Points atypiques et Points influents

Régression Linéaire Multiple

Ricco RAKOTOMALALA

PLAN

1. Points atypiques et points influents
2. Détection des points atypiques
3. Les points mal modélisés
4. Les points influents
5. Conclusion

Points atypiques et points influents

Repérer les observations qui jouent un rôle anormal dans la régression

Elle prend une valeur inhabituelle sur une variable

Ex. Étude de consommation de voitures de tourisme, un des véhicules fait 700cv → Formule 1, erreur de saisie, …

Elle prend une combinaison de valeurs inhabituelles sur plusieurs variables

Ex. Voiture très légère ET très puissante → Voiture de sport ?

Ex. Une personne âgée qui court le 100m en 10 secondes → ??? Erreur de saisie, extra-terrestre ?

Atypique (aberrant)

Elle pèse de manière exagérée dans la régression

c.-à-d. les résultats sont très différents selon que le point est pris en compte ou pas dans la régression

Influent

Elle est très mal reconstituée (expliquée) par la régression

c.-à-d. le résidu observé est très (trop) élevé, le point n'obéit pas à la relation qui a été établie par la régression

Atypique (régression)

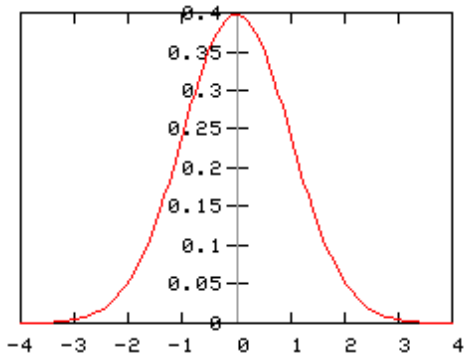
Beaucoup de raisons possibles, une question récurrente : l'observation appartient-elle à la population que nous sommes en train d'étudier ?

Points atypiques

Détection univariée

Une règle simple (simpliste) : la règle des 3-sigmas

- Repose fortement sur l'hypothèse de normalité (symétrie)
- Utilisé à titre indicatif (première approche)



+/- 3 é.t. autour de la moyenne → 99.7% de la population (loi normale)

Une seconde règle simple : la boîte à moustache

- Repose quand même sur l'hypothèse de symétrie
- Moins fruste que la règle des 3-sigmas

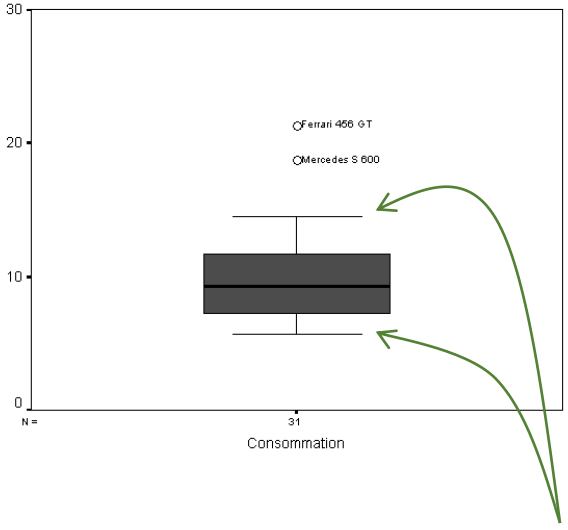
Règles de détection

INNER FENCE

$$LIF = Q1 - 1.5 \times IQ$$
$$UIF = Q3 + 1.5 \times IQ$$

OUTER FENCE

$$LOF = Q1 - 3 \times IQ$$
$$UOF = Q3 + 3 \times IQ$$



Dernières valeurs non-atypiques

Points atypiques

Détection univariée - Exemple

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 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

Manifestement un comportement différent (atypique, *aberrant*) sur plusieurs variables !!!

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

Points atypiques – Détection multivariée

Le levier

La matrice H
Hat matrix

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

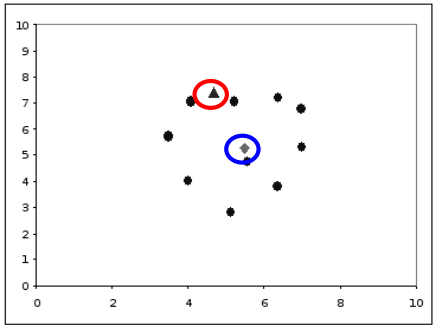
Rôle important : permet de passer de Y vers les projections, de l'erreur théorique vers les résidus observés c.-à-d.

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

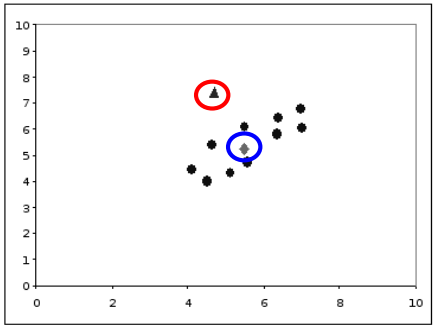
On comprend mieux lorsque les données sont centrées. Il s'agit de la distance au centre de gravité en tenant compte de la forme du nuage de points → **Distance de Mahalanobis**

Sur la diagonale matrice H
Le levier

$$h_i = x_i(X'X)^{-1}x'_i$$



(a)



(b)

Région critique
A partir de quelle valeur du levier faut-il s'inquiéter ?

2 propriétés importantes

$$0 \le h_i \le 1$$

$$\sum_{i=1}^{n} h_i = p+1$$

Règle de détection

$$h_i \ge 2 \times \frac{p+1}{n}$$

Remarque : Plus intéressant est de trier les données selon le levier pour identifier les points problématiques

Points atypiques

Le levier - Exemple

									0.3226
Modèle	const	Prix	Cylindrée	Puissanc	Poids	Consomn	Prédiction	Résidus	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 Hachtback 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 Golt 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

On savait déjà qu'il y avait un problème.

Conjonction de valeurs différentes.
Toyota → Monospace (prix et poids élevés)
Hyundai → Coréenne des années 80, toutes les caractéristiques d'une berline mais prix léger.

Modélisation d'un point

Résidus standardisés (résidus studentisés internes)

Principe : Comparer la valeur de y et la prédiction de y. Normaliser le résidu par l'écart-type.
Résidu élevé → Point mal reconstitué par le modèle. Ne suit pas la liaison qui a été mise en avant par la régression.

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$ Résidu calculé

or $$\sigma^2_{\varepsilon_i} = \sigma^2_{\varepsilon}$$ Écart-type de l'erreur. Par hypothèse : homoscédasticité

mais $$\sigma^2_{\hat{\varepsilon}_i} = \sigma^2_{\varepsilon}(1-h_i)$$ Écart-type du résidu (erreur observée). Estimé par $$\hat{\sigma}^2_{\hat{\varepsilon}_i} = \hat{\sigma}^2_{\varepsilon}(1-h_i)$$

Résidu standardisé →

$$t_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{\varepsilon}\sqrt{1-h_i}} \equiv \mathrm{T}(n-p-1)$$

Loi de Student à (n-p-1) d.d.l.

Région critique →

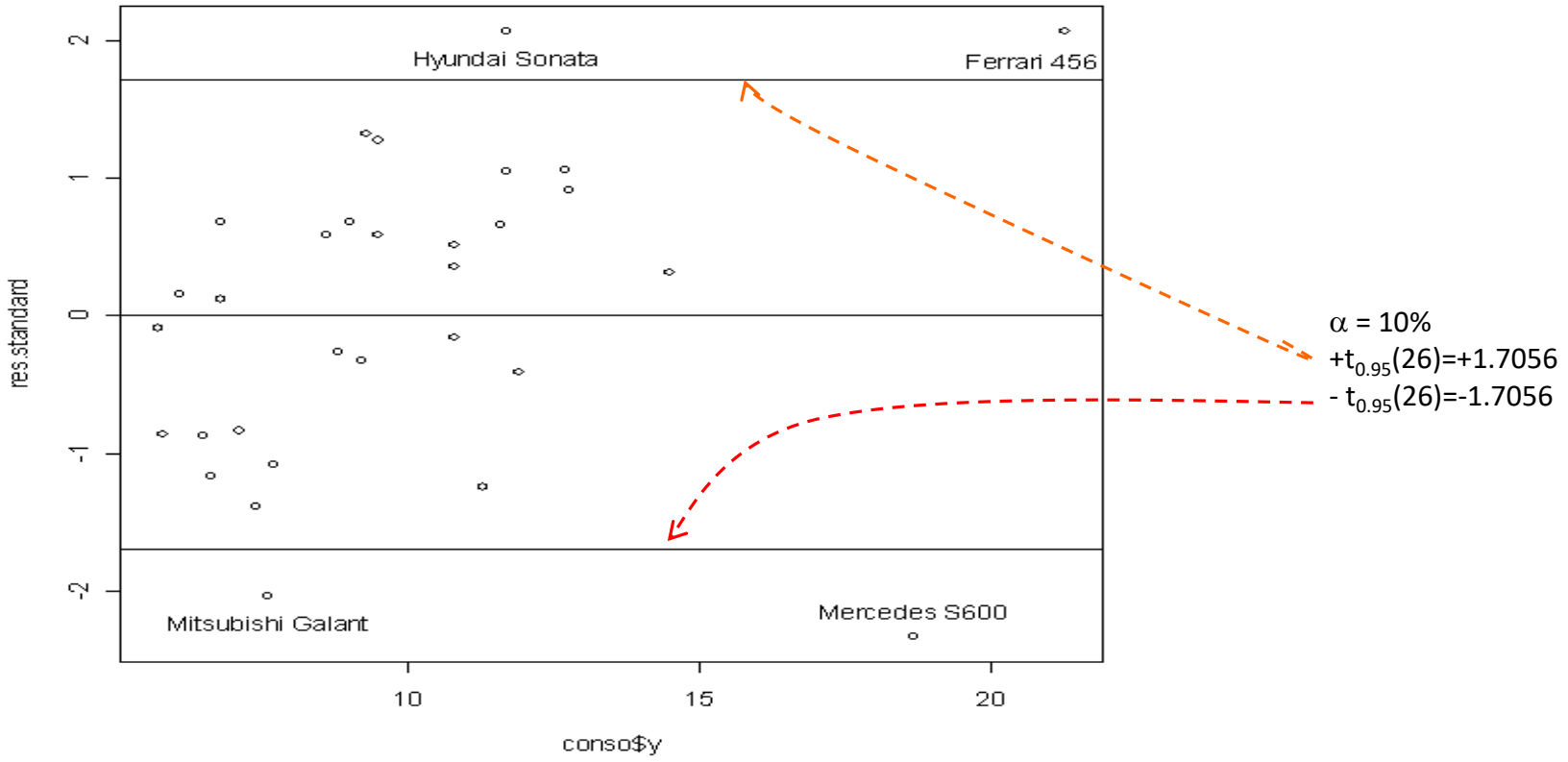
$$|t_i| > t_{1-\alpha/2}(n-p-1)$$

Au risque α
(Test bilatéral)

Modélisation d'un point

Résidu standardisé – Exemple CONSO (1)

La représentation graphique est très avantageuse dans ce cas !!!



Certains points sont mal modélisés car atypiques, d'autres le sont parce qu'ils correspondent à un processus différent (sur-consommation ou sous-consommation par rapport à la modélisation à partir des exogènes).

Modélisation d'un point

Résidu standardisé – Exemple CONSO (2)

Trier le tableau selon le résidu et afficher simultanément les autres informations peut être intéressant aussi

									0.3226	1.7056
Modèle	const	Prix	Cylindrée	Puissanc	Poids	Consomm	Prédiction	Résidus	Leverage	R.Standardisé
Mercedes S 600	1	183900	5987	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 Golt 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

En valeur absolue

Observer le rôle conjoint du levier et du résidu standardisé

Modélisation d'un point

Résidu studentisé (Résidu studentisé externe)

Principe : Comparer toujours la valeur de y et la prédiction de y, mais où le point n°i n'aura pas participé à la régression.

- |RSTUDENT| élevé → Point mal modélisé
- A privilégier par rapport au résidu standardisé (RSTANDARD)
- |RSTUDENT| élevé mais |RSTANDARD| faible → Point exagérément influent (attire les résultats à lui)
- A priori on aura à faire « n » régressions ? Non, dans la pratique

$$\hat{\varepsilon}_i(-i) = y_i - \hat{y}_i(-i)$$

Résidu calculé
$y_i(-i)$ est la prédiction pour le point n°i, ce point n'ayant pas participé à la construction du modèle

Résidu studentisé

$$t_i^* = \frac{y_i - \hat{y}_i(-i)}{\hat{\sigma}_\varepsilon(-i)\sqrt{1-h_i(-i)}} \equiv \mathrm{T}(n-p-2)$$

Loi de Student à (n-p-2) d.d.l.
→ **Pourquoi (n-p-2) ?**
On peut voir la procédure comme le test de nullité du coefficient de la dummy variable associée à l'observation n°i dans la régression à n points.

$$t_i^* = t_i \times \sqrt{\frac{n-p-2}{n-p-1-t_i^2}}$$

Calcul pratique
(à partir du RSTANDARD)

Dans la pratique, nous pouvons la calculer directement à partir de la régression sur les n points.

Modélisation d'un point

Résidu studentisé – Exemple CONSO

									0.3226		1.7081
Modèle	const	Prix	Cylindrée	Puissance	Poids	Consomm	Prédiction	Résidus	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

On retrouve le même groupe. Par d'incohérences véritables entre les deux indicateurs.

Point influent

Distance de COOK

Principe : Comparer globalement les coefficients lorsque le point n°i participe ou pas à la régression. On mesure l'influence d'un point sur les coefficients estimés.

On peut le voir comme un test d'hypothèses

$$H_0 : \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} a_0(-i) \\ a_1(-i) \\ \vdots \\ a_p(-i) \end{pmatrix}$$

Statistique du test

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

Statistique sans avoir à produire n régressions

$$D_i = \frac{t_i^2}{p+1} \times \frac{h_i}{1-h_i}$$

Région critique

$$D_i > \frac{4}{n-p-1}$$

Le plus simple étant toujours de trier les points selon D_i !

Point influent

Distance de Cook – Exemple CONSO

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

Des points manifestement influents : si on les retire de la régression, les coefficients estimés seraient (significativement) différents.

Point influent

DFBETAS

Principe : Comparer le coefficient de la variable X_j lorsque le point n°i participe ou pas à la régression.
On mesure l'influence d'un point sur le coefficient estimé.
DFBETA**S** est normalisé de manière à être comparable d'une variable à l'autre.

Lue sur la j^{ème} position de la diagonale principale de $(X'X)^{-1}$

Statistique du test

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

On peut le voir comme un test d'hypothèses

$$H_0 : a_{j,i} = a_{j,i}(-i)$$

Statistique sans avoir à produire n régressions

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

Région critique

$$\left|DFBETAS_{j,i}\right| > \frac{2}{\sqrt{n}}$$

Trier pour chaque variable devient vite fastidieux ici.

Point influent

DFBETAS – Exemple CONSO

$$\left|DFBETAS_{j,i}\right| > \frac{2}{\sqrt{31}} = 0.3592$$

	DFBETAS				
Modèle	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 Golt 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 Hachtback 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

On observe sur quelles variables pèse chaque observation.
Pas très pratique quand même lorsque le fichier est composé de nombreuses observations et variables.

Conclusion

Analyser les points atypiques et les points influents est important pour identifier les observations qui peuvent fausser les résultats de la régression.

Un point atypique n'est pas forcément influent. Mais les deux situations sont quand même souvent liées.

Attention au traitement des points. La suppression n'est pas la panacée. Elle est licite uniquement lorsque le point est hors population.

Dans notre exemple CONSO, la suppression de la FERRARI et la MERCEDES est évidente. La MASERATI également pose problème, c'est une voiture de sport à hautes performances.

Après, ça devient très difficile. Déjà vérifier les erreurs de saisies serait le bienvenu dans ce type de problème.

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.