Etude des résidus

Régression Linéaire Multiple

Ricco RAKOTOMALALA

PLAN

1. Diagnostic graphique
2. Caractère aléatoire des erreurs (données ordonnées)
3. Test de normalité

Pourquoi étudier les résidus ?

Importance des résidus pour l'inférence statistique

Problème : Les propriétés (biais, convergence) et l'inférence statistique (test de significativité, intervalle de confiance) reposent en grande partie sur des hypothèses sur les erreurs. Il faut s'assurer de la conformité aux hypothèses.

Quelles hypothèses ?

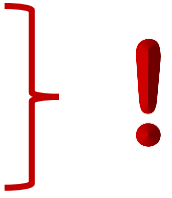
- *$E(\varepsilon) = 0$, en moyenne le modèle est bien spécifié*
- *$E(\varepsilon^2) = \sigma^2_\varepsilon$ la variance de l 'erreur est constante (homoscédasticité)*
- *$E(\varepsilon_i, \varepsilon_j)=0$, les erreurs sont non-corrélés*
- *$Cov(\varepsilon,x)=0$, l 'erreur est indépendante de la variable explicative*
- *$\varepsilon \equiv Normale(0, \sigma^2_\varepsilon)$*

Quelques principes

- On ne dispose pas des erreurs mais des résidus (erreurs observées) → déjà une inférence ici
- Résidus portés en ordonnée, les graphiques diffèrent de ce qu'on met en abscisse
- Traquer toute forme de « régularité » dans les résidus et/ou de dépendance entre les résidus et les variables

→ *Les résidus doivent donc être disséminés « au hasard » dans un certain intervalle*

- Un point s'écartant ostensiblement est la marque d'une observation atypique et/ou mal modélisée



Graphique des résidus

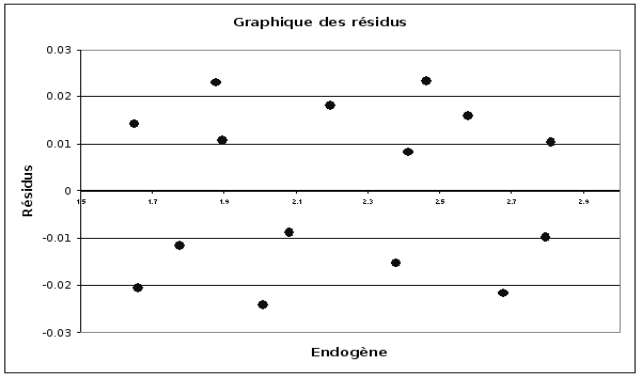
Un petit graphique vaut mieux (souvent) que de longs calculs

Graphiques de base

Résidus vs. Endogène, vs. Exogènes, vs. Temps

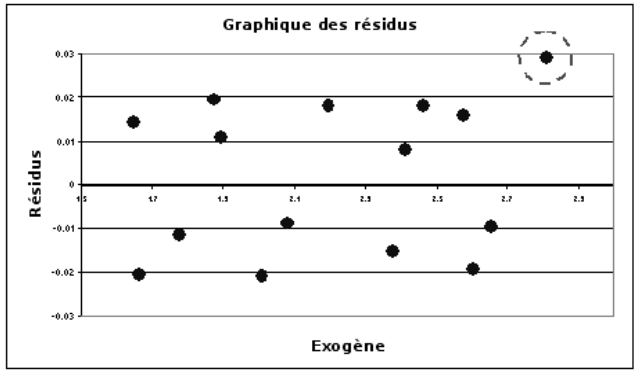
Résidus vs. Endogène

- Vérifier les points atypiques et/ou mal modélisés
- Vérifier si certaines plages de valeurs sont sous ou sur-estimées
- Vérifier la dispersion selon les valeurs de Y



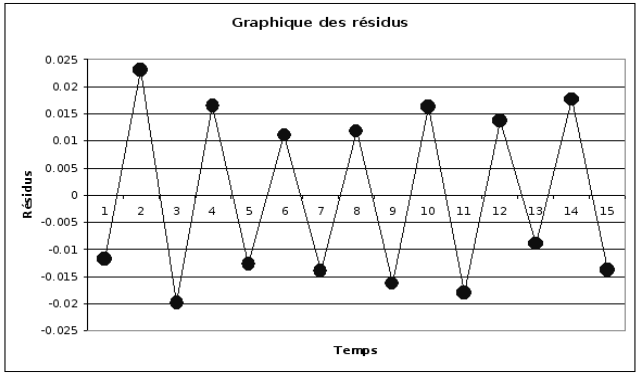
Résidus vs. Exogènes

- Vérifier les points atypiques
- Vérifier les dépendances
- Vérifier la dispersion selon les plages de valeurs de X



Résidus vs. Temps

- Données temporelles
- Tableau ordonné selon le temps
- Vérifier l'existence de « régularités »



Cas pathologiques

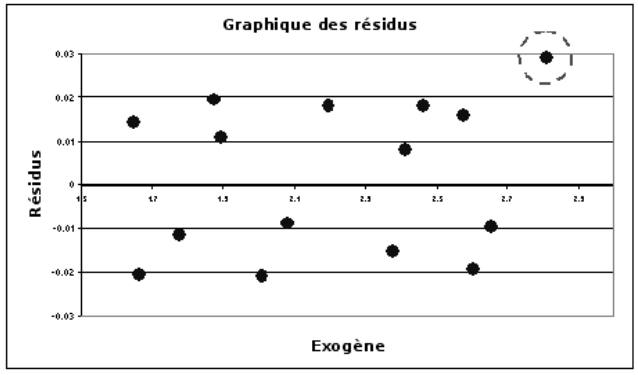
Points atypiques et points influents

Points atypiques : Points qui s'écartent délibérément des autres
Points influents : Points qui pèsent (exagérément) sur les estimations : si on les enlevait, on obtiendrait des résultats (significativement) différents

Point atypique

Une valeur très différente sur l'endogène et/ou sur une ou combinaison d'exogènes. Elle n'est pas forcément mal modélisée (résidu élevé).

Cf. Endogène atypique O/N x Mal/Bien modélisé

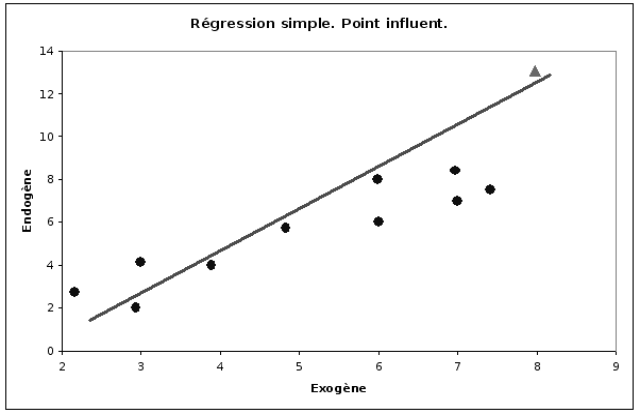


Atypique exogène + Mal modélisé

Point influent

Très difficile à détecter visuellement
→ Peut être atypique ou non
→ Peut être bien modélisé ou non

Cf. Atypique non influent, Non atypique mais influent



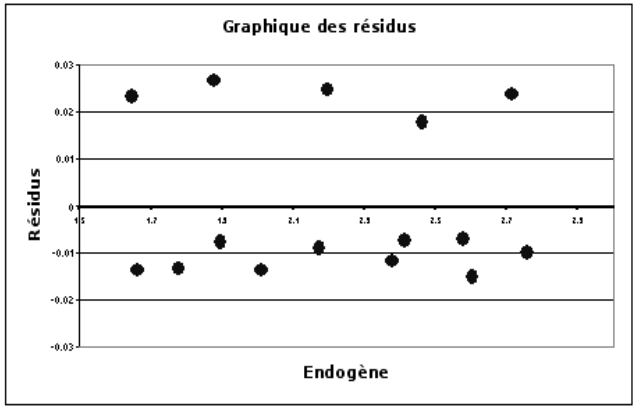
Régression simple : Point manifestement influent
Serait-ce aussi évident dans un graphique des résidus ?

Cas pathologiques

Asymétrie, non linéarité et rupture de structure

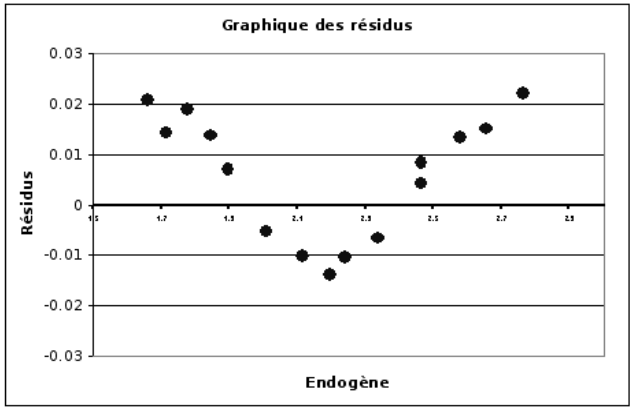
Asymétrie

- Des plages de données de l'endogène mal reconstitués
- Données atypiques
- Mélanges de populations différentes
- Problèmes de spécifications (absence d'exogènes importantes)



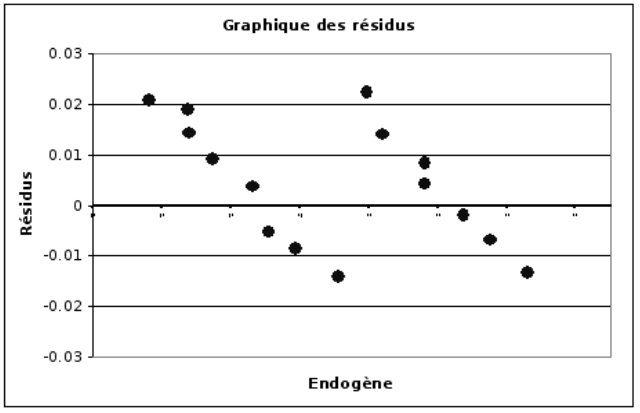
Non linéarité

- Modèle linéaire inadapté, utiliser un modèle non linéaire
- Passer par des transformations de variables (log., carré, racine carrée, produit entre variables : interactions, etc.)



Rupture de structure

- Résidus en « blocs »
- Mélange de populations
- Mutations ou crises dans les séries temporelles

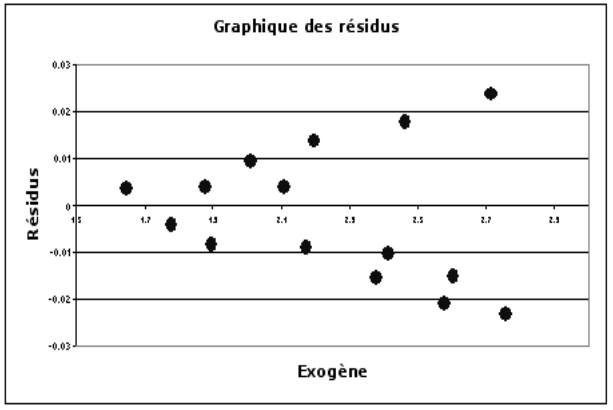


Cas pathologiques

Hétéroscédasticité et autocorrélation des résidus

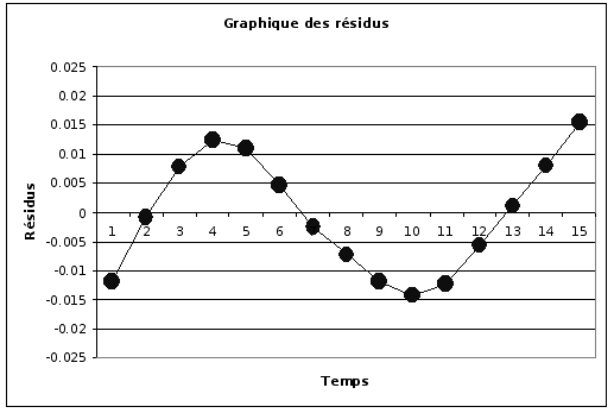
Hétéroscédasticité

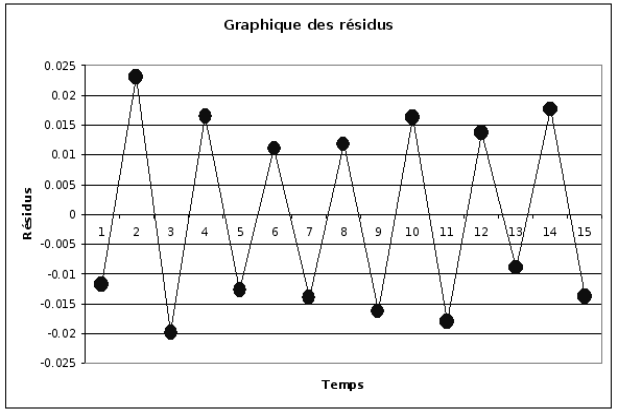
- Variance des résidus non constante
- Exogène en abscisse pour détecter (traiter) dépendance



Autocorrélation

- Associée aux données longitudinales
- Processus particulier (régularité) au cours du temps ?
- Positive (blocs +/-) ou négative (alternance +/-)





Un exemple

Prédiction de la consommation de véhicules

i	Modèle Véhicule	x1 (Frs) 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 Golt 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

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

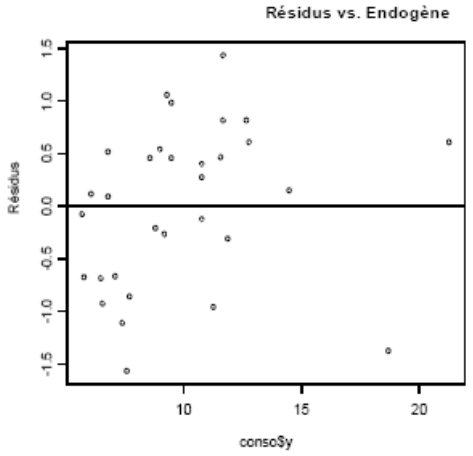
Source	xSS	d.f.	xMS	F	p-value
Regression	364.7719	4	91.1930	136.5413	0.0000
Residual	17.3648	26	0.6679		
Total	382.1368	30			

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

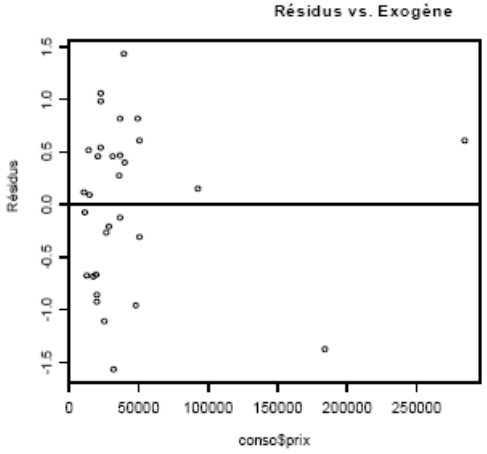
Un exemple

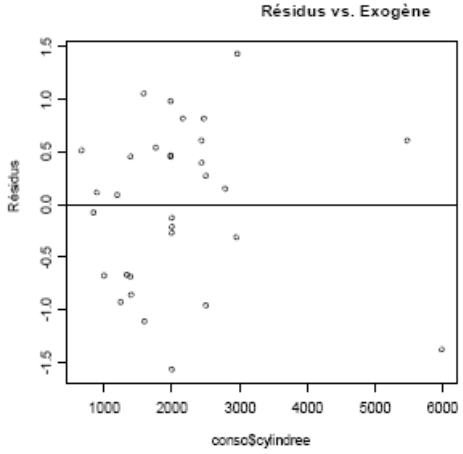
Graphiques des résidus

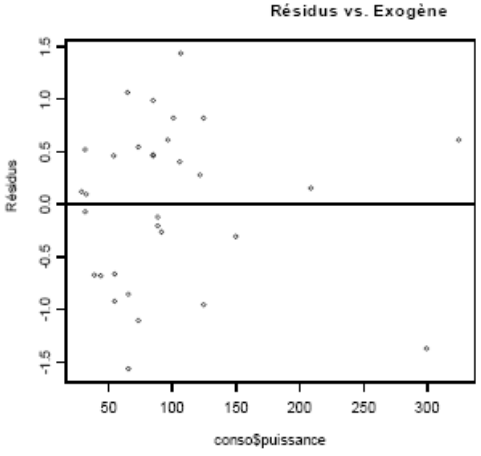


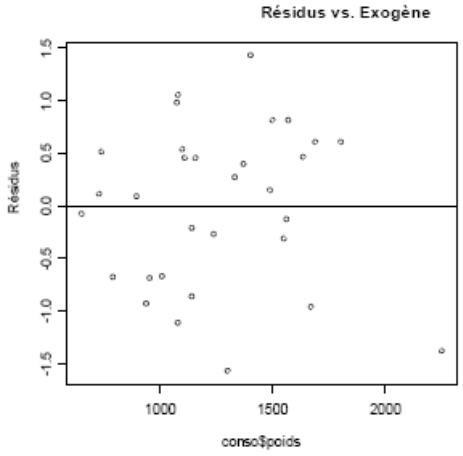
Commentaires

- Globalement, pas de « formes » particulières se dessinent
- Des points manifestement atypiques
- Quelques points très mal modélisés (il faut avoir une idée de l'écart type pour vraiment statuer dessus)
- Des points atypiques bien modélisés et des points atypiques mal modélisés









Trier, filtrer et croiser les données de différentes manières permet d'identifier les points susceptibles de poser problème.

Reste alors à déterminer ce qu'il faut en faire.

Tester le caractère aléatoire des résidus

Pour les données longitudinales (séries chronologiques)…

…mais pas seulement.

Autocorrélation des résidus

Pourquoi c'est important

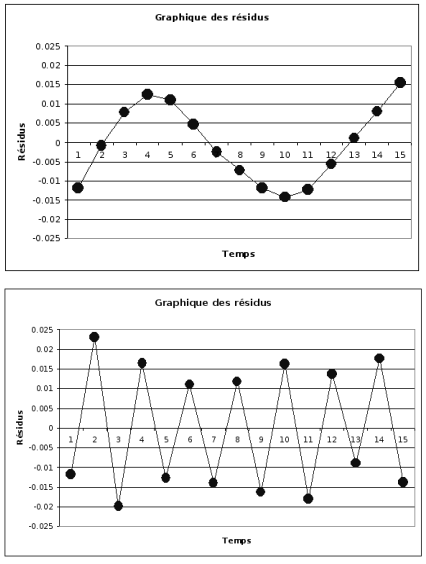
Causes

Problèmes de spécification
Variables importantes manquent
Données déjà manipulées (lissées, moyenne mobile, rétropolées, interpolées, etc. → ex. données fournies par les observatoires statistiques)

Conséquences

MCO quand même non biaisé
Mais MCO n'est plus à variance minimale
Mauvaise estimation de la matrice des VCV
ET (par conséquent) Inférence statistique inopérante

Détection visuelle avec le graphique des résidus



Méthodes numériques

Test de Durbin-Watson

Décrire l'erreur sous la forme

$$\varepsilon_i = \rho.\varepsilon_{i-1} + \upsilon_i \text{ avec } \upsilon_i \equiv N(0;\sigma_\upsilon)$$

Tester → H0 : ρ=0 vs. H1 : ρ≠0

Test spécifique à une forme de l'erreur
Puissant pour cette forme
Mais non opérante pour les autres formes
A voir en M1 (avec les MCG)

Test des séquences (Wald-Wolfowitz)

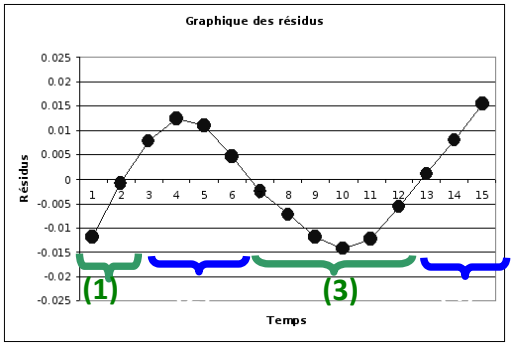
Plus générique
Cherche les régularités sous forme de « séquences »

Test générique, s'applique à toute forme
Moins puissant pour des formes spécifiques
Généralisable pour données transversales (attention, sous certaines conditions uniquement)

Test des séquences

Principe

Les données sont ordonnées (selon le temps)
Compter le nombre de fois où les résidus sont consécutivement au-dessus ou en-dessous de la valeur 0 : on parle de séquences



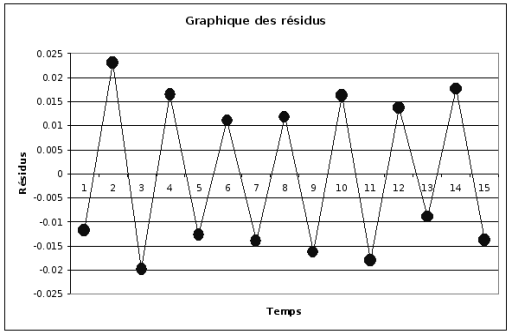
→ r = 4 séquences

Test d'hypothèses

H0 : Les données évoluent de manière aléatoire

Région critique : Un nombre de séquence trop élevé (alternance +/-) est tout aussi suspect qu' un nombre de séquences trop faible (gros blocs de +/-)

(Remarque : quelles sont les valeurs min et max de r ?)



→ r = 15 séquences

Statistique du test et loi asymptotique

Soit n+ (resp. n-), nombre de points positifs (négatifs)

r suit une loi normale de paramètres

$$\mu = \frac{2n_+n_-}{n} + 1$$

$$\sigma = \sqrt{\frac{(\mu-1)(\mu-2)}{n-1}}$$

Statistique centrée réduite

$$z = \frac{r-\mu}{\sigma}$$

Région critique (rejet de H0) :

$$|z| > u_{1-\alpha/2}$$

Test des séquences

Un exemple : Expliquer la consommation en fonction du prix et des revenus

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	-	2
1927	122.2	104.9	86.5	122.452	-0.252	-	
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	+	3
1931	154.2	109.3	70.1	149.804	4.396	+	
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	-	4
1935	136.2	96.4	63.6	145.098	-8.898	-	
1936	168	97.6	52.6	161.584	6.416	+	5
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

Les observations sont compatibles avec H0 : processus aléatoire

Test des séquences

Applicables sur les données transversales ?

Principe

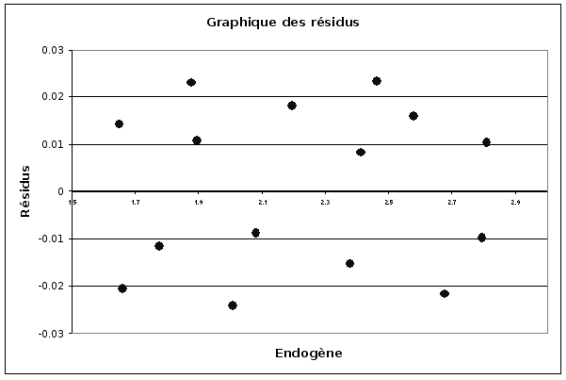
Tester l'Autocorrélation des résidus n'a aucun sens sur les données transversales…
Parce qu'on peut toujours trier (mélanger) les données de manière à ce que les tests concluent H0

Mais on peut exploiter le test des séquences pour détecter les problèmes

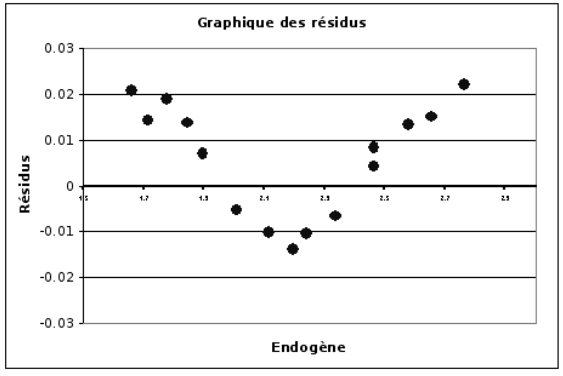
En triant les données selon l'endogène…
Vérifier s'il existe des « zones » où les valeurs de l'endogène sont sur (sous) estimées durablement par le modèle

La nature du test est modifié

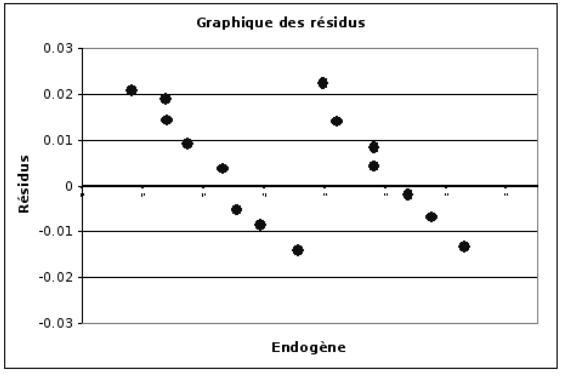
Un nombre élevé de séquences n'est plus un problème dans ce contexte…
Il y a pathologie lorsque le nombre de séquences est anormalement faible
→ On passe sur un test unilatéral



Tout va bien



Non-linéarité : problème



Rupture de structure : problème

Normalité des résidus

Hypothèse nécessaire pour la partie inférentielle
(Tests d'hypothèses sur les coefficients, intervalles de confiance)

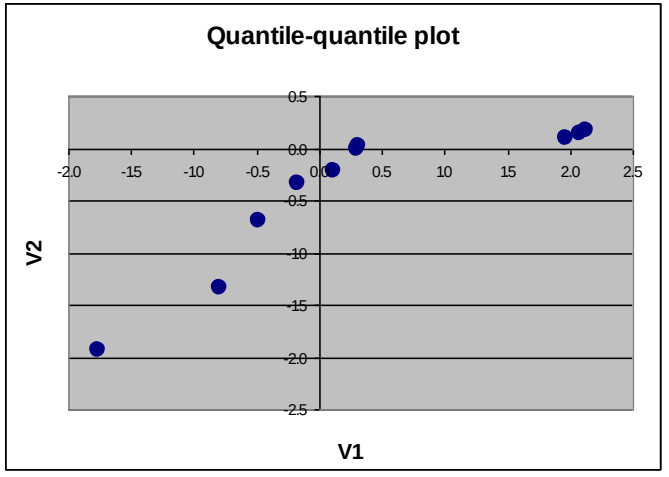
Graphique quantile-quantile

Q-Q norm (Droite de Henry)

Principe Q-Q plot

Confronter les quantiles de 2 distributions (pas nécessairement de même effectif) dans un graphique X-Y…
Si les points forment une droite : les distributions sont identiques (compatibles)

Fréquence	V1 (trié)	V2 (trié)
0.1	-1.764	-1.938
0.2	-0.792	-1.339
0.3	-0.483	-0.694
0.4	-0.171	-0.329
0.5	0.118	-0.221
0.6	0.298	-0.002
0.7	0.317	0.026
0.8	1.962	0.104
0.9	2.079	0.138
1	2.130	0.165



Q-Q plot pour vérifier la compatibilité avec la loi normale : Q-Q norm

En abscisse, les quantiles de la distribution observée…
En ordonnée, les quantiles de la distribution normale (théorique) correspondante (moyenne, écart type estimés)
Si les points forment une droite, la distribution est compatible avec la loi normale

QQ-norm, un exemple

Fréquence (fréquence espérée en accord avec la loi normale)

Quantile de la loi normale (0 ; 1)

Quantile de la loi normale (moyenne ; écart type estimés sur l'échantillon) : quantiles théoriques

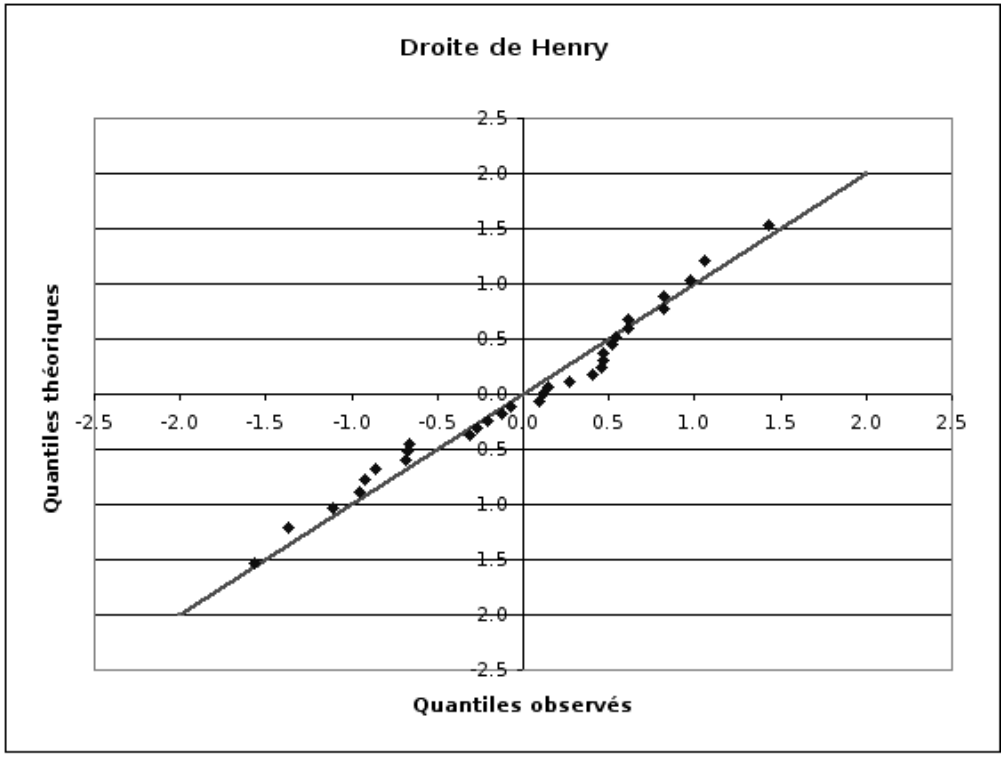
Résidus triés (Quantiles observés)

$$F_i = \frac{i - 0.375}{n + 0.25}$$

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

Moyenne et écart-type estimés



Test de normalité des résidus

Test basé sur l'asymétrie de la distribution

Principe

Si les résidus suivent une loi normale (H0), l'asymétrie = 0
(A contrario) Si asymétrie ≠ 0, alors les résidus ne sont pas compatibles avec la loi normale

Définition du coefficient d'asymétrie

$$\gamma_1 = \frac{\mu_3}{\sigma^3}$$

Coefficient d'asymétrie estimé

$$g_1 = \frac{\frac{1}{n}\sum_i \hat{\varepsilon}_i^3}{\left[\frac{1}{n}\sum_i \hat{\varepsilon}_i^2\right]^{\frac{3}{2}}}$$

Sous H0, g1 suit asymptotiquement une loi normale de paramètres

$$m_1 \approx 0$$

$$s_1 \approx \sqrt{\frac{6}{n}}$$

Test, on forme :

$$c_1 = \frac{g_1 - m_1}{s_1}$$

Région critique :
(Rejet de H0)

$$|c_1| > u_{1-\frac{\alpha}{2}}$$

Application sur les données consommation

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

Test de normalité des résidus

Basé sur aplatissement de la distribution

Principe

Si les résidus suivent une loi normale (H0), l'aplatissement = 0
(A contrario) Si aplatissement ≠ 0, alors les résidus ne sont pas compatibles avec la loi normale

Définition du coefficient d'aplatissement

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

Coefficient d'aplatissement estimé

$$g_2 = \frac{\frac{1}{n}\sum_i \hat{\varepsilon}_i^4}{\left[\frac{1}{n}\sum_i \hat{\varepsilon}_i^2\right]^{\frac{4}{2}}} - 3$$

Sous H0, g2 suit asymptotiquement une loi normale de paramètres

$$m_2 \approx 0$$

$$s_2 \approx \sqrt{\frac{24}{n}}$$

Test, on forme :

$$c_2 = \frac{g_2 - m_2}{s_2}$$

Région critique :
(Rejet de H0)

$$|c_2| > u_{1-\frac{\alpha}{2}}$$

Test de normalité de Jarque-Bera

Combiner les tests basés sur l'asymétrie et aplatissement

Principe

Si les résidus suivent une loi normale (H0), l'asymétrie ET l'aplatissement sont simultanément = 0

Statistique de Jarque-Bera

Sous H0, g1 et g2 sont asymptotiquement indépendants.
On propose la statistique T qui suit une loi du KHI-2 à 2 degrés de liberté (*logique : somme de 2 lois normales au carré indép.*)

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

(n – p – 1) représente les degrés de liberté de la régression c.-à-d. nombre d'observations *moins* nombre de paramètres estimés.

Région critique : $T > \chi^2_{1-\alpha}(2)$

Le test de Jarque Bera est plus puissant (détecte mieux l'écart à la loi normale si elle existe) → à privilégier par rapport aux 2 tests précédents pris individuellement

Application sur les données consommation

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

Conclusion

Analyser les résidus permet de valider ou invalider une régression.

Combiner les techniques numériques et graphiques permettent d'étudier simplement/rapidement les résidus.

En cas d'invalidation, l'analyse graphique des résidus donne une idée des pistes à explorer pour remédier aux problèmes (non-linéarité, rupture de structure, etc.)

Bibliographie

En ligne

R. Rakotomalala, « Pratique de la Régression Linéaire Multiple – Diagnostic et sélection de variables ». Support de cours.
http://eric.univ-lyon2.fr/~ricco/cours/cours/La_regression_dans_la_pratique.pdf

R. Rakotomalala. Portail.
http://eric.univ-lyon2.fr/~ricco/cours/cours_econometrie.html

Wikipédia.
http://fr.wikipedia.org/wiki/Régression_linéaire_multiple

Ouvrages

M. Tenenhaus, « Statistique – Méthodes pour décrire, expliquer et prévoir », Dunod, 2006.

R. Bourbonnais, « Econométrie – Manuel et exercices corrigés », Dunod, 1998.

Y. Dodge, V. Rousson, « Analyse de régression appliquée », Dunod, 2004.