés de liberté. La région critique correspond aux valeurs élevées de W^2 . Nous pouvons aussi calculer la probabilité critique du test (p-value). Dans ce cas, nous rejetons l'hypothèse nulle si la p-value est inférieure au risque α du test.

Application aux données "Éprouvettes" (1)

Nous souhaitons vérifier la nature de la relation entre Y et X pour les données "Éprouvettes" (Figure 6.1). Nous avons constaté graphiquement que l'hypothèse de linéarité n'était pas vraiment crédible. Voyons ce que nous dit la procédure numérique.

Nous avons monté une feuille Excel pour réaliser les calculs (Figure 6.4). Tout d'abord, nous devons calculer le rapport de corrélation :

2. Et de la régression de X sur Y aussi d'ailleurs.

3. Dagnelie, P., *Statistique théorique et appliquée - 2. Inférence statistique à une et deux dimensions*, de Boeck, 2006 ; la présentation est un peu différente mais le principe est le même : on teste la significativité de la différence entre les deux indicateurs.

X	Y
1	13
1	13.3
1	11.8
2	21.9
2	24.5
2	24.7
3	29.8
3	28
3	24.1
3	24.2
3	26.2
7	32.4
7	30.4
7	34.5
7	33.1
7	35.7
28	41.8
28	42.6
28	40.3
28	35.7
28	37.5

Rapport de corrélation			
Valeurs			
Étiquettes de lignes	Moyenne de Y	Nombre de Y	$n_k \cdot (y_{b_k} - \bar{y})^2$
1	12.70	3	780.85
2	23.70	3	79.05
3	26.46	5	28.16
7	33.22	5	96.21
28	39.58	5	577.45
Total général	28.83	21	1561.74

Numérateur η^2	1561.7387
Dénominateur η^2	1642.5267

η^2	0.9508
----------	--------

Coefficient de corrélation	
r de Pearson	0.7873
r^2	0.6199

Test de linéarité	
n	21
K	5
W^2	35.8864
ddl1	3
ddl2	16
p-value	2.4669E-07

Fig. 6.4. Données "Éprouvettes" - Test de linéarité - Variables originales

- Avec l'outil "tableau croisés dynamiques", nous obtenons les moyennes de Y pour chaque valeur distincte de X . Nous calculons le numérateur de $\eta_{y/x}^2$:

$$\sum_{k=1}^K n_k (\bar{y}_k - \bar{y})^2 = 3 \times (12.70 - 28.83)^2 + 3 \times (23.70 - 28.83)^2 + \dots = 1561.7387$$

- Au dénominateur, nous avons la somme des carrés des écarts à la moyenne

$$\sum_{i=1}^n (y_i - \bar{y})^2 = (13 - 28.83)^2 + (13.3 - 28.83)^2 + \dots = 1642.5267$$

- Dès lors,

$$\eta_{y/x}^2 = \frac{\sum_{k=1}^K n_k (\bar{y}_k - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{1561.7387}{1642.5267} = 0.9508$$

Pour obtenir le coefficient de corrélation, nous utilisons la fonction COEFFICIENT.CORRELATION d'Excel. Elle nous fournit $r_{yx} = 0.7873$, nous la passons au carré

$$r_{yx}^2 = (0.7873)^2 = 0.6199$$

Nous pouvons calculer maintenant la statistique de test

$$W^2 = \frac{(0.9508 - 0.6199)/(5 - 2)}{(1 - 0.9508)/(21 - 5)} = 35.8864$$

Avec la loi de Fisher $\mathcal{F}(5-2, 21-5)$, nous avons une probabilité critique (p-value) largement inférieure au niveau de signification $\alpha = 5\%$ que nous nous sommes choisis. Les données ne sont pas compatibles avec l'hypothèse de linéarité.

Application aux données "Éprouvettes" (2)

Réitérons l'analyse après transformation des variables. Pour rappel, nous avons $V = \log_{10}(Y)$ et $Z = \frac{1}{X}$. La structure de la feuille de calcul n'est pas modifiée par rapport à la précédente. Nous obtenons à présent (Figure 6.5) :

$$\eta_{v/z}^2 = 0.9683$$

$$r_{vz}^2 = 0.9612$$

$$W^2 = \frac{(0.9683 - 0.9612)/(5 - 2)}{(1 - 0.9683)/(21 - 5)} = 1.1865$$

Toujours avec un $\mathcal{F}(3, 16)$, nous avons une p-value de 0.3462. Après transformation judicieuse des variables, l'hypothèse de linéarité devient licite. L'analyse graphique est confirmée (Figure 6.3).

Z = 1/X	V = LOG10(Y)
1.00	1.11
1.00	1.12
1.00	1.07
0.50	1.34
0.50	1.39
0.50	1.39
0.33	1.47
0.33	1.45
0.33	1.38
0.33	1.38
0.33	1.42
0.14	1.51
0.14	1.48
0.14	1.54
0.14	1.52
0.14	1.55
0.04	1.62
0.04	1.63
0.04	1.61
0.04	1.55
0.04	1.57

Rapport de corrélation			
Valeurs			
Étiquettes de lignes	Moyenne de V = LOG	Nombre de V = LOG	1 n k*(yb_k - yb)^2
0.04	1.5965	5	0.1313
0.14	1.5207	5	0.0372
0.33	1.4211	5	0.0009
0.50	1.3741	3	0.0109
1.00	1.1032	3	0.3292
Total général	1.4345	21	0.5095

Numérateur eta²	0.5095
Dénominateur eta²	0.5262
eta²	0.9683

Coefficient de corrélation	
r de Pearson	-0.9804
r²	0.9612

Test de linéarité	
n	21
K	5
W²	1.1865
ddl1	3
ddl2	16
p-value	0.3462

Fig. 6.5. Données "Éprouvettes" - Test de linéarité - Variables transformées

6.1.3 Tester l'adéquation d'une spécification

Au-delà du test de non-linéarité, nous pouvons vérifier l'adéquation d'une forme choisie avec les données disponibles. La procédure repose sur la confrontation entre, d'une part, la prédiction $\hat{y}(x_k) = \hat{y}_k$ du modèle pour chaque valeur observée x_k ; et, d'autre part, la prédiction triviale qui consiste à calculer la moyenne de Y pour les n_k observations associées à la valeur x_k .

La statistique du test s'écrit (Aïvazian, page 120) :

$$\nu^2 = \frac{(n - K) \times \sum_{k=1}^K n_k (\bar{y}_k - \hat{y}_k)^2}{(K - g) \times \sum_{k=1}^K \sum_{i=1}^{n_k} (y_{ik} - \bar{y}_k)^2} \quad (6.4)$$

Où n est le nombre d'observations, K est le nombre de valeurs distinctes de X , g est le nombre de paramètres du modèle ($g = 2$ pour la régression simple).

Au numérateur, la somme correspond à la variabilité résiduelle non expliquée par le modèle, dont la forme est contrainte par la spécification choisie. Au dénominateur, nous avons la variabilité de Y non expliquée par les X , sans contrainte sur la forme de la relation. Le rapport est forcément supérieur à 1; s'il s'en écarte significativement, nous pouvons conclure à l'inadéquation de la spécification choisie.

Sous H_0 , la forme choisie est compatible avec les données, ν^2 suit une loi de Fisher à $(K - g, n - K)$ degrés de liberté. La région critique correspond aux valeurs trop élevées de ν^2 .

Modèle logarithmique pour les données "Éprouvettes"

Précédemment, nous avons tenté le modèle logarithmique pour les données "Éprouvettes" (Figure 6.2). Nous reproduisons la courbe de tendance ici en représentant les données dans leur repère initial (X, Y) (Figure 6.6). Manifestement, la courbe d'ajustement n'est pas satisfaisante même si elle introduit une amélioration sensible par rapport au modèle linéaire.

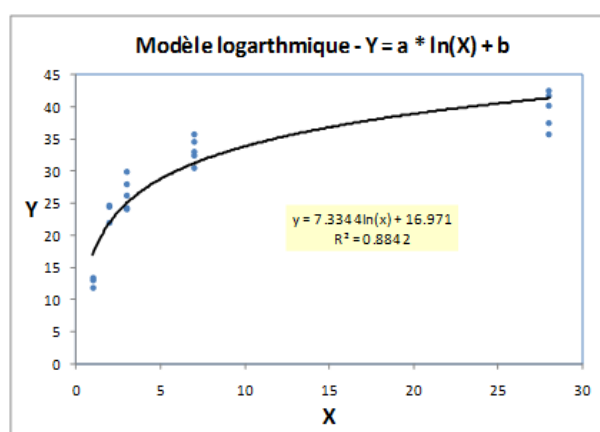


Fig. 6.6. Données "Éprouvettes" - Modèle logarithmique

Voyons si cette intuition est confirmée par le test de spécification développé dans cette section.

X	Y	Z	PRED.Y
1	13	0	16.9715
1	13.3	0	16.9715
1	11.8	0	16.9715
2	21.9	0.6931	22.0553
2	24.5	0.6931	22.0553
2	24.7	0.6931	22.0553
3	29.8	1.0986	25.0292
3	28	1.0986	25.0292
3	24.1	1.0986	25.0292
3	24.2	1.0986	25.0292
3	26.2	1.0986	25.0292
7	32.4	1.9459	31.2436
7	30.4	1.9459	31.2436
7	34.5	1.9459	31.2436
7	33.1	1.9459	31.2436
7	35.7	1.9459	31.2436
28	41.8	3.3322	41.4112
28	42.6	3.3322	41.4112
28	40.3	3.3322	41.4112
28	35.7	3.3322	41.4112
28	37.5	3.3322	41.4112

DROITEREG (Z)	
a * Z	+ b
7.3344	16.9715

Valeurs						
Étiquettes	Moyenne de Y	PRED.Y	Nombre de Y	Varp de Y	NUM	DENOM
1	12.7	16.971	3	0.42	54.737	1.260
2	23.7	22.055	3	1.6267	8.115	4.880
3	26.46	25.029	5	4.8544	10.237	24.272
7	33.22	31.244	5	3.2856	19.531	16.428
28	39.58	41.411	5	6.7896	16.766	33.948
Total général	28.833	28.833	21.000	78.216	109.387	80.788

nu²	7.2213
ddl1	3
ddl2	16
p-value	0.0028

Fig. 6.7. Données "Éprouvettes" - Modèle logarithmique - Test de spécification

Nous disposons des valeurs pour les variables X et Y dans une feuille Excel (Figure 6.7)

- Nous ajoutons la colonne $Z = \ln(X)$ pour former la régression

$$Y = 7.3344 \times Z + 16.9715$$

- Nous formons alors la colonne de prédiction \hat{y}_i (ex. $\hat{y}_1 = 7.3344 \times \ln(1) + 16.9715 = 16.9715$).
- A l'aide de l'outil "tableaux croisés dynamiques", nous calculons les statistiques intermédiaires pour chaque valeur distincte de X (ces valeurs sont $\{1, 2, 3, 7, 28\}$). Nous retrouvons de gauche à droite dans la grille :

$$\bar{y}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} y_i \quad (\text{ex. } \bar{y}_1 = 12.7)$$

$$\hat{y}_k = \hat{a} \times \ln(x_k) + \hat{b} \quad (\text{ex. } \hat{y}_1 = 16.971)$$

$$n_k = \sum_{i:x_i=x_k} 1 \quad (\text{ex. } n_1 = 3)$$

$$s_k^2 = \frac{1}{n_k} \sum_{i=1}^{n_k} (y_i - \bar{y}_k)^2 \quad (\text{ex. } s_1^2 = 0.42)$$

- Nous calculons le numérateur de ν^2

$$\begin{aligned}
 (n - K) \sum_k n_k (\bar{y}_k - \hat{y}_k)^2 &= (21 - 5) \times (3 \times (12.7 - 16.971)^2 + 3 \times (23.7 - 22.055)^2 + \dots \\
 &= 22 \times 109.387 \\
 &= 1750.2
 \end{aligned}$$

- Et le dénominateur

$$\begin{aligned}
(K - g) \times \sum_{k=1}^K \sum_{i=1}^{n_k} (y_{ik} - \bar{y}_k)^2 &= (K - g) \times \sum_{k=1}^K n_k \times s_k^2 \\
&= (5 - 2) \times (3 \times 0.42 + 3 \times 1.6267 + \dots) \\
&= 3 \times 80.788 = 242.364
\end{aligned}$$

- Il reste à former le rapport

$$\nu^2 = \frac{1750.2}{242.364} = 7.2213$$

- Avec un $\mathcal{F}(3, 16)$, nous obtenons une probabilité critique (p-value) de 0.0028.

Les données ne cadrent pas avec un modèle logarithmique. Nous devons trouver une spécification plus appropriée. Ce que nous avons fait avec les transformations $V = \log_{10}(Y)$ et $Z = \frac{1}{X}$ qui donnent pleinement satisfaction (Figure 6.3).

6.2 Non linéarité dans la régression multiple

L'affaire se corse quand il s'agit de passer à la régression linéaire multiple. En effet, nous sommes en présence de plusieurs variables explicatives. Même si les nuages de points dans le repère (X_j, Y) peuvent être intéressants pour analyser le rôle de chaque exogène, ils sont faussés parce que les variables ne sont pas indépendantes, il y a interaction lors de la régression. Nous devons nous tourner vers un outil à la portée plus large pour détecter et traiter la non-linéarité.

6.2.1 Lecture des résidus

Le graphique des résidus est un outil de diagnostic privilégié dans la régression linéaire multiple. Nous avons détaillé sa lecture dans la section 1.1.1. Pour évaluer la (non)linéarité par rapport à l'exogène, nous créons le graphique avec les résidus en ordonnée et les valeurs de X en abscisse. Les points devraient être disposés totalement aléatoirement. Dès qu'une forme de régularité quelconque apparaît, il faut se méfier et approfondir l'analyse.

Concernant notre exemple des données "Éprouvettes", nous avons opéré une régression linéaire, puis calculé la prédiction et l'erreur. Nous avons alors formé le graphique des résidus (Figure 6.8; la prédiction s'écrit $\hat{y}_i = 0.6597 \times x_i + 22.5816$, et le résidu $\hat{\varepsilon}_i = y_i - \hat{y}_i$). Manifestement, il y a un problème. La dispersion des résidus dépend des valeurs de X , c'est le signe d'une hétéroscédasticité. Pire, leur valeur moyenne dépend également des valeurs de X , là nous sommes clairement confrontés à un problème de non-linéarité.

Malheureusement, cette démarche n'est pas transposable à la régression multiple, tout simplement parce que nous avons plusieurs exogènes, elles sont plus ou moins liées. Nous nous tournons alors vers les résidus partiels.

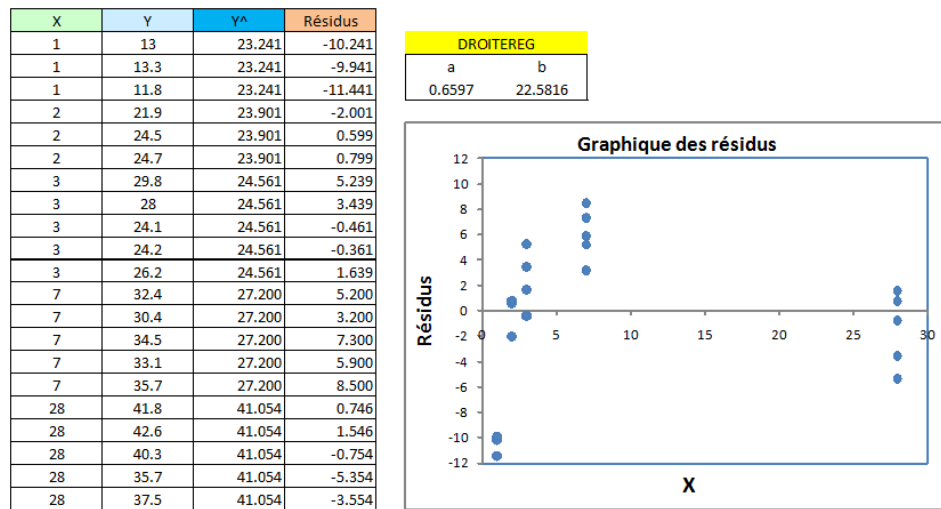


Fig. 6.8. Données "Éprouvettes" - Graphique des résidus de la régression linéaire

6.2.2 Résidus partiels et résidus partiels augmentés

Résidus partiels

Les résidus partiels permettent d'identifier la nature de la relation entre une des exogènes X_j et l'endogène Y , sachant qu'il y a d'autres variables explicatives dans le modèle⁴.

Dans une régression linéaire multiple

$$Y = a_0 + a_1X_1 + \dots + a_pX_p + \varepsilon$$

Les résidus partiels de la variables exogène X_j sont définis comme suit

$$\tilde{\varepsilon}_{i,j} = (y_i - \hat{y}_{i,j}) + \hat{a}_j x_{i,j} \quad (6.5)$$

Où \hat{a}_j est le coefficient estimé relatif à la variable X_j dans la régression ci-dessus.

Si la liaison entre X_j et Y est linéaire, les n couples de points représentés dans le repère $(x_j, \tilde{\varepsilon}_j)$ doivent former une droite⁵. Dans le cas contraire, **le nuage nous donne une indication sur la transformation à opérer pour améliorer l'ajustement**. C'est son principal atout.

Un exemple numérique

Sans restreindre la portée de notre propos, nous décrivons les calculs pour une régression simple. Ils sont directement transposables à la régression multiple.

4. http://en.wikipedia.org/wiki/Partial_residual_plot

5. Dans certains logiciels, une régression permet de souligner l'alignement des points. Ex. la fonction `prplot(.)` du package "faraway" du logiciel R.

Étape 1

Nous disposons de $n = 100$ observations d'un couple de variables Y et X . Nous réalisons la régression linéaire $Y = a \times X + b$. Nous obtenons $\hat{a} = 1218.1841$ et $\hat{b} = -1865.1907$, avec un coefficient de détermination $R^2 = 0.9697$. La régression est de très bonne qualité.

Nous en déduisons les prédictions $\hat{y}_i = 1218.1841 \times x_i - 1865.1907$ et les résidus $\hat{\varepsilon}_i = y_i - \hat{y}_i$.

Pour évaluer la forme de la liaison entre X et Y , nous calculons les résidus partiels. Nous projetons alors les observations dans le repère $(x_j, \tilde{\varepsilon}_j)$ (Figure 6.9).

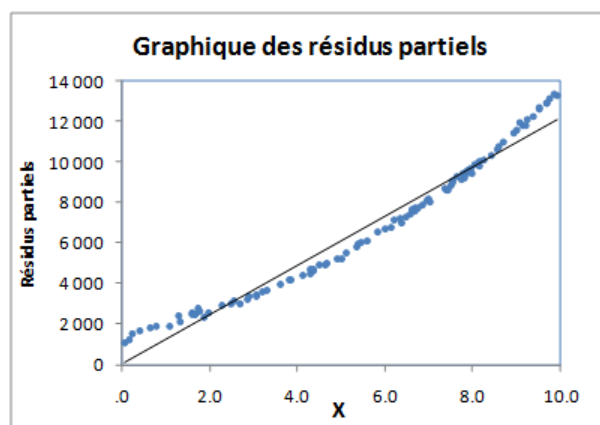


Fig. 6.9. Résidus partiels - Régression $y = ax + b$

Les points sont plus ou moins alignés en formant une courbure assez marquée. Ce constat apparaît clairement lorsque nous ajoutons une courbe de tendance linéaire dans le graphique⁶. Manifestement, il y a une forme de relation entre X et Y que l'on ne prend pas en compte dans la régression linéaire simple.

Étape 2

Sur la base du graphique, nous souhaitons améliorer l'ajustement en ajoutant la variable $Z = X^2$ dans la base. Nous réalisons maintenant la régression $Y = a_0 + a_1X + a_2Z$. Dans notre exemple, il ne s'agit pas de substituer Z à X car, malgré tout, nous avons constaté que la régression linéaire est de très bonne qualité déjà ($R^2 = 0.9697$). Nous voulons vérifier que $Z = X^2$ amène de l'information supplémentaire pertinente dans l'explication de Y .

Nous obtenons le modèle

$$Y = -324.9884 + 350.5567 \times X + 84.1142 \times Z$$

Avec un $R^2 = 0.99861$. X et Z sont tous deux largement significatifs ($t_{\hat{a}_1} = 17.63$ et $t_{\hat{a}_2} = 44.89$). Le modèle s'est bonifié, l'introduction de $Z = X^2$ dans la régression est totalement justifié.

6. Nous avons utilisé l'outil "courbe de tendance" d'Excel.

Voyons ce que nous en annonce les résidus partiels. De nouveau, nous réalisons la prédiction $\hat{y}_i = -324.9884 + 350.5567 \times x_i + 84.1142 \times x_i^2$. Avec la même démarche, nous calculons les résidus partiels $\tilde{\varepsilon}_{i,x} = (y_i - \hat{y}_i) + 350.5567 \times x_i$ et nous construisons le graphique (Figure 6.10).

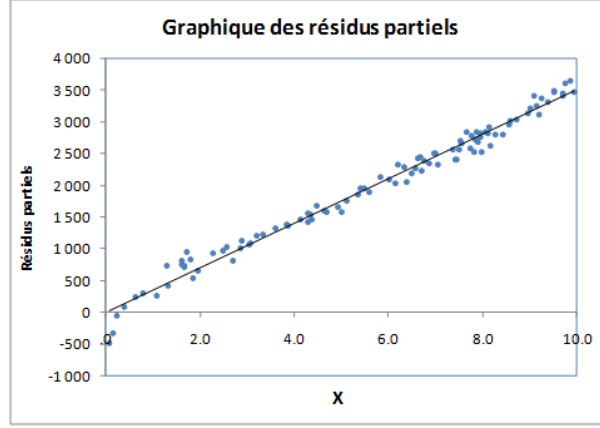


Fig. 6.10. Résidus partiels - Régression $y = a_0 + a_1x + a_2x^2$

La situation est incontestablement meilleure, sauf pour les petites valeurs de X proches de 0 où il reste une distorsion par rapport à la droite. Il faut la prendre en compte. Pour ce faire, nous introduisons la variable supplémentaire $W = \ln(X)$.

Remarque 29 (Résidus partiels par rapport à la variable modifiée). Nous aurions pu également calculer le résidu partiel

$$\tilde{\varepsilon}_i = (y_i - \hat{y}_i) + 84.112 \times x_i^2$$

La conclusion aurait été du même ordre c.-à-d. la nécessité d'introduire une variable de type $W = \ln(X)$.

Étape 3

Nous introduisons la variable $W = \ln(X)$ dans la régression $Y = b_0 + b_1X + b_2Z + b_3W$, nous obtenons :

$$Y = -18.4623 + 109.0188 \times X + 98.8254 \times Z + 317.4565 \times W$$

Tous les coefficients sont significatifs à 5%, avec respectivement $t_{\hat{b}_1} = 3.30$, $t_{\hat{b}_2} = 43.06$, $t_{\hat{b}_3} = 8.24$. La régression est d'excellente qualité avec un coefficient de détermination égal à $R^2 = 0.9919$.

A partir de cette équation, nous calculons les nouvelles prédictions \hat{y}_i et les résidus partiels

$$\tilde{\varepsilon}_{i,x} = (y_i - \hat{y}_i) + 109.0188 \times x_i$$

Il nous reste à construire le graphique $(x_i, \tilde{\varepsilon}_{i,x})$ (Figure 6.11). Les points forment une droite. Nous avons épuisé les différentes formes de X qui permettent d'expliquer linéairement les valeurs de Y .

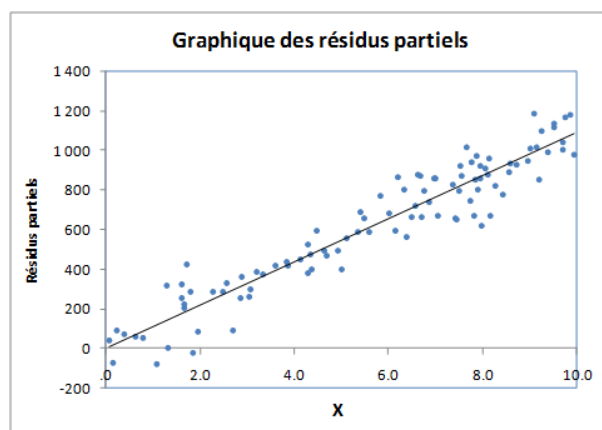


Fig. 6.11. Résidus partiels - Régression $y = b_0 + b_1x + b_2x^2 + b_3 \ln(x)$

Pour être tout à fait honnête, les situations sur données réelles ne sont pas toujours aussi idylliques. Nous avons travaillé sur des variables générées artificiellement dans cette section et, effectivement, nous avons utilisé X , X^2 et $\ln(X)$ pour obtenir Y . Il est heureux que l'on retrouve le bon résultat en nous appuyant sur les résidus partiels.

Résidus partiels augmentés

Dans certains ouvrages, on conseille de passer par les résidus partiels augmentés. Il s'agit simplement d'introduire toutes les expressions de la variable dans le calcul du résidu.

Ainsi, à l'étape 2 de notre exemple précédent, nous avons introduit $Z = X^2$ dans la régression. Les résidus partiels augmentés correspondraient alors à

$$\tilde{\varepsilon}_i = \hat{\varepsilon}_i + \hat{a}_1 \times X + \hat{a}_2 \times X^2 \quad (6.6)$$

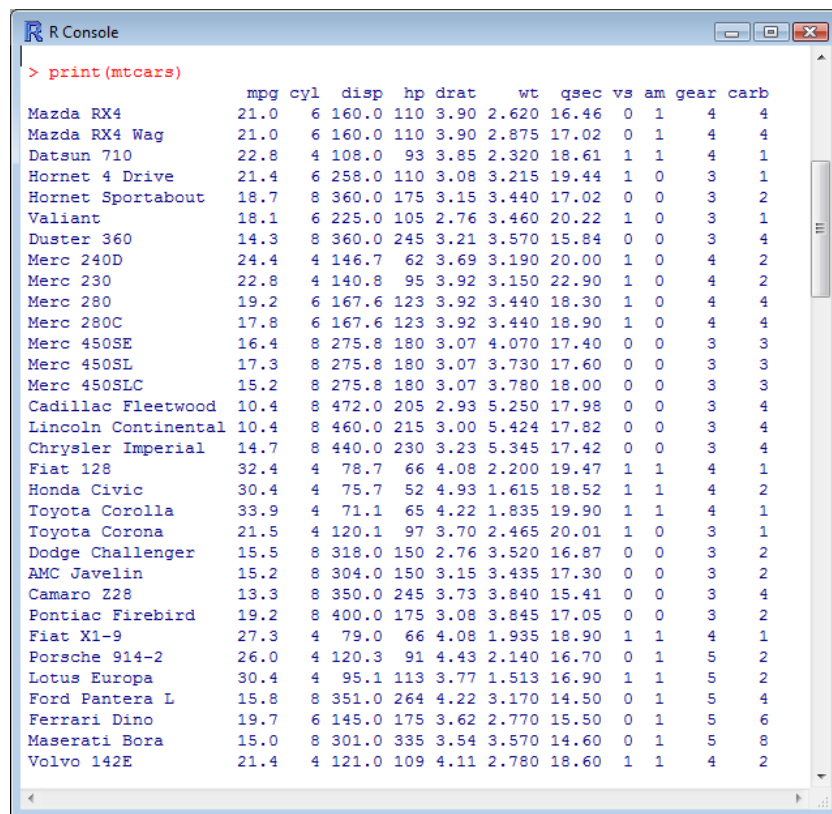
L'intérêt de ce nouvel indicateur n'est pas déterminant dans notre contexte de recherche des différentes transformées possibles des variables exogènes à introduire dans la régression. L'usage des résidus partiels est amplement suffisant.

6.2.3 Un exemple "réaliste" : traitement des données "mtcars" sous R

Tournons-nous maintenant vers des données "réalistes" pour montrer l'intérêt des résidus partiels. Nous utilisons la base **mtcars** livrée en standard avec le logiciel R.

Nous la chargeons à l'aide des commandes suivantes :

```
> data(mtcars)
> print(mtcars)
```



```

> print(mtcars)
      mpg  cyl  disp  hp drat   wt  qsec vs  am  gear  carb
Mazda RX4         21.0    6  160.0  110  3.90  2.620  16.46  0   1    4    4
Mazda RX4 Wag     21.0    6  160.0  110  3.90  2.875  17.02  0   1    4    4
Datsun 710        22.8    4  108.0   93  3.85  2.320  18.61  1   1    4    1
Hornet 4 Drive    21.4    6  258.0  110  3.08  3.215  19.44  1   0    3    1
Hornet Sportabout 18.7    8  360.0  175  3.15  3.440  17.02  0   0    3    2
Valiant           18.1    6  225.0  105  2.76  3.460  20.22  1   0    3    1
Duster 360        14.3    8  360.0  245  3.21  3.570  15.84  0   0    3    4
Merc 240D          24.4    4  146.7   62  3.69  3.190  20.00  1   0    4    2
Merc 230           22.8    4  140.8   95  3.92  3.150  22.90  1   0    4    2
Merc 280           19.2    6  167.6  123  3.92  3.440  18.30  1   0    4    4
Merc 280C          17.8    6  167.6  123  3.92  3.440  18.90  1   0    4    4
Merc 450SE         16.4    8  275.8  180  3.07  4.070  17.40  0   0    3    3
Merc 450SL         17.3    8  275.8  180  3.07  3.730  17.60  0   0    3    3
Merc 450SLC        15.2    8  275.8  180  3.07  3.780  18.00  0   0    3    3
Cadillac Fleetwood 10.4    8  472.0  205  2.93  5.250  17.98  0   0    3    4
Lincoln Continental 10.4    8  460.0  215  3.00  5.424  17.82  0   0    3    4
Chrysler Imperial 14.7    8  440.0  230  3.23  5.345  17.42  0   0    3    4
Fiat 128           32.4    4   78.7   66  4.08  2.200  19.47  1   1    4    1
Honda Civic        30.4    4   75.7   52  4.93  1.615  18.52  1   1    4    2
Toyota Corolla     33.9    4   71.1   65  4.22  1.835  19.90  1   1    4    1
Toyota Corona      21.5    4  120.1   97  3.70  2.465  20.01  1   0    3    1
Dodge Challenger   15.5    8  318.0  150  2.76  3.520  16.87  0   0    3    2
AMC Javelin        15.2    8  304.0  150  3.15  3.435  17.30  0   0    3    2
Camaro Z28         13.3    8  350.0  245  3.73  3.840  15.41  0   0    3    4
Pontiac Firebird   19.2    8  400.0  175  3.08  3.845  17.05  0   0    3    2
Fiat X1-9          27.3    4   79.0   66  4.08  1.935  18.90  1   1    4    1
Porsche 914-2      26.0    4  120.3   91  4.43  2.140  16.70  0   1    5    2
Lotus Europa       30.4    4   95.1  113  3.77  1.513  16.90  1   1    5    2
Ford Pantera L     15.8    8  351.0  264  4.22  3.170  14.50  0   1    5    4
Ferrari Dino       19.7    6  145.0  175  3.62  2.770  15.50  0   1    5    6
Maserati Bora      15.0    8  301.0  335  3.54  3.570  14.60  0   1    5    8
Volvo 142E         21.4    4  121.0  109  4.11  2.780  18.60  1   1    4    2

```

Fig. 6.12. Données `mtcars` - Logiciel R

Elle comporte $n = 32$ observations et 11 variables (Figure 6.12). Dans notre contexte, nous n'utiliserons que 3 variables. Nous essayons d'expliquer la consommation (*mpg*), le nombre de miles que l'on peut parcourir à l'aide d'un gallon de carburant, à l'aide de la puissance (*hp*) et le poids (*wt*). Nous obtenons les résultats la régression.

```
> modele <- lm(mpg ~ hp + wt, data = mtcars)
```

Coefficients:

```

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.22727    1.59879   23.285  < 2e-16 ***
hp           -0.03177    0.00903   -3.519  0.00145 **
wt           -3.87783    0.63273   -6.129  1.12e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 2.593 on 29 degrees of freedom

Multiple R-squared: 0.8268, Adjusted R-squared: 0.8148

F-statistic: 69.21 on 2 and 29 DF, p-value: 9.109e-12

Le modèle est plutôt de bonne qualité avec un coefficient de détermination corrigé⁷ $\bar{R}^2 = 0.8148$. La régression est globalement significative à 5%, les deux variables le sont également, largement même.

Voyons maintenant ce qu'il en est des résidus partiels. Nous utilisons le package **faraway**. Les commandes adéquates sont

```
#librairie pour les résidus partiels
> library(faraway)
#résidus partiels
> par(mfrow=c(1,2))
> prplot(modele,1)
> prplot(modele,2)
```

Les deux graphiques des résidus partiels sont affichés dans la même fenêtre (Figure 6.13). Ici commence les choses délicates. En effet, il faut choisir la transformation appropriée à partir d'informations purement visuelles. Il y a quand même une certaine part de subjectivité là-dedans.

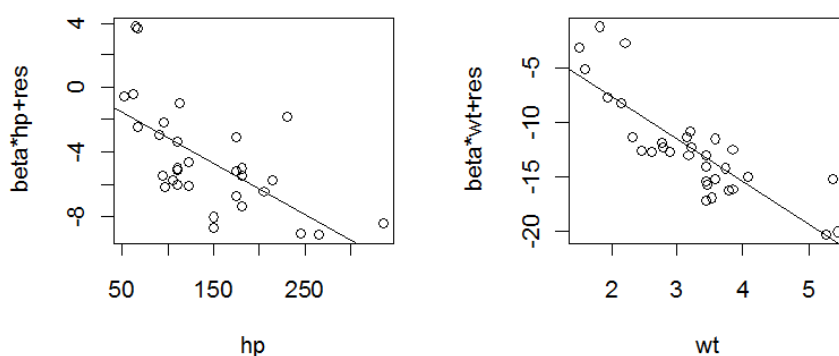


Fig. 6.13. Données **mtcars** - Résidus partiels $mpg = f(hp, wt)$

Compte tenu des formes plus ou moins curvilinéaires des nuages de points, nous tentons les deux transformations suivantes : $zhp = hp^2$ et $zwt = \ln(wt)$. Nous les ajoutons parmi les explicatives. Nous lançons de nouveau la régression.

```
> zhp <- mtcars$hp^2
> zwt <- log(mtcars$wt)
> modele.bis <- lm(mpg ~ hp + wt + zhp + zwt, data = mtcars)
```

7. Nous privilégions cet indicateur car il tient compte des degrés de liberté. Et nous aurons à comparer des modèles avec un nombre d'explicatives différent par la suite.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.217e+01	1.911e+00	22.072	< 2e-16 ***
hp	-9.777e-02	3.110e-02	-3.143	0.00403 **
wt	2.384e+00	1.887e+00	1.263	0.21741
zhp	1.806e-04	7.893e-05	2.287	0.03023 *
zwt	-1.793e+01	5.935e+00	-3.022	0.00545 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.101 on 27 degrees of freedom

Multiple R-squared: 0.8941, Adjusted R-squared: 0.8785

F-statistic: 57.01 on 4 and 27 DF, p-value: 8.922e-13

Le modèle est meilleur que le précédent avec un $\bar{R}^2 = 0.8785$. Les deux variables additionnelles *zhp* et *zwt* sont significatives, *wt* ne l'est plus en revanche. En passant aux résidus partiels,

#résidus partiels

```
> par(mfrow=c(2,2))
```

```
> for (i in 1:4)prplot(modele.bis,i)
```

Nous constatons que les modifications introduites ont permis de réduire les problèmes de non-linéarité. Les formes plus ou moins curvilinéaires constatées précédemment ont été résorbées (Figure 6.14)⁸.

Reste à retirer la variable *wt* rendue inutile.

```
> modele.ter <- lm(mpg ~ hp + zhp + zwt, data = mtcars)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.179e+01	1.906e+00	21.920	< 2e-16 ***
hp	-9.736e-02	3.143e-02	-3.098	0.00441 **
zhp	1.809e-04	7.977e-05	2.268	0.03122 *
zwt	-1.082e+01	1.886e+00	-5.737	3.73e-06 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

8. Ca apparaît plus clairement maintenant, il y a aussi un méchant point atypique sur *hp* (et *zhp*). On passera outre. Mais il est clair que dans une étude réelle, il faudra se pencher attentivement sur ce quidam avant de poursuivre l'analyse.

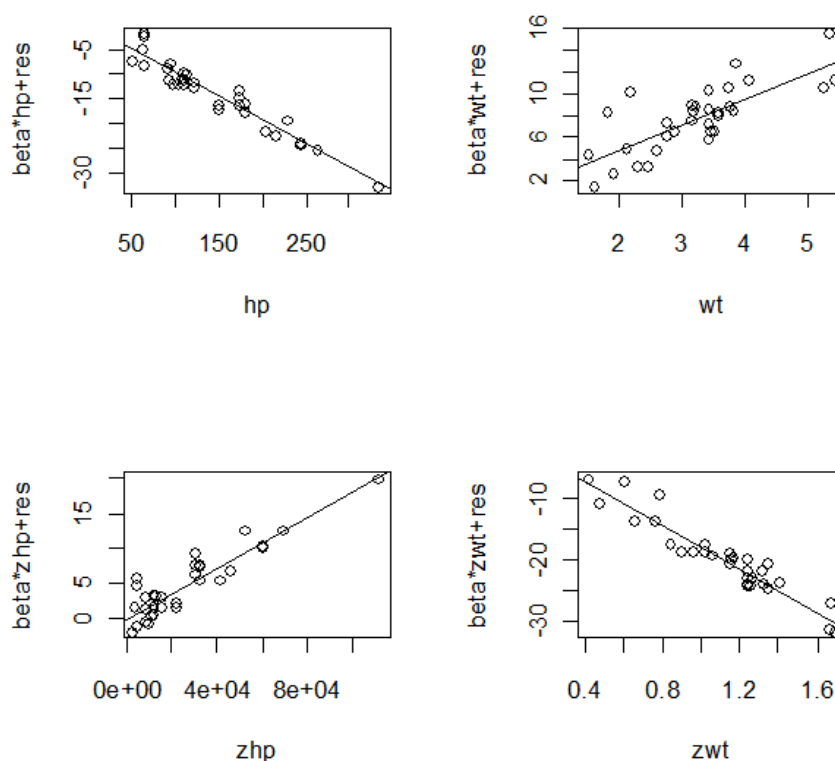


Fig. 6.14. Données `mtcars` - Résidus partiels $mpg = f(hp, wt, hp^2, \ln(wt))$

Residual standard error: 2.123 on 28 degrees of freedom

Multiple R-squared: 0.8879, Adjusted R-squared: 0.8759

F-statistic: 73.91 on 3 and 28 DF, p-value: 2.034e-13

La régression à 3 explicatives donne amplement satisfaction, $\bar{R}^2 = 0.8759$ est équivalente à la régression précédente (on ne va pas commencer à gloser pour une différence à la 3^{ème} décimale). Toutes les variables sont significatives maintenant.

Concernant les résidus partiels,

```
> #résidus partiels
> par(mfrow=c(1,3))
> for (i in 1:3)prplot(modele.ter,i)
```

Nous constatons que l'adjonction d'une transformation supplémentaire ne paraît pas nécessaire. Dans aucun des nuages nous observons une distorsion évidente par rapport à une droite (Figure 6.15). Le modèle à 3 variables explicatives est celui que nous utiliserons pour les interprétations et les prédictions.

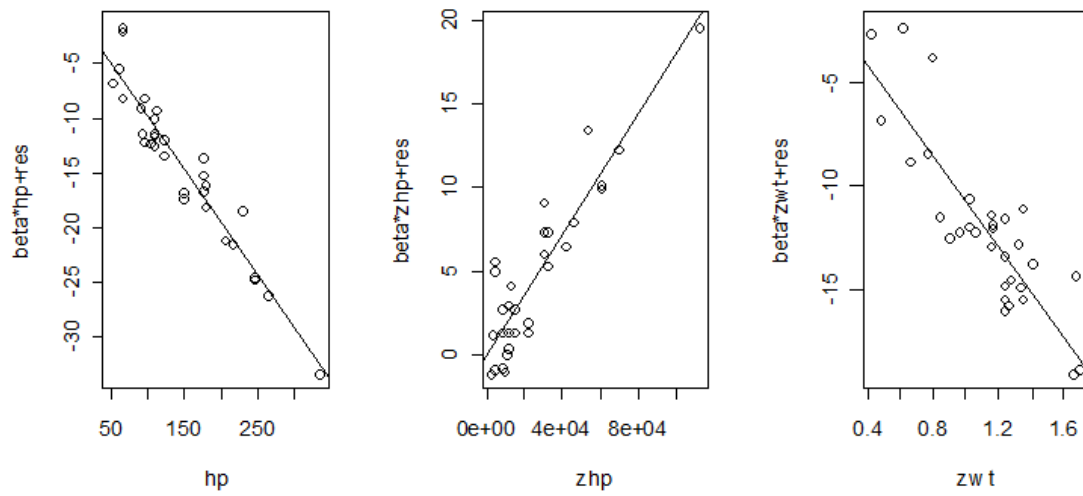


Fig. 6.15. Données *mtcars* - Résidus partiels $\text{mpg} = f(\text{hp}, \text{hp}^2, \ln(\text{wt}))$

Table de Durbin Watson

<http://www.jourdan.ens.fr/~bozio/stats/dw.pdf>

TABLE de DURBIN-WATSON : Test unilatéral de $\rho = 0$ contre $\rho > 0$, au seuil de 5% (test bilatéral : seuil $\alpha = 10\%$)

	k' = 1		k' = 2		k' = 3		k' = 4		k' = 5		k' = 6		k' = 7		k' = 8		k' = 9		k' = 10	
n	d _L	d _u	d _L	d _u	d _L	d _u	d _L	d _u	d _L	d _u	d _L	d _u	d _L	d _u	d _L	d _u	d _L	d _u	d _L	d _u
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21	0,45	2,47	0,34	2,73	0,25	2,98	0,17	3,22	0,11	3,44
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15	0,50	2,40	0,40	2,62	0,30	2,86	0,22	3,09	0,15	3,30
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10	0,55	2,32	0,45	2,54	0,36	2,76	0,27	2,97	0,20	3,20
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06	0,60	2,26	0,50	2,46	0,41	2,67	0,32	2,87	0,24	3,07
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02	0,65	2,21	0,46	2,40	0,46	2,59	0,37	2,78	0,29	2,97
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99	0,69	2,16	0,60	2,34	0,50	2,52	0,42	2,70	0,34	2,88
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96	0,73	2,12	0,64	2,29	0,55	2,46	0,46	2,63	0,38	2,81
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94	0,77	2,09	0,68	2,25	0,59	2,41	0,50	2,57	0,42	2,73
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92	0,80	2,06	0,71	2,21	0,63	2,36	0,54	2,51	0,46	2,67
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90	0,84	2,03	0,75	2,17	0,67	2,32	0,58	2,46	0,51	2,61
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89	0,87	2,01	0,78	2,14	0,70	2,28	0,62	2,42	0,54	2,56
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88	0,90	1,99	0,82	2,12	0,73	2,25	0,66	2,38	0,58	2,51
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86	0,92	1,97	0,84	2,09	0,77	2,22	0,69	2,34	0,62	2,47
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85	0,95	1,96	0,87	2,07	0,80	2,19	0,72	2,31	0,65	2,43
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84	0,97	1,94	0,90	2,05	0,83	2,16	0,75	2,28	0,68	2,40
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83	1,00	1,93	0,93	2,03	0,85	2,14	0,78	2,25	0,71	2,36
31	1,36	1,50	1,30	1,57	1,23	1,65	1,16	1,74	1,09	1,83	1,02	1,92	0,95	2,02	0,88	2,12	0,81	2,23	0,74	2,33
32	1,37	1,50	1,31	1,57	1,24	1,65	1,18	1,73	1,11	1,82	1,04	1,91	0,97	2,00	0,90	2,10	0,84	2,20	0,77	2,31
33	1,38	1,51	1,32	1,58	1,26	1,65	1,19	1,73	1,13	1,81	1,06	1,90	0,99	1,99	0,93	2,08	0,86	2,18	0,79	2,28
34	1,39	1,51	1,33	1,58	1,27	1,65	1,21	1,73	1,15	1,81	1,08	1,89	1,01	1,98	0,95	2,07	0,88	2,16	0,82	2,26
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80	1,10	1,88	1,03	1,97	0,97	2,05	0,91	2,14	0,84	2,24
36	1,41	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,80	1,11	1,88	1,05	1,96	0,99	2,04	0,93	2,13	0,87	2,22
37	1,42	1,53	1,36	1,59	1,31	1,66	1,25	1,72	1,19	1,80	1,13	1,87	1,07	1,95	1,01	2,03	0,95	2,11	0,89	2,20
38	1,43	1,54	1,37	1,59	1,32	1,66	1,26	1,72	1,21	1,79	1,15	1,86	1,09	1,94	1,03	2,02	0,97	2,10	0,91	2,18
39	1,43	1,54	1,38	1,60	1,33	1,66	1,27	1,72	1,22	1,79	1,16	1,86	1,10	1,93	1,05	2,01	0,99	2,08	0,93	2,16
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79	1,17	1,85	1,12	1,92	1,06	2,00	1,01	2,07	0,95	2,14
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78	1,24	1,84	1,19	1,90	1,14	1,96	1,09	2,00	1,04	2,09
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77	1,29	1,82	1,25	1,87	1,20	1,93	1,16	1,99	1,11	2,04
55	1,53	1,60	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77	1,33	1,81	1,29	1,86	1,25	1,91	1,21	1,96	1,17	2,01
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77	1,37	1,81	1,33	1,85	1,30	1,89	1,26	1,94	1,22	1,98
65	1,57	1,63	1,54	1,66	1,50	1,70	1,47	1,73	1,44	1,77	1,40	1,80	1,37	1,84	1,34	1,88	1,30	1,92	1,27	1,96
70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77	1,43	1,80	1,40	1,84	1,37	1,87	1,34	1,91	1,30	1,95
75	1,60	1,65	1,57	1,68	1,54	1,71	1,51	1,74	1,49	1,77	1,46	1,80	1,43	1,83	1,40	1,87	1,37	1,90	1,34	1,94
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77	1,48	1,80	1,45	1,83	1,42	1,86	1,40	1,89	1,37	1,92
85	1,62	1,67	1,60	1,70	1,57	1,72	1,55	1,75	1,52	1,77	1,50	1,80	1,47	1,83	1,45	1,86	1,42	1,89	1,40	1,92
90	1,63	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78	1,52	1,80	1,49	1,83	1,47	1,85	1,44	1,88	1,42	1,91
95	1,64	1,69	1,62	1,71	1,60	1,73	1,58	1,75	1,56	1,78	1,54	1,80	1,51	1,83	1,49	1,85	1,46	1,88	1,44	1,90
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78	1,55	1,80	1,53	1,83	1,51	1,85	1,48	1,87	1,46	1,90
150	1,72	1,75	1,71	1,76	1,69	1,77	1,68	1,79	1,66	1,80	1,65	1,82	1,64	1,83	1,62	1,85	1,60	1,86	1,59	1,88
200	1,73	1,78	1,75	1,79	1,73	1,80	1,73	1,81	1,72	1,82	1,71	1,83	1,70	1,84	1,69	1,85	1,68	1,86	1,66	1,87

Fig. A.1. Table de Durbin-Watson

Gestion des versions

Première version

La première version de ce document a été mis en ligne en septembre 2007. Elle n'a pas été numérotée.

Version 2.0

La seconde version, 2.0, a été mise en ligne fin juin 2011. Elle se distingue (et celles qui suivront) par les graphiques en couleur (hé oui, on apprend à tout âge...).

Plus sérieusement, un chapitre a été ajouté, d'autres ont été complétés. Je distinguerais volontiers :

- **Chapitre 3 - Colinéarité et sélection de variables.** Deux sections ont été ajoutées : régressions partielles (section 3.5), régressions croisées (section 3.6).
- **Chapitre 4 - Régression sur des exogènes qualitatives.** Ce chapitre a été profondément remanié. Je confesse avoir été énormément influencé¹ par la lecture de l'extraordinaire ouvrage de M.A. Hardy, *Regression with dummy variables* [5]. Mon travail a surtout consisté à reprendre les parties qui me paraissaient les plus intéressantes, en l'inscrivant dans mon propre canevas de présentation et en utilisant mes propres exemples. Le fichier LOYER, entres autres, est mis à toutes les sauces dans ce chapitre.
- **Chapitre 6 - Détection et traitement de la non linéarité.** Ce chapitre fait écho à une première approche de la non-linéarité concernant la régression simple développé dans mon premier volume sur la régression ([18], chapitre 6). Des approches plus génériques sont mises en avant dans ce document, dans un premier temps pour la régression simple, dans un second temps pour la régression multiple. Ce chapitre doit beaucoup à l'extraordinaire ouvrage de Aïvazian [1]. Je l'ai depuis plus de 20 ans. A chaque fois que je l'ouvre, je (re)découvre des choses intéressantes. Je l'ai également beaucoup mis à contribution dans mon fascicule consacré à la corrélation [17].

1. Comme je le dis toujours, reprendre en citant, c'est faire honneur au travail des collègues, reprendre sans citer, c'est du plagiat.

Version 2.1

Le chapitre consacré à la régression sur les exogènes qualitatives (**Chapitre 4**) a été remanié suite à la lecture de la page web "R Library : Contrast Coding Systems for categorical variables" - UCLA Statistical Consulting Group (http://www.ats.ucla.edu/stat/r/library/contrast_coding.htm).

Cette référence est particulièrement intéressante. Outre la description relativement complète des différentes stratégies de codage, il y en avait plus que ceux que j'avais moi-même recensé dans la version 2.0 de ce document en tous les cas, l'auteur montre comment les implémenter sous le logiciel R à l'aide des commandes dédiées (ex. *contr.treatment()*, *contr.poly()*, *contrasts()*, etc.). Nous sommes directement opérationnels sous R pour lancer ses régressions avec les différentes stratégies de codage.

Fichiers associés à ce support

Un certain nombre de jeux de données ont servi à illustrer ce support. Ils ont été traités. De nombreuses copies d'écran sont présentées tout le long du texte. Pour que le lecteur puisse accéder aux détails des calculs et, s'il le désire, les reproduire, ces fichiers sont accessibles en ligne.

Les fichiers peuvent être classés en 3 principales catégories :

1. Les classeurs EXCEL contiennent, dans la première feuille, les données ; dans les feuilles suivantes, les traitements associés aux problèmes statistiques. Ils ont contribué à l'élaboration des copies d'écran de ce support de cours.
2. Les fichiers au format CSV contiennent les données destinées à être traités avec le logiciel R.
3. Les scripts R décrivent les traitements relatifs à chaque chapitre du support. *Concernant l'utilisation du logiciel R pour la régression, nous conseillons vivement la lecture du didacticiel de J. Faraway qui est réellement d'une qualité exceptionnelle : il est aussi intéressant pour l'apprentissage de la régression que pour l'apprentissage du logiciel R (Voir la référence en bibliographie).*

Les fichiers et les thèmes rattachés sont décrits dans "_description_des_fichiers.txt", intégré dans l'archive "fichiers_pratique_regression.zip", accessible sur le net - http://eric.univ-lyon2.fr/~ricco/cours/exercices/fichiers_pratique_regression.zip.

Tutoriels

Mes tutoriels relatifs à la pratique de la régression sont sur le site <http://tutoriels-data-mining.blogspot.com/>. Sauf mention contraire, j'utilise principalement les logiciels TANAGRA (<http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html>) et R (<http://www.r-project.org/>).

Plutôt que d'intégrer dans ce fascicule la description de la mise en oeuvre des techniques à l'aide des logiciels, j'ai préféré écrire des tutoriels dans des documents à part. L'idée est de pouvoir construire une histoire complète autour d'une base de données à analyser, en partant de l'importation des données jusqu'au déploiement et l'interprétation des résultats. Chaque tutoriel peut ainsi couvrir plusieurs thèmes de la pratique de la régression. Dans ce qui suit, je recense les sujets abordés dans chacun d'entre eux.

1. **Régression linéaire - Lecture des résultats**, <http://tutoriels-data-mining.blogspot.com/2011/02/regression-lineaire-lecture-des.html>. *Logiciels* : Tanagra, R. *Thèmes* : tests généralisés, tests de conformité, tests de comparaison.
2. **Diagnostic de la régression avec R**, <http://tutoriels-data-mining.blogspot.com/2009/05/diagnostic-de-la-regression-avec-r.html>. *Logiciel* : R. *Thèmes* : analyse des résidus, points atypiques, points aberrants, points influents, colinéarité, critère VIF, sélection de variables.
3. **Points aberrants et influents dans la régression**, <http://tutoriels-data-mining.blogspot.com/2008/04/points-aberrants-et-influents-dans-la.html>. *Logiciels* : Tanagra, R, SAS. *Thèmes* : points influents, points aberrants, points atypiques, résidus standardisés, résidus studentisés, levier (leverage), dffits, distance de cook, covratio, dfbetas.
4. **Colinéarité et régression**, <http://tutoriels-data-mining.blogspot.com/2008/04/colinarit-et-rgression.html>. *Logiciels* : Tanagra, R. *Thèmes* : colinéarité, sélection de variables, analyse en composantes principales, régression pls1.
5. **Sélection forward - Crime dataset**, <http://tutoriels-data-mining.blogspot.com/2008/03/slection-forward-crime-dataset.html>. *Logiciel* : Tanagra. *Thèmes* : sélection de variables, sélection forward, stepwise, colinéarité, corrélation partielle.
6. **REGRESS dans la distribution SIPINA**, <http://tutoriels-data-mining.blogspot.com/2011/05/regress-dans-la-distribution-sipina.html>. *Logiciel* : REGRESS (via la distribution

SIPINA - <http://sipina.over-blog.fr/>). *Thèmes* : points aberrants, points atypiques, points influents, normalité des résidus, test de Jarque-Bera, droite de Henry, q-q plot.

7. **Régression avec le logiciel LazStats (OpenStat)**, <http://tutoriels-data-mining.blogspot.com/2011/05/regression-avec-le-logiciel-lazstats.html>. *Logiciel* : LazStats (<http://www.statprograms4u.com/>) - Malheureusement, je viens de me rendre compte - aujourd'hui 20 juin 2011, que le logiciel est devenu du jour au lendemain commercial. *Thèmes* : sélection de variables, forward, backward, stepwise, régressions croisées.
8. **Régression - Déploiement de modèles**, <http://tutoriels-data-mining.blogspot.com/2011/03/regression-deploiement-de-modeles.html>. *Logiciel* : Tanagra. *Thèmes* : déploiement, régression pls, support vector regression, SVR, arbres de régression, cart, analyse en composantes principales, régression sur axes factoriels.

Littérature

Ouvrages

1. Aïvazian, S., *Étude statistique des dépendances*, Édition de Moscou, 1978.
2. Bourbonnais, R., *Econométrie. Manuel et exercices corrigés*, Dunod, 2^e édition, 1998.
3. Dodge, Y, Rousson, V., *Analyse de régression appliquée*, Dunod, 2^e édition, 2004.
4. Giraud, R., Chaix, N., *Econométrie*, Presses Universitaires de France (PUF), 1989.
5. Hardy, M.A., *Regression with dummy variables*, Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-093, Newbury Park, CA : Sage, 1993.
6. Jacquard, J., Turrisi, R., *Interaction Effects in Multiple Regression (2nd ed)*, Sage University Paper series on Quantitative Applications in the Social Sciences, series no. 07-072, Newbury Park, CA : Sage, 2003.
7. Johnston, J., DiNardo, J., *Méthodes Économétriques*, Economica, 4^e édition, 1999.
8. Labrousse, C., *Introduction à l'économétrie. Maîtrise d'économétrie*, Dunod, 1983.
9. Saporta, G., *Probabilités, Analyse des données et Statistique*, Technip, 2^eème édition, 2006.
10. Tenenhaus, M., *Méthodes Statistiques en Gestion*, Dunod, 1996.
11. Veyseyre R., *Aide mémoire - Statistique et probabilités pour l'ingénieur*, Dunod, 2006.

Supports en ligne

12. Confais, J., Le Guen, M., *Premier pas en régression linéaire avec SAS*, Revue Modulad, numéro 35, 2006 ; <http://www-rocq.inria.fr/axis/modulad/numero-35/Tutoriel-confais-35/confais-35.pdf>
13. , Davidson, R., MacKinnon, J.G., *Estimation et inférence en économétrie*, traduction française de *Estimation and inference in econometrics*, <http://russell.vcharite.univ-mrs.fr/EIE/>
14. Faraway, J., *Practical Regression and ANOVA using R*, July 2002, <http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf>
15. Genest, C., *Modèle de régression linéaire multiple*, sur <http://archimede.mat.ulaval.ca/pages/genest/regression/chap3.pdf>. Voir aussi le chapitre 2 ([chap2.pdf](#)), *Régression linéaire simple*, et le chapitre 4 ([chap4.pdf](#)), *Critères de sélection de modèle*.
16. Haurie, A., *Modèle de régression linéaire*, sur <http://ecolu-info.unige.ch/~haurie/mba05/>
17. Rakotomalala, R., *Analyse de corrélation. Étude des dépendances - Variables quantitatives*, <http://eric.univ-lyon2.fr/~ricco/publications.html>

18. Rakotomalala, R., *Econométrie - La régression linéaire simple et multiple*, <http://eric.univ-lyon2.fr/~ricco/publications.html>
19. *Régression Linéaire Multiple*, sur http://fr.wikipedia.org/wiki/Régression_linéaire_multiple
20. *Xycoon Online Econometrics Textbook*, sur <http://www.xycoon.com/index.htm#econ>