Ricco Rakotomalala

Pratique de la Régression Linéaire Multiple

Diagnostic et sélection de variables

Version du 20 sept. 2009

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 les ouvrages énumérés dans la bibliographie.

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 en ligne sont 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 ; 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 remédier ; 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.

Table des matières

Partie I

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 étapes [1] : 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 :

$$y_i = a_0 + a_1 x_{i,1} + \cdots + a_p x_{i,p} + \epsilon_i \tag{0.1}$$

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 :

$$Y = Xa + \epsilon \tag{0.2}$$

les dimensions de matrices sont respectivement :

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

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.

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

1. http://www.itl.nist.gov/div898/handbook/pmd/section4/pmd41.htm

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 4 variables : le prix, la cylindrée, la puissance et le poids (Figure 0.1). Nous disposons pour cela de 31 observations. Nous connaissons la marque et le modèle de chaque véhicule, cela nous permettra d'affiner certains commentaires.

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

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 ;

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)

– mis à part la variable cylindrée, toutes les variables sont significatives au risque de 10%.

La même régression sous EXCEL donne exactement les mêmes résultats (Figure 0.3) [2]. 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 $\hat{\epsilon}_i = y_i - \hat{y}_i$ de la régression.

i	Modèle	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 Golt 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
R^2	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

Remarque 2 (Interprétation des coefficients). D'ores et déjà, sans trop renter dans les détails, on note des bizarreries dans le rôle des variables. Que le prix et la consommation soient d'une certaine manière liés,

2. Fonction DROITEREG(...)

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.

Logiciels

Nous utiliserons principalement le tableur EXCEL. Mais à plusieurs reprises nous ferons appel à des logiciels gratuits tels que TANAGRA [3] et R [4] ; et à des logiciels commerciaux tels que SPSS [5] et STATISTICA [6]. *Qu'importe le logiciel en réalité, le plus important est de savoir lire correctement les sorties des outils statistiques.*

3. TANAGRA : Un logiciel gratuit de Data Mining pour l'enseignement et la recherche - `http://eric.univ-lyon2.fr/~ricco/tanagra/fr/tanagra.html`
4. The R Project for Statistical Computing - `http://www.r-project.org/`
5. Pour une lecture détaillée des résultats fournis par SPSS, voir `http://www2.chass.ncsu.edu/garson/PA765/regress.htm`
6. Pour une lecture des résultats de STATISTICA, voir `http://www.statsoft.com/textbook/stmulreg.html`

1

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 résultats [1].

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 $(i = 1, \ldots, n)$ sont indépendantes.

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

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

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

Remarque 3 (Moyenne des résidus). Dans un modèle avec constante, la moyenne des résidus $\bar{\hat{\epsilon}} = \frac{1}{n}\sum_i \hat{\epsilon}_i$ 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 appréhendent mal. Toute analyse de régression devrait être immédiatement suivie des graphiques des résidus observés... car il y en a plusieurs.

1. Voir Dodge, pages 113 à 120.

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 loin [2], 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 abcisse.

Résidus en fonction de l'endogène Y

Ce type de graphique permet de se rendre compte de la qualité de la régression. Les résidus $\hat{\epsilon}_i$ 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.

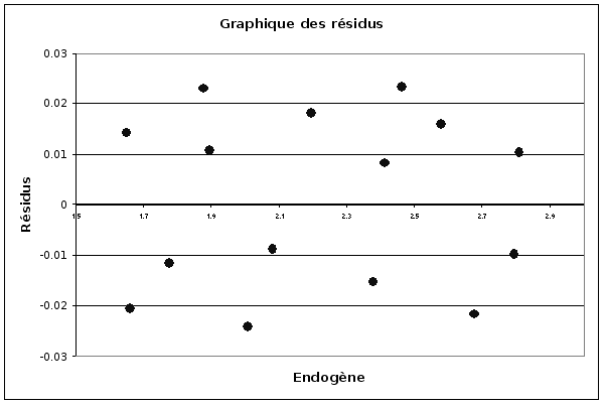


Fig. 1.1. Graphique "normal" des résidus. Endogène vs. Résidus.

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 abcisse la variable temps. Elle permet d'ordonner les valeurs d'une autre manière. Cela peut être utile pour 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 abcisse soit l'endogène ou une des exogènes (Figure 1.2).

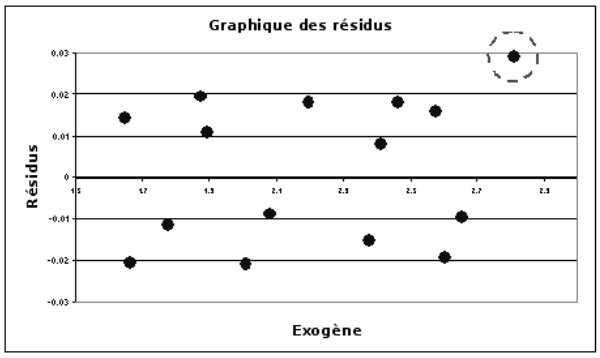


Fig. 1.2. Un point présente une valeur atypique pour une des exogènes. De plus, elle est mal reconstituée par la régression (le résidu est élevé).

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 formés 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.

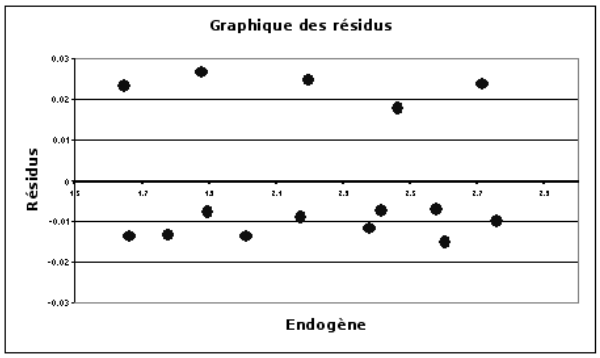


Fig. 1.3. La distribution des résidus est asymétrique.

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-estimé) de la valeur 0 (Figure 1.4). Cela peut être résolu en ajoutant une variable non-linéaire 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 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

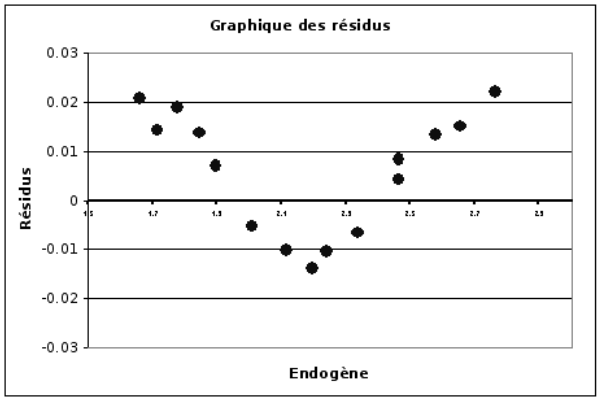


Fig. 1.4. La relation à modéliser est non-linéaire

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).

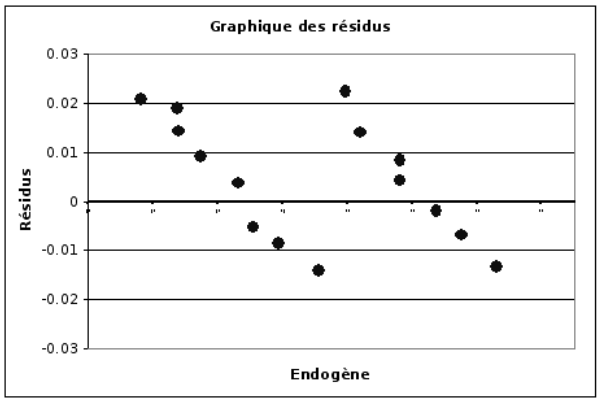


Fig. 1.5. Résidus caractéristiques d'une rupture de structure

Hétéroscédasticité

Souvent associée à une des exogènes en abcisse, 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).

Autocorrélation des résidus

Ce problème est spécifique aux données longitudinales. Dans le graphique des résidus, nous plaçons des dates en abcisse, 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).

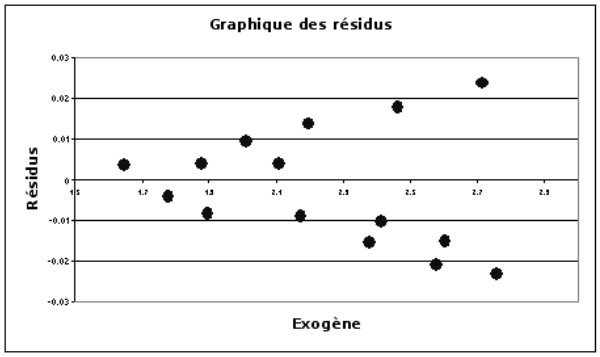


Fig. 1.6. La variance des résidus augmente avec les valeurs d'une des exogènes

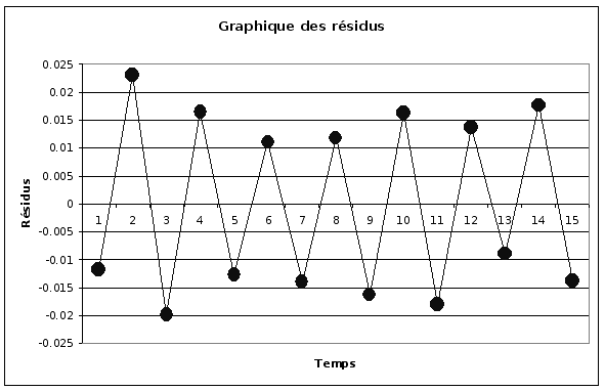


Fig. 1.7. Autocorrélation négative des résidus

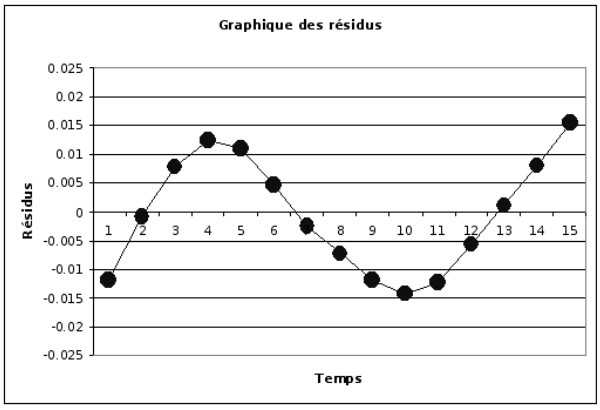


Fig. 1.8. Autocorrélation positive des résidus

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.

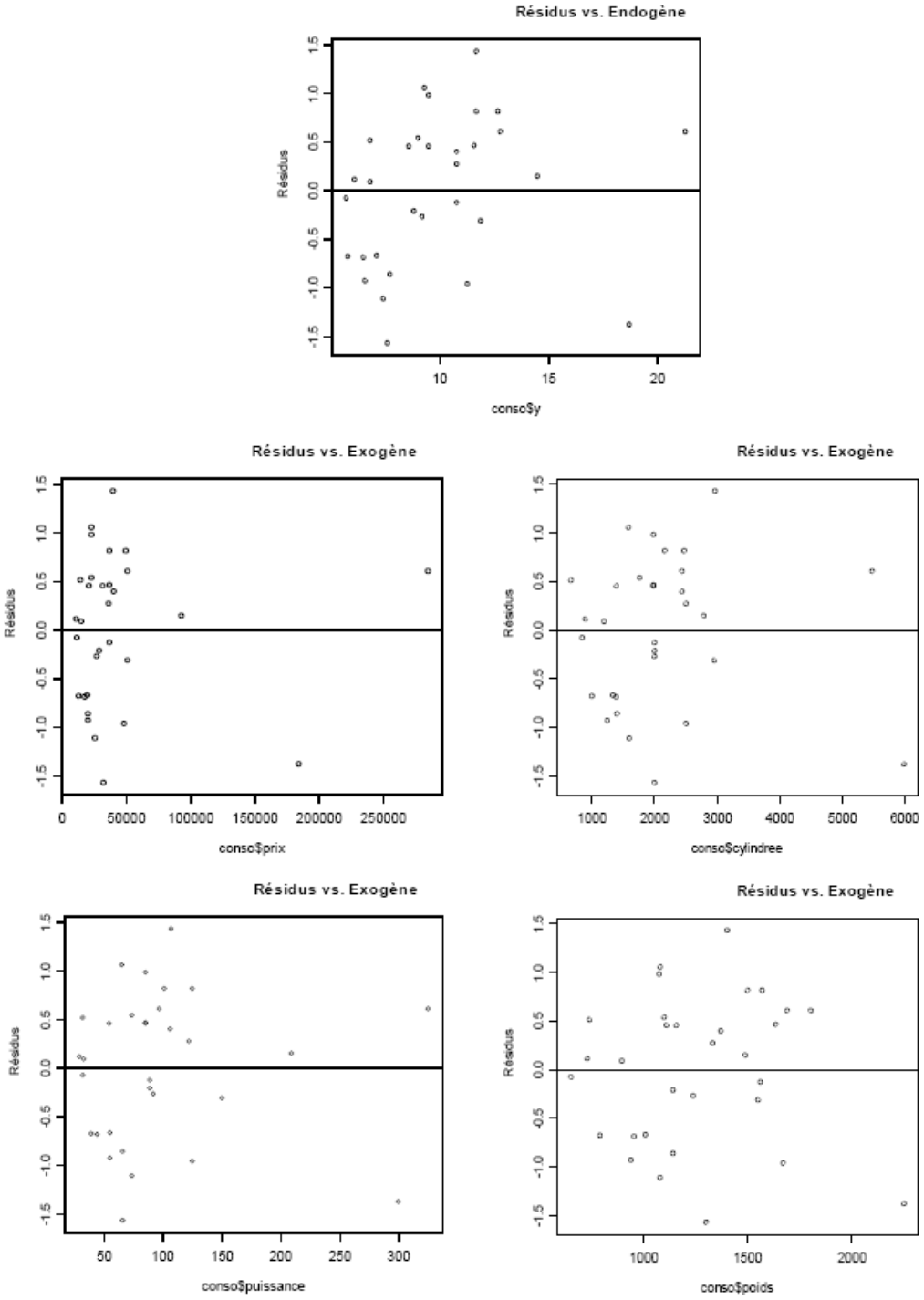


Fig. 1.9. Graphiques des résidus - Données CONSO

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

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 si les résidus sont produits de manière totalement aléatoire. Si l'on conclut au rejet de cette hypothèse, cela indique que 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 BLUE[3] : 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 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 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 :

$$\epsilon_i = \rho.\,\epsilon_{i-1} + \nu_i, \;\; \text{avec } \nu_i \sim \mathcal{N}(0, \sigma_\nu) \tag{1.2}$$

Le test d'hypothèses s'écrit :

3. Best Linear Unbiased Estimator

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

On utilise la statistique de Durbin-Watson

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

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 $H0$ 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 retardée [4].

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).

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).

L'équation de régression à mettre en place est

$$y_i = a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \epsilon_i \ , \ i = 1, \ldots, 17$$

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

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 $a_2 = 1.06$.

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

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

2. Nous formons la prédiction $\hat{y}_i$ avec ces coefficients.
3. Nous calculons l'erreur de prédiction, le résidu de la régression $\hat{\epsilon}_i = e_i = y_i - \hat{y}_i$.
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'autocorrélation d'ordre 1 des résidus peut être rejetée ($\rho = 0$).

1.2.2 Test des séquences

Le test des séquences[6], 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 asymptotiquement[7] une loi normale de paramètres :

$$\mu_r = \frac{2n_+n_-}{n} + 1 \tag{1.4}$$

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

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.

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...

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	-	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
R²	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 $\hat{\epsilon}_i = e_i = y_i - \hat{y}_i$.
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 $\mathcal{N}(0, \sigma_\epsilon)$ du terme d'erreur de l'équation de régression (Équation 0.1). Vérifier cette hypothèse semble incontournable pour obtenir des résultats exacts [8].

Nous disposons des erreurs observés $\hat{\epsilon}_i$, 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 $V(\epsilon_i) = \sigma_\epsilon^2$, la variance du résidu, l'erreur observée, ne l'est pas $V(\hat{\epsilon}_i) = \sigma_\epsilon^2(1 - h_{ii})$, où h_{ii} est lue sur la diagonale principale de la *hat matrix* $H = X(X'X)^{-1}X'$. Et surtout, la covariance $cov(\hat{\epsilon}_i, \hat{\epsilon}_j) = -\sigma_\epsilon^2 h_{ij}$ 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 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 présentons [10].

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.

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/~jravet/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

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

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 $\hat{\epsilon}_i$ 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 $\epsilon_i^* = \hat{\sigma}_{\hat{\epsilon}} * z_i$. Si la distribution empirique cadre parfaitement avec la loi normale, les points devraient être alignés sur la diagonale principale. Ici, $\hat{\sigma}_{\hat{\epsilon}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2}$.

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

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

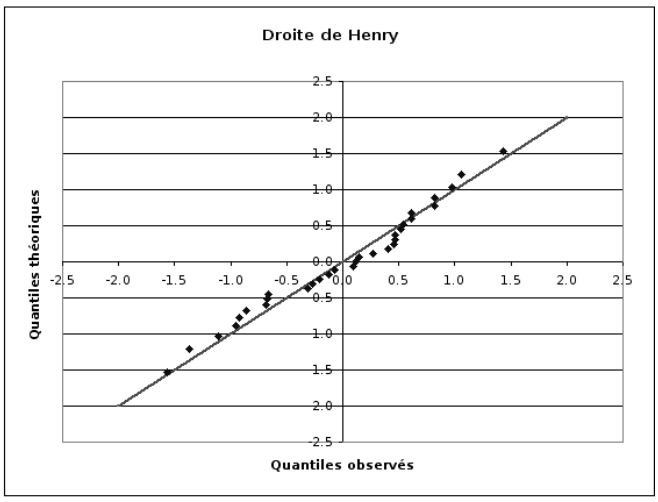


Fig. 1.13. Droite de Henry sur les résidus des MCO – Données CONSO

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

$$\gamma_1 = \frac{\mu_3}{\sigma^3} \tag{1.6}$$

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 :

$$H_0 : \epsilon \text{ suit une loi normale, par conséquent } \gamma_1 = 0$$
$$H_1 : \epsilon \text{ ne suit pas une loi normale, par conséquent } \gamma_1 \neq 0$$

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 :

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

Sous H_0, elle suit asymptotiquement une loi normale d'espérance et d'écart-type [11]

$$\mu_1 \approx 0$$
$$\sigma_1 \approx \sqrt{\frac{6}{n}}$$

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 $\hat{\epsilon}_i$.
2. Nous calculons les colonnes de $\hat{\epsilon}_i^2$ et $\hat{\epsilon}_i^3$.
3. Nous calculons les sommes et formons $g_1 = \frac{-0.1220}{0.5602^{3/2}} = -0.2909$.

11. 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

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 :

$$H_0 : \epsilon \text{ suit une loi normale, par conséquent } \gamma_1 = 0 \text{ et } \gamma_2 = 0$$

$$H_1 : \epsilon \text{ ne suit pas une loi normale, par conséquent } \gamma_1 \neq 0 \text{ ou } \gamma_2 \neq 0$$

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 :

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

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

$$\mu_2 \approx 0$$
$$\sigma_2 \approx \sqrt{\frac{24}{n}}$$

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-Bera [12] :

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

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 libertés 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 :

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

Il s'agit d'un test unilatéral, $\chi^2_{1-\alpha}(2)$ 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) :

12. 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. Nous récupérons la colonne des résidus $\hat{\epsilon}_i$.
2. Nous calculons les colonnes de $\hat{\epsilon}_i^2$, $\hat{\epsilon}_i^3$ et $\hat{\epsilon}_i^4$.
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 $\chi^2_{0.90}(2) = 4.6052$.

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

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.

2

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 pourrons 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.

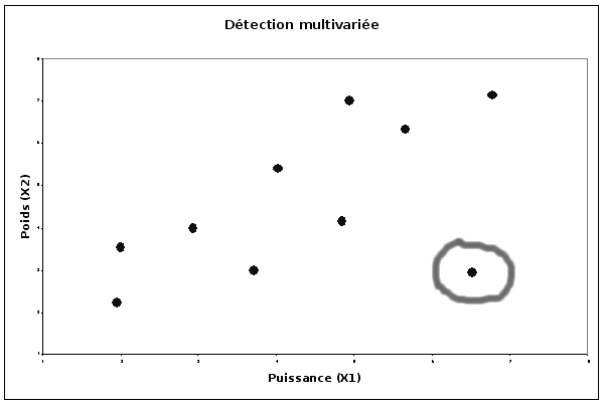


Fig. 2.1. Le point entouré est suspect car la combinaison de valeurs est inhabituelle

Outre les ouvrages enumérés en bibliographie, deux références en ligne complètent à merveille ce chapitre : le document de J. Confais et M. Le Guen [8], 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 ($IQ = Q3 - Q1$).

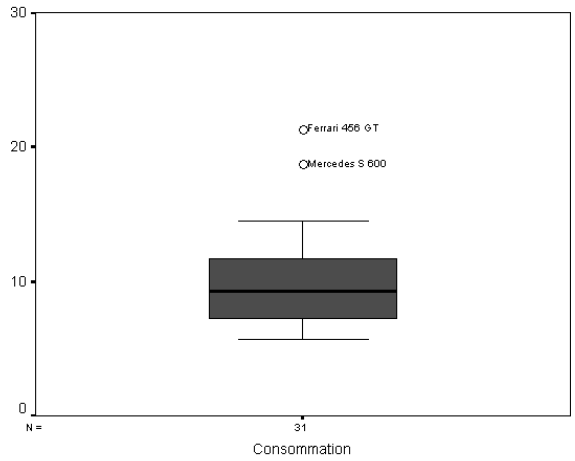


Fig. 2.2. Boxplot de la variable endogène "consommation (y)", 2 observations se démarquent

On pense à tort que les extrémités de la boîte correspond aux valeurs minimales et maximales. En réalité il s'agit des valeurs minimales et maximales non atypiques définies par les règles suivantes[2] :

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

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.

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

1. `http://en.wikipedia.org/wiki/Box_plot`
2. `http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm`

de la moyenne. Si l'on considère que la distribution est normale, 99.7% des observations sont situés 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).

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

Tableau 2.1. Points suspects fichier CONSO : détection univariée

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

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

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).

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

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

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 observés [3].

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

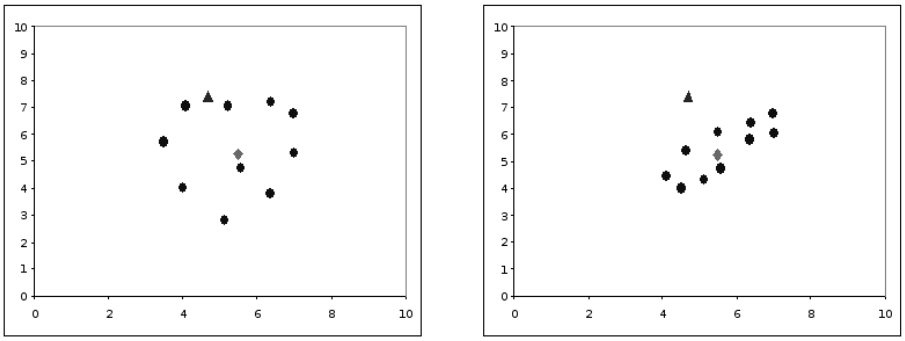


Fig. 2.4. Le point △ et le centre de gravité ◇ sont situés aux mêmes coordonnées dans les graphiques (a) et (b). Pourtant △ apparaît nettement atypique dans (b).

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" ?

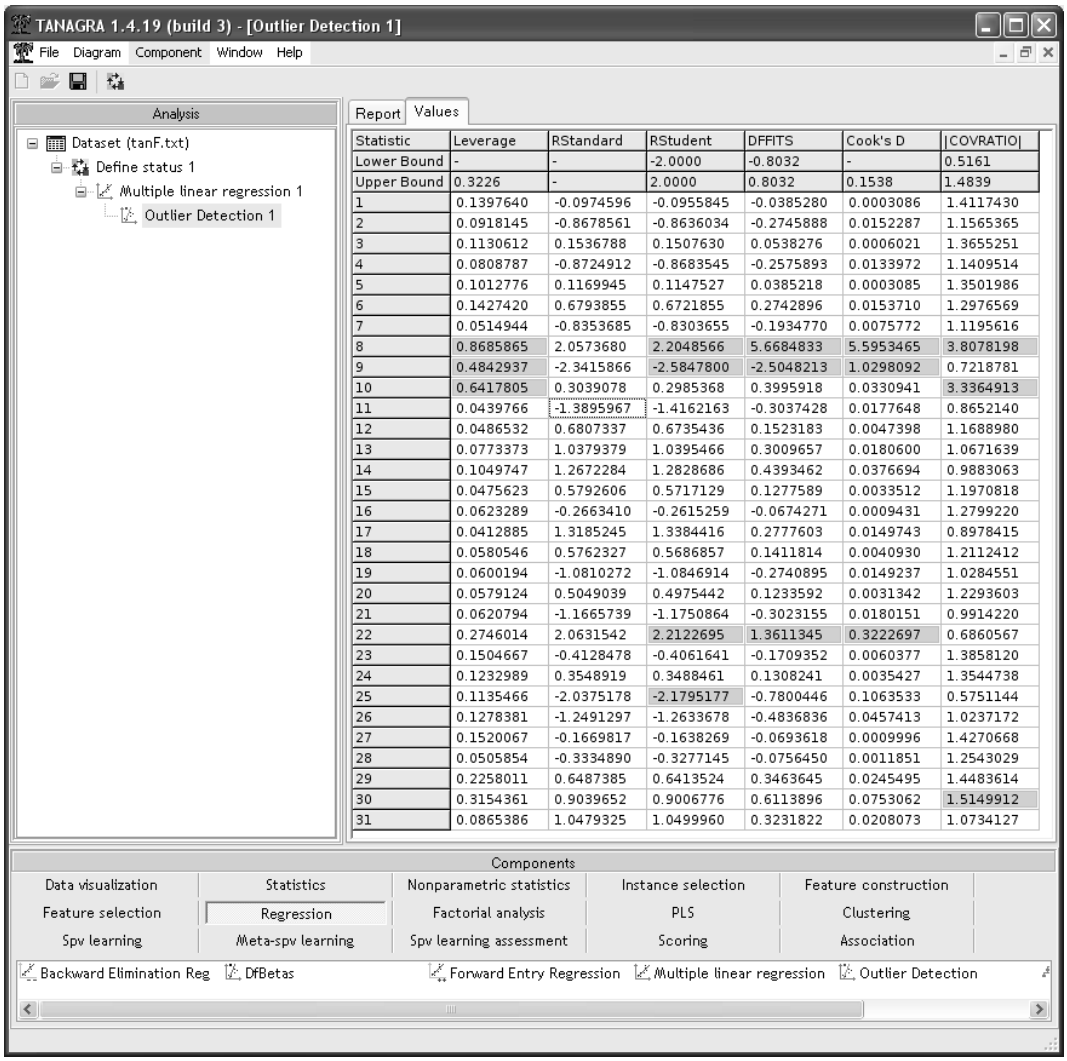
Pour cela, penchons-nous sur quelques propriétés du levier. Par définition $0 \leq h_i \leq 1$, et surtout $\sum_{i=1}^{n} h_i = p+1$, où $p+1$ est le nombre de coefficients à estimer dans une régression avec constante. On considère que le levier d'une observation est anormalement élevé dès lors que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

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

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 de l'écartement semblent évidentes : il s'agit de grosses cylindrées luxueuses, des limousines (Mercedes) ou des véhicules sportifs (Ferrari, Maserati).

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.

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.

									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

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

2.3 Résidu standardisé

Résidu standardisé

Le résidu standardisé, appelé également *résidu studentisé interne* dans certains ouvrages, s'intéresse à l'importance du résidus observé $\hat{\epsilon}_i = y_i - \hat{y}_i$. S'il est anormalement élevé, en valeur absolue, cela indique que 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 $\sigma^2_{\epsilon_i} = \sigma^2_\epsilon$ est constante, il n'en est pas de même du résidu $\sigma^2_{\hat{\epsilon}_i} = \sigma^2_\epsilon(1 - h_i)$. 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^2_ϵ, nous estimons la variance des résidus avec

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

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

Le résidu standardisé est donc définie comme le rapport

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

Région critique

Pour décider qu'un point est aberrant ou pas, il nous faut de nouveau 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 $\epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon)$, nous en déduisons que $\hat{\epsilon}_i \sim \mathcal{N}(0, \sigma_{\hat{\epsilon}_i})$. On peut montrer facilement que $\hat{\sigma}^2_{\hat{\epsilon}_i}$ 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

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

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. $\hat{\epsilon}_i$ est élevé ;
- la régression est précise c.-à-d. $\hat{\sigma}_\epsilon$ 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 résidus standardisé. Il est plus commode de revenir au graphique des résidus en mettant en abcisse 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).

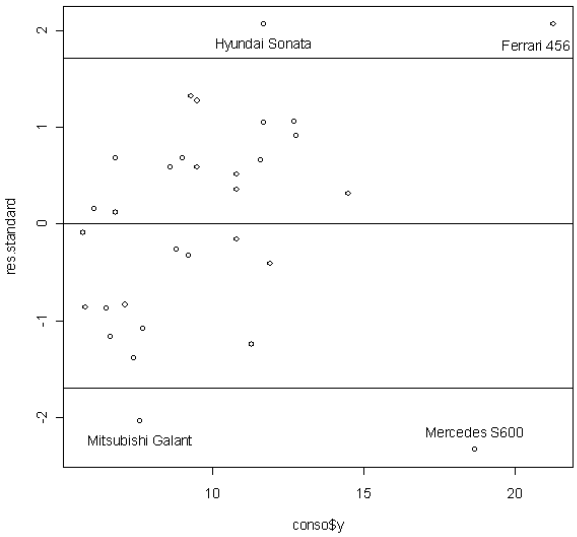


Fig. 2.7. Graphique des résidus standardisés vs. endogène - Données CONSO

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 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.

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.

									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

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

- La "Ferrari" est mal reconstituée parce qu'elle est avant tout très différente des autres $h = 0.8686$. Le résidu brut $\hat{\epsilon} = 0.610$ 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 $\hat{\epsilon} = y - \hat{y} = 11.7 - 10.264 = 1.436$; la Mitsubishi est en revanche particulièrement sobre $\hat{\epsilon} = 7.6 - 9.168 = -1.568$.

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 $\hat{\epsilon}_i$ 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 $\hat{\epsilon} \approx 0$, nous conclurons à tort qu'elle est bien reconstituée et donc ne fausse en rien les résultats de la modélisation (Figure 2.9).

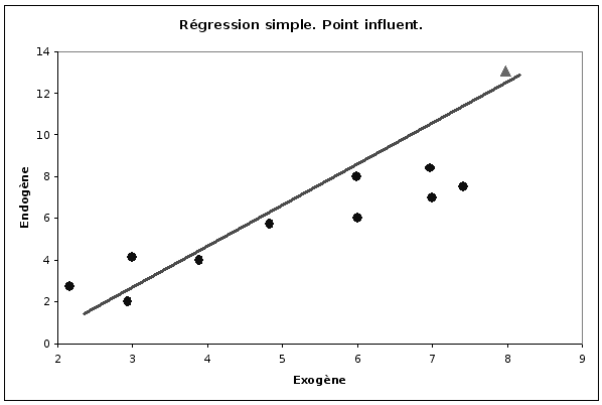


Fig. 2.9. Exemple de régression simple où l'observation $\triangle$ est certes bien modélisée ($\hat{\epsilon} \approx 0$) mais elle fausse totalement les calculs : on parle de point exagérément influent.

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 $\hat{\sigma}_\epsilon(-i)$, 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

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

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).

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 pourrons 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

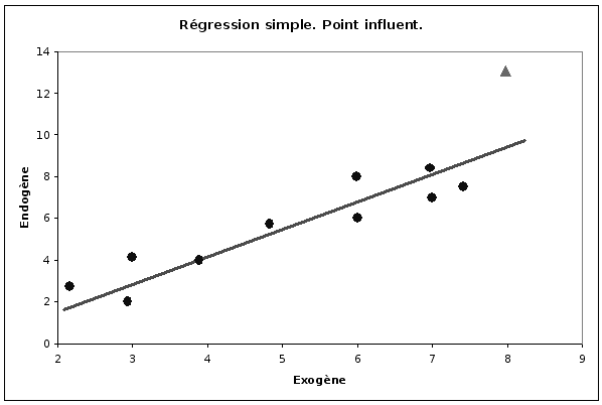


Fig. 2.10. Principe de la donnée supplémentaire : l'observation $\triangle$, exclue du calcul de la droite de régression, devient très mal prédite

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

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

$$y = a_0 + a_1 x_1 + \ldots + a_p x_p + b \times z + \epsilon \tag{2.7}$$

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) [5] :

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

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 de formuler rigoureusement le test d'hypothèses permettant de déterminer si une observation est atypique/influente ou non. Il s'écrit :

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

$$H_0 : b = 0$$
$$H_1 : b \neq 0$$

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.

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).

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 concevoir son influence sur la régression.

									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

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

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 $\triangle$.

Le DFFITS est normalisée de la manière suivante

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

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).

						0.8032		
Observation		Leverage	RStandard	RStudent	DFFITS	|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	Fort Escort 1.4i PT	0.0581	0.5762	0.5687	0.1412	0.1412	0.0041	1.2112
24	Mazda Hachtback V	0.1233	0.3549	0.3488	0.1308	0.1308	0.0035	1.3545
15	VW Golt 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 $DFFITS$

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

$$DFFITS_i = t_i^* \sqrt{\frac{h_i}{1-h_i}} \tag{2.10}$$

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.

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

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

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 :

$$R.C. : D_i > 1 \tag{2.12}$$

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) :

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

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).

						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	Fort 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

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

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

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

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

où $\hat{a}$ est le vecteur des $(p+1)$ coefficients estimés $(\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_p)'$ 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% [6]. 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).

Observation		Leverage	RStandard	RStudent	DFFITS	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 Tempra 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	Fort 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.

6. `http://www-stat.stanford.edu/~jtaylo/courses/stats203/notes/diagnostics.pdf`

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

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

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 ; $\hat{\sigma}_\epsilon(-i)$ 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

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

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

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

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).

	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

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

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

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

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

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 :

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

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

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

que l'on peut simplifier :

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

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).

					< 0.5161 ; > 1.4839	0.4839
	Observation	Leverage	RStandard	RStudent	COVRATIO	|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 Golt 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 $COVRATIO_i$

Calcul pratique

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

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

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 salutaire (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).

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 ?

	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		$\lvert \text{Student residual} \rvert > 2$ alors le résidu i est significativement $\neq 0$	$\lvert \text{RStudent} \rvert > 2$ 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	$\text{CookD} > 1$ alors l'observation i est influente globalement	$\lvert \text{Dfbetas} \rvert > \frac{2}{\sqrt{n}}$ indique une influence de l'obs. i sur l'estimation de β_j	$\lvert \text{Covratio} - 1 \rvert > \frac{3(p+1)}{n}$ nécessite une investigation	$\lvert \text{Dffits} \rvert > 2\sqrt{\frac{(p+1)}{n}}$ 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)

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.

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.

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.

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

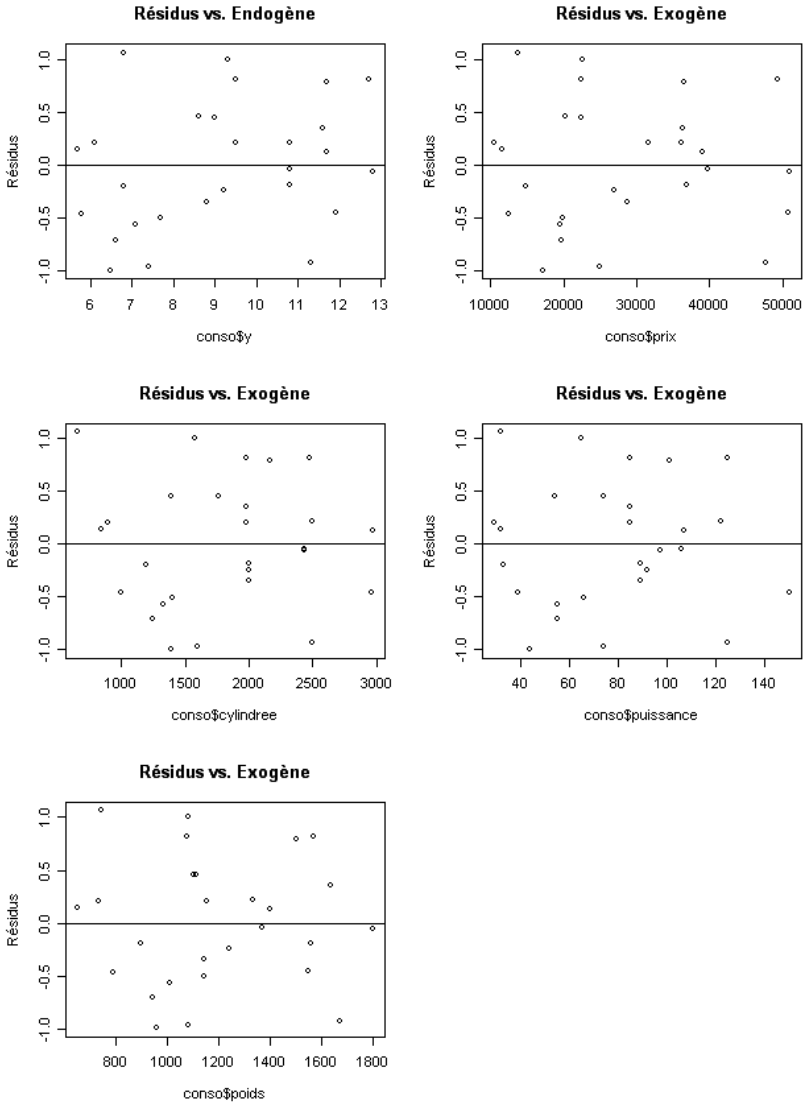


Fig. 2.19. Graphiques des résidus, fichier CONSO après suppression des 4 points atypiques

3

Colinéarité et sélection de variables

L'un des objectifs de la régression est d'essayer de décrire le processus de causalité entre les variables exogènes et la variable 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 unes à unes, 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 les variables exogènes sont redondantes, certaines emmènent le même type d'information : c'est le problème de la colinéarité, certaines exogènes sont linéairement corrélées, elles se gênent mutuellement dans la régression.

Dans ce chapitre, nous présentons tout d'abord 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'indiquer [1] 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://biol10.biol.umontreal.ca/BIO2042/Regr_mult.pdf`

Plusieurs problèmes peuvent surgir[2] :

- 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, $rang(X'X) < p+1 \rightarrow (X'X)^{-1}$ n'existe pas. Le calcul est impossible.
- Si la colinéarité est forte, $\det(X'X) \approx 0$, l'inverse[3] $(X'X)^{-1} = \frac{1}{\det(X'X)} comA'$ contient des valeurs très élevées. Il en est de même pour la matrice de variance covariance des coefficients estimés $\hat{\Omega}_{\hat{a}} = \hat{\sigma}_{\epsilon}^2 (X'X)^{-1}$. 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 $y = a_0 + a_1 x_1 + \cdots + a_p x_p + \epsilon$, nous recueillons le coefficient de détermination R^2.
2. Nous calculons les corrélations croisées entre les variables exogènes X_{j1} et X_{j2} : r_{j_1,j_2} 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 $R^2 < r^2_{j_a,j_b}$.

Dans la pratique, une simple proximité des valeurs R^2 et $r^2_{j_a,j_b}$ 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) :

$$v_j = \frac{1}{1 - R_j^2}$$

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

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

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 et 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 $V(\hat{a}_j) = \frac{\sigma_\epsilon^2}{n}$. 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 préfèrent [5] les seuils 5 ou même 10. A vrai dire, l'essentiel est d'identifier le groupes de variables qui causent le plus de problèmes 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 multiples, chaque variable X_j contre les $(p-1)$ autres, 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}.

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 multicolinéarité

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 à partir de ses valeurs propres (ex. l'indice de multicolinéarité (`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.

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

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 $y = a_0 + a_1x_1 + \cdots + a_px_p + \epsilon$, 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 $signe(\hat{a}_j) \neq signe(r_{y,x_j})$.

Application au 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 écarté 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

$$R^2 = 1 - \frac{SCR}{SCT} \tag{3.1}$$

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

$$\bar{R}^2 = 1 - \frac{CMR}{CMT} = 1 - \frac{SCR/(n-q-1)}{SCT/(n-1)} \tag{3.2}$$

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

$$\bar{R}^2 = 1 - \frac{n-1}{n-q-1}(1-R^2) \tag{3.3}$$

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 3.3). 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 :

$$AIC = n \ln \frac{SCR}{n} + 2(q+1) \tag{3.4}$$

Avec le critère BIC de Schwartz, nous cherchons à optimiser

$$BIC = n \ln \frac{SCR}{n} + \ln(n)(q+1) \tag{3.5}$$

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 l'interaction entre les variables. Cela est fait de manière implicite avec la pénalisation de la complexité : deux variables redondantes 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 disions-nous plus haut. 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'observation, 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 supplémentaire [6]. On déduit alors un indicateur similaire au SCR que l'on appelle PRESS (Predicted Residual Sum of Squares) [7] :

$$PRESS = \sum_{i=1}^{n} (y_i - \hat{y}_i(-i))^2 \tag{3.6}$$

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

$$y_i - \hat{y}_i(-i) = \frac{y_i - \hat{y}_i}{1 - h_i} \tag{3.7}$$

Procédure de sélection basée sur le PRESS

A la différence du R^2, nous disposons d'un critère *honnête* de 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, si le R^2 peut s'améliorer (fallacieusement), le PRESS lui en revanche se dégrade, indiquant par là l'inutilité de la variable.

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

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

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	UISSANC	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 Golt 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 Hachtback 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

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 $\hat{\epsilon}_i = y_i - \hat{y}_i$;
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* $y_i - \hat{y}_i(-i) = \frac{\hat{\epsilon}_i}{1-h_i}$;
5. Nous en déduisons le $PRESS = \sum_{i=1}^{n}[y_i - \hat{y}_i(-i)]^2 = 13.54$.

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.eduzSz~ronnykzSzwrappers.pdf/kohavi97wrappers.pdf`

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.

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).

Appliqué sur le fichier CONSO de 27 observations, nous obtenons la séquence de calculs[9] :

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.

Le modèle optimal au sens du critère AIC est

$$y = 1.392276 + 0.01311 \times cylindree + 0.004505 \times poids$$

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

Remarque 19 (Recherche forward). Si nous avions 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 correspond 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 offre 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	$t^2_{\hat{a}_j} = F$ (p-value)
1	y = cte	-	poids → 207.63 (0.0000) prix → 199.19 (0.0000) cylindrée → 118.60 (0.0000) puissance → 93.53 (0.0000)
2	y = poids + cte	0.8925	cylindrée → 11.66 (0.0023) puissance → 7.42 (0.0118) prix → 6.32 (0.0190)
2	y = poids + cylindrée + cte	0.9277	prix → 0.53 (0.4721) puissance → 0.01 (0.9288)

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. A 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. A 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 elles [10]. La valeur $\alpha = 10\%$ est proposée par défaut dans la logiciel SPSS par exemple. La plupart des logiciels procèdent ainsi.

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.

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 = prix + cylindrée + puissance + poids + cte	0.9295	puissance → 0.0620 (0.8057) prix → 0.5666 (0.4596) cylindrée → 2.7978 (0.1086) poids → 8.2287 (0.0089)
2	y = prix + cylindrée + poids + cte	0.9293	prix → 0.5344 (0.4721) cylindrée → 4.6779 (0.0412) poids → 9.4345 (0.0054)
3	y = cylindrée + poids + cte	0.9277	cylindrée → 11.6631 (0.0023) poids → 33.7761 (0.0000)

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 suivante, à 3 exogènes, a 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.

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

$$y = 1.392276 + 0.01311 \times cylindree + 0.004505 \times poids$$

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 ; Dodge[11], 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 test[12]. On utilise
$$t_b = \frac{r}{\sqrt{\frac{1-r^2}{n-3}}}.$$
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.

11. La description donnée par Dodge est un peu différente, il utilise 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

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 différentes 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 $p - value = 0.0000$. 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 $y = 1.0353 + 0.0068 \times poids$; (b) nous calculons le résidus $e_1 = y - (1.0353 + 0.0068 \times poids)$.

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 $r_{e_1,poids} = 0$ 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* :

$$y = 1.0353 + 0.0068 \times poids$$

3.4 Coefficient de corrélation partielle et sélection de variables

3.4.1 Corrélation brute et partielle

3.4.2 Coefficient de corrélation brute

Le coefficient de corrélation[13] quantifie le degré de liaison **linéaire** entre deux variables continues Y et X. Elle est définie par

$$\rho_{y,x} = \frac{cov(y,x)}{\sigma_y \cdot \sigma_x} \tag{3.8}$$

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 :

$$r_{y,x} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \cdot \sqrt{\sum_i (y_i - \bar{y})^2}} \tag{3.9}$$

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.

Quelques exemples sur les données CONSO

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

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

13. `http://en.wikipedia.org/wiki/Correlation`

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

3.4.3 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 fichiers : les hommes et les femmes (Figure 3.5). 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 3.6 ; 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

$$r_{y,x/z} = \frac{r_{y,x} - r_{y,z}r_{x,z}}{\sqrt{1-r_{y,z}^2} \cdot \sqrt{1-r_{x,z}^2}} \tag{3.10}$$

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'un corrélation partielle, nous adoptons la même démarche que pour la corrélation brute. Les hypothèses à tester sont :

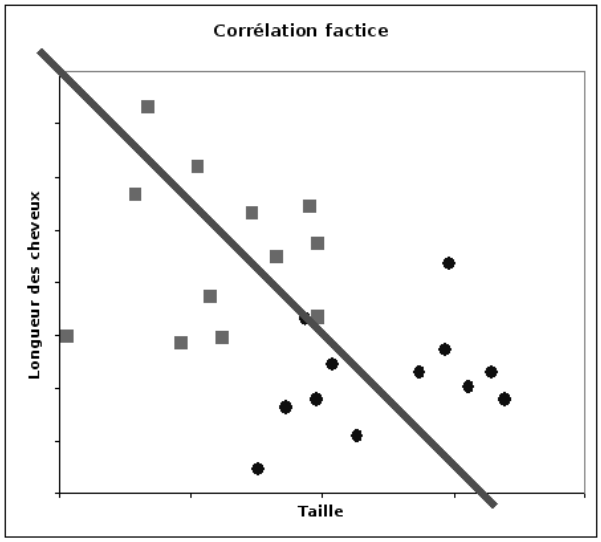


Fig. 3.5. La corrélation est la conséquence de la présence de 2 populations distinctes dans le fichier

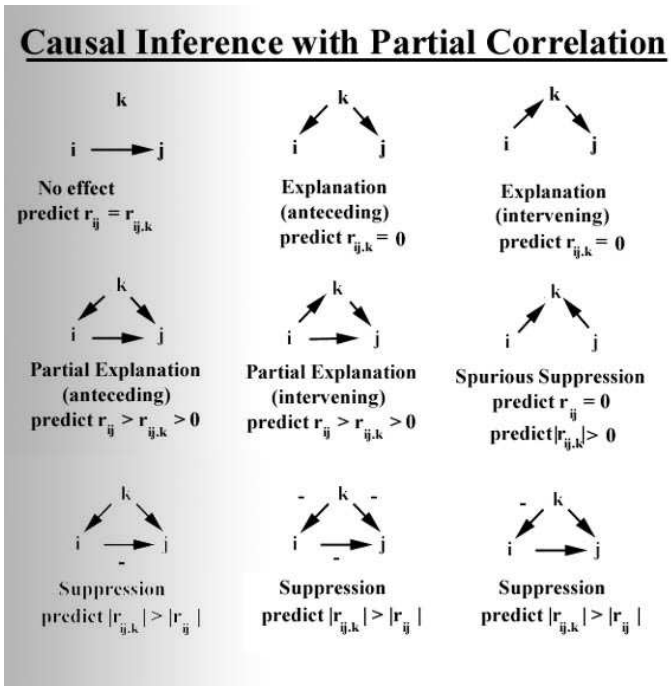


Fig. 3.6. Différentes interactions dans la mesure de la corrélation partielle

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

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

La statistique du test s'écrit :

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

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 bien 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 3.10) en utilisant les corrélations brutes calculées précédemment (Tableau 3.2) :

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

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.

3.4.4 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 $(z_1, z_2, \ldots)$ variables.

Il existe une formule de passage qui permet de généraliser la première expression (Équation 3.10). 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 :

$$H_0 : \rho_{y,x/z_1,\ldots,z_k} = 0$$
$$H_1 : \rho_{y,x/z_1,\ldots,z_k} \neq 0$$

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 :

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

3.4.5 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

Appliquée sur les données CONSO, le modèle choisi comporte les exogènes *poids* et *cylindrée* (Figure 3.7). 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 $y = a_0 + a_1 poids + a_2 cylindree$ 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.

Forward Selection Process

partial corr. F (p-value)	Step 1	Step 2	Step 3
d.f.	25	24	23
r(Y,Xj*/Xj1,Xj2...)	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. 3.7. Sélection de variables fondée sur la corrélation partielle - Données CONSO

4. Au final, le modèle définitif comprend les variables *poids* et *cylindrée*.

3.4.6 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 3.7) 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 $(q-1)$, $r_{y,x_q/x_1,\ldots,x_{q-1}}$, 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) :

$$r^2_{y,x_q/x_1,\ldots,x_{q-1}} = \frac{t^2_{\hat{a}_q}}{t^2_{\hat{a}_q} + (n-q-1)} \tag{3.11}$$

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 faisions 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 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. Même s'ils sont performants, 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.

4

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 largement notre propos, nous ne l'aborderons pas dans ce chapitre.

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

1. définir quel type de codage utiliser ;
2. donner un sens aux résultats et tester 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 technique [1].

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 moyenne à k populations. Pour fixer les idées, travaillons sur un jeu de données.

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.

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/`

Loyer (Euro au m²)	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 suivante : 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

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu$$
$$H_1 : \text{une des moyennes au moins diffère des autres}$$

où μ_j est la moyenne de la variable d'intérêt Y pour la population j.

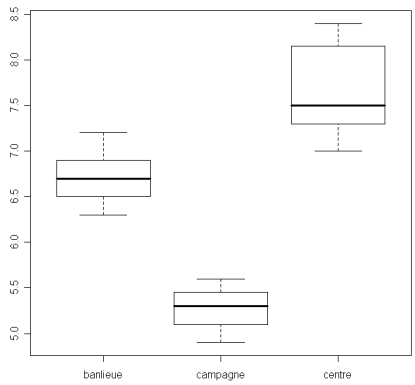


Fig. 4.2. Boîtes à moustaches des loyers selon le lieu d'habitation - Fichier LOYER

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.

Statistique du test

On passe par l'équation de décomposition de la variance pour construire la statistique du test. Elle s'écrit

$$SCT = SCE + SCR$$
$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{i,j} - \bar{y})^2 = \sum_{j=1}^{k} n_j(\bar{y}_j - \bar{y})^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{i,j} - \bar{y}_j)^2$$

où $y_{i,j}$ représente la valeur de Y pour l'individu i du groupe j ; $\bar{y}$ est la moyenne globale de Y, $\bar{y}_j$ est la moyenne conditionnelle, la moyenne de Y dans le groupe j.

Cette décomposition se lit comme suit :
- SCT représente la somme des carrés des écarts totaux, elle indique la variabilité totale de la variable d'intérêt ;
- SCE représente la somme des carrés des écarts inter-groupes c.-à-d. expliqués par l'appartenance aux groupes ;
- SCR représente la somme des carrés des écarts intra-groupes c.-à-d. 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 la valeur de la variable d'intérêt.

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 à $(k-1, n-k)$ degrés de liberté.

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.

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 sur les données LOYER

Nous appliquons ces calculs sur les données LOYER (Figure 4.3), nous procédons selon les étapes suivantes :

1. Recenser les effectifs n_j et les moyennes $\bar{y}_j$ conditionnelles ;
2. Calculer la moyenne globale $\bar{y} = 6.88$;
3. Former $SCT = 15.02$ et $SCE = 5(6.72-6.8)^2 + 3(5.27-6.88)^2 + 7(7.69-6.88)^2 = 12.48$;
4. En déduire $SCR = 15.02 - 12.48 = 2.54$;
5. Calculer la statistique du test $F = \frac{12.48/2}{2.54/12} = 29.446$;
6. Obtenir enfin la p-value à l'aide de la loi de Fisher à $(2, 12)$ degrés de liberté, $p - value < 0.0001$.

Au risque de 5%, l'hypothèse d'égalité des moyennes peut être 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 ne détaillerons pas ces techniques mais nous garderons quand même 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 :

Lieu Habitation	Moyenne	n	n x Ecart moyenne²
banlieue	6.72	5	0.13
campagne	5.27	3	7.81
centre	7.69	7	4.54
Globale	6.88	15	

Tableau ANOVA			
Source	ddl	SC	CM
SCE	2	12.48	6.24
SCR	12	2.54	0.21
SCT	14	15.02	-

F	29.4446
p-value	0.0000

Loyer (Euro au m²)	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

$$y_{i,j} = \mu + \alpha_j + \epsilon_{i,j}$$

où α_j est l'effet du facteur j, $\epsilon_{i,j} \sim \mathcal{N}(0, \sigma)$.

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

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$$

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

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 j de X, on associe une variable Z_j telle que

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

Sur l'exemple LOYER, cela nous emmènerait à produire un nouveau tableau de données (Figure 4.4), et nous définirions naturellement la régression de la manière suivante

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

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*

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

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

Il y a interférence avec la constante de la régression, la matrice $(Z'Z)$ n'est pas inversible car la première colonne de Z est composée de la valeur 1, la somme des 3 colonnes suivantes est égale à 1.

Régression sans constante et lecture des coefficients

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

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

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.	7.69	5.27	6.72
std.dev	0.17	0.27	0.21
R^2	0.83	0.46	#N/A
	19.63	12	#N/A
	12.48	2.54	#N/A

Fig. 4.5. Régression sans constante - Données LOYER

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

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 $\bar{y}_j$ de la manière suivante, selon qu'on utilise ou non la variable recodée

$$\begin{aligned}\bar{y}_j &= \frac{1}{n_j}\sum_{i:z_{i,j}=1} y_i \\ &= \frac{1}{n_j}\sum_{i:x_i=j} y_i\end{aligned}$$

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.***

Vers des solutions plus générales

Malgré son intérêt, cette technique n'est pas généralisable : 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 continues. 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

Partant de l'idée que la dernière modalité k peut être déduite des autres dans le codage disjonctif complet

$$Z_{i,k} = 1 - (Z_{i,1} + Z_{i,2} + \cdots + Z_{i,k-1})$$

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

$$X_i = k \Leftrightarrow Z_{i,1} = Z_{i,2} = \cdots = Z_{i,k-1} = 0$$

Lorsque la variable X prend la modalité k, toutes les indicatrices $Z_1, \ldots, Z_{k-1}$ prennent la valeur zéro. L'équation de régression estimée à l'aide des MCO pour les données LOYER en omettant la variable Z_{centre} devient

$$loyer = \hat{a}_0 + \hat{a}_1 Z_{banlieue} + \hat{a}_2 Z_{campagne} \tag{4.1}$$

Reste à interpréter les coefficients de la régression.

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 $\hat{y}_{i^*} = \hat{a}_0$, le loyer prédit est $\hat{a}_0$.
- 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 $\hat{y}_{i^*} = \hat{a}_0 + \hat{a}_1$.

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 peuvent s'interpréter comme une moyenne conditionnelle de l'endogène (la constante) ou des écarts à la moyenne conditionnelle de référence (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.
3. Du coup, réaliser des tests de significativité des coefficients a_j $(j \geq 1)$

$$H_0 : a_j = 0$$
$$H_1 : a_j \neq 0$$

   s'apparente[2] à un test de comparaison de la moyenne conditionnelle $\bar{y}_j$ avec la moyenne de référence $\bar{y}_k$.
4. Et le test de significativité globale de la régression correspond **exactement** à une ANOVA à 1 facteur.

4.3.3 Application aux 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 toujours les sorties du tableur EXCEL :

2. "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 différences dans les degrés de liberté.

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.	-2.42	-0.97	7.69
	0.32	0.27	0.17
	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. 4.6. Régression avec données codées "cornered effect" - Données LOYER

Indicateur	Valeur
F	29.44
$ddl1 = p$	2
$ddl2 = n - p - 1$	12
p-value	< 0.0001

Ces résultats correspondent exactement à ceux de l'ANOVA à 1 facteur (Figure 4.3). Les deux approches sont totalement équivalentes.

4.4 Codage "Centered effect" de l'exogène qualitative

4.4.1 Principe

Nous comprenons maintenant 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 $(j = 1, \ldots, k-1)$:

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

La modalité k (*centre-ville*) sert toujours de référence. A la différence que cette fois-ci, nous signalons explicitement sa présence pour l'individu i en attribuant la valeur -1 à toutes les variables recodées Z_j. Nous estimons avec les MCO les coefficients de la régression :

$$loyer = \hat{b}_0 + \hat{b}_1 Z_{banlieue} + \hat{b}_2 Z_{campagne} \tag{4.2}$$

Comment lire ces coefficients ?

4.4.2 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".
- 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 lorsque les effectifs dans les groupes sont 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.

4.4.3 Application aux données LOYER

Nous effectuons la régression sur les données LOYER recodées (Figure 4.7). Nous obtenons les coefficients $\hat{b}$ (Équation 4.2) et nous en déduisons les moyennes conditionnelles :

- $\hat{b}_2 = -1.29 \Rightarrow \bar{y}_{campagne} = \hat{b}_2 + \hat{b}_0 = -1.29 + 6.259 = 5.27$;
- $\hat{b}_1 = 0.16 \Rightarrow \bar{y}_{banlieue} = \hat{b}_1 + \hat{b}_0 = 0.16 + 5.56 = 6.72$;
- $\bar{y}_{centre} = \hat{b}_0 - (\hat{b}_1 + \hat{b}_2) = 6.56 - (0.16 + (-1.29)) = 7.69$;
- 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.

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. 4.7. Régression avec données codées "centered effect" - Données LOYER

4.5 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 $\{1, 2, 3, \ldots\}$ des variables qualitatives.

4.5.1 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

$$Z_i = \begin{cases} 1 & \text{si } X_i = \text{"}banlieue\text{"} \\ 2 & \text{si } X_i = \text{"}campagne\text{"} \\ 3 & \text{si } X_i = \text{"}centre\text{"} \end{cases}$$

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 2 types de codages décrits plus haut (*cornered et centered effect*) sont plus adaptés, à charge au statisticien de choisir celui qui paraît le mieux répondre au problème traité.

4.5.2 Codage numérique d'une variable discrète ordinale

On parle de variable discrète ordinale lorsque (1) la variable prend un nombre fini de modalités (de valeurs) ; (2) il y a une relation d'ordre entre les modalités. L'exemple typique est la *satisfaction* d'un client par rapport à un service, il peut être mécontent, satisfait, très satisfait, etc.

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

$$Z_i = \begin{cases} 1 & \text{si } X_i = \text{"}centre\text{"} \\ 2 & \text{si } X_i = \text{"}banlieue\text{"} \\ 3 & \text{si } X_i = \text{"}campagne\text{"} \end{cases}$$

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".

En réalité, nous ne savons rien de tout cela. Peut-être est-ce vrai, peut être est-ce erroné. Quoi qu'il en soit, le pire serait de lancer les calculs sans être conscient de ce qu'on manipule.

Remarquons que le codage disjonctif (ou dérivés) peut fonctionner pour les variables ordinales. Dans ce cas, nous perdons irrémédiablement l'information sur le caractère ordinal des données. La régression n'en tiendra pas compte pour produire les coefficients.

4.6 Conclusion

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. Ce codage est primordial car il détermine les informations que nous retenons des données initiales, il détermine également l'interprétation des coefficients fournis par la régression.

Dans ce support, nous avons exclusivement décrit le modèle avec une variable exogène qualitative. Rien ne nous empêche de généraliser la démarche ci-dessus à plusieurs exogènes qualitatives. Nous pouvons étudier leurs effets indépendamment, mais surtout, et c'est un des intérêts de l'approche, nous pouvons analyser finement les interactions (ex. étudier les effets conjoints du cannabis et de l'alcool sur les temps de réaction au volant). La technique est très riche et ses applications sont multiples.

5

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 abcisse, 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.

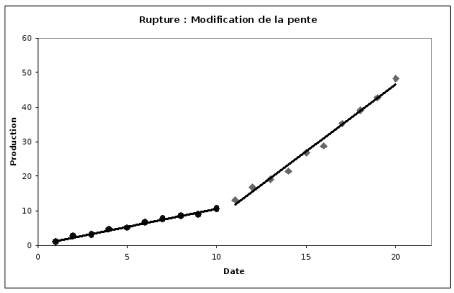


Fig. 5.1. Rupture de structure : modification de la pente à la date $t = 11$

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).

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. A partir d'un certain

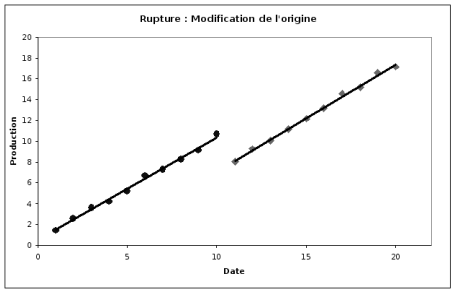


Fig. 5.2. Rupture de structure : modification de l'origine à la date $t = 11$

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.

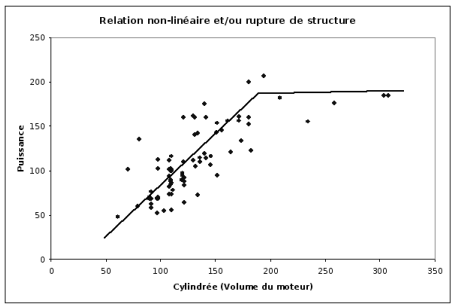


Fig. 5.3. Relation non-linéaire ou linéaire par morceaux ?

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. A 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.

$$y_i = a_0 + a_1 x_{i,1} + \cdots + a_p x_{i,p} + \epsilon_i, i = 1, \ldots, n \tag{5.1}$$

(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).

$$\begin{aligned} y_i &= a_{0,1} + a_{1,1} x_{i,1} + \cdots + a_{p,1} x_{i,p} + \epsilon_i, i = 1, \ldots, n_1 \\ y_i &= a_{0,2} + a_{1,2} x_{i,1} + \cdots + a_{p,2} x_{i,p} + \epsilon_i, i = n_1 + 1, \ldots, n \ (n_2 \text{ obs.}) \end{aligned}$$

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

   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".

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.

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 Chow [2] 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 avions bien $(p + 1)$ contraintes linéaires dans l'hypothèse nulle de notre test de comparaison des coefficients.

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.

2. Gregory C. Chow (1960). *Tests of Equality Between Sets of Coefficients in Two Linear Regressions.* in Econometrica 28(3) : 591-605.

5.1.2 Un exemple

Nous reprenons un exemple décrit dans Johnston (pages 134 à 138). Nous voulons effectuer une régression linéaire simple $Y = aX + b + \epsilon$. Les données (fichier CHOW) peuvent être subdivisées en 2 sous-parties (sous-périodes) correspondant à une variable supplémentaire[3] (Figure 5.4).

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

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) :

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

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 $ddl_1 = 3$;

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.

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

$$F = \frac{[6.56 - (0.69 + 2.47)]/2}{(0.69 + 2.47)/11} = 5.91$$

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 :

$$d_{i,1} = \begin{cases} 1 \ , \ i = 1, \ldots, n_1 \\ 0 \ , \ i = n_1 + 1, \ldots n \end{cases}$$
$$d_{i,2} = \begin{cases} 0 \ , \ i = 1, \ldots, n_1 \\ 1 \ , \ i = n_1 + 1, \ldots, n \end{cases}$$

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.

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.

$$y_i = a_{0,1}d_{i,1} + a_{0,2}d_{i,2} + a_1x_{i,1} + \cdots + a_px_{i,p} + \epsilon_i \tag{5.2}$$

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-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 :

$$z_{i,1} = \begin{cases} x_{i,1} & , i = 1, \ldots, n_1 \\ 0 & , i = n_1 + 1, \ldots n \end{cases}$$

$$z_{i,2} = \begin{cases} 0 & , i = 1, \ldots, n_1 \\ x_{i,1} & , i = n_1 + 1, \ldots, n \end{cases}$$

Nous pourrions alors être tenté de proposer comme formulation non-contrainte de la régression :

$$y_i = a_0 + a_{1,1} z_{i,1} + a_{1,2} z_{i,2} + \cdots + a_p x_{i,p} + \epsilon_i \tag{5.3}$$

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).

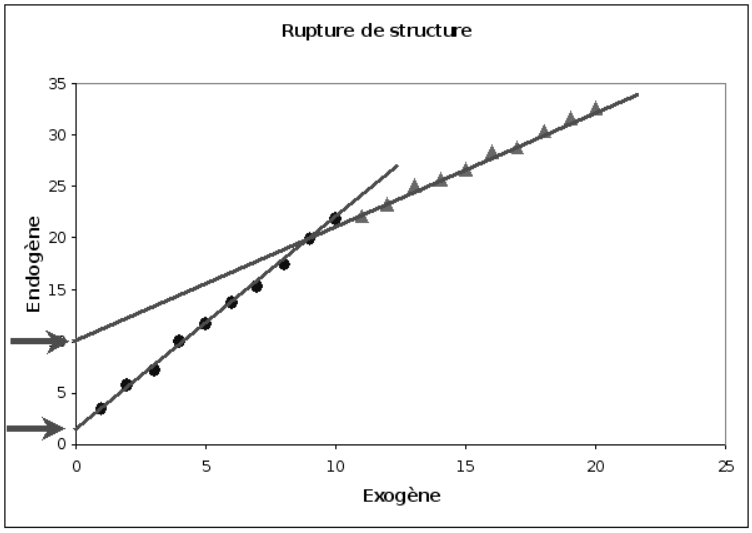


Fig. 5.7. Un changement de pente entraîne automatiquement une modification de l'origine

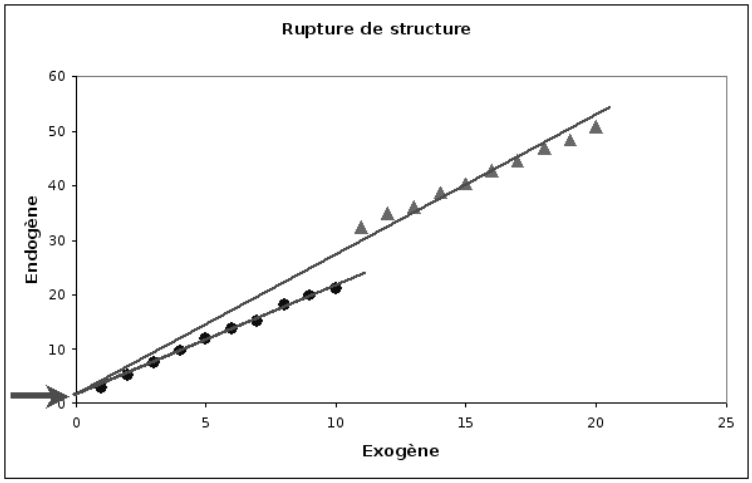


Fig. 5.8. En imposant la même origine aux deux régressions, on fausse l'appréciation des pentes

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.

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 :

$$y_i = a_{0,1}d_{i,1} + a_{0,2}d_{i,2} + a_{1,1}z_{i,1} + a_{1,2}z_{i,2} + \cdots + a_p x_{i,p} + \epsilon_i \tag{5.4}$$

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 avions 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.

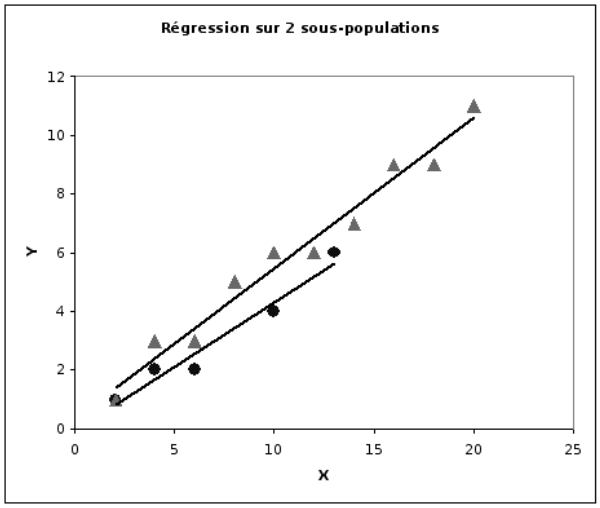


Fig. 5.10. Nuage de points (X,Y) et droites de régression pour les deux sous-populations de notre fichier exemple (Figure 5.4)

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 évènements (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 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.

A

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

B

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 la page web `http://eric.univ-lyon2.fr/~ricco/cours/supports_data_mining.html`.

Littérature

Ouvrages

1. Bourbonnais, R., *Econométrie. Manuel et exercices corrigés*, Dunod, 2è édition, 1998.
2. Dodge, Y, Rousson, V., *Analyse de régression appliquée*, Dunod, 2è édition, 2004.
3. Giraud, R., Chaix, N., *Econométrie*, Presses Universitaires de France (PUF), 1989.
4. Johnston, J., DiNardo, J., *Méthodes Econométriques*, Economica, 4è édition, 1999.
5. Labrousse, C., *Introduction à l'économétrie. Maîtrise d'économétrie*, Dunod, 1983.
6. Saporta, G., *Probabilités, Analyse des données et Statistique*, Technip, 2ème édition, 2006.
7. Tenenhaus, M., *Méthodes Statistiques en Gestion*, Dunod, 1996.

Supports en ligne

8. 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`
9. , 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/`
10. Faraway, J., *Practical Regression and ANOVA using R*, July 2002, `http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf`
11. 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*.
12. Haurie, A., *Modèle de régression linéaire*, sur `http://ecolu-info.unige.ch/~haurie/mba05/`
13. *Régression Linéaire Multiple*, sur `http://fr.wikipedia.org/wiki/Régression_linéaire_multiple`
14. *Xycoon Online Econometrics Textbook*, sur `http://www.xycoon.com/index.htm#econ`