Ricco Rakotomalala

Econométrie

La régression linéaire simple et multiple

Version 1.1

Université Lumière Lyon 2

Avant-propos

Lorsqu'on m'a demandé si je voulais assurer le cours d'économétrie en Licence L3-IDS (`http://dis.univ-lyon2.fr/`), j'ai ressenti une grande joie mais aussi une certaine inquiétude.

D'une part une grande joie car c'est à travers l'économétrie que je suis venu au traitement statistique des données. Lorsque j'ai vu un nuage de point avec une forme plus ou moins affirmée, et que j'ai compris qu'on pouvait en déduire une liaison fonctionnelle représentée par une courbe passant au milieu de ces points, je me suis dit qu'il y avait là quelque chose de magique. Je trouvais formidable l'idée que des données recèlent une vérité que l'on est capable de reconstituer ou bien, inversement, que l'on s'imagine une certaine forme de vérité que l'on peut confirmer ou infirmer à travers des données observées, totalement objectives. Par la suite, de fil en aiguille, j'ai découvert une très vaste littérature autour de ces principes. Les appellations sont différentes selon les cultures : on parle d'analyse de données, de data mining, etc. Mais qu'importe finalement, pour ma part je sais très bien ce que je fais. Et ce qui était initialement une sorte de loisir (*ah, le temps passé sur mon Thomson M05 à programmer des petites procédures statistiques...*[1]) est devenu mon métier.

D'autre part, je ressentais quand même un certaine inquiétude car c'était la première fois que je passais de l'autre côté de la barrière dans ce domaine. A priori, je connais bien la régression. Je l'ai beaucoup étudiée jusqu'en DEA (l'équivalent d'un Master 2 Recherche de nos jours). Trouver mes repères ne devait pas poser de problèmes particuliers. Mais comme la grande majorité des étudiants (j'imagine), j'avais surtout étudié dans l'optique de restituer, pour préparer les examens quoi (un peu pour la programmer aussi, d'où le logiciel REGRESS qui a près de 20 ans aujourd'hui, et qui est toujours en ligne – `http://eric.univ-lyon2.fr/~ricco/regress.html` – même si, honnêtement, il doit y avoir très peu d'utilisateurs je pense). Ici, l'affaire est autrement plus corsée. Il s'agit d'expliquer à d'autres personnes. La différence est énorme. C'est donc non sans inquiétude que j'ai sorti mes anciennes notes de cours (entres autres les fameux polycopiés de Patrick Sylvestre-Baron de la Faculté de Sciences Économiques de l'Université Lyon 2) et que j'ai fait l'acquisition de plusieurs ouvrages qui allaient me servir de base de préparation.

1. La courbe bleue tracée à une allure d'escargot au milieu des points verts (on n'avait droit qu'à 16 couleurs en mode graphique), c'était jouissif !

Je me suis rendu compte que la régression linéaire est toujours aussi passionnante. Plus même, les années post DEA passées à étudier les techniques de Data Mining, en particulier l'apprentissage supervisé, m'ont apporté un recul que je n'avais pas (quelques années en plus, il faut bien que ça serve à quelque chose aussi). Tout de suite, j'ai pu raccrocher ce que je lisais à ce que je savais par ailleurs. Quand même, ils avaient vraiment découvert beaucoup de choses ces économètres. Par exemple, pouvoir calculer une erreur de prédiction en *leave-one-out* sans avoir à construire explicitement le modèle sur les $(n-1)$ observations grâce au concept de *levier* est tout bonnement fabuleux. En retour, j'ai mieux compris certains aspects de l'apprentissage supervisé en étudiant les techniques économétriques. Bref c'est tout bonus. Ce travail m'a d'ailleurs permis par la suite de monter mon cours de régression logistique, et de rédiger le support associé [14].

Reste une question. A quoi peut bien servir un polycopié supplémentaire sur la régression linéaire simple et multiple. En effet, ils sont légions sur internet (tapez "économétrie" dans Google pour voir). S'il s'agit de reproduire ce qui est déjà (très bien) écrit par ailleurs, on ne voit pas vraiment où est l'intérêt.

La première raison est mon cours de licence. Au fil des années, le nombre d'heures dont je dispose pour le faire a été réduit comme une peau de chagrin. Ce qui ne manque pas de me chagriner d'ailleurs (ok, ok, elle est facile celle-là). Comme je ne souhaite absolument pas diminuer le nombre des séances TD sur machine, je suis obligé de rogner sur les CM. De fait, il ne m'est plus possible de détailler certaines démonstrations au tableau comme je pouvais le faire naguère. De même, en utilisant de plus en plus des slides pour le cours, je fais des ellipses à de nombreux endroits. Je me suis dit que la seule manière de donner des repères identiques à tous les étudiants est de leur épargner la prise de notes en fournissant le cours rédigé. En cela, mon cours d'économétrie se rapproche de plus en plus de mon cours de Data Mining où je parle de beaucoup de choses en très peu de temps en me focalisation sur les aspects opérationnels (en cours tout du moins), mais en donnant accès aux étudiants à une abondante documentation gratuite.

La seconde raison est que cela me permet tout simplement de présenter les choses à ma manière, en donnant la part belle aux exemples traités sur tableur [2]. Ce qui est une de mes principales marques de fabrique. Parfois, je ferais le parallèle avec les résultats fournis par les logiciels de statistique, en privilégiant toujours les outils libres (Tanagra, Regress et R principalement) [3]. Ainsi, le lecteur pourra refaire tous les calculs décrits dans ce document. A cet effet, les fichiers de données qui ont servi à sa préparation sont également accessibles en ligne. Ils sont énumérés en annexes.

Bien évidemment, selon l'expression consacrée, ce support n'engage que son auteur. Toutes suggestions ou commentaires qui peuvent en améliorer le contenu sont bienvenus.

2. Excel, mais sous Open Office les traitements sont identiques.

3. Parfois je m'autoriserai des digressions sur des outils un peu moins gratuits, mais ayant pignon sur rue (SAS, SPAD, SPSS et STATISTICA pour ne pas les nommer). Parce que certains d'entre vous les rencontreront en entreprise. Je ne suis pas sectaire non plus.

Table des matières

Partie I

Régression Linéaire Simple

1

Modèle de régression linéaire simple

1.1 Modèle et hypothèses

1.1.1 Régression linéaire simple

Nous cherchons à mettre en avant une relation de dépendance entre les variables Y et X. Y est celle que l'on cherche à expliquer (à prédire), on parle de variable **endogène** (dépendante) ; X est la variable explicative (prédictive), on parle de variable **exogène** (indépendante).

Le modèle de régression linéaire simple s'écrit :

$$y_i = a \times x_i + b + \varepsilon_i \tag{1.1}$$

a et b sont les paramètres (les coefficients) du modèle. Dans le cas spécifique de la régression simple, a est la *pente*, b est la *constante*.

Nous disposons d'un échantillon de n observations i.i.d (indépendantes et identiquement distribuées) pour estimer ces paramètres.

Le terme aléatoire ε, que l'on appelle l'erreur du modèle, tient un rôle très important dans la régression. Il permet de résumer toute l'information qui n'est pas prise en compte dans la relation linéaire que l'on cherche à établir entre Y et X c.-à-d. les problèmes de spécifications, l'approximation par la linéarité, résumer le rôle des variables explicatives absentes, etc. Comme nous le verrons plus bas, les propriétés des estimateurs reposent en grande partie sur les hypothèses que nous formulerons à propos de ε. En pratique, après avoir estimé les paramètres de la régression, les premières vérifications portent sur l'erreur calculée sur les données (on parle de "résidus") lors de la modélisation [13] (Chapitre 1).

Exemple - Rendement de maïs et quantité d'engrais. Dans cet exemple tiré de l'ouvrage de Bourbonnais (page 12), nous disposons de $n = 10$ observations (Figure 1.1) [1]. On cherche à expliquer Y le rendement en maïs (en quintal) de parcelles de terrain, à partir de X la quantité d'engrais (en kg) que l'on y a épandu. L'objectif est de modéliser le lien à travers une relation linéaire. Bien évidemment, si l'on ne

1. `regression_simple_rendements_agricoles.xlsx` - "data"

met pas d'engrais du tout, il sera quand même possible d'obtenir du maïs, c'est le sens de la constante b de la régression. Sa valeur devrait être positive. Ensuite, plus on mettra de l'engrais, meilleur sera le rendement. On suppute que cette relation est linéaire, d'où l'expression $a \times x$, on imagine à l'avance que a devrait être positif.

i	Y	X
1	16	20
2	18	24
3	23	28
4	24	22
5	28	32
6	29	28
7	26	32
8	31	36
9	32	41
10	34	41

Fig. 1.1. Tableau de données "Rendements Agricoles" - Bourbonnais, page 12

Le graphique nuage de points associant X et Y semble confirmer cette première analyse (Figure 1.2) [2]. Dans le cas contraire où les coefficients estimés contredisent les valeurs attendues (b ou/et a sont négatifs), cela voudrait dire que nous avons une perception faussée du problème, ou bien que les données utilisées ne sont pas représentatives du phénomène que l'on cherche à mettre en exergue, ou bien... On entre alors dans une démarche itérative qui peut durer un moment avant d'obtenir le modèle définitif [3]. C'est le *processus de modélisation*.

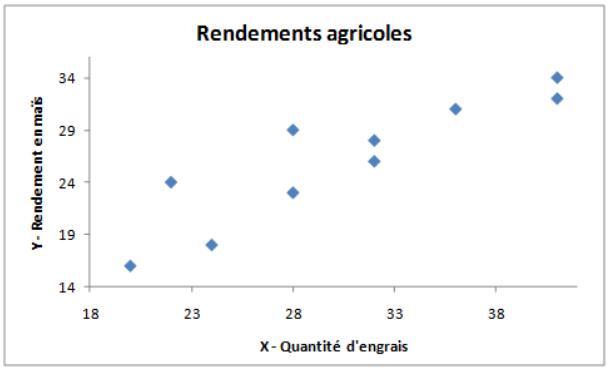


Fig. 1.2. Graphique nuage de points "Rendements Agricoles" - Bourbonnais, page 12

2. `regression_simple_rendements_agricoles.xlsx` - "data"
3. Voir l'excellent site du NIST – `http://www.itl.nist.gov/div898/handbook/pmd/pmd.htm` – au sujet du processus de modélisation : les terminologies utilisées, les principales étapes, la lecture des résultats. Avec des études de cas complètes.

1.1.2 Hypothèses

Ces hypothèses pèsent sur les propriétés des estimateurs (biais, convergence) et l'inférence statistique (distribution des coefficients estimés).

H1 – Hypothèses sur Y et X. X et Y sont des grandeurs numériques mesurées sans erreur. X est une donnée exogène dans le modèle. Elle est supposée non aléatoire. Y est aléatoire par l'intermédiaire de ε c.-à-d. la seule erreur que l'on a sur Y provient des insuffisances de X à expliquer ses valeurs dans le modèle.

H2 – Hypothèses sur le terme aléatoire ε. Les ε_i sont i.i.d (indépendants et identiquement distribués).

H2.a – $E(\varepsilon_i) = 0$, en moyenne les erreurs s'annulent c.-à-d. le modèle est bien spécifié.

H2.b – $V(\varepsilon_i) = \sigma_\varepsilon^2$, la variance de l'erreur est constante et ne dépend pas de l'observation. C'est l'hypothèse d'homoscédasticité.

H2.c – En particulier, l'erreur est indépendante de la variable exogène c.-à-d. $COV(x_i, \varepsilon_i) = 0$

H2.d – Indépendance des erreurs. Les erreurs relatives à 2 observations sont indépendantes c.-à-d. $COV(\varepsilon_i, \varepsilon_j) = 0$. On parle de "non auto-corrélation des erreurs".

Remarque : Cette hypothèse est toujours respectée pour les coupes transversales. En effet l'échantillon est censé construit de manière aléatoire et les observations i.i.d. Nous pouvons donc intervertir aléatoirement les lignes sans porter atteinte à l'intégrité des données. En revanche, la question se pose pour les données temporelles. Il y a une contrainte qui s'impose à nous (contrainte temporelle - les données sont ordonnées) dans le recueil des données.

H2.e – $\varepsilon_i \equiv \mathcal{N}(0, \sigma_\varepsilon)$. L'hypothèse de normalité des erreurs est un élément clé pour l'inférence statistique.

1.2 Principe de l'ajustement des moindres carrés

1.2.1 Estimateur des moindres carrés ordinaires (MCO)

Notre objectif est de déterminer les valeurs de a et b en utilisant les informations apportées par l'échantillon. Nous voulons que l'estimation soit la meilleure possible c.-à-d. la droite de régression doit approcher *au mieux* le nuage de points.

Si graphiquement, la solution semble intuitive. Il nous faut un critère numérique qui réponde à cette spécification pour réaliser les calculs sur un échantillon de données.

Le critère des **moindres carrés** consiste à minimiser la somme des carrés des écarts (des erreurs) entre les vraies valeurs de Y et les valeurs prédites avec le modèle de prédiction (Figure 1.3). L'estimateur des moindres carrées ordinaires (MCO) des paramètres a et b doit donc répondre à la minimisation de

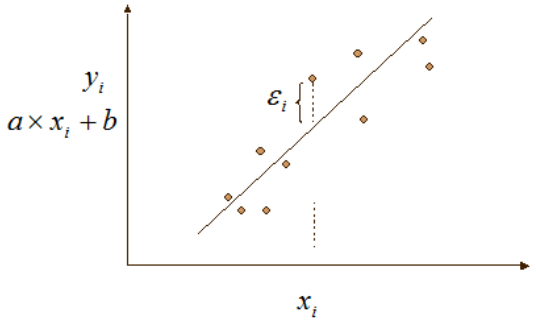


Fig. 1.3. Comptabilisation de l'erreur : écart entre Y observé et Y prédit par le modèle linéaire

$$\begin{aligned} S &= \sum_{i=1}^{n} \varepsilon_i^2 \\ &= \sum_{i=1}^{n} [y_i - (ax_i + b)]^2 \\ &= \sum_{i=1}^{n} [y_i - ax_i - b]^2 \end{aligned}$$

Pour déterminer les valeurs de a et b, les conditions suivantes sont nécessaires :

$$\begin{cases} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \end{cases}$$

En appliquant ces dérivées partielles, nous obtenons les **équations normales** (Giraud et Chaix, page 25 ; Bourbonnais, page 21 ; Johnston et DiNardo, page 22) :

$$\begin{cases} \sum_i x_i y_i - a \sum_i x_i^2 - b \sum_i x_i = 0 \\ \bar{y} - a\bar{x} - b = 0 \end{cases} \tag{1.2}$$

Que l'on retrouve également sous la forme suivante dans la littérature (Tenenhaus, page 70).

$$\begin{cases} \sum_i x_i \varepsilon_i = 0 \\ \sum_i \varepsilon_i = 0 \end{cases} \tag{1.3}$$

En appelant $\hat{a}$ et $\hat{b}$ les solutions de ces équations normales, nous obtenons les **estimateurs des moindres carrés** :

$$\hat{a} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{1.4}$$

$$\hat{b} = \bar{y} - \hat{a}\bar{x} \tag{1.5}$$

Détail des calculs

Quelques pistes pour obtenir ces résultats. Voyons tout d'abord la dérivée partielle $\frac{\partial S}{\partial b}$

$$\frac{\partial S}{\partial b} = 0$$
$$\sum_i 2(-1)(y_i - ax_i - b) = 0$$
$$-2[\sum_i y_i - a\sum_i x_i - n \times b] = 0$$

En multipliant le tout par $-\frac{2}{n}$, nous avons :

$$b = \bar{y} - a\bar{x}$$

Occupons-nous maintenant de $\frac{\partial S}{\partial a}$

$$\frac{\partial S}{\partial a} = \sum_i 2(-x_i)(y_i - ax_i - b) = 0$$

En introduisant le résultat relatif à b ci-dessus, nous obtenons :

$$a = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

1.2.2 Calculs pour les données "Rendements agricoles"

Revenons à notre exemple des "Rendements agricoles" (Figure 1.1). Nous montons la feuille Excel permettant de réaliser les calculs (Figure 1.4) [4].

i	Y	X	(Y-YB)	(X-XB)	(Y-YB)*(X-XB)	(X-XB)²
1	16	20	-10.1	-10.4	105.04	108.16
2	18	24	-8.1	-6.4	51.84	40.96
3	23	28	-3.1	-2.4	7.44	5.76
4	24	22	-2.1	-8.4	17.64	70.56
5	28	32	1.9	1.6	3.04	2.56
6	29	28	2.9	-2.4	-6.96	5.76
7	26	32	-0.1	1.6	-0.16	2.56
8	31	36	4.9	5.6	27.44	31.36
9	32	41	5.9	10.6	62.54	112.36
10	34	41	7.9	10.6	83.74	112.36
Moyenne	26.1	30.4		Somme	351.6	492.4
				a^	0.7141	
				b^	4.3928	

Fig. 1.4. Estimation des coefficients "Rendements agricoles" - Feuille de calcul Excel

Voici les principales étapes :

- Nous calculons les moyennes des variables, $\bar{y} = 26.1$ et $\bar{x} = 30.4$.
- Nous formons alors les valeurs de $(y_i - \bar{y})$, $(x_i - \bar{x})$, $(y_i - \bar{y}) \times (x_i - \bar{x})$ et $(x_i - \bar{x})^2$.
- Nous réalisons les sommes $\sum_i (y_i - \bar{y}) \times (x_i - \bar{x}) = 351.6$ et $\sum_i (x_i - \bar{x})^2 = 492.4$.

4. `regression_simple_rendements_agricoles.xlsx` - "reg.simple.1"

– Nous déduisons enfin les estimations :

$$\hat{a} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{351.6}{492.4} = 0.7141$$
$$\hat{b} = \bar{y} - \hat{a}\bar{x} = 26.1 - 0.7141 \times 30.4 = 4.3928$$

La droite de régression peut être représentée dans le graphique nuage de points. Nous avons utilisé l'outil "Courbe de tendance" d'Excel (Figure 1.5) [5].

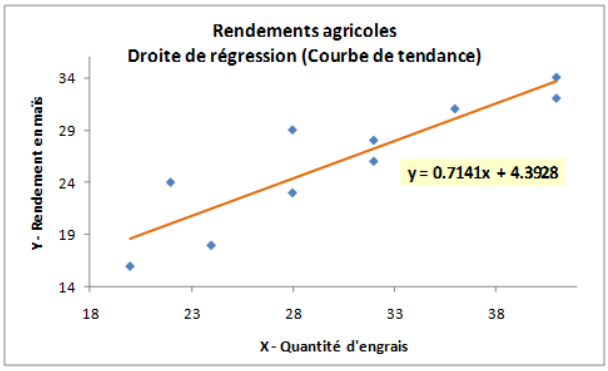


Fig. 1.5. Droite de régression - "Rendements agricoles"

Nous constatons que la droite passe peu ou prou au milieu du nuage de points. Mais nous ne saurions pas dire dans quelle mesure notre modélisation est suffisamment *intéressante*. La simple évaluation visuelle ne suffit pas. La seule manière d'obtenir une réponse rigoureuse est de produire un critère quantitatif que l'on saura interpréter. Nous nous pencherons sur cette question dans la section consacrée à l'évaluation du modèle (section 1.3).

1.2.3 Quelques remarques

Autre écriture de l'estimateur de la pente. Il y a une relation directe entre l'estimateur de la pente et le coefficient de corrélation linéaire de Pearson r_{yx}.

$$\begin{aligned}\hat{a} &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \frac{\widehat{COV}(Y,X)}{\hat{\sigma}_X^2} \\ &= r_{yx} \times \frac{\hat{\sigma}_Y}{\hat{\sigma}_X}\end{aligned}$$

De fait, nous le verrons dans la partie inférentielle, tester la significativité de la pente revient à tester la significativité de la corrélation entre Y et X.

5. `regression_simple_rendements_agricoles.xlsx` - "reg.simple.1"

Erreur et résidu. ε est l'erreur inconnue introduite dans la spécification du modèle. Nous avons alors estimé les paramètres $\hat{a}$ et $\hat{b}$ à partir de l'échantillon et nous appuyant sur le principe des moindres carrés. Nous pouvons obtenir la valeur prédite de l'endogène Y pour l'individu i avec

$$\begin{aligned}\hat{y}_i &= \hat{y}(x_i) \\ &= \hat{a} \times x_i + \hat{b}\end{aligned}$$

On peut en déduire l'erreur observée, appelée "résidu" de la régression

$$\hat{\varepsilon}_i = y_i - \hat{y}_i \tag{1.6}$$

La distinction "erreur vs. résidu" est importante car, comme nous le verrons par la suite, les expressions de leurs variances ne sont pas les mêmes.

Toujours concernant le résidu, notons une information importante :

$$\sum_i \hat{\varepsilon}_i = 0 \tag{1.7}$$

La somme (et donc la moyenne) des résidus est nulle *dans une régression avec constante*. En effet :

$$\begin{aligned}\sum_i \hat{\varepsilon}_i &= \sum_i [y_i - (\hat{a}x_i + \hat{b})] \\ &= n\bar{y} - n\hat{a}\bar{x} - n\hat{b} \\ &= n\bar{y} - n\hat{a}\bar{x} - n \times (\bar{y} - \hat{a}\bar{x}) \\ &= 0\end{aligned}$$

Centre de gravité du nuage de points. La droite de régression *avec constante* passe forcément par le centre de gravité du nuage de points. Pour le vérifier simplement, réalisons la projection pour le point $\bar{x}$:

$$\begin{aligned}\hat{y}(\bar{x}) &= \hat{a}\bar{x} + \hat{b} \\ &= \hat{a}\bar{x} + (\bar{y} - \hat{a}\bar{x}) \\ &= \bar{y}\end{aligned}$$

Dans notre exemple des "Rendements agricoles", nous constatons effectivement que la droite passe le point $G(x, y)$ de coordonnées $(\bar{x} = 30.4, \bar{y} = 26.1)$ (Figure 1.6).

1.3 Décomposition de la variance et coefficient de détermination

1.3.1 Décomposition de la variance - Équation d'analyse de variance

L'objectif est de construire des estimateurs qui minimisent la somme des carrés des résidus

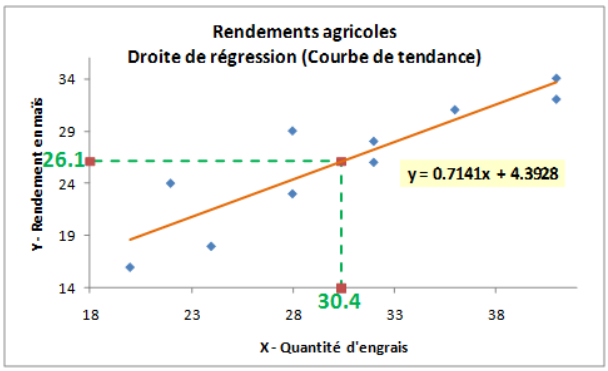


Fig. 1.6. La droite de régression passe par le barycentre - "Rendements agricoles"

$$\begin{aligned} SCR &= \sum_i \hat{\varepsilon}_i^2 \\ &= \sum_i (y_i - \hat{y}_i)^2 \end{aligned}$$

Lorsque la prédiction est parfaite, tout naturellement $SCR = 0$. Mais dans d'autre cas, qu'est-ce qu'une bonne régression ? A partir de quelle valeur de SCR peut-on dire que la régression est mauvaise ?

Pour répondre à cette question, il faut pouvoir comparer la SCR avec une valeur de référence. Pour cela, nous allons décomposer la variance de Y.

On appelle *somme des carrés totaux* (SCT) la quantité suivante :

$$\begin{aligned} SCT &= \sum_i (y_i - \bar{y})^2 \\ &= \sum_i (y_i - \hat{y}_i + \hat{y}_i + \bar{y})^2 \\ &= \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2 + 2\sum_i (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) \end{aligned}$$

Dans la régression avec constante, et uniquement dans ce cas, on montre que

$$2\sum_i (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0$$

En s'appuyant sur deux éléments :

$$\begin{aligned} \bar{\hat{y}} &= \frac{1}{n}\sum_i (\hat{a}x_i + \hat{b}) \\ &= \frac{1}{n}[\hat{a}\sum_i x_i + n \times \hat{b}] \\ &= \hat{a}\bar{y} + \hat{b} \\ &= \bar{y} \end{aligned}$$

et

$$\frac{\partial S}{\partial a} = \sum_i 2(-x_i)(y_i - ax_i - b) = 0$$

On obtient dès lors l'équation d'analyse de variance :

$$SCT = SCE + SCR \tag{1.8}$$

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2 \tag{1.9}$$

Comment interpréter ces quantités ?

- **SCT** est la somme des carrés totaux. Elle indique la variabilité totale de Y c.-à-d. l'information disponible dans les données.
- **SCE** est la somme des carrés expliqués. Elle indique la variabilité expliquée par le modèle c.-à-d. la variation de Y expliquée par X.
- **SCR** est somme des carrés résiduels. Elle indique la variabilité non-expliquée (résiduelle) par le modèle c.-à-d. l'écart entre les valeurs observées de Y et celles prédites par le modèle.

Deux situations extrêmes peuvent survenir :

- Dans le meilleur des cas, $SCR = 0$ et donc $SCT = SCE$: les variations de Y sont complètement expliquées par celles de X. On a un modèle parfait, la droite de régression passe exactement par tous les points du nuage($\hat{y}_i = y_i$).
- Dans le pire des cas, $SCE = 0$: X n'apporte aucune information sur Y. Ainsi, $\hat{y}_i = \bar{y}$, la meilleure prédiction de Y est sa propre moyenne.

A partir de ces informations, nous pouvons produire une première version du **tableau d'analyse de variance** (Tableau 1.1). La version complète nous permettra de mener le test de significativité globale de la régression comme nous le verrons plus loin (section 3.1).

Source de variation	Somme des carrés
Expliquée	$SCE = \sum_i (\hat{y}_i - \bar{y})^2$
Résiduelle	$SCR = \sum_i (y_i - \hat{y}_i)^2$
Totale	$SCT = \sum_i (y_i - \bar{y})^2$

Tableau 1.1. Tableau simplifié d'analyse de variance

1.3.2 Coefficient de détermination

Il est possible de déduire un indicateur synthétique à partir de l'équation d'analyse de variance. C'est le **coefficient de détermination** R^2.

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} \tag{1.10}$$

Il indique la proportion de variance de Y expliquée par le modèle.

- Plus il sera proche de la valeur 1, meilleur sera le modèle, la connaissance des valeurs de X permet de deviner avec précision celle de Y.
- Lorsque R^2 est proche de 0, cela veut dire que X n'apporte pas d'informations utiles (intéressantes) sur Y, la connaissance des valeurs de X ne nous dit rien sur celles de Y.

Remarque 1 (Une autre lecture du coefficient de détermination.). Il existe une lecture moins usuelle, mais non moins intéressante, du coefficient de détermination.

On définit le modèle par défaut comme la régression qui n'utilise pas X pour prédire les valeurs de Y c.-à-d. le modèle composé uniquement de la constante.

$$y_i = b + \varepsilon_i \tag{1.11}$$

On montre très facilement dans ce cas que l'estimateur des MCO de la constante est

$$\hat{b} = \bar{y} \tag{1.12}$$

Dès lors, on peut considérer que R^2 confronte la prédiction du modèle s'appuyant sur X ($\hat{y}_i = \hat{a} \times x_i + \hat{b}$) avec le pire modèle possible, celui qui n'utilise pas l'information procurée par X c.-à-d. basée uniquement sur Y ($\hat{y}_i = \bar{y}$).

Par construction, dans la régression avec constante, on sait que $SCR \leq SCT$, le coefficient de détermination nous indique donc dans quelle mesure X permet d'améliorer nos connaissances sur Y.

Cette lecture nous permet de mieux comprendre les pseudo-R^2 calculés dans des domaines connexes telles que la régression logistique [14] (Section 1.6) où l'on confronte la vraisemblance du modèle complet (ou le taux d'erreur), incluant toutes les exogènes, avec celle du modèle réduit à la constante.

1.3.3 Coefficient de corrélation linéaire multiple

Le coefficient de corrélation linéaire multiple est la racine carrée du coefficient de détermination.

$$R = \sqrt{R^2} \tag{1.13}$$

Dans le cas de la régression simple (et uniquement dans ce cas), on montre aisément qu'il est égal au coefficient de corrélation r_{yx} entre Y et X. Son signe est défini par la pente $\hat{a}$ de la régression.

$$r_{yx} = signe(\hat{a}) \times R \tag{1.14}$$

La démonstration est relativement simple.

$$
\begin{aligned}
r^2_{yx} &= \hat{a}^2 \times \frac{\hat{\sigma}^2_x}{\hat{\sigma}^2_y} \\
&= \frac{\hat{a}^2 \sum_i (x_i - \bar{x})^2}{\sum_i (y_i - \bar{y})^2} \\
&= \frac{\sum_i [(\hat{a}x_i + \hat{b}) - (\hat{a}\bar{x} + \hat{b})]^2}{\sum_i (y_i - \bar{y})^2} \\
&= \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2} \\
&= \frac{SCE}{SCT} \\
&= R^2
\end{aligned}
$$

1.3.4 L'exemple des rendements agricoles

Nous nous appuyons sur les coefficients estimés précédemment (section 1.2.2), à savoir $\hat{a} = 0.71405$ et $\hat{b} = 4.39277$ pour construire la colonne des valeurs prédites $\hat{y}_i$, en déduire le résidu $\hat{\varepsilon}_i$ et finalement obtenir les sommes des carrés. Le tableau de calcul est organisé comme suit (Figure 1.7) [6] :

i	Y	X
1	16	20
2	18	24
3	23	28
4	24	22
5	28	32
6	29	28
7	26	32
8	31	36
9	32	41
10	34	41

Moyenne	26.1

a^	0.71405
b^	4.39277

Y^	epsilon^
18.674	-2.674
21.530	-3.530
24.386	-1.386
20.102	3.898
27.242	0.758
24.386	4.614
27.242	-1.242
30.099	0.901
33.669	-1.669
33.669	0.331

(Y-YB)²	(Y^-YB)²	(Y-Y^)²
102.010	55.148	7.149
65.610	20.884	12.461
9.610	2.937	1.922
4.410	35.977	15.195
3.610	1.305	0.574
8.410	2.937	21.286
0.010	1.305	1.544
24.010	15.990	0.812
34.810	57.289	2.785
62.410	57.289	0.110

314.900	251.061	63.839
SCT	SCE	SCR

R²	0.797273

Racine(R²)	0.892901

Correl(y,x)	0.892901

Fig. 1.7. Décomposition de la variance - "Rendements agricoles"

- Nous calculons $\hat{y}_i$. Par exemple, pour le 1^{er} individu : $\hat{y}_1 = \hat{a} \times x_1 + \hat{b} = 0.71405 \times 20 + 4.39277 = 18.674$.
- Sur la colonne suivante, nous en déduisons le résidu $\hat{\varepsilon}_i$ (ex. $\hat{\varepsilon}_1 = y_1 - \hat{y}_1 = 16 - 18.674 = -2.674$).
- Pour obtenir la SCT, nous réalisons la somme des $(y_i - \bar{y}_i)$ passées au carré : $SCT = (16 - 26.1)^2 + \cdots = 102.010 + \cdots = 314.900$
- Pour la SCE, nous sommons $(\hat{y}_i - \bar{y})^2$ c.-a-d. $SCE = (18.674 - 26.1)^2 + \cdots = 55.148 + \cdots = 251.061$
- Nous pouvons obtenir la SCR par différence, en faisant $SCR = SCT - SCE = 314.900 - 251.061 = 63.839$.

6. `regression_simple_rendements_agricoles.xlsx` - "reg.simple.decomp.variance"

– Nous pouvons aussi la former explicitement en sommant les $(y_i - \hat{y}_i)^2$, soit $SCR = (16 - 18.674)^2 + \cdots = 7.149 + \cdots = 63.839$. Les deux résultats coïncident, il ne peut pas en être autrement (dans la régression avec constante tout du moins).

Le coefficient de détermination est obtenu avec sa forme usuelle (Équation 1.10) :

$$R^2 = \frac{SCE}{SCT} = \frac{251.061}{314.900} = 0.797273$$

Puis, le coefficient de corrélation linéaire multiple

$$R = \sqrt{0.797273} = 0.892901$$

$\hat{a} = 0.71405$ étant positif, on vérifiera aisément dans notre exemple que ce dernier est identique au coefficient de corrélation de Pearson entre Y et X :

$$R = r_{yx} = 0.892901$$

2

Propriétés des estimateurs

Ce chapitre est assez théorique. Sa lecture n'est pas nécessaire pour la compréhension de la mise en pratique de la régression linéaire. J'invite donc les lecteurs surtout intéressés par les aspects opérationnels à se reporter au chapitre suivant (chapitre 3).

Ce chapitre est essentiel en revanche pour la compréhension des propriétés des estimateurs des MCO. Il permet notamment de circonscrire les hypothèses qui conditionnent leur efficacité. Sa lecture est conseillée pour ceux qui s'intéressent à ces aspects théoriques.

Pour les étudiants de la licence L3-IDS, vous devez lire ce chapitre !

Deux propriétés importantes sont mises en avant dans l'évaluation d'un estimateur. (1) Est-ce qu'il est sans biais c.-à-d. est-ce qu'en moyenne nous obtenons la vraie valeur du paramètre ? (2) Est-ce qu'il est convergent c.-à-d. à mesure que la taille de l'échantillon augmente, l'estimation devient de plus en plus précise ?

2.1 Biais

On dit que $\hat{\theta}$ est un estimateur sans biais de θ si $E[\hat{\theta}] = \theta$.

Comment procéder à cette vérification pour $\hat{a}$ et $\hat{b}$?

Voyons ce qu'il en est pour $\hat{a}$. Il y a deux étapes principalement dans la démonstration : dans un premier temps, il faut exprimer $\hat{a}$ en fonction de a ; dans un deuxième temps, en passant à l'espérance mathématique, il faut souhaiter que tout ce qui ne dépend pas de a devienne nul, au besoin en s'appuyant sur quelques hypothèses – pour le coup bien commodes – énoncées en préambule de notre présentation (section 1.1).

Nous reprenons ici la démarche que l'on retrouve dans la plupart des références citées en bibliographie (Bourbonnais, page 24 pour la régression simple ; Giraud et Chaix, page 25, qui a servi de base pour les calculs ci-dessous ; Labrousse, page 24 pour la régression multiple ; Dodge et Rousson, page 25).

Soit $y_i = ax_i + b + \varepsilon_i$, nous pouvons calculer :

$$\frac{1}{n}\sum_i y_i = a(\frac{1}{n}\sum_i x_i) + \frac{1}{n}(nb) + \frac{1}{n}\sum_i \varepsilon_i$$
$$\bar{y} = a\bar{x} + b\bar{\varepsilon}$$

Formons la différence

$$\begin{array}{r} \left\{ \begin{array}{l} y_i = ax_i + b + \varepsilon_i \\ \bar{y} = a\bar{x} + b + \bar{\varepsilon} \end{array} \right. \\ \hline y_i - \bar{y} = a(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon}) \end{array}$$

Rappelons que

$$\hat{a} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

Ainsi

$$\begin{aligned} \hat{a} &= \frac{\sum_i (x_i - \bar{x})[a(x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})]}{\sum_i (x_i - \bar{x})^2} \\ &= \frac{a\sum_i (x_i - \bar{x})^2 + \sum_i (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum_i (x_i - \bar{x})^2} \\ &= a + \frac{\sum_i (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})}{\sum_i (x_i - \bar{x})^2} \end{aligned}$$

On montre facilement que $\bar{\varepsilon}\sum_i (x_i - \bar{x}) = 0$, nous obtenons ainsi

$$\hat{a} = a + \frac{\sum_i (x_i - \bar{x})\varepsilon_i}{\sum_i (x_i - \bar{x})^2} \tag{2.1}$$

Il nous reste à démontrer que la partie après l'addition est nulle en passant à l'espérance mathématique. Nous devrons introduire les hypothèses adéquates pour ce faire.

$$\begin{aligned} E(\hat{a}) &= E(a) + E\left[\frac{\sum_i (x_i - \bar{x})\varepsilon_i}{\sum_i (x_i - \bar{x})^2}\right] \\ &= a + E\left[\sum_i \frac{(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2}\varepsilon_i\right] \end{aligned}$$

Pour simplifier les écritures, posons

$$\omega_i = \frac{(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2}$$

Nous avons :

$$E(\hat{a}) = a + E\left[\sum_i \omega_i \varepsilon_i\right]$$

La variable exogène X n'est pas stochastique par hypothèse. Donc

$$E(\hat{a}) = a + \sum_i \omega_i \times E(\varepsilon_i)$$

Autre hypothèse, $E(\varepsilon_i) = 0$. A la sortie nous obtenons

$$E(\hat{a}) = a$$

Conclusion. L'estimateur des moindres carrés ordinaires (EMCO) est sans biais, si et seulement si les deux hypothèses suivantes sont respectées :

1. (H1) L'exogène X n'est pas stochastique (X est non aléatoire) ;
2. (H2.a) $E(\varepsilon_i) = 0$, l'espérance de l'erreur est nulle.

Concernant la constante

De manière analogue, en partant de $\hat{b} = b + \bar{\varepsilon} - (\hat{a} - a)\bar{x}$, on montre sous les mêmes hypothèses que

$$E(\hat{b}) = b$$

2.2 Variance - Convergence

Un petit rappel : Un estimateur $\hat{\theta}$ sans biais de θ est convergent si et seulement si

$$V(\hat{\theta}) \underset{n\to\infty}{\longrightarrow} 0 \tag{2.2}$$

Nous devons donc d'abord produire une expression de la variance de l'estimateur, et montrer qu'il tend vers 0 quand l'effectif n tend vers ∞.

2.2.1 Variance de la pente

La variance est définie de la manière suivante :

$$V(\hat{a}) = E[(\hat{a} - a)^2]$$

Or, dans la section précédente, nous avons montré que l'estimateur pouvait s'écrire

$$\hat{a} = a + \sum_i \omega_i \varepsilon_i$$

Exploitons cela

$$\begin{aligned} V(\hat{a}) &= E\left[\left(\sum_i \omega_i \varepsilon_i\right)^2\right] \\ &= E\left[\sum_i \omega_i^2 \varepsilon_i^2 + 2\sum_{i<i'} \omega_i \omega_{i'} \varepsilon_i \varepsilon_{i'}\right] \\ &= \sum_i \omega_i^2 E(\varepsilon_i^2) + 2\sum_{i<i'} \omega_i \omega_{i'} E\left(\varepsilon_i \varepsilon_{i'}\right) \end{aligned}$$

Or, par hypothèse :

1. (H2.b) $E(\varepsilon_i^2) = V(\varepsilon_i) = \sigma_\varepsilon^2$, la variance de l'erreur est constante. C'est l'hypothèse d'homoscédasticité.
2. (H2.d) $COV(\varepsilon_{i'}\varepsilon_i) = E(\varepsilon_{i'}\varepsilon_i) = 0$. Les erreurs sont deux à deux indépendantes. C'est l'hypothèse de non-autocorrélation des erreurs.

A la sortie, nous pouvons simplifier grandement l'expression de la variance :

$$V(\hat{a}) = \sigma_\varepsilon^2 \sum_i \omega_i^2$$

Sachant que le terme ω_i correspond à

$$\omega_i = \frac{x_i - \bar{x}}{\sum_j (x_j - \bar{x})^2}$$

la somme de ces termes au carré devient

$$\begin{aligned}\sum_i \omega_i^2 &= \sum_i \left[\frac{x_i - \bar{x}}{\sum_j (x_j - \bar{x})^2}\right]^2 \\ &= \frac{1}{\left(\sum_j (x_j - \bar{x})^2\right)^2} \sum_i (x_i - \bar{x})^2 \\ &= \frac{1}{\sum_j (x_j - \bar{x})^2}\end{aligned}$$

A la sortie, nous avons la variance de l'estimation de la pente

$$V(\hat{a}) = \frac{\sigma_\varepsilon^2}{\sum_i (x_i - \bar{x})^2} \tag{2.3}$$

2.2.2 Convergence de la pente

Qu'en est-il de la convergence alors ?

Nous observons que :

- σ_ε^2 est une valeur qui ne dépend pas de n, c'est la variance de l'erreur définie dans la population.
- En revanche, lorsque $n \to \infty$, on constate facilement que $\sum_i (x_i - \bar{x})^2 \to \infty$. En effet, c'est une somme de valeurs toutes positives ou nulles.

Nous pouvons donc affirmer que $\hat{a}$ est un estimateur convergent de a, parce que

$$V(\hat{a}) \underset{n \to \infty}{\longrightarrow} 0 \tag{2.4}$$

Conclusion. Récapitulons tout ça. Nous avons introduit plusieurs hypothèses pour montrer la convergence de l'estimateur de la pente :

1. (H2.b) $E(\varepsilon_i^2) = V(\varepsilon_i) = \sigma_\varepsilon^2$. C'est l'hypothèse d'homoscédasticité.
2. (H2.d) $COV(\varepsilon_{i'}\varepsilon_i) = E(\varepsilon_{i'}\varepsilon_i) = 0$. C'est l'hypothèse de non-autocorrélation des erreurs.

2.2.3 Variance et convergence de la constante

En suivant la même démarche, nous pouvons produire l'expression de la variance de l'estimateur de la constante :

$$V(\hat{b}) = \sigma_\varepsilon^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right] \tag{2.5}$$

$\hat{b}$ est convergent, aux mêmes conditions (hypothèses) que l'estimateur de la pente.

2.2.4 Quelques remarques sur la précision des estimateurs

En scrutant un peu les formules de la variance produites dans les sections précédentes, nous remarquons plusieurs éléments. Les estimateurs seront d'autant plus précis, les variances seront d'autant plus petites, que :

- La variance de l'erreur est faible c.-à-d. la régression est de bonne qualité.
- La dispersion des X est forte c.-à-d. les points recouvrent bien l'espace de représentation.
- Le nombre d'observations n est élevé.

Nous pouvons illustrer cela à l'aide de quelques graphiques caractérisant les différentes situations (Figure 2.1).

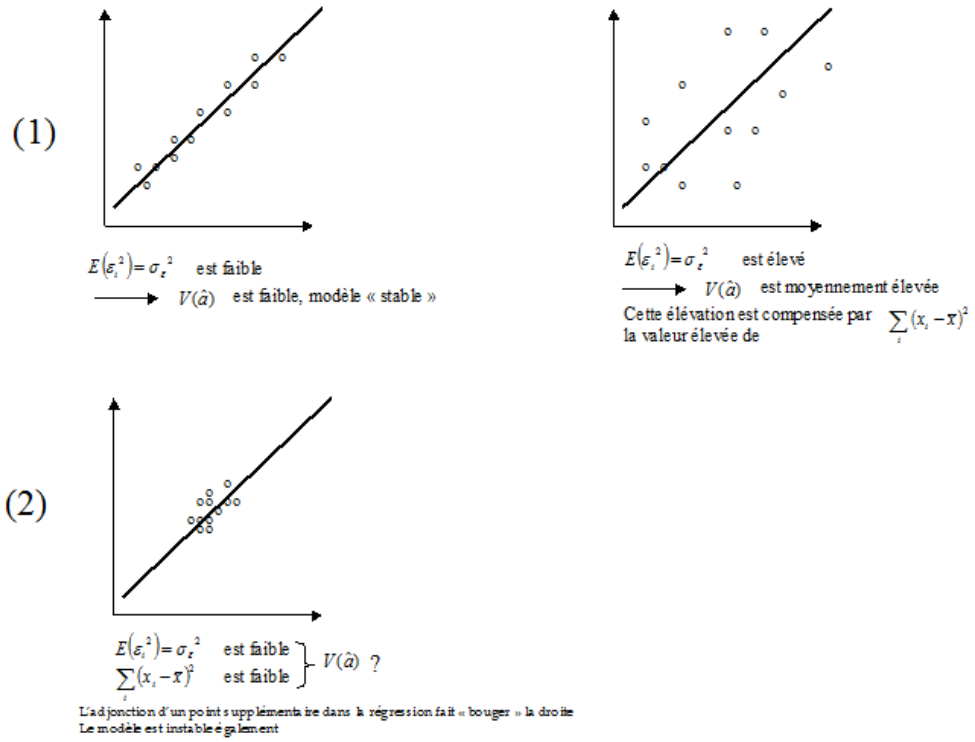


Fig. 2.1. Quelques situations caractéristiques - Influence sur la variance de la pente

2.3 Théorème de Gauss-Markov

Les estimateurs des MCO de la régression sont sans biais et convergents. On peut même aller plus loin et prouver que parmi les estimateurs linéaires sans biais de la régression, les estimateurs MCO sont à variance minimale c.-à-d. il n'existe pas d'autres estimateurs linéaires sans biais présentant une plus petite variance. **Les estimateurs des MCO** sont BLUE (best linear unbiased estimator). On dit qu'ils **sont efficaces** (pour les démonstrations montrant qu'il est impossible d'obtenir des variances plus faibles, voir Johnston, page 27 et pages 40-41 ; Labrousse, page 26).

3

Inférence statistique

3.1 Évaluation globale de la régression

Nous avions mis en avant la décomposition de la variance et le coefficient de détermination R^2 pour évaluer la qualité de l'ajustement (section 1.3). Le R^2 indiquait dans quelle proportion la variabilité de Y pouvait être expliquée par X. En revanche, il ne répond pas à la question : est-ce que la régression est globalement significative ? En d'autres termes, est-ce que les X (il n'y en a qu'un seul pour l'instant dans la régression simple) emmènent significativement de l'information sur Y, représentative d'une relation linéaire réelle dans la population, et qui va au-delà des simples fluctuations d'échantillonnage ?

Un autre point de vue est de considérer le test d'évaluation globale comme un test de significativité du R^2 : dans quelle mesure s'écarte-t-il réellement de la valeur 0 ? On a des réticences à le présenter ainsi dans la littérature francophone car le R^2 n'est pas un paramètre de la population estimée sur l'échantillon ; on a moins de scrupules dans la littérature anglo-saxonne (cf. par exemple D. Garson, *Multiple Regression*, `http://faculty.chass.ncsu.edu/garson/PA765/regress.htm#significance` – *"... The F test is used to test the significance of R, which is the same as testing the significance of R2, which is the same as testing the significance of the regression model as a whole...*; ou encore D. Mc Lane, *HyperStat Online Contents*, `http://davidmlane.com/hyperstat/B142546.html` – *... The following formula (le test F) is used to test whether an R2 calculated in a sample is significantly different from zero...*) [1].

Quoiqu'il en soit, l'hypothèse nulle correspond bien à l'absence de liaison linéaire entre l'endogène et les exogènes.

3.1.1 Tableau d'analyse de Variance - Test de significativité globale

Pour répondre à cette question, nous allons étendre l'étude de la décomposition de la variance en complétant le tableau d'analyse de variance par les degrés de liberté (Tableau 3.1).

1. Note : Tout le monde aura remarqué que je blinde mon discours avec des références facilement vérifiables pour éviter que les puristes me tombent dessus à coups de hache.

Source de variation	Somme des carrés	Degrés de liberté	Carrés moyens
Expliquée	$SCE = \sum_i (\hat{y}_i - \bar{y})^2$	1	$CME = \frac{SCE}{1}$
Résiduelle	$SCR = \sum_i (y_i - \hat{y}_i)^2$	$n-2$	$CMR = \frac{SCR}{n-2}$
Totale	$SCT = \sum_i (y_i - \bar{y})^2$	$n-1$	-

Tableau 3.1. Tableau d'analyse de variance pour la régression simple

Un petit mot sur les **degrés de liberté**, on peut les voir de différentes manières. La définition la plus accessible est de les comprendre comme le nombre de termes impliqués dans les sommes (le nombre d'observations) moins le nombre de paramètres estimés dans cette somme (Dodge et Rousson, page 41). Ainsi :

- Nous avons besoin de l'estimation de la moyenne $\bar{y}$ pour calculer la somme SCT.
- Nous avons besoin des coefficients estimés $\hat{a}$ et $\hat{b}$ pour obtenir la projection $\hat{y}_i$ et former la SCR.
- Concernant la SCE, le plus simple est de l'obtenir par déduction c.-à-d. $(n-1)-(n-2)=1$.

Pour tester la significativité globale de la régression, nous nous basons sur **la statistique F**,

$$F = \frac{CME}{CMR} = \frac{\frac{SCE}{1}}{\frac{SCR}{n-2}} \tag{3.1}$$

Interprétation. Cette statistique indique si la variance expliquée est significativement supérieure à la variance résiduelle. Dans ce cas, on peut considérer que l'explication emmenée par la régression traduit une relation qui existe réellement dans la population (Bourbonnais, page 34).

Écriture à partir du coefficient de détermination. D'aucuns considèrent le test F comme un test de significativité du coefficient de détermination, on peut le comprendre dans la mesure où il peut s'écrire en fonction du R^2

$$F = \frac{\frac{R^2}{1}}{\frac{(1-R^2)}{n-2}} \tag{3.2}$$

Distribution sous H0. Sous $H0$, SCE est distribué selon un $\chi^2(1)$ et SCR selon un $\chi^2(n-2)$, de fait pour F nous avons

$$F \equiv \frac{\frac{\chi^2(1)}{1}}{\frac{\chi^2(n-2)}{n-2}} \equiv \mathcal{F}(1, n-2) \tag{3.3}$$

Sous H0, F est donc distribué selon une loi de Fisher à $(1, n-2)$ degrés de liberté.

La région critique du test, correspondant au rejet de H0, au risque α est définie pour les valeurs anormalement élevées de F c.-à-d.

$$R.C. : F > F_{1-\alpha}(1, n-2) \tag{3.4}$$

Décision à partir de la p-value. Dans la plupart des logiciels de statistique, on fournit directement la probabilité critique (p-value) α', elle correspond à la probabilité que la loi de Fisher dépasse la statistique calculée F.

Ainsi, la règle de décision au risque α devient :

$$R.C. : \alpha' < \alpha \tag{3.5}$$

3.1.2 Exemple : les rendements agricoles

Revenons à notre exemple des rendements agricoles. Nous complétons notre feuille de calcul précédente (Figure 1.7) de manière à mettre en exergue le tableau d'analyse de variance complet et le test F de significativité globale (Figure 3.1) [2].

i	Y	X
1	16	20
2	18	24
3	23	28
4	24	22
5	28	32
6	29	28
7	26	32
8	31	36
9	32	41
10	34	41

Y^	epsilon^
18.674	-2.674
21.530	-3.530
24.386	-1.386
20.102	3.898
27.242	0.758
24.386	4.614
27.242	-1.242
30.099	0.901
33.669	-1.669
33.669	0.331

(Y-YB)²	(Y^-YB)²	(Y-Y^)²
102.010	55.148	7.149
65.610	20.884	12.461
9.610	2.937	1.922
4.410	35.977	15.195
3.610	1.305	0.574
8.410	2.937	21.286
0.010	1.305	1.544
24.010	15.990	0.812
34.810	57.289	2.785
62.410	57.289	0.110

Moyenne	26.1

a^	0.71405
b^	4.39277

314.900	251.061	63.839
SCT	SCE	SCR

Tableau d'analyse de variance

Source	SC	DDL	Carrés Moyens
Expliquée	251.061	1	251.061
Résiduelle	63.839	8	7.980
Totale	314.900	9	

F	31.462

ddl1	1
ddl2	8

F_0.95	5.318

p-value	0.00050487

Conclusion : Le modèle est globalement significatif au risque 5%

Fig. 3.1. Tableau d'analyse de variance et Test de significativité globale - "Rendements agricoles"

Voici le détail des calculs :

- Nous avions expliqué précédemment l'obtention des SCT, SCE et SCR (section 1.3.4).
- Nous réorganisons les valeurs pour construire le tableau d'analyse de variance. Nous en déduisons les carrés moyens expliqués $CME = \frac{SCE}{1} = \frac{251.061}{1} = 251.061$ et les carrés moyens résiduels $CMR = \frac{SCR}{n-2} = \frac{63.839}{10-2} = 7.980$

2. `regression_simple_rendements_agricoles.xlsx` - "reg.simple.test.global"

- Nous en déduisons la statistique de test $F = \frac{CME}{CMR} = \frac{251.061}{7.980} = 31.462$
- Que nous comparons au quantile d'ordre $(1-\alpha)$ de la loi $\mathcal{F}(1, n-2)$. Pour $\alpha = 5\%$, elle est égale [3] à $F_{0.95}(1,8) = 5.318$. Nous concluons que le modèle est globalement significatif au risque 5%. La relation linéaire entre Y et X est représentatif d'un phénomène existant réellement dans la population.
- En passant par la probabilité critique, nous avons [4] $\alpha' \approx 0.00050$, inférieure à $\alpha = 5\%$. La conclusion est la même. Il ne peut pas y avoir de contradictions entre ces deux visions de toute manière.

3.2 Distribution des coefficients estimés

Pour étudier les coefficients estimés, il importe d'en calculer les paramètres (l'espérance et la variance essentiellement) et de déterminer la loi de distribution. Nous pourrons dès lors mettre en oeuvre les outils usuels de la statistique inférentielle : la définition des intervalles de variation à un niveau de confiance donné ; la mise en place des tests d'hypothèses, notamment les tests de significativité.

3.2.1 Distribution de $\hat{a}$ et $\hat{b}$

Dans un premier temps, concentrons-nous sur la pente de la régresion. Rappelons que $\hat{a}$ est égal à

$$\hat{a} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

X est non stochastique, Y l'est par l'intermédiaire du terme d'erreur ε. Nous introduisons l'hypothèse selon laquelle :

$$\varepsilon_i \equiv \mathcal{N}(0, \sigma_\varepsilon)$$

De fait, $y_i = ax_i + b + \varepsilon_i$ suit aussi une loi normale, et $\hat{a}$ étant une combinaison linéaire des y_i, il vient

$$\frac{\hat{a} - a}{\sigma_{\hat{a}}} \equiv \mathcal{N}(0,1) \tag{3.6}$$

Rappelons que la variance de $\hat{a}$ s'écrit (section 2.2) :

$$\sigma_{\hat{a}}^2 = \frac{\sigma_\varepsilon^2}{\sum_i (x_i - \bar{x})^2} \tag{3.7}$$

Ce résultat est très intéressant mais n'est pas utilisable en l'état, tout simplement parce que nous ne disposons pas de l'estimation de la variance de l'erreur σ_ε^2. Pour obtenir une estimation calculable sur un échantillon de données de l'écart-type $\hat{\sigma}_{\hat{a}}$ du coefficient $\hat{a}$, nous devons produire une estimation de l'écart type de l'erreur $\hat{\sigma}_\varepsilon$. La variance estimée s'écrirait alors

$$\hat{\sigma}_{\hat{a}}^2 = \frac{\hat{\sigma}_\varepsilon^2}{\sum_i (x_i - \bar{x})^2} \tag{3.8}$$

3. INVERSE.LOI.F(0.05 ;1 ;8) dans Excel
4. LOI.F(31.462 ;1 ;8) dans Excel.

La suite logique de notre exposé consiste donc à proposer une estimation sans biais de la variance de l'erreur σ_ε^2.

Le cas de la constante. La situation est identique pour ce qui est de l'estimation de la constante $\hat{b}$. Nous avons :

$$\frac{\hat{b}-b}{\sigma_{\hat{b}}} \equiv \mathcal{N}(0,1) \tag{3.9}$$

Avec pour variance de $\hat{b}$ (section 2.2) :

$$\sigma_{\hat{b}}^2 = \sigma_\varepsilon^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2}\right]$$

De nouveau, si nous souhaitons obtenir son estimation c.-à-d. mettre un chapeau sur le σ de $\hat{b}$ comme j'ai coutume de le dire en cours, il faut mettre un chapeau sur le σ de ε. C'est ce que nous faisons dans la section suivante.

3.2.2 Estimation de la variance de l'erreur

Estimateur sans biais de la variance de l'erreur

Le résidus $\hat{\varepsilon}_i$ est l'erreur observée, on peut la ré-écrire de la manière suivante :

$$\begin{aligned}\hat{\varepsilon}_i &= y_i - \hat{y}_i \\ &= ax_i + b + \varepsilon_i - (\hat{a}x_i + \hat{b}) \\ &= \varepsilon_i - (\hat{a}-a)x_i - (\hat{b}-b)\end{aligned}$$

Remarque 2 (Espérance des résidus). On note au passage que l'espérance du résidu est nulle ($E[\hat{\varepsilon}_i] = 0$) si les estimateurs sont sans biais.

On montre que (Giraud et Chaix, page 31) :

$$E\left[\sum_i \hat{\varepsilon}_i^2\right] = (n-2)\sigma_\varepsilon^2 \tag{3.10}$$

On propose comme estimateur sans biais de la variance de l'erreur :

$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_i \hat{\varepsilon}_i^2}{n-2} = \frac{SCR}{n-2} \tag{3.11}$$

Quelques commentaires :

- Au numérateur, nous avons la somme des carrés des résidus. Nous l'obtenons facilement comme nous avons pu le constater dans notre exemple des "Rendements agricoles".
- Au dénominateur, nous avons les degrés de liberté de la régression. La valeur **2** dans $(n-2)$ représente le nombre de paramètres estimés. De fait, la généralisation de cette formule au cadre de la régression linéaire multiple avec p variables exogènes ne pose aucun problème. Le nombre de degrés de liberté sera $n-(p+1) = n-p-1$.

Distribution de l'estimation de la variance de l'erreur

Il nous faut connaître la distribution de l'estimation de la variance de l'erreur pour pouvoir déterminer la distribution des coefficients estimés lorsque nous introduirons $\hat{\sigma}_\varepsilon^2$ dans les expressions de leur variance.

On sait par hypothèse que $\frac{\varepsilon_i}{\sigma_\varepsilon} \equiv \mathcal{N}(0,1)$. Comme $\hat{\varepsilon}_i$ est une réalisation de ε_i, il vient

$$\frac{\hat{\varepsilon}_i}{\sigma_\varepsilon} \equiv \mathcal{N}(0,1) \tag{3.12}$$

En passant au carré, nous avons un $\chi^2(1)$. Il ne nous reste plus qu'à former la somme des termes :

$$\sum_i \left(\frac{\hat{\varepsilon}_i}{\sigma_\varepsilon}\right)^2 = \frac{\sum_i \hat{\varepsilon}_i^2}{\sigma_\varepsilon^2} \equiv \chi^2(n-2) \tag{3.13}$$

Ou, de manière équivalente, en se référant à l'estimateur de la variance de l'erreur (Équation 3.11) :

$$\frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} \equiv \frac{\chi^2(n-2)}{n-2} \tag{3.14}$$

Nous pouvons maintenant revenir sur la distribution des coefficients calculés lorsque toutes ses composantes sont estimées à partir des données.

3.2.3 Distribution des coefficients dans la pratique

Voyons dans un premier temps la pente, la transposition à la constante ne pose aucun problème.

Avec les équations 3.7 et 3.8, nous pouvons écrire :

$$\frac{\hat{\sigma}_{\hat{a}}^2}{\sigma_{\hat{a}}^2} = \frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2}$$

En reprenant l'équation 3.14, nous déduisons :

$$\frac{\hat{\sigma}_{\hat{a}}^2}{\sigma_{\hat{a}}^2} = \frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} \equiv \frac{\chi^2(n-2)}{n-2} \tag{3.15}$$

De fait, la distribution réellement exploitable pour l'inférence statistique est la loi de Student à $(n-2)$ degrés de liberté.

$$\frac{\hat{a}-a}{\hat{\sigma}_{\hat{a}}} \equiv \mathcal{T}(n-2) \tag{3.16}$$

Comment ?

N'oublions pas que la loi de Student est définie par un rapport entre une loi normale et la racine carrée d'un loi du χ^2 normalisée par ses degrés de liberté. Ainsi,

$$\frac{\frac{\hat{a}-a}{\sigma_{\hat{a}}}}{\frac{\hat{\sigma}_{\hat{a}}}{\sigma_{\hat{a}}}} \equiv \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2(n-2)}{n-2}}}$$
$$\frac{\hat{a}-a}{\hat{\sigma}_{\hat{a}}} \equiv \mathcal{T}(n-2)$$

De manière complètement analogue, pour la constante estimée $\hat{b}$

$$\frac{\hat{b}-b}{\hat{\sigma}_{\hat{b}}} \equiv \mathcal{T}(n-2) \tag{3.17}$$

Nous disposons maintenant de tous les éléments pour analyser les paramètres estimés de la régression.

3.3 Étude de la pente de la droite de régression

3.3.1 Test de significativité de la pente

Le test de significativité de la pente consiste à vérifier l'influence réelle de l'exogène X sur l'endogène Y. Les hypothèses à confronter s'écrivent :

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

Nous formons la statistique de test

$$t_{\hat{a}} = \frac{\hat{a}}{\hat{\sigma}_{\hat{a}}} \tag{3.18}$$

Elle suit une loi de Student à $(n-2)$ degrés de liberté. La région critique (de rejet de H0) au risque α s'écrit :

$$R.C. : |t_{\hat{a}}| > t_{1-\frac{\alpha}{2}} \tag{3.19}$$

Où $t_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $(1-\frac{\alpha}{2})$ de la loi de Student. Il s'agit d'un test bilatéral.

Test de significativité de la pente pour les "Rendements agricoles"

Testons la significativité de la pente pour la régression sur les "Rendements agricoles". Nous construisons la feuille Excel pour les calculs intermédiaires (Figure 3.2) [5] :

- Nous calculons les projections pour chaque individu de l'échantillon. Pour le 1^{er} individu, nous avons $\hat{y}_1 = \hat{a} \times x_1 + \hat{b} = 0.71405 \times 20 + 4.39277 = 18.674$.
- Nous en déduisons le résidu (ex. $\hat{\varepsilon}_1 = y_1 - \hat{y}_1 = 16 - 18.674 = -2.674$), que nous passons au carré (ex. $\hat{\varepsilon}_1^2 = (-2.674)^2 = 7.149$).
- Nous réalisons la somme des résidus au carré, soit $SCR = \sum_i \hat{\varepsilon}_i^2 = 7.149 + \cdots = 63.839$

5. `regression_simple_rendements_agricoles.xlsx` - "reg.simple.test.pente"

i	Y	X	Y^	epsilon^	(epsilon^)²	(X-XB)²
1	16	20	18.674	-2.674	7.149	108.16
2	18	24	21.530	-3.530	12.461	40.96
3	23	28	24.386	-1.386	1.922	5.76
4	24	22	20.102	3.898	15.195	70.56
5	28	32	27.242	0.758	0.574	2.56
6	29	28	24.386	4.614	21.286	5.76
7	26	32	27.242	-1.242	1.544	2.56
8	31	36	30.099	0.901	0.812	31.36
9	32	41	33.669	-1.669	2.785	112.36
10	34	41	33.669	0.331	0.110	112.36
						Somme
Moyenne	26.1	30.4		SCR	63.839	492.4
a^	0.71405			(sigma^)²_eps	7.980	
b^	4.39277			sigma^(eps)	2.825	

Fig. 3.2. Calculs intermédiaires pour les tests relatifs à la pente - "Rendements agricoles"

A ce stade, nous obtenons l'estimation de la variance de l'erreur, soit

$$\hat{\sigma}_\varepsilon^2 = \frac{SCR}{n-2} = \frac{63.839}{8} = 7.980$$

L'écart-type estimé de l'erreur correspond à la racine carrée, il est bien de le préciser car de nombreux logiciels (la fonction DROITEREG d'Excel par exemple) l'affichent plutôt que la variance.

$$\hat{\sigma}_\varepsilon = \sqrt{7.980} = 2.825$$

Pour obtenir l'estimation de l'écart-type de la pente, nous avons besoin de la somme des écarts à la moyenne au carré des X c.-à-d. $\sum_i (x_i - \bar{x})^2 = (20 - 30.4)^2 + \cdots = 108.16 + \cdots = 492.4$. Nous avons alors :

$$\begin{aligned}
\hat{\sigma}_{\hat{a}} &= \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_i (x_i - \bar{x})^2}} \\
&= \sqrt{\frac{7.980}{492.4}} \\
&= \sqrt{0.01621} \\
&= 0.12730
\end{aligned}$$

Nous formons la statistique de test

$$t_{\hat{a}} = \frac{\hat{a}}{\hat{\sigma}_{\hat{a}}} = \frac{0.71405}{12730} = 5.60909$$

Au risque $\alpha = 5\%$, le seuil critique pour la loi de Student à $(n-2)$ degrés de liberté pour un test bilatéral [6] est $t_{1-\frac{\alpha}{2}} = 2.30600$. Puisque $|5.60909| > 2.30600$, nous concluons que la pente est significativement non nulle au risque 5%.

6. LOI.STUDENT.INVERSE(0.05 ;8) sous Excel. Attention, la fonction renvoie directement le quantile pour un test bilatéral!

Si nous étions passés par le calcul de la p-value, nous aurions obtenu [7] $\alpha' = 0.00050$. Puisque $\alpha' < \alpha$, nous rejetons de même l'hypothèse nulle.

3.3.2 Test de conformité à un standard

Nous pouvons aller plus loin que le simple test de significativité. En effet, la distribution de $\hat{a}$ (section 3.2.3, équation 3.16) est valable sur tout le domaine de définition de a et non pas seulement dans le voisinage ($a = 0$). Ainsi, nous pouvons définir tout type de test de conformité à un standard, où l'hypothèse nulle s'écrirait $H_0 : a = c$; c étant une valeur de référence quelconque.

Exemple sur les "Rendements agricoles"

On souhaite mettre en oeuvre le test d'hypothèses suivant pour les "Rendements agricoles"

$$\begin{cases} H_0 : a = 0.5 \\ H_1 : a > 0.5 \end{cases}$$

Il s'agit d'un test de conformité à un standard unilatéral. La région critique au risque α du test s'écrit

$$R.C. : \frac{\hat{a} - 0.5}{\hat{\sigma}_{\hat{a}}} > t_{1-\alpha}$$

Voyons ce qu'il en est sur nos données,

$$\frac{\hat{a} - 0.5}{\hat{\sigma}_{\hat{a}}} = \frac{0.71405 - 0.5}{0.12730} = 1.68145$$

A comparer avec $t_{0.95}(8) = 1.85955$ pour un test à 5% [8]. Nous sommes dans la région d'acceptation c.-à-d. nous ne pouvons pas rejeter l'hypothèse nulle. La valeur du paramètre a n'est pas significativement supérieur à la référence 0.5 au risque 5%.

3.3.3 Intervalle de confiance

Toujours parce que la distribution de $\hat{a}$ est définie sur tout l'intervalle de définition de a, nous pouvons construire des intervalles de variation (ou intervalle de confiance) au niveau de confiance $(1 - \alpha)$.

Elle est définie par

$$\hat{a} \pm t_{1-\frac{\alpha}{2}} \times \hat{\sigma}_{\hat{a}} \tag{3.20}$$

7. LOI.STUDENT(ABS(5.60909) ;8 ;2) sous Excel. Le paramètre **2** pour spécifier que nous souhaitons obtenir la p-value pour un test bilatéral.

8. Attention, comme il s'agit d'un test unilatéral, le seuil critique est modifié par rapport à l'exemple du test de significativité précédent.

Exemple sur les "Rendements agricoles"

Reprenons la pente du fichier "Rendements agricoles". Nous disposons de toutes les informations nécessaires pour produire l'intervalle de confiance au niveau 95% :

$$\left[\hat{a} - t_{1-\frac{\alpha}{2}} \times \hat{\sigma}_{\hat{a}} \ ; \ \hat{a} + t_{1-\frac{\alpha}{2}} \times \hat{\sigma}_{\hat{a}}\right]$$
$$[0.71405 - 2.30600 \times 0.12730 \ ; \ 0.71405 + 2.30600 \times 0.12730]$$
$$[0.42049 \ ; \ 1.00761]$$

Le résultat est cohérent avec le test de significativité de la pente, l'intervalle de confiance ne contient pas la valeur **0**.

3.4 Intervalle de confiance de la droite de régression

Les coefficients formant le modèle sont entachées d'incertitude, il est normal que la droite de régression le soit également. L'objectif dans cette section est de produire un intervalle de confiance de la droite de régression (Bressoux, page 76).

Pour formaliser cela, n'oublions pas que notre objectif est de modéliser au mieux les valeurs de Y en fonction des valeurs prises par X c.-à-d. $\mu_{Y/X} = E[Y/X]$. Dans la régression linéaire, on fait l'hypothèse que la relation est linéaire

$$\mu_{Y/X} = a \times X + b \tag{3.21}$$

C'est pour cette raison que dans la plupart des ouvrages, on présente les résultats décrits dans cette section comme le calcul de l'intervalle de confiance de la prédiction de la moyenne de Y conditionnellement X (Dodge et Rousson, page 34 ; Johnston et DiNardo, page 36 ; Tenenhaus, page 92). Mais il s'agit bien de l'intervalle de confiance de ce que l'on a modélisé avec la droite, à ne pas confondre avec l'intervalle de confiance d'une prédiction lorsque l'on fourni la valeur x_{i*} pour un nouvel individu $i*$ n'appartenant pas à l'échantillon.

J'avoue que pendant longtemps, cette distinction ne me paraissait pas très claire. Je ne voyais pas très bien quelle était la différence entre l'intervalle de confiance de la prédiction l'espérance de Y sachant X et la prédiction ponctuelle de Y. Dans les deux cas, nous avions la même valeur ponctuelle calculée $\hat{a} \times x_i + \hat{b}$. Le passage de l'un à l'autre dans Jonhston et DiNardo – livre que j'avais beaucoup lu quand j'étais étudiant – pages 35 et 36, formules (1.67) et (1.68), est particulièrement périlleux.

Bref, la terminologie "intervalle de confiance de la droite de régression" (Bressoux, page 76) me sied mieux.

Pour un individu donné, nous obtenons l'estimation de sa moyenne conditionnelle :

$$\hat{\mu}_{Y/x_i} = \hat{a} \times x_i + \hat{b} \tag{3.22}$$

Et l'estimation de la variance de cette moyenne conditionnelle estimée s'écrit :

$$\hat{\sigma}^2_{\hat{\mu}_{Y/x_i}} = \hat{\sigma}^2_\varepsilon \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2} \right) \tag{3.23}$$

Enfin, la moyenne conditionnelle estimée suit une loi de Student à $(n-2)$ degrés de libertés.

Tous ces éléments nous permettent de construire l'intervalle de confiance au niveau $(1-\alpha)$ de la droite de régression (Bressoux, page 76 ; équation 2.17) :

$$\hat{a} \times x_i + \hat{b} \pm t_{1-\frac{\alpha}{2}} \times \hat{\sigma}_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}} \tag{3.24}$$

Levier. L'expression

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_j (x_j - \bar{x})^2} \tag{3.25}$$

est appelée levier de l'observation i dans la littérature. Il tient une place très importante dans la régression, notamment dans la détection des points atypiques (voir [13], chapitre 2).

Intervalle de confiance de la droite "Rendements agricoles"

i	Y	X
1	16	20
2	18	24
3	23	28
4	24	22
5	28	32
6	29	28
7	26	32
8	31	36
9	32	41
n = 10	34	41

Y^	epsilon^	(epsilon^)²
18.674	-2.674	7.149
21.530	-3.530	12.461
24.386	-1.386	1.922
20.102	3.898	15.195
27.242	0.758	0.574
24.386	4.614	21.286
27.242	-1.242	1.544
30.099	0.901	0.812
33.669	-1.669	2.785
33.669	0.331	0.110

(X-XB) ^2
108.16
40.96
5.76
70.56
2.56
5.76
2.56
31.36
112.36
112.36

b.basse	b.haute
14.99	22.36
18.74	24.32
22.21	26.56
16.89	23.32
25.13	29.36
22.21	26.56
25.13	29.36
27.46	32.73
29.94	37.40
29.94	37.40

Moyenne	30.4

SCR	63.8387	Somme	492.4
sigma^(epsilon)	2.8249		

n	10

t_0.975(8)	2.30600

a^	0.71405
b^	4.39277

Fig. 3.3. Calculs pour l'intervalle de confiance à 95% de droite - "Rendements agricoles"

Reprenons notre exemple des "Rendements agricoles". Nous formons la feuille Excel permettant de calculer les bornes basses et hautes de la droite de régression au niveau de confiance 95% (Figure 3.3) [9] :

- Une grande partie des informations ont déjà été calculées dans les précédents exemples, nous savons que $n = 10$, $\hat{a} = 0.71405$, $\hat{b} = 4.39277$, $\hat{\sigma}_\varepsilon = 2.8249$, $\bar{x} = 30.4$, la somme $\sum_j (x_j - \bar{x})^2 = 492.4$.
- Pour un niveau de confiance 95%, la loi de Student nous fournit le quantile $t_{0.975}(8) = 2.30600$

9. `regression_simple_rendements_agricoles.xlsx` - "reg.simple.intv.confiance"

– Nous sommes parés pour construire les intervalles de confiance. Pour le 1-er individu, nous avons :

$$b.b.(\mu_{Y/X=x_1}) = 18.674 - 2.30600 \times 2.8249 \times \sqrt{\frac{1}{10} + \frac{(20-30.4)^2}{492.4}} = 14.99$$

$$b.h.(\mu_{Y/X=x_1}) = 18.674 + 2.30600 \times 2.8249 \times \sqrt{\frac{1}{10} + \frac{(20-30.4)^2}{492.4}} = 22.36$$

Dans la régression simple, la représentation graphique est très intuitive (Figure 3.4). Il y a 95% de chances que la droite soit comprise entre les deux courbes bleues. Attention, la droite ne peut être placée n'importe où dans la zone délimitée, *elle pivote forcément autour du barycentre.*

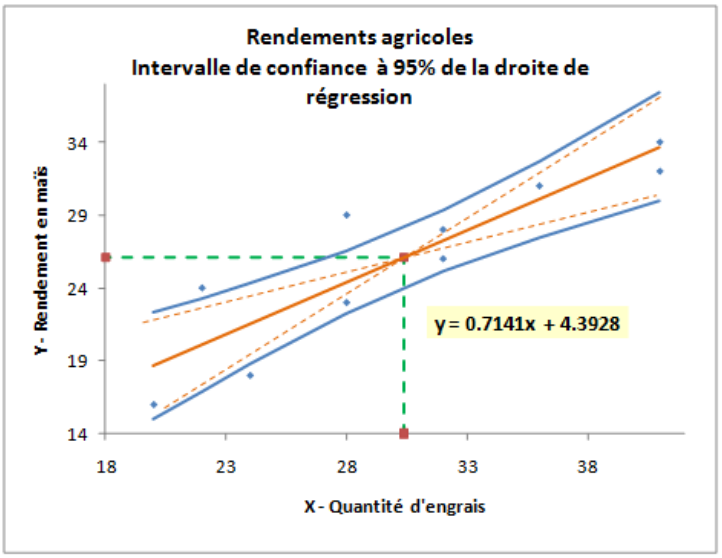


Fig. 3.4. Représentation de l'intervalle de confiance à 95% de la droite - "Rendements agricoles"

3.5 La régression avec la fonction DROITEREG d'EXCEL

Tous les résultats mis en avant dans ce support (du moins jusqu'à ce stade) peuvent être obtenus facilement en partant des valeurs fournies par la fonction DROITEREG d'Excel. Nous avons donc élaboré une feuille de calcul où, à partir des sorties de DROITEREG, nous avons établi les principaux indicateurs d'évaluation globale et individuelle des coefficients (Figure 3.5) [10]. Nous avons volontairement affiché les en-têtes des lignes et colonnes de la feuille Excel pour rendre la lecture plus facile.

Nous disposons du tableau de données de 10 observations en **B2 :C11**. Nous insérons la fonction DROITEREG sur la plage **F3 :G7**. Elle fournit les coefficients estimés sur la première ligne, nous réservons autant de colonnes qu'il y a de coefficients (2 dans notre cas, la pente et la constante de la régression) ; et, si nous souhaitons consulter les statistiques intermédiaires relatifs à la régression, nous devons réserver

10. `regression_simple_rendements_agricoles.xlsx` - "droitereg"

4 lignes supplémentaires (5 lignes en tout). Attention, il s'agit d'une fonction matricielle, elle complète directement plusieurs cellules, nous devons donc valider en appuyant simultanément sur les touches CTRL + MAJ + ENTREE.

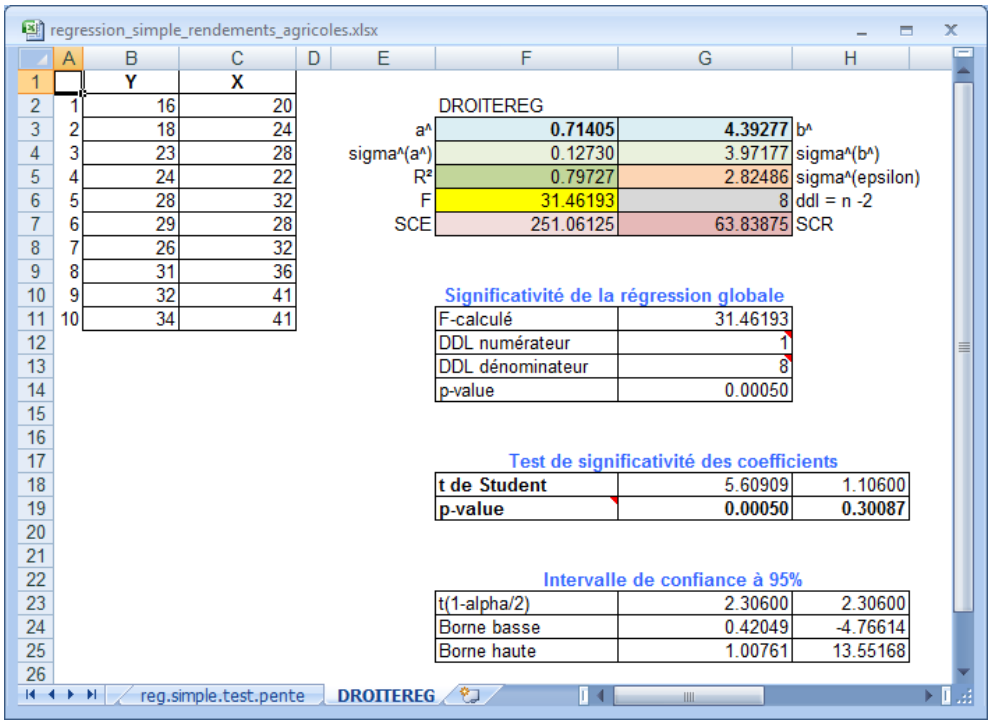


Fig. 3.5. Sorties de la fonction DROITEREG d'Excel - "Rendements agricoles"

Décrivons les valeurs fournies par la fonction DROITEREG en les énumérant (de gauche à droite, du haut vers le bas) (Figure 3.5) :

Ligne 1 – Les coefficients de la régression. A gauche $\hat{a} = 0.71405$; en dernière colonne (ce sera toujours la place de la constante, y compris dans la régression multiple), $\hat{b} = 4.39277$.

Ligne 2 – Nous avons les estimations des écarts-type des coefficients estimés, soit $\hat{\sigma}_{\hat{a}} = 0.12730$ et $\hat{\sigma}_{\hat{b}} = 3.97177$.

Ligne 3 – Nous avons sur la première colonne le coefficient de détermination $R^2 = 0.79727$, sur la seconde l'estimation de l'écart-type de l'erreur, $\hat{\sigma}_{\varepsilon} = 2.82486$.

Ligne 4 – A gauche la statistique de test d'évaluation globale de la régression (test F) $F = 31.46193$; à droite, le degré de liberté de la régression $n - 2 = 8$.

Ligne 5 – Nous avons respectivement, la $SCE = 251.06125$ et la $SCR = 63.83875$.

A partir de ces informations, nous pouvons établir tous les résultats mis en avant dans ce support (jusqu'à ce stade, précisons le bien). Nous avons ainsi construit (Figure 3.5, partie basse) : le tableau pour l'évaluation globale de la régression, avec le calcul de la probabilité critique ; les tests de significativité

individuelle des coefficients ; et leurs intervalles de confiance à 95%. Toutes les valeurs sont identiques à celles que nous avions établies dans les chapitres précédents.

3.6 Quelques équivalences concernant la régression simple

La régression simple ne faisant intervenir qu'une seule variable explicative, on montre facilement que le test de significativité de la pente – c.-à-d. tester la nullité du coefficient associé à l'exogène – équivaut d'une part, au test de significativité globale de la régression et, d'autre part, au test de significativité de la corrélation entre Y et X.

3.6.1 Équivalence avec le test de significativité globale

Revenons sur la statistique F du test de significativité globale, elle s'écrit (Tenenhaus, page 83) :

$$\begin{aligned} F &= \frac{SCE/1}{SCR/(n-2)} \\ &= \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\hat{\sigma}_\varepsilon^2} \\ &= \frac{\sum_i (\hat{a}x_i + \hat{b} - \bar{y})^2}{\hat{\sigma}_\varepsilon^2} \\ &= \frac{\sum_i [\hat{a}x_i + (\bar{y} - \hat{a}\bar{x}) - \bar{y}]^2}{\hat{\sigma}_\varepsilon^2} \\ &= \frac{\hat{a}^2 \sum_i (x_i - \bar{x})^2}{\hat{\sigma}_\varepsilon^2} = \frac{\hat{a}^2}{\frac{\hat{\sigma}_\varepsilon^2}{\sum_i (x_i - \bar{x})^2}} \\ &= \frac{\hat{a}^2}{\hat{\sigma}_{\hat{a}}^2} = \left(\frac{\hat{a}}{\hat{\sigma}_{\hat{a}}}\right)^2 \\ &= t_{\hat{a}}^2 \end{aligned}$$

Ainsi, tester la significativité de la pente dans la régression simple avec constante revient à tester la significativité globale. Les statistiques de test sont cohérentes. Il en est de même en ce qui concerne les distributions car il y a une équivalence entre la loi de Student et la loi de Fisher.

$$\left(\mathcal{T}(n-2)\right)^2 \equiv \mathcal{F}(1, n-2) \tag{3.26}$$

Vérification sur les données "Rendements agricoles". Nous le constatons après coup sur notre exemple. Nous avons $t_{\hat{a}} = 5.60909$ (section 3.3.1). En passant au carré, nous obtenons la valeur de statistique de test $F = 31.462 = (5.60909)^2$ (section 3.1).

3.6.2 Équivalence avec le test de significativité de la corrélation

De la même manière, nous pouvons relier $t_{\hat{a}}$ avec la statistique de test utilisée pour tester la significativité de la corrélation (Giraud, page 57 ; Tenenhaus, page 84).

Développons de nouveau l'expression de F :

$$\begin{aligned} F &= \frac{SCE/1}{SCR/(n-2)} \\ &= \frac{(n-2) \times SCE}{SCR} \\ &= \frac{(n-2) \times SCE}{SCT - SCE} \\ &= \frac{(n-2) \times R^2}{1 - R^2} \\ &= t_{\hat{a}}^2 \end{aligned}$$

Or, concernant la régression linéaire simple (avec constante), le carré du coefficient de corrélation entre Y et X est égal au coefficient de détermination de la régression c.-à-d. $r_{yx}^2 = R^2$ (section 1.3.3). Nous constatons dès lors que :

$$t_{\hat{a}}^2 = \frac{r_{yx}^2}{\frac{1-r_{yx}^2}{n-2}}$$

Qui correspond au carré de la statistique t utilisée pour tester la significativité du coefficient de corrélation linéaire (cf. Rakotomalala, [12], section 2.4, page 16). Les distributions de t et $t_{\hat{a}}$ sont identiques, à savoir un Student à $(n-2)$ degrés de liberté.

Vérification sur les données "Rendements agricoles". Nous avons calculé le coefficient de corrélation entre Y et X précédemment (Figure 1.7), nous avions $r_{yx} = 0.892901$. Formons la statistique pour le test de significativité du coefficient de corrélation :

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.892901}{\sqrt{\frac{1-0.892901^2}{8}}} = 5.60909 = t_{\hat{a}}$$

Nous obtenons effectivement la valeur de $t_{\hat{a}}$ utilisée pour tester la significativité de la pente.

4

Prédiction et intervalle de prédiction

Outre l'analyse structurelle et l'interprétation des coefficients, la régression est beaucoup utilisée pour la prédiction (ou prévision, on utilise plutôt ce terme quand on manipule des données longitudinales). Pour un nouvel individu donné, à partir de la valeur de l'exogène X, nous voulons connaître la valeur que prendrait l'endogène Y.

4.1 Prédiction ponctuelle

Pour un nouvel individu $i*$, qui n'appartient pas à l'échantillon de données ayant participé à l'élaboration du modèle, connaissant la valeur de x_{i*}, on cherche à obtenir la prédiction $\hat{y}_{i*}$. On applique directement l'équation de régression :

$$\begin{aligned}\hat{y}_{i*} &= \hat{y}(x_{i*}) \\ &= \hat{a} \times x_{i*} + \hat{b}\end{aligned}$$

On vérifie facilement que **la prédiction est sans biais** c.-à-d. $E[\hat{y}_{i*}] = y_{i*}$. Pour ce faire, on forme l'erreur de prédiction $\hat{\varepsilon}_{i*} = \hat{y}_{i*} - y_{i*}$ et on montre qu'elle est d'espérance nulle.

Voyons voir :

$$\begin{aligned}\hat{\varepsilon}_{i*} &= \hat{y}_{i*} - y_{i*} \\ &= \hat{a} \times x_{i*} + \hat{b} - y_{i*} \\ &= \hat{a} \times x_{i*} + \hat{b} - (a \times x_{i*} + b + \varepsilon_{i*}) \\ &= (\hat{a} - a)x_{i*} + (\hat{b} - b) - \varepsilon_{i*}\end{aligned}$$

Passons à l'espérance mathématique,

$$\begin{aligned}E\left[\hat{\varepsilon}_{i*}\right] &= E\left[(\hat{a} - a)x_{i*} + (\hat{b} - b) - \varepsilon_{i*}\right] \\ &= x_{i*} \times E(\hat{a} - a) + E(\hat{b} - b) - E(\varepsilon_{i*}) \\ &= 0\end{aligned}$$

Cette espérance est nulle si l'on se réfère aux hypothèses et aux résultats des moindres carrés ordinaires. En effet, les estimateurs $\hat{a}$ et $\hat{b}$ sont sans biais ($E(\hat{a}) = a$ et $E(\hat{b}) = b$), et l'espérance de l'erreur est nulle $E[\varepsilon_{i*}] = 0$. Par conséquent, la prédiction est non biaisée c.-à-d.

$$E[\hat{y}_{i*}] = y_{i*}$$

4.2 Prédiction par intervalle

Une prédiction ponctuelle est intéressante. Mais nous ne savons pas quel degré de confiance nous pouvons lui accorder. Il est donc plus intéressant de s'intéresser à une intervalle de prédiction (fourchette de prédiction) en lui associant une probabilité de recouvrir la vraie valeur y_{i*}.

Pour construire la fourchette, nous avons besoin de connaître d'une part la variance de l'erreur de prédiction et, d'autre part, sa loi de distribution.

4.2.1 Variance de l'erreur de prédiction

Puisque l'erreur de prédiction est non biaisée c.-à-d. $E[\varepsilon_{i*}] = 0$, nous savons que $V(\varepsilon_{i*}) = E[\varepsilon_{i*}^2]$.

Pour calculer la variance, nous devons donc développer ε_{i*}^2 et calculer son espérance (la démarche est détaillée dans Giraud, page 44). Nous obtenons à la sortie la variance de l'erreur de prédiction (Bourbonnais, page 38; Dodge et Rousson, page 36; Johnston, page 35) :

$$\sigma^2_{\hat{\varepsilon}_{i*}} = \sigma^2_{\varepsilon}\left[1 + \frac{1}{n} + \frac{(x_{i*} - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right] \tag{4.1}$$

Estimation. On obtient une estimation $\left(\hat{\sigma}^2_{\hat{\varepsilon}_{i*}}\right)$ de cette variance en introduisant l'estimation de la variance de l'erreur dans la régression $\hat{\sigma}^2_{\varepsilon}$, à savoir :

$$\hat{\sigma}^2_{\hat{\varepsilon}_{i*}} = \hat{\sigma}^2_{\varepsilon}\left[1 + \frac{1}{n} + \frac{(x_{i*} - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right] \tag{4.2}$$

Quelques remarques

La variance sera d'autant plus petite, et par conséquent la fourchette d'autant plus étroite, que :

- $\hat{\sigma}_{\varepsilon}$ est faible, c.-à-d. la régression est de bonne qualité.
- n est élevé c.-à-d. la taille de l'échantillon ayant servi à la construction du modèle est élevé.
- $(x_{i*} - \bar{x})$ est faible c.-à-d. l'observation est proche du centre de gravité du nuage de points (en abcisse, sur l'axe des X). De fait, l'intervalle de prédiction s'évase à mesure que x_{i*} s'éloigne de $\bar{x}$.
- La somme $\sum_i (x_i - \bar{x})^2$ est élevée c.-à-d. la dispersion des points ayant servi à la construction du modèle est grande, ils couvrent bien l'espace de représentation. En réalité, c'est surtout le rapport $\frac{(x_{i*} - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}$ qui joue.

4.2.2 Loi de distribution de l'erreur de prédiction

Pour définir la loi de distribution de l'erreur de prédiction, nous devons nous référer à l'hypothèse de gaussienne du terme d'erreur dans le modèle de régression $\varepsilon_i \equiv \mathcal{N}(0, \sigma_\varepsilon)$. De fait,

$$\frac{\hat{\varepsilon}_{i*}}{\sigma_{\hat{\varepsilon}_{i*}}} = \frac{\hat{y}_{i*} - y_{i*}}{\sigma_{\hat{\varepsilon}_{i*}}} \equiv \mathcal{N}(0,1) \tag{4.3}$$

Lorsque l'on passe à l'estimation de la variance de l'erreur $\hat{\sigma}_\varepsilon^2$, à l'instar de ce que nous avions établi lors de la définition de la distribution des coefficients estimés (section 3.2.3), sachant que $(n-2)\frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} \equiv \chi^2(n-2)$, nous pouvons écrire (remarquez bien l'adjonction du "chapeau" sur le σ) :

$$\frac{\hat{\varepsilon}_{i*}}{\hat{\sigma}_{\hat{\varepsilon}_{i*}}} = \frac{\hat{y}_{i*} - y_{i*}}{\hat{\sigma}_{\hat{\varepsilon}_{i*}}} \equiv \mathcal{T}(n-2) \tag{4.4}$$

4.2.3 Intervalle de prédiction

Nous disposons d'une prédiction non biaisée, de la variance et de la loi de distribution, nous pouvons dès lors définir l'intervalle de prédiction au niveau de confiance $(1-\alpha)$:

$$\hat{y}_{i*} \pm t_{1-\frac{\alpha}{2}} \times \hat{\sigma}_\varepsilon \sqrt{1 + \frac{1}{n} + \frac{(x_{i*} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \tag{4.5}$$

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

4.2.4 Application numérique - Rendements agricoles

Nous désirons construire l'intervalle de prédiction pour l'individu $x_{i*} = 38$ au niveau de confiance $(1-\alpha) = 95\%$. Nous partons des résultats fournis par la fonction DROITEREG d'Excel (Figure 4.1) [1].

Dans un premier temps, nous calculons la prédiction ponctuelle

$$\hat{y}_{i*} = 0.71405 \times 38 + 4.39277 = 31.5$$

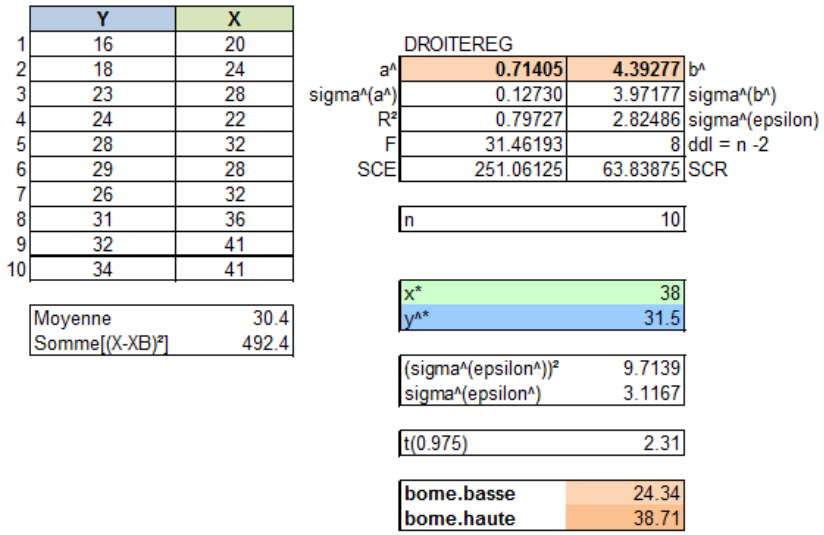
Dans un deuxième temps, nous calculons l'écart-type estimé de l'erreur de prédiction :

- Nous disposons d'un échantillon d'apprentissage avec $n = 10$ observations.
- L'écart-type de l'erreur estimée durant la régression est $\hat{\sigma}_\varepsilon = 2.82486$
- La somme des carrés des écarts à la moyenne de X sur cet échantillon est $\sum_i (x_i - \bar{x})^2 = 492.4$
- L'écartement du point à prédire par rapport à la moyenne des X est $(x_{i*} - \bar{x})^2 = (38 - 30.4)^2 = 57.76$
- Nous déduisons alors l'estimation de l'écart-type de l'erreur

$$\hat{\sigma}_{\hat{\varepsilon}_{i*}} = 2.82486 \times \sqrt{1 + \frac{1}{10} + \frac{57.76}{492.4}} = 3.1167$$

Enfin, pour un intervalle de confiance à 95% :

1. `regression_simple_rendements_agricoles.xlsx` - "prediction"



	Y	X
1	16	20
2	18	24
3	23	28
4	24	22
5	28	32
6	29	28
7	26	32
8	31	36
9	32	41
10	34	41

Moyenne	30.4	
Somme[(X-XB)²]	492.4	

DROITEREG

a^	**0.71405**	**4.39277**	b^
sigma^(a^)	0.12730	3.97177	sigma^(b^)
R²	0.79727	2.82486	sigma^(epsilon)
F	31.46193	8	ddl = n -2
SCE	251.06125	63.83875	SCR

n	10

x*	38
y^*	31.5

(sigma^(epsilon^))²	9.7139
sigma^(epsilon^)	3.1167

t(0.975)	2.31

bome.basse	24.34
bome.haute	38.71

Fig. 4.1. Calculs - Intervalle de prédiction pour ($x_{i*} = 38$) - "Rendements agricoles"

- Nous utilisons le quantile d'ordre $1 - \frac{\alpha}{2}$ de la loi de Student à 8 degrés de liberté, soit $t_{0.975} = 2.31$
- Nous obtenons la borne basse de l'intervalle de prédiction

$$bb(y_{i*}) = 31.5 - 2.31 \times 3.1167 = 24.34$$

- Et la borne haute

$$bh(y_{i*}) = 31.5 + 2.31 \times 3.1167 = 38.71$$

Nous représentons ces informations graphiquement (Figure 4.2). La prédiction ponctuelle est forcément située sur la droite de régression. Ensuite, l'intervalle de prédiction est définie par rapport à l'axe des ordonnées (des Y). Il y a 95% de chances qu'elle couvre la vraie valeur de y_{i*}. On notera que la fourchette est relativement large. Il faut y voir la conjonction de plusieurs éléments défavorables : le point est plutôt éloignée de la moyenne ($\bar{x} = 30.4$, et la valeur max dans l'échantillon est égale à 41) ; l'effectif ayant servi à la construction du modèle est très faible ($n = 10$, on peut difficilement faire quelque chose de bon avec ça) ; et la régression elle-même n'est pas de qualité mirifique (avec un $R^2 = 0.792$).

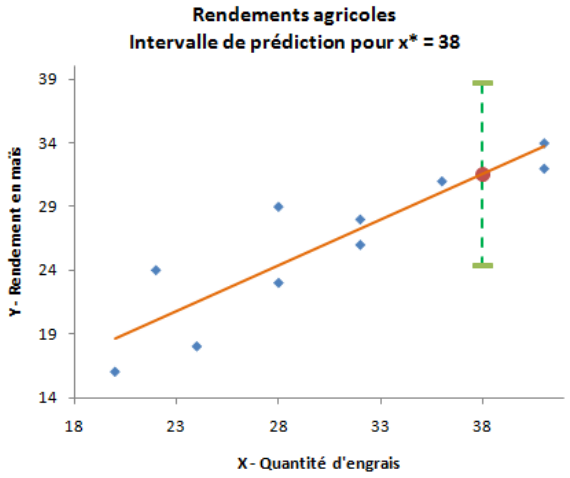


Fig. 4.2. Graphique - Intervalle de prédiction pour ($x_{i*} = 38$) - "Rendements agricoles"

5

Étude de cas - Consommation des véhicules vs. Poids

Récapitulons tous les éléments étudiés jusqu'à présent en réalisant une étude de cas. On souhaite expliquer la consommation des véhicules (en l/100km) (Y) à partir de leur poids (en kg) (X). Nous disposons d'un échantillon de $n = 28$ observations.

Le modèle s'écrit classiquement

$$y_i = ax_i + b + \varepsilon_i$$

Le graphique nuage de points (Figure 5.1) laisse à penser qu'il y a effectivement une relation entre les deux variables. Elle est plutôt positive c.-à-d. lorsque le poids augmente, la consommation a tendance à augmenter également. Sans être un grand expert en automobile, on imagine bien que la causalité est dans ce sens : c'est le poids qui influe sur la consommation, et non l'inverse. On conçoit mal qu'en faisant baisser la consommation par un moyen quelconque, on arriverait par magie à réduire le poids des véhicules.

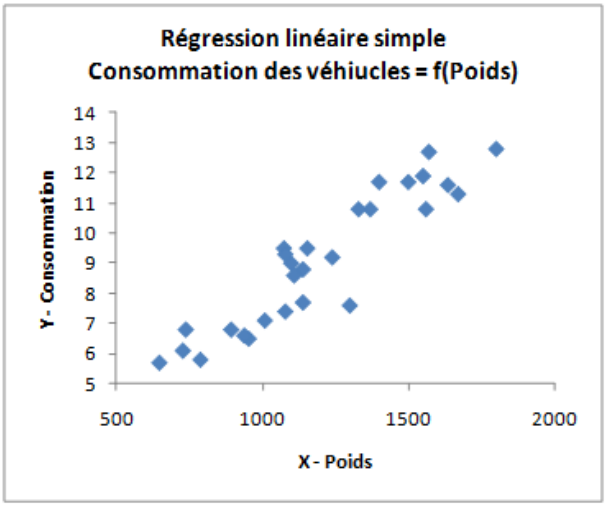


Fig. 5.1. Consommation des véhicules vs. Poids

Nous avons construit la feuille Excel pour la totalité des calculs (Figure 5.2) [1]. Nous énumérons les principaux résultats.

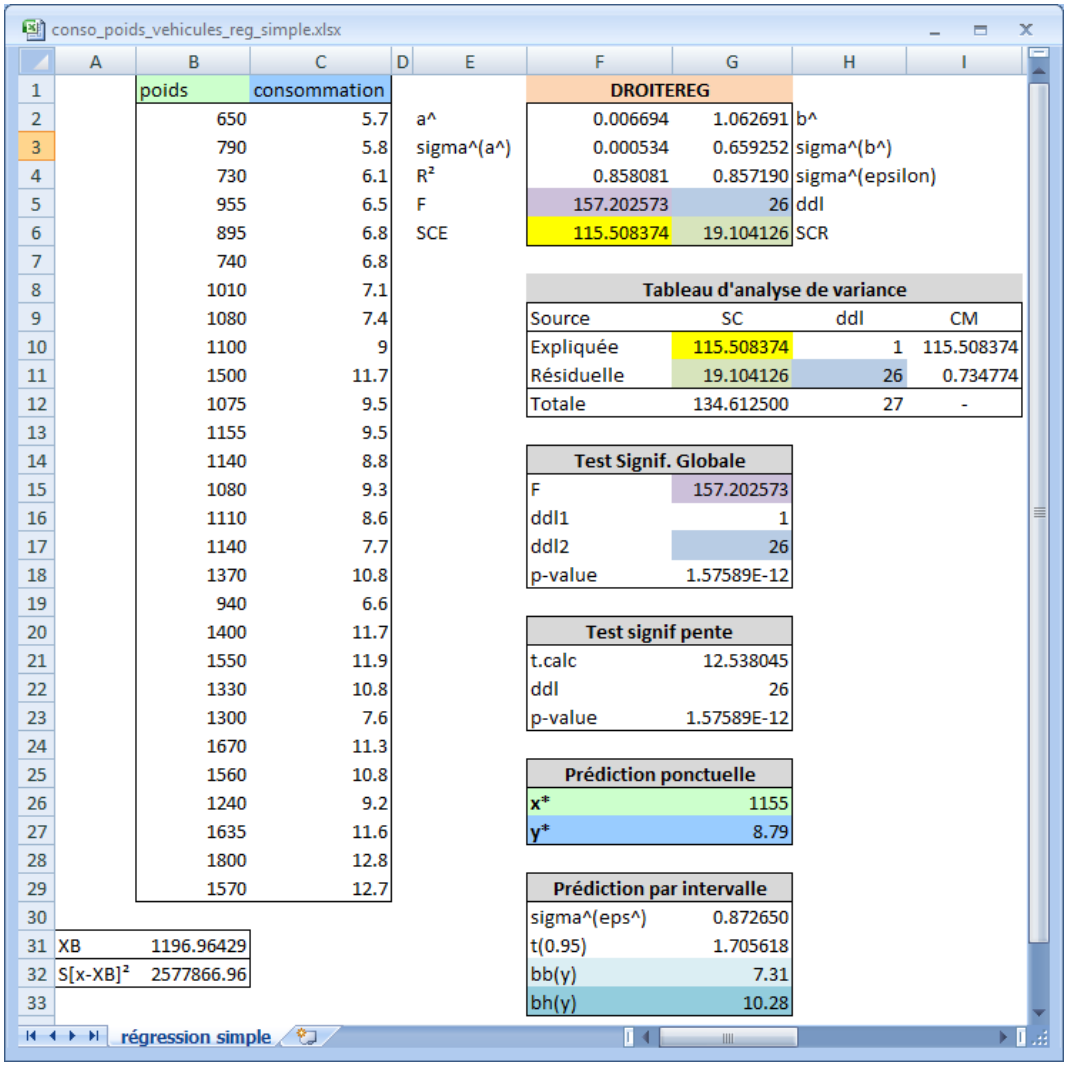


Fig. 5.2. Consommation des véhicules vs. Poids - DROITEREG et calculs subséquents

Coefficients estimés. La fonction DROITEREG nous fournit directement les coefficients estimés

$$\hat{a} = 0.006694$$

$$\hat{b} = 1.062691$$

Tableau d'analyse de variance et coefficient de détermination. DROITEREG nous fournit la $SCE = 115.508374$ et la $SCR = 19.104126$. Nous pouvons en déduire la $SCT = SCE + SCR = 134.612500$ et donc recalculer le coefficient de détermination $R^2 = \frac{SCE}{SCT} = 0.858081$ qui est en réalité

1. `conso_poids_vehicules_reg_simple.xlsx`

directement fourni par Excel. La régression est plutôt de bonne qualité. Ce qui est confirmé par le tracé de la droite de régression au sein du nuage de points (Figure 5.3).

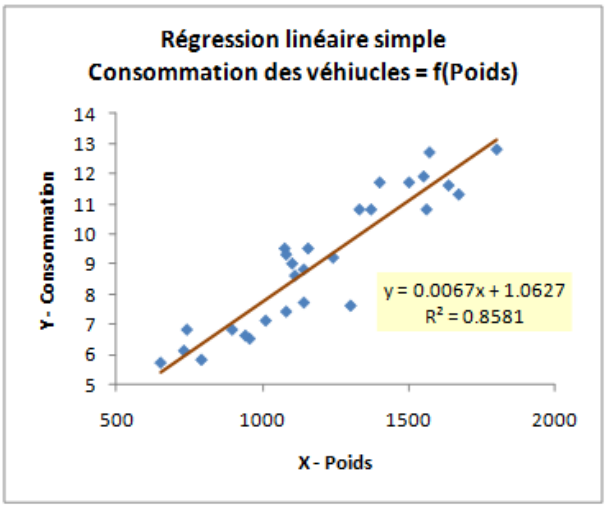


Fig. 5.3. Consommation des véhicules vs. Poids - Tracé de la droite de régression

Test de significativité globale de la régression. La statistique $F = 157.202573$ est aussi fournie. Avec les degrés de libertés adéquates, 1 au numérateur, $n - 2 = 26$ au dénominateur, nous obtenons une probabilité critique très faible (1.57589×10^{-12}). Le modèle est globalement significatif au risque $\alpha = 5\%$.

Test de significativité de la pente. Sans surprise, la pente est aussi significative à 5%. La statistique de test est formée par le rapport de valeurs toutes deux proposée par Excel, $t_{\hat{a}} = \frac{\hat{a}}{\hat{\sigma}_{\hat{a}}} = \frac{0.006694}{0.000534} = 12.538045$. La probabilité critique est identique à celle du test global.

Prédiction ponctuelle et par intervalle. Nous souhaitons prédire la consommation pour un véhicule présentant un poids de $x* = 1155$ kg. Nous calculons la prédiction ponctuelle de la consommation :

$$y* = \hat{a} \times x* + \hat{b} = 0.006694 \times 1155 + 1.062691 = 8.79$$

Pour construire l'intervalle de prédiction, nous avons besoin de l'estimation l'écart-type de l'erreur de prédiction

$$\hat{\sigma}_{\hat{\varepsilon}*} = \hat{\sigma}_{\varepsilon}\sqrt{1 + \frac{1}{n} + \frac{(x* - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}} = 0.857190\sqrt{1 + \frac{1}{28} + \frac{(1195 - 1196.96429)^2}{2577866.96}} = 0.872650$$

Au niveau de confiance 90%, nous prenons le quantile $t_{0.95}(26) = 1.705618$, nous avons ainsi les bornes

$$[8.79 - 1.705618 \times 0.872650 \ ; \ 8.79 + 1.705618 \times 0.872650]$$

$$[7.31 \ ; \ 10.28]$$

Un véhicule pesant 1155 kg a 90% de chances de consommer entre 7.31 et 10.28 litres au 100 km. Nous visualisons la fourchette de prédiction dans le graphique nuage de points (Figure 5.4).

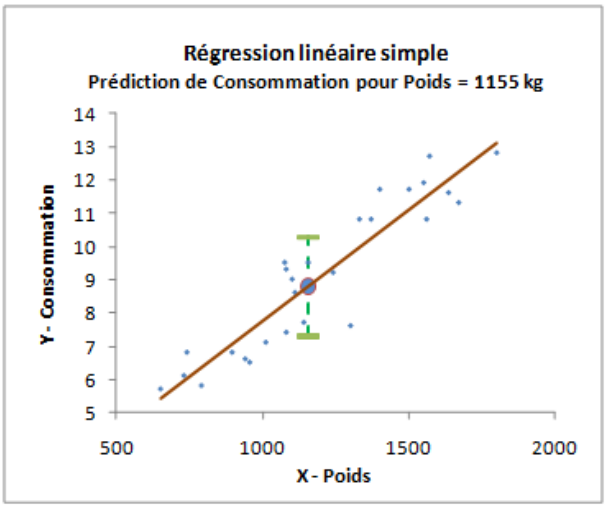


Fig. 5.4. Consommation des véhicules vs. Poids - Intervalle de prédiction

L'amplitude de la fourchette semble visuellement moindre par rapport celle que nous avions calculée pour les rendements agricoles (Figure 4.2). Ce n'est pas qu'une impression. Si on rapporte l'étendue des intervalles à l'écart-type de l'endogène, on se rend compte que le second [consommation = f(poids)] est (presque) deux fois moins large que le premier [rendement = f(engrais)]. Il y a plusieurs raisons à cela : la régression est de meilleure qualité (R^2) ; l'individu à prédire est plus proche du centre de gravité du nuage de points ; la taille n de l'échantillon est plus élevée ; et... nous avons spécifié un niveau de confiance moindre (ah le coquin, la comparaison est forcément avantageuse). Après coup, ce résultat n'est pas étonnant du tout.

6

Non linéarité - Modèles dérivés et interprétation des coefficients

6.1 Interprétation de la droite de régression

On peut lire la régression de 2 manières. La première est une interprétation par niveaux c.-à-d. à une valeur de X, on associe une valeur de Y en appliquant l'équation de régression. Par exemple, dans une équation

$$ventes = -12 \times prix + 1000$$

Lorsque $prix = 10$ euros alors $ventes = 980$ unités.

Mais on peut aussi produire une interprétation selon l'évolution. On se concentre sur la pente de la droite de régression dans ce cas. En effet,

$$\frac{\partial y}{\partial x} = a$$

Dans notre exemple, nous dirons : lorsque le prix augmente d'un euro, les ventes baissent de 12 unités.

Le modèle est linéaire, la variation de Y est proportionnelle à la variation de X. Son principal atout est la simplicité. On l'utilise souvent dans un premier temps pour apprécier l'existence d'une relation (dont on ne cerne pas très bien la nature) entre Y et X. Les paramètres peuvent être estimés directement à l'aide de la méthode des moindres carrés comme nous avons pu le constater dans ce fascicule.

6.2 Modèles non-linéaires mais linéarisables

Parfois nous savons que la liaison n'est pas linéaire, soit parce que nous avons des connaissances expertes sur le problème sur nous traitons, soit parce que nous le constatons visuellement en construisant le nuage de points. Nous sommes alors confrontés à un double problème : déterminer la forme de la liaison, la fonction reliant Y à X ; en estimer les paramètres éventuels à partir des données disponibles. L'affaire est plus que compliquée.

Il existe cependant une classe de fonctions que nous pouvons linéariser en appliquant les transformations adéquates. Dans ce cas, l'estimation des paramètres devient possible. L'interprétation des résultats est modifiée cependant, notamment en ce qui concerne la pente.

Dans cette section, nous allons décrire quelques modèles très utilisés en économétrie.

6.2.1 Modèle log-linéaire - Schéma à élasticité constante

La liaison log-linéaire (dite "transformation log-log" dans Johnston et DiNardo, page 46) est définie de la manière suivante (Figure 6.1)

$$Y = b \times X^a \tag{6.1}$$

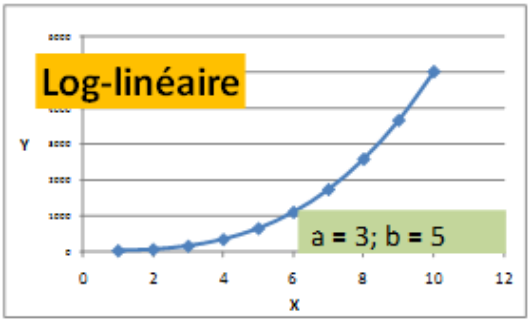


Fig. 6.1. Liaison log-linéaire - $Y = b \times X^a$, $(a = 3, b = 5)$

En termes d'interprétation, le coefficient de la pente est lue de la manière suivante

$$a = \frac{\frac{\partial y}{y}}{\frac{\partial x}{x}} \tag{6.2}$$

Nous avons un modèle à élasticité constante, c'est la favori des économistes [ex. emploi = f(production), demande = f(prix)].

Nous linéarisons en passant par les logarithmes. Nous pouvons ainsi obtenir facilement une estimation des paramètres a et b avec la méthode des MCO.

$$\ln(Y) = \ln(b) + a \times \ln(X) \tag{6.3}$$

6.2.2 Modèle exponentiel (géométrique)

Dans le modèle exponentiel, la relation s'écrit

$$Y = e^{aX+b} \tag{6.4}$$

Le coefficient de la pente se lit

$$a = \frac{\frac{\partial y}{y}}{\partial x} \tag{6.5}$$

Le taux de variation de Y est proportionnelle à la variation de X. Ce type de modèle est surtout utilisé quand X correspond au temps, ainsi $\partial x = 1$. Dans ce cas, la croissance (ou décroissance) de Y

est constante dans le temps. Ce type d'évolution (exponentielle) ne dure pas longtemps (Figure 6.2). On linéarise la relation de la manière suivante

$$\ln(Y) = a \times X + \ln(b) \tag{6.6}$$

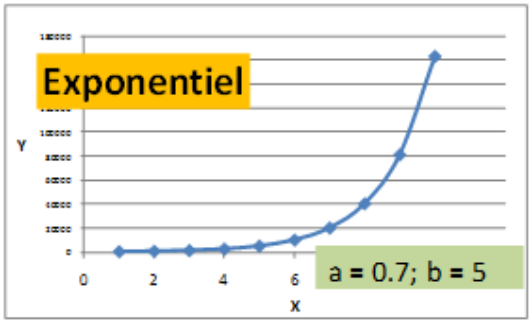


Fig. 6.2. Liaison exponentielle - $Y = e^{aX+b}$, $(a = 0.7, b = 5)$

6.2.3 Modèle logarithmique

Le modèle logarithmique s'écrit

$$Y = a \times \ln(X) + b \tag{6.7}$$

Dans ce cas, la variation de Y est proportionnelle au taux de variation de X c.-à-d.

$$a = \frac{\partial y}{\frac{\partial x}{x}} \tag{6.8}$$

C'est l'archétype de la croissance (ou décroissance) qui s'épuise (Figure 6.3)[ex. salaire = f(ancienneté) ; vente = f(publicité)].

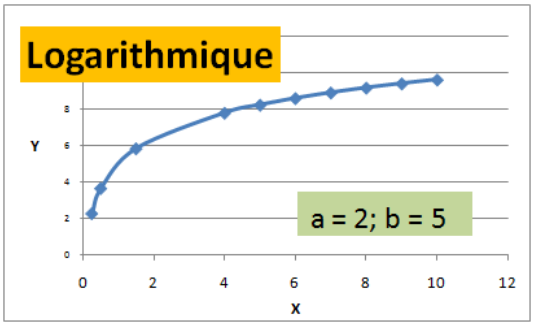


Fig. 6.3. Liaison exponentielle - $Y = a \times \ln(X) + b$, $(a = 2, b = 5)$

6.2.4 Le modèle logistique

Tous les liaisons que nous avons étudiées jusqu'ici sont à concavité constante. Dans certaines situations, nous avons besoin d'une modélisation intégrant plusieurs phases (Figure 6.4). Pour la vente d'un produit dans le temps par exemple, nous distinguons 3 phases : le décollage, le produit est mal connu, les ventes progressent doucement ; la croissance accélérée, le produit connaît une diffusion importante, c'est la période des vaches grasses ; le freinage, les consommateurs se lassent, le marché est saturé, la concurrence a réagi.

Le modèle logistique permet de traduire cette idée, elle s'écrit :

$$Y = y_{min} + \frac{y_{max} - y_{min}}{1 + e^{aX+b}} \tag{6.9}$$

Les valeurs y_{min} et y_{max} peuvent être estimées à partir des données. Mais le plus souvent, elles sont fournies par les connaissances du domaine.

Nous obtenons une forme linéaire dont les paramètres peuvent être estimées par les MCO via l'écriture suivante

$$\ln\left(\frac{y_{max} - Y}{Y - y_{min}}\right) = aX + b \tag{6.10}$$

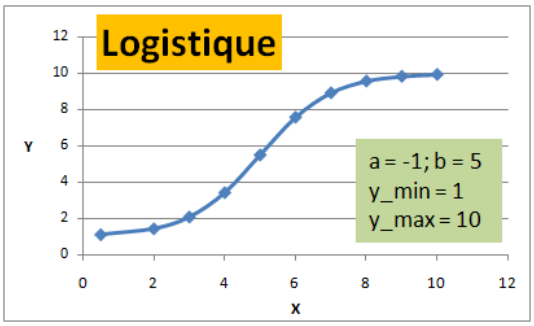


Fig. 6.4. Liaison logistique - $Y = y_{min} + \frac{y_{max} - y_{min}}{1+e^{aX+b}}$, $(a = 2, b = 5, y_{min} = 1, y_{max} = 10)$

Les modèles ci-dessus sont intéressants parce qu'ils correspondent à des phénomènes économiques connus et reconnus. La lecture des résultats, l'analyse des coefficients principalement, est bien cadrée. L'utilisation qui en découle l'est également. C'est leur principal intérêt.

Dans certains cas, nous sommes plus intéressés par les capacités prédictives que par l'interprétation. Nous souhaitons produire le modèle le plus performant possible en termes de proportion de variance expliquée (R^2). La meilleure piste consiste alors à tenter diverses transformations tant sur l'endogène Y que sur l'exogène X. Si l'idée est simple, trouver la solution adéquate est loin d'être évidente tant les possibilités sont innombrables. Nous approfondirons cette piste dans un chapitre dédié de notre second support consacré à la "Pratique de la régression linéaire multiple" ([13], chapitre 6).

6.3 Un exemple de modèle logistique : taux d'équipement en magnétoscope des ménages

Cet exemple est tiré de l'ouvrage de Bourbonnais (pages 160 à 163). Il s'agit de modéliser l'évolution du taux d'équipement en magnétoscope des ménages (Y) sur la période 1979 - 1997. Le temps (X) est la variable explicative. La courbe des points laisse à penser que le modèle logistique semble approprié (Figure 6.5). On notera également que nous sommes dans la phase de freinage en 1997, l'inflexion ayant eu lieu vers (à vue d'oeil) 1989.

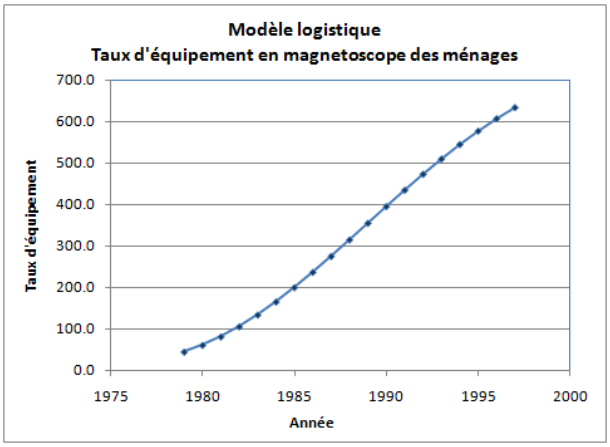


Fig. 6.5. Taux d'équipement en magnétoscope des ménages

L'expression générique du modèle logistique est la suivante :

$$\ln\left(\frac{y_{max}-y}{y-y_{min}}\right)=ax+b$$

Dans notre cas, $y_{min}=0$, le magnétoscope n'existait pas il fut un temps ; et $y_{max}=0.800$ par analogie avec les États-Unis. Ces informations permettent de simplifier le modèle dont il faudra estimer les paramètres a et b

$$\ln\left(\frac{y_{max}}{y}-1\right)=ax+b$$

Dans notre feuille de calcul (Figure 6.6) [1],

- nous construisons la colonne des valeurs $z=\ln\left(\frac{y_{max}}{y}-1\right)$ (ex. $z_1=\ln(\frac{800}{44.7}-1)=2.82714$;
- puis nous estimons les paramètres de $z_i=ax_i+b+\varepsilon_i$.
- Nous obtenons via DROITEREG

$$\begin{cases}\hat{a}=-0.22457\\ \hat{b}=446.98081\end{cases}$$

- La régression est d'excellente qualité avec un $R^2=0.99229$. Elle est bien évidemment globalement significative avec $F=2187.39514$ et une p-value très faible.

1. `equipementmagnetoscope.xlsx` - "régression"

Année (X)	Taux (/1000)	Z
1979	44.7	2.82714
1980	61.0	2.49442
1981	81.3	2.17930
1982	105.8	1.88121
1983	134.0	1.60345
1984	165.6	1.34310
1985	200.1	1.09795
1986	236.9	0.86582
1987	275.4	0.64441
1988	315.0	0.43158
1989	355.2	0.22494
1990	395.3	0.02350
1991	434.8	-0.17444
1992	473.3	-0.37069
1993	510.1	-0.56507
1994	544.9	-0.75895
1995	577.3	-0.95254
1996	607.0	-1.14584
1997	633.9	-1.33930

	a	b	
coef^	-0.22457	446.98081	
sigma^(coef^)	0.00480	9.54551	
R^2	0.99229	0.11464	sigma^(epsilon)
F	2187.39514	17	ddl
SCE	28.74516	0.22340	SCR

Significativité globale

F	2187.39514
ddl1	1
ddl2	17
p-value	2.10229E-19

Signif. des coefficients

t-calculé	-46.76960	46.82627
p-value	2.10229E-19	2.05976E-19

Fig. 6.6. Taux d'équipement en magnétoscope des ménages - DROITEREG

- Les deux paramètres a et b sont significatifs.
- Représentée dans le graphique, nous constatons que la courbe d'évolution du taux d'équipement est plutôt bien reconstituée (Figure 6.7). Ce n'est guère étonnant avec un R^2 aussi élevé.

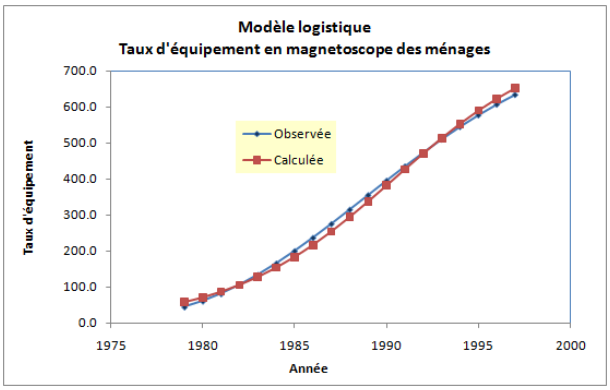


Fig. 6.7. Taux d'équipement en magnétoscope des ménages - Courbes observée et estimée

Essayons de voir quel serait le taux d'équipement en 1998 ? Pour ce faire, nous appliquons directement le modèle pour obtenir $\hat{z}_{1998}$,

$$\hat{z}_{1998} = 0.22457 \times 1998 + 446.98081 = -1.7030$$

Puis nous appliquons la transformation inverse.

$$\hat{y}_{1998} = \frac{y_{max}}{1 + e^{\hat{z}_{1998}}} = \frac{800}{1 + e^{-1.7030}} = 676.74$$

L'autre solution aurait été d'utiliser directement le modèle sous sa forme originelle :

$$\hat{y}_{1998} = y_{min} + \frac{y_{max} - y_{min}}{1 + e^{ax+b}} = 0 + \frac{800 - 0}{1 + e^{-0.22457 \times 1998 + 446.98081}} = 676.74$$

Si on veut produire une fourchette de prédiction, la première solution est préférable. Nous calculons tout d'abord l'intervalle de prédiction pour z_{1998}, puis nous appliquons la transformation inverse sur les bornes pour obtenir la fourchette pour y_{1998}.

Estimation de y_{max}. Dernier point avant de conclure cette section, nous avions considéré $y_{max} = 800$ comme acquise dans notre démarche. Elle était le fruit d'une information exogène au processus modélisation (en référence à une autre population).

En réalité, nous pouvons également intégrer son estimation dans les calculs. Bourbonnais (page 162) décrit une procédure de balayage : elle tente plusieurs valeurs probables comprises entre 680 et 990 (des valeurs crédibles bien évidemment, il ne s'agit pas de tester n'importe quoi), la valeur sélectionnée est celle qui minimise la SCR du modèle final. Avec le logiciel Rats, il obtient sur notre exemple la valeur de $\hat{y}_{max} = 710$ [2].

Nous avons voulu réitérer la même expérimentation en utilisation **la table de simulation à deux entrées d'Excel** [3] (nous n'utilisons qu'une seule entrée en l'occurrence). y_{max} est devenu un paramètre dans la feuille de calcul, utilisé pour construire la variable intermédiaire z. Pour chaque valeur de y_{max} allant de 680 à 990 avec un pas de 10, Excel a relancé Droitereg et nous avons collecté la somme des carrés des résidus de la régression. Au final, la valeur qui minimise la SCR ($SCR = 0.08892$) est bien $\hat{y}_{max} = 710$ (Figure 6.8) [4].

2. La forme qu'il utilise est un peu différente de la notre, elle s'écrit $y = \frac{y_{max}}{1+b \times a^x}$. Mais cela ne modifie pas la nature du modèle.

3. Voilà pourquoi j'adore les tableurs. Avec un peu de réflexion et trois clics, on peut mener des analyses assez complexes. La feuille Excel est autrement plus simple que le code source rapporté dans Bourbonnais (page 162), pourtant particulièrement limpide si on sait un tant soit peu coder (une boucle DO avec un condition à l'intérieur). Mais c'est le genre de choses à faire fuir les étudiants pourtant friands de statistique mais réfractaires à toute idée de programmation.

4. `equipementmagnetoscope.xlsx` - "estimation y.max"

	Y_max	710
Année (X)	Taux (/1000)	Z
1979	44.7	2.70026
1980	61.0	2.36456
1981	81.3	2.04551
1982	105.8	1.74235
1983	134.0	1.45827
1984	165.6	1.19011
1985	200.1	0.93540
1986	236.9	0.69167
1987	275.4	0.45620
1988	315.0	0.22631
1989	355.2	-0.00113
1990	395.3	-0.22803
1991	434.8	-0.45739
1992	473.3	-0.69294
1993	510.1	-0.93679
1994	544.9	-1.19405
1995	577.3	-1.47027
1996	607.0	-1.77380
1997	633.9	-2.11984

	a	b	
coef\	-0.25386	504.93023	
sigma^(coef^)	0.00303	6.02232	
R^2	0.99759	0.07232	
F	7022.50109	17	ddl
SCE	36.73312	0.08892	SCR

Y_MAX	SCR
680	0.20198
690	0.13270
700	0.10025
710	0.08892
720	0.09021
730	0.09917
740	0.11283
750	0.12930
760	0.14737
770	0.16626
780	0.18544
790	0.20456
800	0.22340
810	0.24181
820	0.25968
830	0.27697
840	0.29364
850	0.30969
860	0.32510
870	0.33990
880	0.35409
890	0.36769
900	0.38073
910	0.39323
920	0.40521
930	0.41669
940	0.42770
950	0.43826
960	0.44839
970	0.45811
980	0.46745
990	0.47641

MIN SCR	0.08892

Fig. 6.8. Taux d'équipement en magnétoscope des ménages - Détection de la valeur "optimale" de y_{max}

7

Régression sans constante

Jusqu'à présent dans tous les exemples décrits dans ce support, nous n'avions jamais tenté de tester la significativité de la constate. La raison est que nous serions bien embêtés si elle s'avérait non significative. En effet, la supprimer de l'équation de régression modifie (un peu beaucoup) la nature de l'affaire. Le modèle s'écrit

$$y_i = ax_i + \varepsilon_i \tag{7.1}$$

Nous devons faire face à plusieurs phénomènes :

- Nous introduisons une contrainte dans la régression. La droite passe forcément par l'origine c.-à-d. lorsque $x = 0$, $\hat{y}(0) = 0$. Et, sauf cas particulier des données centrées que nous aborderons plus bas (section 7.1), elle ne passe pas forcément par le barycentre $G(\bar{x}, \bar{y})$ du nuage de points.
- La décomposition de la variance telle que nous l'avons décrite précédemment (équation 1.9) n'est plus valable. La tableau d'analyse de variance n'a plus de sens. Le coefficient de détermination R^2 ne peut plus être lue en termes de proportion de variance expliquée par la régression. Il peut même prendre *des valeurs négatives*. C'est très gênant pour un indicateur qui présente un carré dans son expression.
- La pente de la régression peut être interprétée d'une autre manière. Elle représente directement le rapport entre les variables c.-à-d. $a = \frac{Y}{X}$. Nous exploiterons cette propriété dans l'exemple que nous détaillerons dans la section 7.2. La lecture en termes de rapport de variation reste valable cependant.

7.1 Cas des données centrées

Dans le cas des données centrées, on montre que la constante de la régression est par construction égale à zéro. En effet, posons $\dot{y}_i = y_i - \bar{y}$ et $\dot{x}_i = x_i - \bar{x}$, l'estimation de la constante s'écrit

$$\hat{b} = \bar{\dot{y}} - \hat{a} \times \bar{\dot{x}}$$

Or, par définition $\bar{\dot{y}} = \bar{\dot{x}} = 0$. On constate facilement que $\hat{b} = 0$.

C'était logique dans la mesure où l'on sait que la droite de régression passe toujours par le centre de gravité des points. Lorsque les données sont centrées, le barycentre est le point de coordonnées $(0, 0)$, il est normal donc qu'elle passe par l'origine sur Y et sur X.

Régression sur "Rendements agricoles" - Données centrées. Nous reprenons notre feuille de calcul des "Rendements agricoles". Nous avons centré les données à l'aide des moyennes empiriques $\bar{y} = 26.1$ et $\bar{x} = 30.4$. Nous avons construit le nuage de points puis, à l'aide de l'outil "Courbe de tendance" d'Excel, nous avons tracé la droite de régression (Figure 7.1) [1]. Elle passe bien par l'origine du repère, la constante estimée $\hat{b} = 0$. Par rapport aux résultats obtenus dans la régression avec constante (section 1.2.2), nous remarquons que la pente de la droite n'est pas modifiée, $\hat{a} = 0.71405$.

	Y	X
	16	20
	18	24
	23	28
	24	22
	28	32
	29	28
	26	32
	31	36
	32	41
	34	41
Moyenne	26.1	30.4

Y°	X°
-10.1	-10.4
-8.1	-6.4
-3.1	-2.4
-2.1	-8.4
1.9	1.6
2.9	-2.4
-0.1	1.6
4.9	5.6
5.9	10.6
7.9	10.6

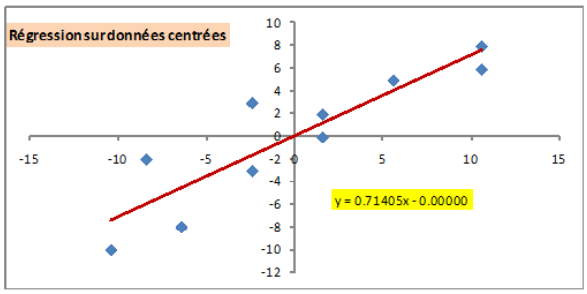


Fig. 7.1. Régression sur données centrées - Rendements agricoles

7.2 Cas des données quelconques

7.2.1 Problématique

Dans le cas des données quelconques, pas forcément centrées, la contrainte faisant passer la droite par l'origine modifie l'estimation de la pente. Reprenons notre exemple des "Rendements agricoles" avec les données originelles. Nous réalisons une régression sans constate, la pente devient $\hat{a} = 0.85124$ (Figure 7.2 ; nuage de points, courbe de tendance et résultats de la fonction DROITEREG), différente de celle de la régression avec constante.

De manière générale, la régression sans constante, du fait de l'introduction d'une contrainte supplémentaire dans la construction du modèle, est moins performante en termes de SCR c.-à-d. $SCR_{(ax)} \leq SCR_{(ax+b)}$. Lorsqu'elle est totalement inadaptée, sa SCR peut même être supérieure à la SCT. Le modèle est moins bon que la simple prédiction à l'aide de la moyenne de l'endogène. D'où la possibilité d'obtenir des coefficients de détermination R^2 négatifs. C'est la raison pour laquelle nous avons hachuré le R^2 fourni par Excel dans les sorties de DROITEREG (Figure 7.2).

1. `regression_sans_constante.xlsx` - "rendements agricoles"

DROITEREG			
a^	0.85124	0	
sigma^(a^)	0.02898	#N/A	
R^2	0.98967	2.85968	se
F	862.50864	9	ddl
SCE	7053.40004	73.59996	SCR

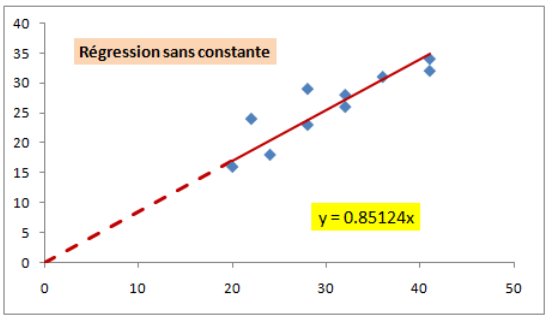


Fig. 7.2. Régression sans constante - Rendements agricoles

Dans notre exemple des Rendements agricoles, nous avons $SCR_{(ax)} = 73.59996$ (Figure 7.2) contre $SCR_{(ax+b)} = 63.83875$ (Figure 3.5).

Le second point important est le calcul des degrés de libertés. Nous n'estimons plus qu'un seul paramètre dans la régression, il est donc égal à $(n-1)$ [nous avons $(n-1=9)$ **ddl** pour l'exemple des Rendements agricoles, figure 7.2]. Il faudra en tenir compte lors de la mise en oeuvre des tests d'hypothèses.

7.2.2 Formules

Les férus de calculs pourront aisément reproduire la démarche des moindres carrés ordinaires pour obtenir $\hat{a}$. Nous donnons directement les principaux résultats sans démonstration dans cette section.

L'estimateur des MCO de la pente de la régression sans constante s'écrit

$$\hat{a} = \frac{\sum_i y_i x_i}{\sum_i x_i^2} \tag{7.2}$$

On remarque l'analogie avec l'estimateur de la pente pour la régression avec constante, surtout en tenant compte du fait que la droite passe forcément par l'origine.

L'estimateur de la variance de l'erreur doit tenir compte des degrés de liberté, c.-à-d.

$$\hat{\sigma}_\varepsilon^2 = \frac{SCR}{n-1} \tag{7.3}$$

Et l'estimation de la variance de la pente estimée devient

$$\hat{\sigma}_{\hat{a}}^2 = \frac{\hat{\sigma}_\varepsilon^2}{\sum_i x_i^2} \tag{7.4}$$

Enfin, la quantité

$$\frac{\hat{a}-1}{\hat{\sigma}_{\hat{a}}} \equiv \mathcal{T}(n-1) \tag{7.5}$$

Suit une loi de Student à $(n-1)$ degrés de liberté.

Dans la régression sans constante également, plus que jamais puisqu'il n'y a qu'un seul paramètre dans le modèle, tester la significativité de la pente équivaut à tester la significativité globale de la régression.

7.3 Un exemple d'application : comparaison de salaires

Nous avons une régression qui introduit une contrainte supplémentaire et qui s'avère être moins performante (en termes de SCR). Quel est l'intérêt de ce type d'approche ? La réponse la plus convaincante je pense est la possibilité d'élargir le spectre des analyses que nous pouvons mener à l'aide de la régression. Voyons un exemple pour donner un tour concret à notre discours.

Nous étudions un échantillon de $n = 50$ ménages composés de couples hommes-femmes actifs. Nous connaissons leurs salaires respectifs. Nous souhaitons montrer qu'en moyenne le salaire de l'élément masculin du ménage est supérieur à celui de l'élément féminin. Nous avions déjà étudié ce fichier dans un de nos supports [2], nous avions utilisé alors une comparaison de moyennes pour échantillons appariés. Il s'est avéré que l'hypothèse nulle d'égalité des salaires a été rejetée au risque 5%. Le même problème aurait pu être traité avec une approche non paramétrique d'ailleurs. Le test des signes par exemple [3], la conclusion est identique.

Comment faire avec la régression ? Nous utilisons la régression sans constante pour réaliser la comparaison. Si Y est le salaire de l'homme, X celui de la femme, le rapport $\frac{Y}{X} = a$ devrait être supérieur à 1. Nous modélisons la relation avec

$$y_i = ax_i + \varepsilon_i$$

Et nous mettons en oeuvre le test d'hypothèses au risque $\alpha = 5\%$

$$\begin{cases} H_0 : a = 1 \\ H_1 : a > 1 \end{cases}$$

Nous utilisons la statistique :

$$t_{(a>1)} = \frac{\hat{a} - 1}{\hat{\sigma}_{\hat{a}}}$$

La région critique du test est définie pour les valeurs "anormalement" élevées de $\hat{a}$ par rapport à 1 :

$$R.C. : t_{(a>1)} > t_{1-\alpha} \tag{7.6}$$

Le test est unilatéral, nous comparons la statistique avec la valeur critique $t_{1-\alpha}$.

La fonction DROITEREG [4] nous fournit $\hat{a} = 1.02083$, avec un écart-type estimé $\hat{\sigma}_{\hat{a}} = 0.00547$ (Figure 7.3) [5]. La statistique de test est donc égal à

$$t_{(a>1)} = \frac{\hat{a} - 1}{\hat{\sigma}_{\hat{a}}} = \frac{1.02083 - 1}{0.00547} = 3.80528$$

2. Rakotomalala, *Comparaison de populations - Tests paramétriques*, chapitre 4, `http://eric.univ-lyon2.fr/~ricco/cours/cours/Comp_Pop_Tests_Parametriques.pdf`.

3. Rakotomalala, *Comparaison de populations - Tests non paramétriques*, chapitre 6, `http://eric.univ-lyon2.fr/~ricco/cours/cours/Comp_Pop_Tests_Nonparametriques.pdf`.

4. `regression_sans_constante.xlsx` - "salaire H.F dans les ménages"

5. Contrairement à ce que laisse croire le graphique, la droite de régression passe bien par l'origine $(0, 0)$.

Que nous comparons au seuil critique fournie par la loi de Student à $(n-1=49)$ degrés de liberté, $t_{0.95}(49) = 1.67655$. Nous nous situons dans la région critique. Les données confirment l'idée selon laquelle le salaire de l'homme a tendance à être supérieur à celui de sa conjointe au sein des ménages.

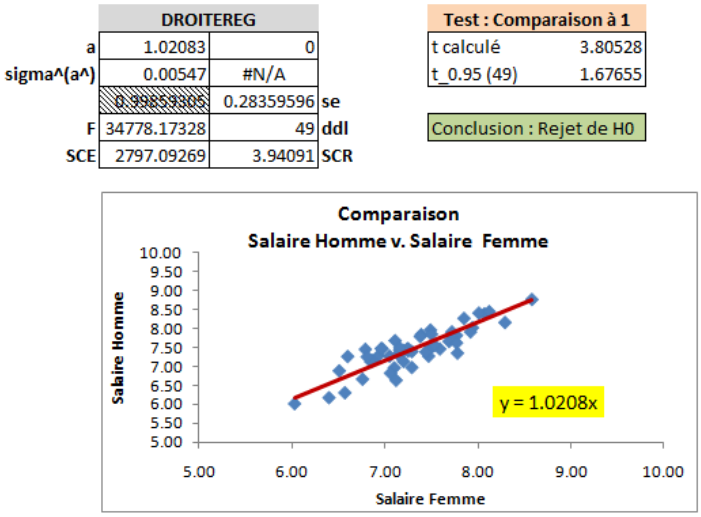


Fig. 7.3. Comparaison des salaires H/F via la régression sans constante

8

Comparaison des régressions

L'objectif de la comparaison des régressions est de vérifier que la liaison existant entre X et Y est de la même nature dans différentes sous-populations.

Prenons un exemple simple dont nous détaillerons l'analyse plus loin (section 8.5.1). On pense que le montant du salaire mensuel des employés est fonction de leur niveau d'études. Cela semble logique : plus la personne est qualifiée, plus élevée sera sa rémunération. Mais est-ce que la liaison est la même chez les hommes et chez les femmes ? Valorise-t-on de la même manière la qualification ? Dans cette configuration, la variable endogène Y est le salaire ; le nombre d'années d'études est l'explicative X ; les sous-populations sont définies par la sexe Z, avec $(K = 2)$ groupes.

Dans ce chapitre, même si nos exemples porterons sur le cas particulier de $(K = 2)$ groupes pour faciliter les interprétations, l'exposé et les formules seront valables pour un nombre quelconque de sous-populations $(K \geq 2)$.

Cette configuration n'est pas sans rappeler un autre type de problème que nous avons étudié dans notre second polycopié [13] (chapitre 5). Nous y abordons la comparaison de modèles sous l'angle de la rupture de structure dans la régression multiple. Nous cherchons à savoir dans un premier temps si, dans deux sous-périodes (ou deux sous-populations), la relation entre les exogènes et l'endogène est la même. Dans un deuxième temps, nous essayons de détecter la source de la différence, si elle existe évidemment.

L'idée est la même dans ce chapitre. Sauf que nous nous plaçons dans le cadre de la régression simple et que nous pouvons traiter un nombre quelconque de groupes.

Ainsi, dans les exemples que nous détaillerons dans ce chapitre : régression simple et comparaison de $K = 2$ groupes, les deux approches sont applicables. C'est le genre de situations que j'apprécie tout particulièrement. Nous disposons de deux prismes différents pour traiter le même problème. A priori, les approches devraient converger. C'est ce que nous ne manquerons pas de vérifier bien évidemment.

Ce chapitre doit beaucoup à Aïvazian (pages 151 à 156, [1]), Dagnelie (pages 486 à 494, [5]) et Scherrer (pages 713 à 717, [16]).

8.1 Comparaison des régressions dans leur globalité

8.1.1 Principe du test

La première étape consiste à vérifier si les deux régressions simples sont globalement identiques dans les K groupes. Si l'hypothèse d'égalité est rejetée, nous essayerons de détecter la nature de la différence (la pente ou la constante) dans la section suivante.

Le test d'hypothèses oppose : (H_0) l'égalité des coefficients dans les sous-populations ; contre (H_1), les coefficients sont différents dans au moins un des groupes. Il repose sur une confrontation entre plusieurs régressions.

1. Dans un premier temps, nous réalisons la "régression contrainte" sous H_0, elle considère que les coefficients sont les mêmes quels que soient les groupes. Dans ce cas, on procède à la modélisation sur la totalité des n observations :

   $$y_i = ax_i + b + \varepsilon_i \ , \ i = 1, \ldots, n.$$

   A partir de cette droite, nous calculons la somme des carrés des résidus SCR_T.

2. Dans un deuxième temps, nous réalisons les "régressions non contraintes", hors H_0 c.-à-d. pour les K groupes, nous calculons les paramètres (a_k, b_k) du modèle sur des échantillons de taille n_k :

   $$y_{i,k} = a_k x_{i,k} + b_k + \varepsilon_{i,k} \ , \ i = 1, \ldots, n_k, \ k = 1, \ldots, K.$$

   Pour chaque régression nous avons la somme des carrés des résidus SCR_k. Nous formons la somme

   $$SCR_W = \sum_{k=1}^{K} SCR_k$$

   Qui correspond en quelque sorte à la somme des carrés des résidus intra-groupes.

Ayant retirer la contrainte d'égalité des coefficients dans les groupes pour les secondes régressions, nous sommes certains de la propriété suivante

$$SCR_W \leq SCR_T$$

Toute la problématique revient alors à poser la question : est-ce que l'écart est suffisamment important pour qu'il ne soit pas imputable aux simples fluctuations d'échantillonnage ? Auquel cas, la contrainte d'égalité des coefficients dans les groupes (H_0) est trop forte, inappropriée.

On devine aisément que la statistique de test est basée sur l'opposition entre les SCR, elle s'écrit :

$$F = \frac{(SCR_T - SCR_W)/(2(K-1))}{SCR_W/(n-2K)} \tag{8.1}$$

Un petit mot sur les degrés de liberté. Au dénominateur nous avons :

$$\sum_k (n_k - 2) = \sum_k n_k - 2K$$
$$= n - 2K$$

Et au numérateur :

$$(n-2) - (n-2K) = 2K - 2$$
$$= 2(K-1)$$

La démarche est totalement cohérente avec les tests sur les changements structurels dans la régression linéaire multiple que nous exposons par ailleurs [13] (chapitre 5).

Sous H_0, F suit une loi de Fisher à $[2(K-1), n-2K]$ degrés de liberté. La région critique au risque α est définie pour les valeurs exceptionnellement grandes de F

$$R.C. : F > F_{1-\alpha}[2(K-1), n-2K]$$

8.1.2 Un exemple numérique

Nous reprenons l'exemple décrit dans Johnston et DiNardo (page 135) utilisé pour illustrer le test de Chow pour les changements structurels. Il correspond à des données longitudinales, les sous-groupes sont en réalité des périodes. Mais qu'importe, cela n'affecte pas l'applicabilité du test. Le principal intérêt pour nous est de vérifier que les résultats sont identiques même si les prismes utilisés sont différents.

Obs	Groupes	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.	0.524	-0.070
	0.03	0.37
	0.95	0.71
	252.71	13
	127.44	**6.5561**
		SCR_T

Régression groupe 1

	X	const.
coef.	0.438	-0.063
	0.05	0.43
	0.96	0.48
	66.82	3
	15.31	**0.6875**
		SCR1

Régression groupe 2

	X	const.
coef.	0.509	0.400
	0.03	0.38
	0.97	0.56
	276.71	8
	85.53	**2.4727**
		SCR2

SCR_W	**3.1602**
ddl_n	2
ddl_d	11
SCR_T - SCR_W	3.3959
F	**5.9101**
p-value	0.0181

Fig. 8.1. Comparaison des régressions dans des sous-populations

Nous avons $K = 2$ groupes, avec $n_1 = 5$ et $n_2 = 10$. Nous avons construit le modèle sur la totalité des données ("Régression globale") et dans les sous-populations ("Régression groupe k") (Figure 8.1) [1] :

1. `comparaisondesregressions.xls` - "comp.groupes"

- Sur la totalité de l'échantillon, nous obtenons le modèle :

$$y_i = 0.524x_i - 0.070,\ SCR_T = 6.5561$$

- Sur le premier groupe, nous avons

$$y_i = 0.438x_i - 0.063,\ SCR_1 = 0.6875$$

- Et sur le second

$$y_i = 0.509x_i + 0.400,\ SCR_2 = 2.4727$$

- Nous calculons la SCR intra-groupes

$$SCR_W = SCR_1 + SCR_2 = 0.6875 + 2.4727 = 3.1602$$

- Il ne nous reste plus qu'à former la statistique de test

$$F = \frac{(6.5561 - 3.1602)/(2(2-1))}{3.1602/(15 - 2 \times 2)} = 5.9101$$

- Avec une loi $\mathcal{F}(2, 11)$, nous avons une probabilité critique de 0.0181
- Au risque $\alpha = 5\%$, nous pouvons rejeter l'hypothèse d'égalité des régression dans les sous-groupes.
- Ce résultat n'est guère étonnant si l'on considère le nuage des points (X, Y) mettant en exergue l'appartenance aux groupes (Figure 8.2).

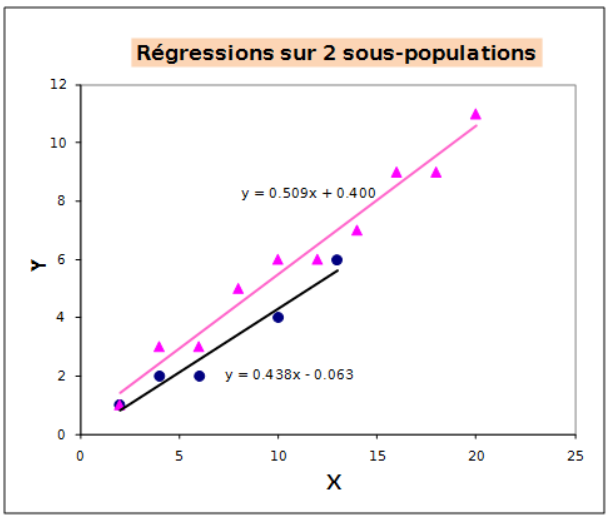


Fig. 8.2. Comparaison des régressions dans des sous-populations - Nuage de points

Reste à détecter maintenant la nature de la différence. On le devine un peu (beaucoup) à la lumière du nuage de points. Mais c'est quand même mieux lorsque l'intuition est confirmée par les calculs statistiques.

8.2 Détecter la nature de la différence

8.2.1 Différences entre les pentes

Les hypothèses à confronter s'écrivent :

$$\begin{cases} H_0 : a_1 = a_2 = \cdots = a_K = 0 \\ H_1 : \exists k, k' \texttt{ tel que } a_k \neq a_{k'} \end{cases}$$

Pour répondre à la question, nous devons calculer l'estimation commune aux K groupes de la pente de la droite de régression :

$$\hat{a}_c = \frac{\sum_{k=1}^{K}(n_k - 1)s_{yx,k}}{\sum_{k=1}^{K}(n_k - 1)s_{x,k}^2} \tag{8.2}$$

Nous nous servons d'une séries de statistiques définies dans les sous-échantillons de taille n_k relatifs aux K groupes :

- $s_{yx,k} = \frac{1}{n_k - 1}\sum_{i=1}^{n_k}(y_i - \bar{y}_k)(x_i - \bar{x}_k)$ est la covariance entre Y et X dans le groupe k.
- $\bar{y}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} y_i$ (resp. $\bar{x}_k$) est la moyenne de Y (resp. X) dans le groupe k.
- $s_{x,k}^2 = \frac{1}{n_k - 1}\sum_{i=1}^{n_k}(x_i - \bar{x}_k)^2$ (resp. $s_{y,k}^2$) est la variance estimée de X (resp. Y) dans le groupe k.

On déduit une somme des carrés des résidus associés aux K droites parallèles :

$$SCR_C = \sum_{k=1}^{K}(n_k - 1)s_{y,k}^2 - \hat{a}_c^2 \sum_{k=1}^{K}(n_k - 1)s_{x,k}^2 \tag{8.3}$$

La contrainte de "parallélisme" des droites, exprimée à travers une estimation commune de la pente $\hat{a}_c$, font que $SCR_C \geq SCR_W$ (issu des estimations séparées dans la groupes, sans contraintes). La question est : est-ce que l'écart est suffisamment significatif ? Auquel cas, l'hypothèse d'égalité des pentes ne tiendrait pas la route.

A partir de cette idée, on propose la statistique de test suivante :

$$F = \frac{(SCR_C - SCR_W)/(K - 1)}{SCR_W/(n - 2K)} \tag{8.4}$$

Sous H_0 (égalité des pentes), elle suit une loi de Fisher à $(K - 1, n - 2K)$ degrés de liberté. La région critique correspond aux fortes valeurs de F.

Application numérique

Revenons sur notre exemple (section 8.1.2). Nous avions conclu que les régressions étaient différentes dans les $K = 2$ groupes. Mais nous n'avions pas déterminé le paramètre (pente ou constante) responsable de cette différence. Nous allons vérifier maintenant le rôle de la pente.

A partir des données et des résultats des précédentes régressions (Figure 8.1), nous calculons les nouveaux indicateurs nécessaires au test (Figure 8.3) [2] :

2. `comparaisondesregressions.xls` - "comp.groupes"

– Nous calculons les covariances et variances conditionnelles

$$
\begin{aligned}
s_{yx,1} &= \frac{1}{n_1-1}\sum_{i=1}^{n_1}(y_i-\bar{y}_1)(x_i-\bar{x}_1) = \frac{1}{5-1}\times 35 = 8.75\\
s_{yx,2} &= \frac{1}{9}\times 168 = 18.6667\\
s^2_{x,1} &= \frac{1}{n_1-1}\sum_{i=1}^{n_1}(x_i-\bar{x}_1)^2 = \frac{1}{4}\times 80 = 20.0\\
s^2_{x,2} &= \frac{1}{9}\times 330 = 36.6667\\
s^2_{y,1} &= 4.0\\
s^2_{y,2} &= 9.7778
\end{aligned}
$$

– La pente commune aux régressions conditionnelles est obtenue avec

$$a_c = \frac{\sum_{k=1}^{K}(n_k-1)s_{yx,k}}{\sum_{k=1}^{K}(n_k-1)s^2_{x,k}} = \frac{4\times 8.75 + 9\times 18.6667}{4\times 20 + 9\times 36.6667} = 0.4951$$

– Nous en tirons la SCR_C, l'erreur résiduelle associée aux K droites parallèles

$$SCR_C = (4\times 4.0 + 9\times 9.7778) - 0.4951\times(4\times 20.0 + 9\times 36.6667) = 3.4902$$

– La statistique de test est basée sur l'écart entre cette quantité et la somme des erreurs résiduelles des régressions conditionnelles (SCR_W)

$$F = \frac{(SCR_C - SCR_W)/(K-1)}{SCR_W/(n-2K)} = \frac{(3.4902-3.1602)/(2-1)}{3.1602/(15-2\times 2)} = 1.1487$$

– Avec un $\mathcal{F}(1,11)$, nous avons une probabilité critique de 0.3068.

– Au risque 5%, la différence entre les régressions n'est pas imputable à une inégalité des pentes.

Test d'égalité des pentes

Covariances (Y,X)

Groupe 1	8.7500
Groupe 2	18.6667

Variances X

Groupe 1	20.0000
Groupe 2	36.6667

Variances Y

Groupe 1	4.0000
Groupe 2	9.7778

Pente commune

a^_c	0.4951

Erreur résiduelle pentes parallèles

SCR_C	3.4902

Ecart Erreur résiduelles

SCR_C	3.4902
SCR_W	3.1602
Ecart	0.3300

Comparaison des pentes

F-calculé	1.1487
ddl1	1
ddl2	11
p-value	0.3068

Fig. 8.3. Comparaison des pentes des régressions conditionnelles

8.2.2 Différences entre les constantes

Si l'égalité entre les pentes est établie, les divergences (si divergences il y a) seraient alors imputables aux constantes des régressions.

Pour les comparer, il suffit de confronter la somme des carrés des résidus de la régression opérée sur la totalité des données (SCR_T) et celle obtenue à partir de l'estimation commune des pentes (SCR_C). De nouveau, si la différence est trop forte, elle serait due ici à un décalage entre les constantes des régressions (Scherrer, page 715) :

$$F = \frac{(SCR_T - SCR_C)/(K-1)}{SCR_C/(n-2K)} \tag{8.5}$$

Sous H_0, $F \equiv \mathcal{F}(K-1, n-2K)$. La région critique correspond aux valeurs élevées de F.

Application numérique

Toujours sur notre exemple (section 8.1.2), l'égalité entre les pentes a été établie dans la section précédente. Voyons maintenant ce qu'il en est concernant les constantes. Tous les éléments intermédiaires sont déjà prêts (Figures 8.1 et 8.3), il ne nous reste plus qu'à calculer la statistique de test (Figure 8.4) [3] :

$$F = \frac{(SCR_T - SCR_C)/(K-1)}{SCR_C/(n-2K)} \frac{(6.5561 - 3.4902)/(2-1)}{3.4902/(15-2\times 2)} = 10.6716$$

Test d'égalité des constantes

SCR_T	6.5561
SCR_C	3.4902

Comparaison des constantes

F-calculé	10.6716
ddl1	1
ddl2	11
p-value	0.007509

Fig. 8.4. Comparaison des constantes des régressions conditionnelles

Avec un $F\mathcal{F}(1, 11)$, la probabilité critique est $\alpha' = 0.007509$, en deçà de notre risque $\alpha = 5\%$.

Conclusion : l'écart entre les régressions est due à une disparité entre les constantes.

*Remarque 3 (**Différence avec le test de Chow**).* Dans notre polycopié sur la pratique de régression, sur les mêmes données, en comparant les constantes dans les sous-groupes, nous obtenons certes la même conclusion mais avec des valeurs numériques légèrement différentes [13] (chapitre 5, section 5.2.1). Après avoir étudié de près la question, la divergence s'explique essentiellement par la comptabilisation des degrés de liberté. Dans le test de Chow (traité dans Johnston et DiNardo, pages 134 et 135), nous estimons directement la pente sur la totalité des données, le degré de liberté dans la régression non contrainte est

3. `comparaisondesregressions.xls` - "comp.groupes"

égale à $n-3=12$ (3 parce que 2 constantes et 1 pente commune). Dans la procédure que nous décrivons ici, nous tirons les résultats à partir des régressions opérées sur les sous groupes, les degrés de liberté deviennent $n-4=11$ (4 parce que 2 constantes et 2 pentes). Si les SCR sont identiques, le degré de liberté au dénominateur qui entre dans le calcul de F et de la probabilité critique n'est pas le même.

8.3 Un récapitulatif des différentes SCR

Récapitulons les différentes sommes des carrés résiduels pour bien situer leur positionnement :

- SCR_T, nous réalisons la régression sur la totalité des données, nous posons la contrainte d'égalité des paramètres à la fois sur la pente et sur la constante.
- SCR_C, la contrainte d'égalité des pentes d'un groupe à l'autre est posée, les constantes en revanche sont laissées libres. De fait, l'écart $(SCR_T - SCR_C)$ permet de vérifier si l'hypothèse d'égalité des constantes dans les groupes est licite ou non.
- SCR_W, les contraintes d'égalité, tant sur la pente que sur la constante, sont relâchées. De fait, le passage $(SCR_C - SCR_W)$ permet d'éprouver l'hypothèse d'égalité des pentes, sachant que nous laissons libres les constantes.
- Enfin, la différence $(SCR_T - SCR_W)$ permet simplement de tester l'existence d'une différence entre les régressions dans les sous-populations, quel qu'en soit la nature.

Une manière simple de comprendre le test d'égalité des modèles dans les sous-populations consiste donc à opposer les sommes des carrés résiduels des régressions sur lesquelles nous posons différents types de contraintes d'égalité des coefficients. Les écarts permettent de mettre en évidence le paramètre (pente ou constante ou les deux) à l'origine des divergences, si elles existent bien évidemment.

8.4 Le cas particulier de $K = 2$ groupes

Dans le cas de deux groupes, Aïvazian (pages 151 à 156) propose une procédure qui s'apparente au test paramétrique de comparaison de moyennes. Rappelons-en le principe : nous vérifions dans un premier temps que les variances conditionnelles sont identiques. Si c'est le cas, nous calculons une estimation commune de la variance, et nous procédons au très connu test de Student de comparaison de moyennes. Si les variances sont différentes, on utilise le test (moins connu) d'Aspin-Welch [4].

Dans le cas de régression, le schéma est analogue sauf que (1) nous vérifions l'égalité des variances de l'erreur de la régression dans les groupes ; (2) et ce sont les coefficients du modèle, en particulier la pente, que nous comparons par la suite.

4. Rakotomalala R., *Comparaison de populations - Tests paramétriques*, chapitres 1 et 2, `http://eric.univ-lyon2.fr/~ricco/cours/cours/Comp_Pop_Tests_Parametriques.pdf`

8.4.1 Tester l'égalité des variances de l'erreur dans les 2 groupes

Après les régressions dans les 2 groupes, nous obtenons une estimation des variances des erreurs $\left(\hat{\sigma}^2_{\varepsilon,k}\right)$. Si les variances sont identiques, leur rapport doit être égal à 1 ; s'il s'en écarte significativement, la disparité va au-delà des fluctuations d'échantillonnage, elles sont différentes dans les sous-groupes.

Nous utilisons la statistique de test suivante :

$$\nu^2 = \frac{\hat{\sigma}^2_{\varepsilon,1}}{\hat{\sigma}^2_{\varepsilon,2}} \tag{8.6}$$

Sous H_0, égalité des régressions dans les 2 sous-populations, ν^2 suit une loi de Fisher $\mathcal{F}(n_1-2, n_2-2)$. La région critique au risque α est située sur les valeurs anormalement faibles ou anormalement élevée par rapport à l'unité c.à-d.

$$R.C. : (\nu^2 < F_{\alpha/2}) \text{ ou } (\nu^2 > F_{1-\alpha/2}) \tag{8.7}$$

Cette procédure n'est pas sans rappeler le test de Fisher de comparaison de variances de deux sous-populations. Elle est séduisante par son principe, on peut faire le rapprochement avec des techniques que l'on connaît bien. Mais elle en partage également les défauts, à savoir une très faible robustesse par rapport à un écart à l'hypothèse de normalité des données (des résidus en l'occurrence).

8.4.2 Comparaison des coefficients - Cas des variances identiques

Si l'hypothèse d'égalité des variances résiduelles conditionnelles est confirmée, nous pouvons passer à une estimation de la variance commune, une sorte de variance intra-classes en quelque sorte.

$$s^2_\varepsilon = \frac{(n_1-2)\hat{\sigma}^2_{\varepsilon,1} + (n_2-2)\hat{\sigma}^2_{\varepsilon,2}}{n_1+n_2-4} \tag{8.8}$$

Munis de cette estimation, nous pouvons procéder aux comparaisons de coefficients.

Comparaison des pentes

Nous opposons les deux pentes

$$\begin{cases} H_0 : a_1 = a_2 \\ H_1 : a_1 \neq a_2 \end{cases}$$

La statistique de test est formée par la différence entre les coefficients estimés, soit

$$D_a = \hat{a}_1 - \hat{a}_2 \tag{8.9}$$

Dont l'estimation de l'écart-type est obtenu avec

$$\hat{\sigma}_{D_a} = s_\varepsilon \times \sqrt{\frac{1}{(n_1-1)s^2_{x,1}} + \frac{1}{(n_2-1)s^2_{x,2}}} \tag{8.10}$$

Sous H_0, D suit une loi de Student à $(n1 + n_2 - 4)$ degrés de liberté. La région critique au risque α, conduisant au rejet de l'hypothèse de l'égalité des pentes, est définie par :

$$R.C : \frac{|D_a|}{\hat{\sigma}_{D_a}} \geq t_{1-\alpha/2}(n_1+, n_2 - 4) \tag{8.11}$$

Comparaison des constantes

Si l'égalité des pentes est établie, nous passons à la comparaison des constantes. Curieusement, nous n'utilisons pas directement les coefficients estimés $\hat{b}_1$ et $\hat{b}_2$. Pour réaliser le test, nous opposons deux estimations de la pente. La première correspond à l'estimation conjointe de la pente dans les deux sous-populations (c'est un cas particulier de la pente commune pour K groupes, équation 8.2) :

$$\hat{a}_c = \frac{(n_1 - 1)s^2_{x,1}\hat{a}_1 + (n_2 - 1)s^2_{x,2}\hat{a}_2}{(n_1 - 1)s^2_{x,1} + (n_2 - 1)s^2_{x,2}} \tag{8.12}$$

Et la seconde, l'estimation de la pente sous l'hypothèse nulle d'égalité des constantes :

$$\hat{a}_0 = \frac{\bar{y}_1 - \bar{y}_2}{\bar{x}_1 - \bar{x}_2} \tag{8.13}$$

Soit $(D_b = \hat{a}_c - \hat{a}_0)$ l'écart entre ces deux valeurs, son écart-type est égal à

$$\hat{\sigma}_{D_b} = s_\varepsilon \times \sqrt{\frac{1}{(n_1 - 1)s^2_{x,1} + (n_2 - 1)s^2_{x,2}} + \frac{\frac{1}{n_1} + \frac{1}{n_2}}{(\bar{x}_1 - \bar{x}_2)^2}} \tag{8.14}$$

Et la région critique au risque α devient

$$R.C. : \frac{|D_b|}{\hat{\sigma}_{D_b}} \geq t_{1-\alpha/2}(n_1 + n_2 - 4) \tag{8.15}$$

8.4.3 Comparaison des coefficients - Cas des variances différentes

Lorsque les variances des erreurs sont différents dans les groupes, l'affaire devient nettement plus compliquée. Nous n'avons que des résultats asymptotiques, de mauvaise qualité sur les petits effectifs, mais qui deviendront de plus en plus précis à mesure que la taille des échantillons augmente.

Comparaison des pentes

Les variances des erreurs $\sigma^2_{\varepsilon,1}$ et $\sigma^2_{\varepsilon,2}$ sont différentes. Une nouvelle estimation de l'écart-type de la différence D_a entre les pentes est produite :

$$s_{D_a} = \sqrt{\frac{\hat{\sigma}^2_{\varepsilon,1}}{(n_1 - 1)s^2_{x,1}} + \frac{\hat{\sigma}^2_{\varepsilon,2}}{(n_1 - 1)s^2_{x,2}}} \tag{8.16}$$

La région critique devient :

$$R.C. : \frac{|D_a|}{s_{D_a}} \geq t_{1-\alpha/2}(l) \tag{8.17}$$

A l'instar du test d'Aspin-Welch pour la comparaison de moyennes, la difficulté réside dans le calcul des degrés de liberté. La formule est particulièrement tarabiscotée (Aïvazian, page 153) [5] :

$$l = \left[\frac{C^2}{n_1 - 2} + \frac{(1-C)^2}{n_2 - 2} \right]^{-1}$$

où

$$C = \frac{\frac{\hat{\sigma}^2_{\varepsilon,1}}{(n_1-1)s^2_{x,1}}}{\frac{\hat{\sigma}^2_{\varepsilon,1}}{(n_1-1)s^2_{x,1}} + \frac{\hat{\sigma}^2_{\varepsilon,2}}{(n_2-1)s^2_{x,2}}}$$

Comparaison des constantes

Comme pour le cas des variances résiduelles égales, si l'égalité des pentes est établie, nous vérifions l'égalité des constantes b_1 et b_2. La procédure repose toujours sur une confrontation entre deux estimations de la pente.

L'estimation de la pente sous H_0 reste la même, à savoir

$$\hat{a}_0 = \frac{\bar{y}_1 - \bar{y}_2}{\bar{x}_1 - \bar{x}_2}$$

En revanche, l'estimation conjointe de la pente doit tenir du fait que les variances des erreurs sont différentes dans les groupes :

$$\hat{a}_{c'} = \frac{\hat{a}_1 \frac{(n_1-1)s^2_{x,1}}{\hat{\sigma}^2_{\varepsilon,1}} + \hat{a}_2 \frac{(n_2-1)s^2_{x,2}}{\hat{\sigma}^2_{\varepsilon,2}}}{\frac{(n_1-1)s^2_{x,1}}{\hat{\sigma}^2_{\varepsilon,1}} + \frac{(n_2-1)s^2_{x,2}}{\hat{\sigma}^2_{\varepsilon,2}}} \tag{8.18}$$

Nous rejetons l'hypothèse d'égalité des constantes au risque α si

$$R.C. : |\hat{a}_{c'} - \hat{a}_0| \geq u_{1-\alpha/2} \times \sqrt{\frac{n_2\hat{\sigma}^2_{\varepsilon,1} + n_1\hat{\sigma}^2_{\varepsilon,2}}{n_1 n_2 (\bar{x}_1 - \bar{x}_2)^2} + \frac{\hat{\sigma}^2_{\varepsilon,1}\hat{\sigma}^2_{\varepsilon,2}}{n_1 s^2_{x,1}\hat{\sigma}^2_{\varepsilon,2} + n_2 s^2_{x,2}\hat{\sigma}^2_{\varepsilon,1}}} \tag{8.19}$$

Il s'agit bien d'une procédure approximative, nous utilisons la loi normale : $u_{1-\alpha/2}$ correspond au quantile de la loi normale centrée et réduite.

8.4.4 Application numérique

Reprenons notre exemple de la section précédente (section 8.1.2).

Nous désirons dans un premier temps **vérifier l'égalité des variances des erreurs conditionnellement aux groupes**. Nous modifions la feuille Excel de manière à obtenir la statistique de test (Figure 8.5) [6] :

5. NDA : J'ai du vérifier 20 fois les écritures. J'espère seulement ne pas avoir introduit des erreurs en recopiant les équations, particulièrement alambiquées il faut dire. Malheureusement, je n'ai pas trouvé d'autres références bibliographiques pour croiser les formules, comme je le fais habituellement. Et la procédure n'est implémentée nulle part, je n'ai pas pu contrôler non plus sur des jeux de données... Bon, on retiendra surtout et avant tout l'idée qu'il est possible de procéder à des comparaisons des paramètres des modèles dans le cas où les variances des erreurs sont différentes. Les formulations sont un peu plus compliquées simplement.

6. `comparaisondesregressions.xls` - "comp.2.groupes"

Obs	Groupes	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 groupe 1	
X	const.
0.438	-0.063
0.05	0.43
0.96	**0.4787**
66.82	3
15.31	0.6875

coef.

Régression groupe 2	
X	const.
0.509	0.400
0.03	0.38
0.97	**0.5560**
276.71	8
85.53	2.4727

coef.

V1	0.2292
V2	0.3091
Rapport	**0.7414**
ddl1	3
ddl2	8
Seuil.bas	0.0688
Seuil.haut	5.4160

Conclusion : Variances des erreurs égales dans les deux groupes

Fig. 8.5. Comparaison des variances des erreurs des régressions dans 2 sous-populations

- Pour la première régression, DROITEREG fournit $\hat{\sigma}_{\varepsilon,1} = 0.4787$
- Pour la seconde, $\hat{\sigma}_{\varepsilon,2} = 0.5560$
- Nous formons le rapport de leurs carrés

$$\nu^2 = \frac{\hat{\sigma}^2_{\varepsilon,1}}{\hat{\sigma}^2_{\varepsilon,2}} = \frac{0.4787^2}{0.5560^2} = \frac{0.2292}{0.3091} = 0.7414$$

- Les valeurs délimitant la région critique au risque 5% sont

$$F_{0.025}(3,8) = 0.0688$$
$$F_{0.975}(3,8) = 5.4160$$

- Nous ne sommes pas dans la région critique (équation 8.7), l'hypothèse nulle d'égalité des variances de l'erreur dans les deux groupes ne peut être rejetée.

A partir de là, nous pouvons produire une estimation de la variance commune de l'erreur dans les deux régressions

$$s^2_\varepsilon = \frac{(n_1-2)\hat{\sigma}^2_{\varepsilon,1} + (n_2-2)\hat{\sigma}^2_{\varepsilon,2}}{n_1+n_2-4} = \frac{4\times 0.2292 + 9\times 0.3091}{5+10-4} = 0.2873$$

Comparaison des pentes. Pour comparer les pentes, nous calculons leur différence (Figure 8.6) [7]

$$D_a = \hat{a}_1 - \hat{a}_2 = 0.4375 - 0.5091 = -0.0716$$

Et son écart-type

$$\hat{\sigma}_{D_a} = s_\varepsilon \times \sqrt{\frac{1}{(n_1-1)s^2_{x,1}} + \frac{1}{(n_2-1)s^2_{x,2}}} = \sqrt{0.2873} \times \sqrt{\frac{1}{(5-1)\times 20} + \frac{1}{(10-1)\times 36.6667}} = 0.0668$$

Nous formons le rapport

7. `comparaisondesregressions.xls` - "comp.2.groupes"

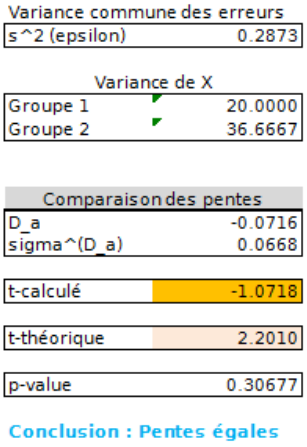


Fig. 8.6. Comparaison des pentes des régressions dans 2 sous-populations

$$t = \frac{D_a}{\hat{\sigma}_{D_a}} = \frac{-0.0716}{0.0668} = -1.0718$$

Puisque $|t| = 1.0718 < 2.2010 = t_{0.975}(11)$ au risque $\alpha = 5\%$, nous ne pouvons pas rejeter l'hypothèse selon laquelle les pentes sont identiques. La probabilité critique est $\alpha' = 0.30677$. Elle est exactement la même que celle produite par le test des pentes valable pour $K \geq 2$ groupes décrit dans la section précédente (Figure 8.3). D'ailleurs, concernant les statistiques de test, nous constatons également que $t^2 = (-1.0718)^2 = 1.1487 = F$.

C'est plutôt rassurant. Les deux approches, l'une valable pour un nombre quelconque de groupes ($K \geq 2$), l'autre spécifique au traitement de ($K = 2$) sous-populations, fournissent des résultats identiques lorsque l'on traite la situation ($K = 2$).

Comparaison des constantes. L'égalité des pentes étant établie, on s'interroge maintenant sur les différences entre les constantes (Figure 8.7) [8]. Tout d'abord, nous calculons la pente commune aux droites

$$\hat{a}_c = \frac{(n_1 - 1)s_{x,1}^2\hat{a}_1 + (n_2 - 1)s_{x,2}^2\hat{a}_2}{(n_1 - 1)s_{x,1}^2 + (n_2 - 1)s_{x,2}^2} = \frac{(5-1) \times 20 \times 0.4375 + (10-1) \times 36.6667 \times 0.5091}{(5-1) \times 20 + (10-1) \times 36.6667} = 0.4951$$

Puis la pente dans le cas où l'hypothèse nulle d'égalité des constantes serait vraie

$$\hat{a}_0 = \frac{\bar{y}_1 - \bar{y}_2}{\bar{x}_1 - \bar{x}_2} = \frac{3.0 - 6.0}{7.0 - 11.0} = 0.75$$

Nous calculons la statistique de test

$$D_b = \hat{a}_c - \hat{a}_0 = 0.4951 - 0.7500 = -0.2549$$

Et son écart-type

8. `comparaisondesregressions.xls` - "comp.2.groupes"

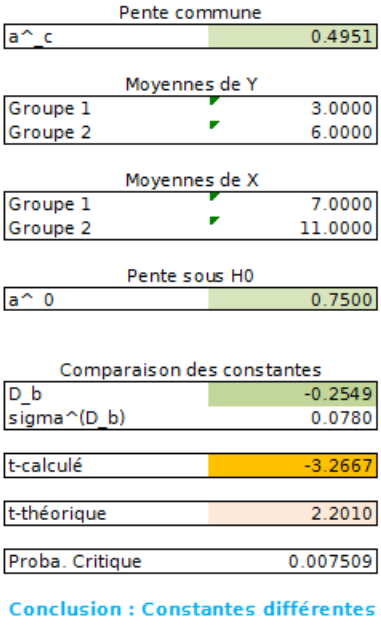


Fig. 8.7. Comparaison des constantes des régressions dans 2 sous-populations

$$\begin{aligned}\hat{\sigma}_{D_b} &= s_\varepsilon \times \sqrt{\frac{1}{(n_1-1)s_{x,1}^2+(n_2-1)s_{x,2}^2}+\frac{\frac{1}{n_1}+\frac{1}{n_2}}{(\bar{x}_1-\bar{x}_2)^2}} \\ &= \sqrt{0.2873} \times \sqrt{\frac{1}{(5-1)20+(10-1)36.6667}+\frac{\frac{1}{5}+\frac{1}{10}}{(7.0-11.0)^2}} \\ &= 0.0780\end{aligned}$$

Nous comparons la valeur absolue du rapport

$$t = \frac{D_b}{\hat{\sigma}_{D_b}} = \frac{-0.2549}{0.0780} = -3.2667$$

Avec le seuil critique au risque $\alpha = 5\%$, $t_{0.975}(11) = 2.2010$. Comme $|t| > t_{0.975}(11)$, nous rejetons l'hypothèse d'égalité des constantes. La probabilité critique est égale à $\alpha' = 0.007509$. Ici aussi, le résultat est complètement cohérent $[t^2 = (-3.2667)^2 = 10.6716 = F]$ avec l'approche générique pour un nombre de groupes quelconques (Figure 8.4).

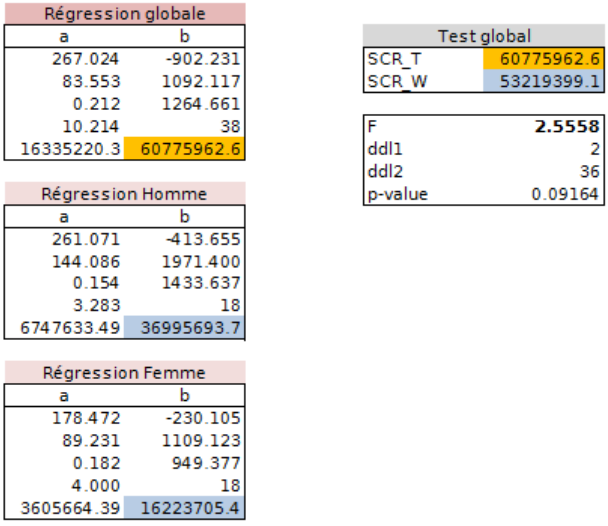
8.5 Deux études de cas

8.5.1 Le salaire selon le niveau d'études

Nous souhaitons expliquer le salaire (Y) des individus à partir de leur niveau d'études (X). Une qualification d'autant plus élevée devrait induire une rémunération plus élevée. Après cette première étape, nous souhaitons savoir si la relation est la même chez les hommes et chez les femmes. Ou bien y a-t-il une disparité ? Et si c'est le cas, de quelle nature serait-elle ?

Nous utiliserons ces mêmes données dans un autre contexte (régression sur des exogènes qualitatives) plus loin dans ce fascicule. Le fichier provient du site `http://www.cabannes.net/`.

Tester la différence globale. Nous disposons de $n = 40$ observations, dont $n_1 = 20$ hommes et $n_2 = 20$ femmes. Nous sommes en présence de $K = 2$ groupes. Les effectifs étant relativement faibles, nous réaliserons nos tests à $\alpha = 10\%$.



Régression globale

a	b
267.024	-902.231
83.553	1092.117
0.212	1264.661
10.214	38
16335220.3	60775962.6

Régression Homme

a	b
261.071	-413.655
144.086	1971.400
0.154	1433.637
3.283	18
6747633.49	36995693.7

Régression Femme

a	b
178.472	-230.105
89.231	1109.123
0.182	949.377
4.000	18
3605664.39	16223705.4

Test global

SCR_T	60775962.6
SCR_W	53219399.1

F	**2.5558**
ddl1	2
ddl2	36
p-value	0.09164

Fig. 8.8. Comparaison des régressions - Salaire = f(années d'études) / sexe

– La régression sur la totalité des données indique (Figure 8.8) [9] :

$$y = 267.024x - 902.231, \ SCR_T = 60775962.6$$

– Chez les hommes, nous avons

$$y = 261.071x - 413.655, \ SCR_1 = 36995693.7$$

– Et chez les femmes,

$$y = 178.472x - 230.105, \ SCR_2 = 16223705.4$$

– La somme des erreurs résiduelles intra-groupes est égale à

$$SCR_W = SCR_1 + SCR_2 = 36995693.7 + 16223705.4 = 53219399.1$$

– Nous formons la statistique de test

$$F = \frac{(SCR_T - SCR_W)/(2(K-1))}{SCR_W/(n-2K)} = \frac{(60775962.6 - 53219399.1)/(2 \times (2-1))}{53219399.1/(40 - 2 \times 2)} = 2.5558$$

9. `comparaisondesregressions.xls` - "salaires-ed-sexe"

- Avec la distribution $\mathcal{F}(2 \times (2-1) = 2,40 - 2 \times 2 = 36)$, nous avons une probabilité critique de $\alpha' = 0.09164$. Au risque $\alpha = 10\%$, nous pouvons considérer que les régressions sont différentes c.-à-d. la liaison entre les années d'études et le salaire n'est pas la même selon le sexe de l'employé.
- Visuellement, les nuages de points et les courbes de tendance associées confirment cette conclusion (Figure 8.9).

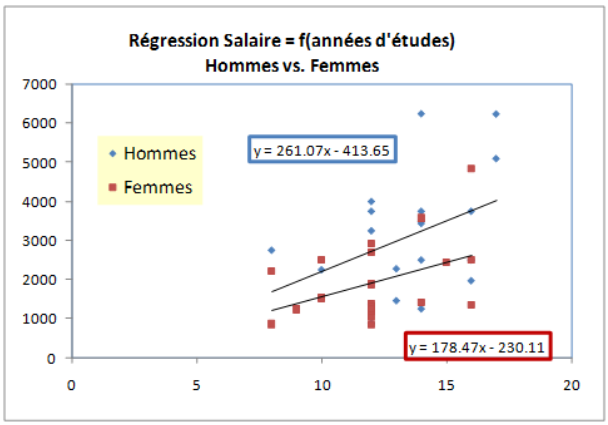


Fig. 8.9. Comparaison des régressions - Nuages de points - Salaire = f(années d'études) / sexe

Comparaison des pentes	
Covariance (y,x)	
H	1360.3158
F	1063.3158
Variance (x)	
H	5.2105
F	5.9579
Variance (y)	
H	2302280.379
F	1043651.039
Pente commune	
a^_c	217.0075
Erreur résiduelle	
SCR_C	53579716.74
F	**0.2437**
ddl1	1
ddl2	36
p-value	**0.6245**

Comparaison des constantes	
F	**4.8351**
ddl1	1
ddl2	36
p-value	**0.0344**

Fig. 8.10. Comparaison des pentes et des constantes - Salaire = f(années d'études) / sexe

A quel paramètre alors serait imputable cette divergence ? Penchons-nous sur le rôle de la pente.

Tester la différence entre les pentes. Pour élaborer le test, nous avons besoin des covariances et des variances de Y et X, conditionnellement aux groupes (Figure 8.10) [10] :

10. `comparaisondesregressions.xls` - "salaires-ed-sexe"

$$
\begin{aligned}
s_{yx,1} &= \frac{1}{n_1 - 1}\sum_{i=1}^{20}(y_i - \bar{y}_1)(x_i - \bar{x}_1) = 1360.3158 \\
s_{yx,2} &= 1063.3158 \\
s^2_{x,1} &= 5.2105 \\
s^2_{x,2} &= 5.9579 \\
s^2_{y,1} &= 2302280.379 \\
s^2_{y,2} &= 1043651.039
\end{aligned}
$$

Nous pouvons en extraire la pente commune

$$\hat{a}_c = \frac{\sum_{k=1}^{K}(n_k - 1)s_{yx,k}}{\sum_{k=1}^{K}(n_k - 1)s^2_{x,k}} = \frac{19 \times 1360.3158 + 19 \times 1063.3158}{19 \times 5.2105 + 19 \times 5.9579} = 217.0075$$

Et la somme des erreurs résiduelles des $K = 2$ droites parallèles

$$
\begin{aligned}
SCR_C &= \sum_{k=1}^{K}(n_k - 1)s^2_{y,k} - \hat{a}^2_c \sum_{k=1}^{K}(n_k - 1)s^2_{x,k} \\
&= (19 \times 2302280.379 + 19 \times 1043651.039) - (217.0075)^2 \times (19 \times 5.2105 + 19 \times 5.9579) \\
&= 53579716.74
\end{aligned}
$$

Il ne reste plus qu'à former la statistique de test

$$F = \frac{(SCR_C - SCR_W)/(K-1)}{SCR_W/(n-2K)} = \frac{(53579716.74 - 53219399.1)/(2-1)}{53219399.1/(40 - 2 \times 2)} = 0.2437$$

Avec un $\mathcal{F}(1, 36)$, nous avons une p-value de $\alpha' = 0.6245$. Les données ne contredisent pas l'hypothèse d'égalité des pentes des deux régressions.

Tester la différence entre les constantes. Si les pentes sont censées être identiques (*hum, ça ne paraît pas très évident sur le graphique nuage de points, on y reviendra plus loin...*), voyons ce qu'il en est concernant les constantes (Figure 8.9).

Nous disposons de tous les éléments nécessaires au calcul déjà, il ne reste plus qu'à former la statistique de test

$$F = \frac{(SCR_T - SCR_C)/(K-1)}{SCR_C/(n-2K)} = \frac{(60775962.6 - 53579716.74)/(2-1)}{53579716.74/(40 - 2 \times 2)} = 4.8351$$

Avec un $\mathcal{F}(1, 36)$, nous avons une p-value de $\alpha' = 0.0344$. Au risque 10%, nous concluons à une différence significative des constantes. La divergence constatée globalement est essentiellement due à un décalage sur l'axe des ordonnées entre les droites de régression

Conclusion : L'évolution des salaires selon la qualification est la même chez les hommes et chez les femmes. En revanche, il y a une différence intrinsèque du niveau de rémunération selon le sexe, en faveur des hommes.

Aller plus loin dans notre étude

Jusqu'à ce stade, nous nous sommes scrupuleusement (de manière très scolaire je dirais) conformés à la démarche décrite dans ce chapitre. Pourtant, au delà des conclusions de numériques, on ne manquera pas de remarquer plusieurs choses dans le graphique ci-dessus (Figure 8.9 - *un graphique vaut souvent tous les calculs du monde...*) : les droites ne sont pas si parallèles que ça contrairement à ce que semble affirmer le test d'égalité des pentes ; et surtout, la dispersion des salaires est plus forte à mesure que niveau d'études augmente.

Nous avons essayé d'introduire une transformation log-log pour stabiliser la variance c.-à-d. réaliser les régressions sur les variables transformées endogène = ln(salaire) vs. exogène = ln(années d'études).

Le résultat est particulière édifiant (Figure 8.11) [11]. La nature de la divergence est confirmée, elle est manifestement du à un décalage entre les droites qui sont quasi-parfaitement parallèles. Et ce décalage correspond en réalité à un **rapport constant entre les salaires hommes/femmes, quel que soit le niveau d'études**.

Comme quoi, des transformations de variables judicieusement choisies peuvent transfigurer les résultats de la régression. Il ne faut jamais l'oublier.

Une autre information importante découle de cette nouvelle analyse : la relation entre le salaire et les années d'études est à élasticité constante, une augmentation relative des années d'études entraîne une augmentation relative proportionnelle du salaire.

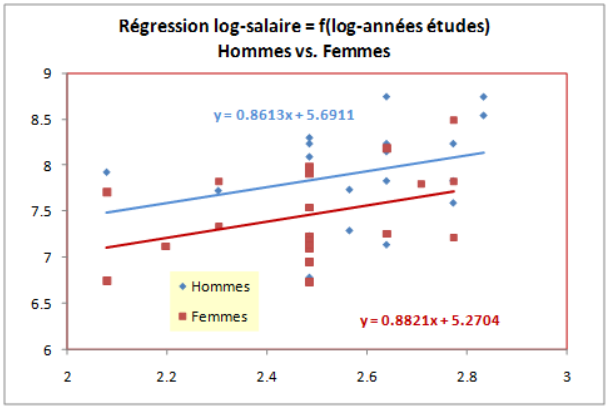


Fig. 8.11. Comparaison des régressions - ln(Salaire) = f[ln(années d'études)] / sexe

11. `comparaisondesregressions.xls` - "salaires-ed-sexe-loglog

8.5.2 Taille des méduses

Dans cette seconde étude, nous voulons expliquer la largeur des méduses à partir de leur longueur [12]. Elles ont été pêchées sur deux sites différentes. On souhaite savoir si la relation entre la largeur et la longueur est la même sur ces deux groupes.

largeur	longueur	Site
6.5	8	1
6	9	1
6.5	9	1
7	9	1
8	9.5	1
7	10	1
8	10	1
8	10	1
7	11	1
8	11	1
9	11	1
10	13	1
11	13	1
12	13	1
11	14	1
11	14	1
13	14	1
14	16	1
15	16	1
15	16	1
15	19	1
16	16	1
12	14	2
15	16	2
14	16.5	2
13	17	2
15	17	2
15	18	2
15	18	2
16	18	2
14	19	2
15	19	2
16	19	2
16.5	19	2
18	19	2
18	19	2
16	20	2
16	20	2
17	20	2
18	20	2
19	20	2
15	21	2
16	21	2
21	21	2
19	22	2
20	22	2

Régression globale	
0.9351	-1.4423
0.0460	0.7520
0.9036	1.2873
412.5389	44.0000
683.6151	72.9121

Régression Groupe 1	
1.0392	-2.6433
0.0760	0.9638
0.9034	1.0464
187.0877	20.0000
204.8716	21.9011

Régression Groupe 2	
0.8478	0.1385
0.1591	3.0339
0.5636	1.4771
28.4093	22.0000
61.9871	48.0025

Erreurs résiduelles	
SCR_T	72.9121
SCR_W	69.9036

Test	
F	**0.9038**
ddl1	2
ddl2	42
p-value	**0.4128**

Fig. 8.12. Comparaison des régressions - Largeur vs. longueur des méduses

12. Ou l'inverse, qu'importe, cet exemple vaut surtout pour la singularité des résultats que l'on obtient. Les données proviennent du site *Datasets for Statistical Analysis*, `http://www.sci.usq.edu.au/staff/dunn/Datasets/Books/Hand/Hand-R/jelly-R.html`

Nous disposons de $n = 46$ observations, avec $n_1 = 22$ et $n_2 = 24$. Nous réalisons la régression globale et les régressions conditionnelles (Figure 8.12) [13]. Nous en déduisons les informations pour réaliser la comparaison globale :

- A partir de la régression sur les $n = 46$ observations, nous avons

$$SCR_T = 72.9121$$

- A partir des deux régressions dans les groupes,

$$SCR_W = SCR_1 + SCR_2 = 21.9011 + 48.0025 = 69.9036$$

- Nous formons la statistique de test

$$F = \frac{(SCR_T - SCR_W)/(2(K-1))}{SCR_W/(n-2K)} = \frac{(72.9121 - 69.9036)/(2 \times (2-1))}{69.9036/(46 - 2 \times 2)} = 0.9038$$

- Avec une distribution $\mathcal{F}(2, 42)$, nous obtenons une probabilité critique de $\alpha' = 0.4128$.
- Au risque $\alpha = 5\%$, nous pouvons affirmer que la relation entre la longueur et la largeur est la même pour les méduses en provenance des deux sites. On *pourrait être emmené à penser* que les méduses proviennent de la même population.

Peut-on s'en tenir à cette conclusion ? Toujours un petit graphique, surtout dans le cadre de la régression simple, pour vérifier qu'il n'y a pas une entourloupe quelque part. On ne sait jamais.

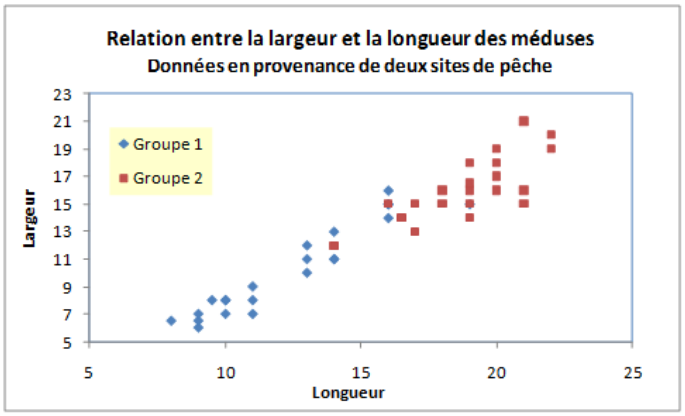


Fig. 8.13. Comparaison des régressions - Largeur vs. longueur des méduses - Nuages de points

Grand bien nous en a pris (Figure 8.13). Les résultats obtenus à travers la procédure statistique masquaient en réalité un problème de taille (si je puis dire). Effectivement, la relation entre la largeur et la longueur semblent identique dans les deux sous-populations. En revanche, les méduses ne sont pas de même taille. Les méduses du second groupe sont plus larges et plus longs que ceux du premier. Le test de comparaison des régressions, dont l'objectif est de détecter les disparités sur les coefficients a et b de la

13. `comparaisondesregressions.xls` - "comp.meduses"

droite, n'est absolument pas armé pour déceler ce type de phénomène. Alors qu'une simple comparaison de moyennes, tant sur X que sur Y, l'aurait immédiatement mis en évidence.

Moralité, il ne faut jamais demander aux tests plus que ce qu'ils savent faire. Il nous appartient de délimiter précisément leur champ d'action.

Partie II

Régression Linéaire Multiple

9

Régression linéaire multiple

9.1 Formulation - Hypothèses

La régression linéaire multiple est la généralisation multivariée de la régression simple. Nous cherchons à expliquer les valeurs prises par la variable endogène Y à l'aide de p variables exogènes X_j, $(j = 1, \ldots, p)$. L'équation de régression s'écrit :

$$y_i = a_0 + a_1 x_{i,1} + \cdots + a_p x_{i,p} + \epsilon_i \tag{9.1}$$

Nous devons estimer les valeurs des $(p+1)$ paramètres $(a_0, a_1, \ldots, a_p)$ à partir d'un échantillon de n observations. Nous remarquons dans le modèle (Équation 9.1) :

- $i = 1, \ldots, n$ correspond au numéro des observations ;
- 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 (ex. problème de spécification, valeurs exogènes manquantes, etc.).

Les étapes **processus de modélisation** sont les suivantes (Tenenhaus, pages 104 et 105) :

1. Estimer les valeurs des coefficients $(a_0, a_1, \ldots, a_p)$ à partir d'un échantillon de données (estimateur des moindres carrés ordinaires).
2. Évaluer la précision de ces estimations (biais, variance des estimateurs).
3. Mesurer le pouvoir explicatif du modèle dans sa globalité (tableau d'analyse de variance, coefficient de détermination).
4. Tester la réalité de la relation entre Y et les exogènes X_j (test de significativité globale de la régression).
5. Tester l'apport marginal de chaque variable explicative dans l'explication de Y (test de significativité de chaque coefficient).

6. Tester l'apport d'un groupe de variables explicatives dans l'explication de Y (test de significativité simultanée d'un groupe de coefficient).
7. Pour un nouvel individu $i*$ pour lequel on fournit la description $(x_{i*,1}, \ldots, x_{i*,p})$, calculer la valeur prédite $\hat{y}_{i*}$ et la fourchette de prédiction.
8. Interpréter les résultats en mettant en avant notamment l'impact des exogènes sur l'endogène (interprétation des coefficients, analyse structurelle).

La modélisation est un processus itératif. Lorsqu'on essaie réellement d'approfondir, on se rend compte que le processus de modélisation est très complexe. Il nécessite parfois plusieurs aller-retour pour vérifier la validité des résultats que l'on essaie d'établir. Quelques outils de diagnostic de la régression sont décrits dans un second support en ligne [13]. Y sont étudiés notamment :

- L'étude des résidus, graphiquement mais aussi numériquement avec les tests de normalité, les tests du caractère aléatoire des erreurs.
- La détection des points aberrants et influents, ces points qui peuvent peser de manière indue sur les résultats de la régression.
- Les problèmes de colinéarité et la sélection de variables.
- Les ruptures de structure c.-à-d. la vérification de l'existence de plusieurs sous-populations dans les données, avec des relations de nature différente entre les exogènes et l'endogène (ex. le lien entre le poids et la taille n'est pas le même chez les hommes et chez les femmes).
- Les problèmes de non linéarité que nous avons commencé à aborder dans la partie consacrée à la régression simple.

Lecture des coefficients. Chaque coefficient se lit comme un propension marginale : $\frac{\partial y}{\partial x_j} = a_j$.

Mais, à la différence de la régression linéaire simple, on prend en compte le rôle des autres variables lors de son calcul. On dit alors que c'est un coefficient partiel : il indique l'impact de la variable en contrôlant l'effet des autres variables, c'est la fameux "toutes choses égales par ailleurs". Nous approfondirons cette notion dans un chapitre dédié à l'interprétation des coefficients (chapitre 13).

Enfin, l'effet des variables est additif c.-à-d. toutes les autres étant constantes, si x_j et $x_{j'}$ sont tous deux augmentés d'une unité, alors y est augmenté $(a_j + a_{j'})$.

Régression sans constante. Les remarques émises concernant le modèle sans constante dans la régression simple (section 7.2) restent valables. Il faut faire attention aux degrés de liberté puisque nous n'estimons plus que p paramètres. Le coefficient de détermination R^2 n'est plus interprétable en termes de proportion de variance expliquée.

9.2 Notation matricielle

Pour simplifier les notations, on retrouve souvent une écriture matricielle du modèle dans la littérature (Equation 9.2).

$$Y = Xa + \varepsilon \tag{9.2}$$

Les dimensions des matrices sont respectivement :

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

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

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

9.3 Hypothèses

Comme pour la régression simple, les hypothèses permettront de déterminer les propriétés des estimateurs (biais, convergence) et les lois de distribution (loi de Student pour chaque coefficient pris individuellement, loi de Fisher dès que l'on traite un groupe de coefficients).

Nous distinguons (Bourbonnais, page 51 ; Labrousse, page 19 ; Giraud et Chaix, pages 22 et 23) :

- **Les hypothèses stochastiques**
  - **H1** – Les X_j sont non aléatoires c.-à-d. les $x_{i,j}$ sont observés sans erreur.
  - **H2** – $E[\varepsilon_i] = 0$, l'espérance de l'erreur est nulle. En moyenne, le modèle est bien spécifié.
  - **H3** – $E[\varepsilon_i^2] = \sigma_\varepsilon^2$, la variance de l'erreur est constante, c'est l'hypothèse de homoscédasticité.
  - **H4** – $COV(\varepsilon_i, \varepsilon_{i'}) = 0$ pour $i \neq i'$, les erreurs sont indépendantes, c'est l'hypothèse de non-autocorrélation des résidus.
  - **H5** – $COV(x_{i,j}, \varepsilon_i) = 0$, l'erreur est indépendante des variables exogènes.
  - **H6** – $\varepsilon_i \equiv \mathcal{N}(0, \sigma_\varepsilon)$, les erreurs sont distribués selon une loi normale.
- **Les hypothèses structurelles**
  - **H7** – La matrice $(X'X)$ est régulière c.-à-d. $det(X'X) \neq 0$ et $(X'X)^{-1}$ existe. Elle indique l'absence de colinéarité entre les exogènes. Nous pouvons aussi voir cette hypothèse sous l'angle $rang(X) = p+1$ et $rang(X'X) = p+1$.
  - **H8** – $\frac{(X'X)}{n}$ tend vers une matrice finie non singulière lorsque $n \rightarrow +\infty$.

H9 – $n > p+1$, le nombre d'observations est supérieur au nombre de paramètres à estimer. Dans le cas où $n = p+1$, nous avons un interpolation, la droite passe exactement par tous les points. Lorsque $n < p+1$, la matrice $(X'X)$ n'est plus inversible.

9.4 Ajustement des moindres carrés ordinaires (MCO)

9.4.1 Minimisation de la somme des carrés des erreurs

Comme pour la régression simple, on cherche les coefficients qui permettent de minimiser la quantité suivante

$$S = \sum_{i=1}^{n} \varepsilon_i^2 \tag{9.3}$$

où $\varepsilon_i^2 = [y_i - (a_0 + a_1 x_{i,1} + \cdots + a_p x_{i,p}]^2$

On passe de nouveau par les dérivées partielles que l'on annule pour obtenir les $(p+1)$ **équations normales**.

$$\begin{cases} \frac{\partial S}{\partial a_0} = 0 \\ \vdots \\ \frac{\partial S}{\partial a_p} = 0 \end{cases} \Leftrightarrow \begin{cases} -2\sum_i \varepsilon_i = 0 \\ \vdots \\ -2\sum_i x_{i,p} \times \varepsilon_i = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} a_0 + a_1\bar{x}_1 + \cdots + a_p\bar{x}_p = \bar{y} \\ \vdots \\ a_0\sum_i x_{i,p} + a_1\sum_i x_{i,1}x_{i,p} + \cdots + a_p\sum_i x_{i,p}x_{i,p} = x_{i,p}y_i \end{cases}$$

Nous avons $(p+1)$ équations à $(p+1)$ inconnues. Nous pouvons en extraire les estimations $(\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_p)$. Mais cette écriture est difficile à manipuler. Passons aux matrices.

9.4.2 Écriture matricielle

Avec l'écriture matricielle, nous pouvons produire une écriture condensée. Soit ε le vecteur des erreurs, avec $\varepsilon' = (\varepsilon_1, \ldots, \varepsilon_n)$. La somme des carrés des erreurs devient

$$S = \sum_i \varepsilon_i^2 = \varepsilon'\varepsilon$$

Développons l'expression

$$\begin{aligned} \varepsilon'\varepsilon &= (Y - Xa)'(Y - Xa) \\ &= Y'Y - Y'Xa - a'X'Y + a'X'Xa \\ &= Y'Y - 2a'X'Y + a'X'Xa \\ S &= Y'Y - 2a'X'Y + a'X'Xa \end{aligned}$$

Quelques éléments sur les calculs matriciels pour comprendre les développements ci-dessus :

- $(Xa)' = a'X'$
- $(Y'Xa)' = a'X'Y$
- La transposée d'un scalaire est égal à lui même. Or en se référant aux dimensions des vecteurs et matrice, on constate que $(a'X'Y)$ est de dimension $(1, 1)$, un scalaire.

Pour déterminer le minimum de S, nous réalisons la dérivation matricielle que nous annulons (Labrousse, page 22) :

$$\frac{\partial S}{\partial a} = -2(X'Y) + 2(X'X)a = 0$$
$$(X'X)a = X'Y$$

L'estimateur des moindres carrés ordinaires (MCO) des coefficients du modèle s'écrit :

$$\hat{a} = (X'X)X'Y \tag{9.4}$$

9.4.3 Un exemple : consommation des véhicules

Nous reprenons l'exemple que nous décrivons dans un de nos supports [13]. Il s'agit d'expliquer la consommation des véhicules (en L/100 km) à partir de $p = 3$ variables exogènes : la cylindrée (taille du moteur, en cm^3), la puissance (en kw) et le poids (en kg). Par rapport au fichier original, nous avons éliminé les 3 points atypiques qui posaient problèmes. Nous disposons donc de $n = 28$ observations.

Nous avons élaboré une feuille Excel qui reconstitue tous les calculs intermédiaires permettant d'obtenir le vecteur $\hat{a}$ (Figure 9.1) [1] :

- Nous distinguons les valeurs des exogènes (X_1, X_2, X_3), et celles de l'endogène Y.
- Nous accolons au tableau des exogènes une colonne de constante, avec la valeur 1. Nous obtenons ainsi la matrice X.

$$\mathrm{X} = \begin{pmatrix} 1 & 846 & 32 & 650 \\ 1 & 993 & 39 & 790 \\ \vdots & & & \\ 1 & 2473 & 125 & 1570 \end{pmatrix}$$

- Nous pouvons élaborer la matrice $(X'X)$, avec

$$(X'X) = \begin{pmatrix} 28 & 50654 & 2176 & 33515 \\ \vdots & & & \\ 33515 & 65113780 & 2831550 & 42694125 \end{pmatrix}$$

  Nous devrions obtenir $n = \sum_{i=1}^{28} 1 \times 1 = 28$ dans la première cellule de la matrice. C'est le cas.
- Nous inversons cette matrice pour obtenir $(X'X)^{-1}$ (*attention, certains chiffres de la matrice sont en notation scientifique dans la figure 9.1*).

1. `reg_multiple_consommation_automobiles.xlsx` - "EMCO"

X				Y
constante	cylindree	puissance	poids	consommation
1	846	32	650	5.7
1	993	39	790	5.8
1	899	29	730	6.1
1	1390	44	955	6.5
1	1195	33	895	6.8
1	658	32	740	6.8
1	1331	55	1010	7.1
1	1597	74	1080	7.4
1	1761	74	1100	9
1	2165	101	1500	11.7
1	1983	85	1075	9.5
1	1984	85	1155	9.5
1	1998	89	1140	8.8
1	1580	65	1080	9.3
1	1390	54	1110	8.6
1	1396	66	1140	7.7
1	2435	106	1370	10.8
1	1242	55	940	6.6
1	2972	107	1400	11.7
1	2958	150	1550	11.9
1	2497	122	1330	10.8
1	1998	66	1300	7.6
1	2496	125	1670	11.3
1	1998	89	1560	10.8
1	1997	92	1240	9.2
1	1984	85	1635	11.6
1	2438	97	1800	12.8
1	2473	125	1570	12.7

(X'X)			
28	50654	2176	33515
50654	102138444	4451219	65113780
2176	4451219	197200	2831550
33515	65113780	2831550	42694125

(X'X)^-1			
0.70598604	-0.00014708	0.005586344	-0.00070038
-0.00014708	1.07417E-06	-1.58914E-05	-4.6883E-07
0.00558634	-1.5891E-05	0.000358366	-3.9165E-06
-0.00070038	-4.6883E-07	-3.91645E-06	1.54799E-06

(X'Y)
254.1
493218.1
21473.7
321404.5

	a^
constante	1.7020
cylindree	0.0005
puissance	0.0183
poids	0.0042

Fig. 9.1. Calculs matriciels - Consommation des véhicules

– Ensuite, nous calculons la matrice des produits croisés entre X et Y, soit $(X'Y)$, nous avons

$$(X'Y) = \begin{pmatrix} 254.1 \\ 493218.1 \\ 21473.7 \\ 321404.5 \end{pmatrix}$$

– Enfin, il ne nous reste plus qu'à calculer $\hat{a} = (X'X)^{-1}(X'Y)$. Nous obtenons les estimations des paramètres de la régression

$$\hat{a} = \begin{pmatrix} \hat{a}_0 \\ \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \end{pmatrix} = \begin{pmatrix} 1.7020 \\ 0.0005 \\ 0.0183 \\ 0.042 \end{pmatrix}$$

Les coefficients sont dans l'ordre des colonnes de la matrice X.

– Le modèle s'écrit

$$CONSO = 1.7020 + 0.0005 \times cylindree + 0.0183 \times puissance + 0.042 \times poids$$

Toutes les variables semblent jouer positivement sur la consommation c.-à-d. lorsque la cylindrée, la puissance ou le poids augmentent, la consommation a tendance à augmenter.

9.4.4 Quelques remarques sur les matrices

Les matrices $(X'X)^{-1}$ et $(X'Y)$ qui entrent dans la composition de $\hat{a}$ peuvent être interprétées d'une manière qui nous éclaire sur l'influence des variables dans l'estimation.

Matrice $(X'X)$

Chaque case de la matrice $(X'X)$, de dimension $(p+, 1, p+1)$, est formée par la somme du produit croisé entre les exogènes, en effet :

$$(X'X) = \begin{pmatrix} n & \sum_i x_{i,1} & \cdots & \sum_i x_{i,p} \\ \sum_i x_{i,1} & \sum_i x_{i,1}^2 & \cdots & \sum_i x_{i,1}x_{i,p} \\ \vdots & & & \\ \sum_i x_{i,p} & \sum_i x_{i,1}x_{i,p} & \cdots & \sum_i x_{i,p}^2 \end{pmatrix}$$

$(X'X)$ est une matrice symétrique. Elle indique le degré de liaison entre les exogènes.

Matrice $(X'Y)$

Chaque case du vecteur $(X'Y)$, de dimension $(p+1, 1)$, est composée du produit croisé entre les exogènes et l'endogène.

$$(X'X) = \begin{pmatrix} \sum_i y_i \\ \sum_i x_{i,1}y_i \\ \vdots \\ \sum_i x_{i,p}y_i \end{pmatrix}$$

Le vecteur indique le degré de liaison entre chaque exogène et Y.

Ainsi le coefficient associé à une variable explicative sera d'autant plus élevée en valeur absolue, relativement aux autres (nonobstant les disparités dues aux unités de mesures), qu'elle est fortement liée avec l'endogène et, dans le même temps, faiblement liée avec les autres exogènes.

Cas des variables centrées

Lorsque les variables sont centrées, nous retrouvons des concepts que nous connaissons bien. Soient

$$\dot{x}_{i,j} = x_{i,j} - \bar{x}_j$$
$$\dot{y}_i = y_i - \bar{y}$$

les variables centrées. Alors les matrices

$$\frac{1}{n}(\dot{X}'\dot{X}) = cov(X_j, X_{j'})$$
$$\frac{1}{n}(\dot{X}'\dot{Y}) = cov(X_j, Y)$$

représentent respectivement la matrice des variances covariances des exogènes, et le vecteur des covariances entre les exogènes et l'endogène.

Cas des variables centrées et réduites

De la même manière, lorsque les variables sont centrées et réduites c.à-d.

$$\overset{cr}{x}_{i,j} = \frac{x_{i,j} - \bar{x}_j}{\sigma_{x_j}}$$
$$\overset{cr}{y}_i = \frac{y_i - \bar{y}}{\sigma_y}$$

Les matrices

$$\frac{1}{n}(\overset{cr}{X}{}' \overset{cr}{X}) = r(X_j, X_{j'})$$
$$\frac{1}{n}(\overset{cr}{X}{}' \overset{cr}{Y}) = r(X_j, Y)$$

représentent respectivement les corrélations croisées entre les X_j et les corrélations des X_j avec Y.

9.5 Propriétés des estimateurs

De nouveau, cette section est surtout intéressante pour les férus de théorie. Sa lecture n'est pas primordiale si vous êtes avant tout intéressés par la mise en oeuvre de la régression sur des problèmes réels.

A l'attention des étudiants de la Licence IDS : vous par contre, vous devez bien la lire, en détail même, et comprendre si possible. Désolé.

Deux questions reviennent toujours lorsque l'on souhaite étudier les propriétés d'un estimateur : est-il sans biais ? est-il convergent ?

Nous allons directement à l'essentiel dans cette partie. Le détail de la démarche a déjà été exposé dans le cadre de la régression simple (chapitre 2).

9.5.1 Biais

L'estimateur $\hat{a}$ est sans biais si $E(\hat{a}) = a$. Voyons à quelles conditions cette propriété est respectée.

Développons $\hat{a}$:

$$\begin{aligned}\hat{a} &= (X'X)^{-1}X'Y \\ &= (X'X)^{-1}X'(Xa+\varepsilon) \\ &= (X'X)^{-1}X'Xa + (X'X)^{-1}X'\varepsilon \\ \hat{a} &= a + X'X)^{-1}X'\varepsilon\end{aligned}$$

Ainsi, en passant à l'espérance mathématique :

$$E(\hat{a}) = a + E[(X'X)^{-1}X'\varepsilon]$$

On sait que X est non aléatoire, nous avons $E[(X'X)^{-1}X'\varepsilon] = (X'X)^{-1}X'E(\varepsilon)$; de plus $E(\varepsilon) = 0$ par hypothèse. Au final, nous avons bien

$$E(\hat{a}) = a$$

L'estimateur des MCO est sans biais sous les deux hypothèses suivantes (section 9.3) : (H1) X est non aléatoire, les exogènes sont mesurées sans erreur ; (H2) la moyenne de l'erreur est nulle $E(\varepsilon) = 0$.

9.5.2 Variance - Convergence

Soit $\Omega_{\hat{a}}$, de dimension $(p+1, p+1)$ la matrice de variance covariance des coefficients c.-à-d.

$$\Omega_{\hat{a}} = \begin{pmatrix} V(\hat{a}_0) & COV(\hat{a}_0,\hat{a}_1) & \cdots & COV(\hat{a}_0,\hat{a}_p) \\ \cdots & V(\hat{a}_1) & \cdots & COV(\hat{a}_1,\hat{a}_p) \\ \vdots & & & \\ \cdots & \cdots & \cdots & V(\hat{a}_p) \end{pmatrix}$$

La matrice est symétrique, sur la diagonale principale nous observons les variances des coefficients estimés.

Comment obtenir cette matrice ?

Elle est définie de la manière suivante

$$\Omega_{\hat{a}} = E[(\hat{a}-a)(\hat{a}-a)']$$

Or

$$\begin{aligned}\hat{a} - a &= (X'X)^{-1}X'\varepsilon \\ (\hat{a}-a)' &= \varepsilon'X[(X'X)^{-1}]' \\ &= \varepsilon'X(X'X)^{-1} \texttt{ car } \texttt{(X'X)}^{-1} \texttt{ est symetrique}\end{aligned}$$

Ainsi

$$(\hat{a}-a)(\hat{a}-a)' = (X'X)^{-1}X'\varepsilon\varepsilon'X(X'X)^{-1}$$

En passant à l'espérance mathématique, et sachant que les X sont non-stochastiques (H1),

$$E[(\hat{a}-a)(\hat{a}-a)'] = (X'X)^{-1}X'E[\varepsilon\varepsilon']X(X'X)^{-1}$$

La quantité $E[\varepsilon\varepsilon']$, de dimension (n,n), représente la matrice de variance covariance des erreurs, en voici le détail

$$E[\varepsilon\varepsilon'] = \begin{pmatrix} E(\varepsilon_1^2) & E(\varepsilon_1\varepsilon_2) & \cdots & E(\varepsilon_1\varepsilon_n) \\ \vdots & & & \\ \cdots & \cdots & \cdots & E(\varepsilon_n^2) \end{pmatrix}$$

Nous observons les variances des erreurs sur la diagonale principale, et les covariances sur les autres cases. Or, par hypothèse (section 9.3), (H3) la variance de l'erreur est constante $V(\varepsilon_i) = E(\varepsilon_i^2) = \sigma_\varepsilon^2$ et, (H4) leurs covariances nulles $COV(\varepsilon_i, \varepsilon_{i'}) = 0$. De fait,

$$E[\varepsilon\varepsilon'] = \sigma_\varepsilon^2 I$$

Où I est la matrice unité de dimension (n,n).

La matrice de variance covariance des estimateurs s'en retrouve grandement simplifiée. En effet,

$$\begin{aligned} E[(\hat{a}-a)(\hat{a}-a)'] &= (X'X)^{-1}X'E[\varepsilon\varepsilon']X(X'X)^{-1} \\ &= \sigma_\varepsilon^2(X'X)^{-1}X'IX(X'X)^{-1} \\ &= \sigma_\varepsilon^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma_\varepsilon^2(X'X)^{-1} \end{aligned}$$

Nous trouvons ainsi la matrice de variance covariance des coefficients estimés :

$$\Omega_{\hat{a}} = \sigma_\varepsilon^2(X'X)^{-1} \tag{9.5}$$

On montre qu'une condition nécessaire et suffisante pour que $\hat{a}$ soit un estimateur convergent de a est que les variables exogènes ne tendent pas à devenir colinéaires lorsque n tend vers l'infini, autrement dit que l'hypothèse (H8) reste valable lorsque n tend vers l'infini. (Giraud et Chaix, page 65 ; que l'on retrouve sous des formes plus ou moins analogues chez Bourbonnais, page 53, et Labrousse, page 26).

9.5.3 L'estimateur des MCO est BLUE

Théorème de Gauss-Markov. Exactement comme pour la régression simple, on montre pour la régression multiple qu'il n'existe pas d'estimateurs sans biais avec une variance plus faible que celle des moindres carrés ordinaires (Labrousse, page 26). Les estimateurs des MCO sont BLUE (*best linear unbiased estimator*).

9.6 Estimation de la variance de l'erreur

9.6.1 Estimation de la variance de l'erreur

L'expression de la variance covariance des coefficients estimés (Équation 9.5) est très jolie mais inutilisable tant que l'on ne dispose pas d'une estimation de la variance de l'erreur $\hat{\sigma}^2_\varepsilon$.

Par analogie avec la régression simple (section 3.2.2), nous la comprenons comme le rapport entre la somme des carrés des résidus (SCR) et le nombre de degrés de liberté de la régression, soit le nombre d'observations moins le nombre de paramètres estimés : $[n-(p+1) = n-p-1]$. Ainsi, nous écrirons

$$\hat{\sigma}^2_\varepsilon = \frac{SCR}{n-p-1} = \frac{\sum_i \hat{\varepsilon}^2_i}{n-p-1} \tag{9.6}$$

Où $\hat{\varepsilon}_i$ est le résidu de la régression pour l'observation $n^o i$.

Le lecteur désireux d'approfondir la question, notamment le détail de la démarche, trouvera une démonstration plus rigoureuse dans les ouvrages listés en bibliographie (Labrousse, pages 28 à 33 ; Dodge et Rousson, pages 65 à 67 ; Giraud et Chaix, pages 67 à 69 ; etc.).

9.6.2 Estimation de la matrice de variance covariance des coefficients

Disposant maintenant d'une estimation de la variance de l'erreur, nous pouvons produire une estimation de la matrice de variance covariance des coefficients estimés.

$$\hat{\Omega}_{\hat{a}} = \hat{\sigma}^2_\varepsilon (X'X)^{-1} \tag{9.7}$$

Sur la diagonale principale de cette matrice, nous disposons de l'estimation de la variance des coefficients et, en passant à la racine carrée, de leur écart-type. Leur rôle sera très important dans l'inférence statistique.

9.6.3 Détails des calculs pour les données "Consommation des véhicules"

Nous reprenons notre exemple des véhicules (section 9.4.3). Nous avons reconstruit la feuille de calcul de manière à obtenir les éléments nécessaires à l'estimation de la variance de l'erreur et de la matrice de variance covariance des coefficients estimés (Figure 9.2) [2].

Nous reprenons des résultats précédents (Figure 9.1) la matrice $(X'X)^{-1}$ et les coefficients estimés $\hat{a}$. Nous formons alors :

- La valeur prédite de l'endogène $\hat{y}_i$ pour chaque individu (ex. $\hat{y}_1 = 1.070205 + 0.00049 \times 846 + 0.01825 \times 32 + 0.00423 \times 650 = 5.4523$).
- Le résidu $\hat{\varepsilon}_i = y_i - \hat{y}_i$ (ex. $\hat{\varepsilon}_1 = y_1 - \hat{y}_1 = 5.7 - 5.4523 = 0.2477$.

2. `reg_multiple_consommation_automobiles.xlsx` - "variance erreur"

X				Y	Y^	epsilon^	(epsilon^)²
constante	cylindree	puissance	poids	consommation			
1	846	32	650	5.7	5.4523	0.2477	0.0613
1	993	39	790	5.8	6.2447	-0.4447	0.1978
1	899	29	730	6.1	5.7621	0.3379	0.1142
1	1390	44	955	6.5	7.2296	-0.7296	0.5324
1	1195	33	895	6.8	6.6789	0.1211	0.0147
1	658	32	740	6.8	5.7402	1.0598	1.1233
1	1331	55	1010	7.1	7.6339	-0.5339	0.2850
1	1597	74	1080	7.4	8.4079	-1.0079	1.0159
1	1761	74	1100	9	8.5734	0.4266	0.1820
1	2165	101	1500	11.7	10.9571	0.7429	0.5519
1	1983	85	1075	9.5	8.7780	0.7220	0.5212
1	1984	85	1155	9.5	9.1168	0.3832	0.1468
1	1998	89	1140	8.8	9.1333	-0.3333	0.1111
1	1580	65	1080	9.3	8.2353	1.0647	1.1336
1	1390	54	1110	8.6	8.0676	0.5324	0.2834
1	1396	66	1140	7.7	8.4164	-0.7164	0.5133
1	2435	106	1370	10.8	10.6319	0.1681	0.0283
1	1242	55	940	6.6	7.2939	-0.6939	0.4815
1	2972	107	1400	11.7	11.0420	0.6580	0.4329
1	2958	150	1550	11.9	12.4542	-0.5542	0.3072
1	2497	122	1330	10.8	10.7853	0.0147	0.0002
1	1998	66	1300	7.6	9.3902	-1.7902	3.2047
1	2496	125	1670	11.3	12.2774	-0.9774	0.9553
1	1998	89	1560	10.8	10.9094	-0.1094	0.0120
1	1997	92	1240	9.2	9.6105	-0.4105	0.1685
1	1984	85	1635	11.6	11.1467	0.4533	0.2055
1	2438	97	1800	12.8	12.2875	0.5125	0.2626
1	2473	125	1570	12.7	11.8432	0.8568	0.7342

	a^	sigma^(a^)
constante	1.70205	0.63205
cylindree	0.00049	0.00078
puissance	0.01825	0.01424
poids	0.00423	0.00094

n	28
p	3

ddl	24

SCR	13.5807

(sigma^)²eps	0.56586
(sigma^)eps	0.75224

(X'X)^-1			
0.705986044	-0.00014708	0.00558634	-0.00070038
-0.000147084	1.0742E-06	-1.5891E-05	-4.6883E-07
0.005586344	-1.5891E-05	0.00035837	-3.9165E-06
-0.000700376	-4.6883E-07	-3.9165E-06	1.548E-06

SIGMA^(a^)			
0.399490226	-8.3229E-05	0.0031611	-0.00039632
-8.32291E-05	6.0783E-07	-8.9923E-06	-2.6529E-07
0.003161096	-8.9923E-06	0.00020279	-2.2162E-06
-0.000396316	-2.6529E-07	-2.2162E-06	8.7595E-07

Fig. 9.2. Estimation de la variance de l'erreur et des coefficients estimés - Consommation des véhicules

- Que nous passons au carré $\hat{\varepsilon}_i^2$ (ex. $\hat{\varepsilon}_1^2 = (0.2477)^2 = 0.0613$).
- Nous sommons pour obtenir la $SCR = \sum_i \hat{\varepsilon}_i^2$ (dans notre exemple, $SCR = \sum_i \hat{\varepsilon}_i^2 = 0.0613 + 0.1978 + \cdots = 13.5807$).
- L'estimation de la variance de l'erreur s'écrit

$$\hat{\sigma}_\varepsilon^2 = \frac{SCR}{n-p-1} = \frac{13.5807}{28-3-1} = 0.56586$$

- L'estimation de son écart-type en est déduite, valeur souvent automatiquement retournée par les logiciels de statistique

$$\hat{\sigma}_\varepsilon = \sqrt{0.56586} = 0.75224$$

- Reste la dernière multiplication pour obtenir l'estimation de la matrice de variance covariance des coefficients :

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

  Elle est forcément symétrique parce que la covariance est un opérateur symétrique.

Comme nous l'avons souligné précédemment, nous disposons sur la diagonale de cette matrice de l'estimation de la variance des coefficients. Dans notre exemple,

$$\begin{cases} \hat{\sigma}_{\hat{a}_0}^2 = 0.399490226 \\ \hat{\sigma}_{\hat{a}_1}^2 = 6.0783 \times 10^{-7} \\ \hat{\sigma}_{\hat{a}_2}^2 = 0.00020279 \\ \hat{\sigma}_{\hat{a}_3}^2 = 8.7595 \times 10^{-7} \end{cases}$$

En passant à la racine carrée, nous avons l'estimation de l'écart-type, valeurs directement fournies par la majorité des logiciels de statistique. C'est pour cette raison que nous avons affiché les $\hat{\sigma}_{\hat{a}_j}$ à côté des estimations $\hat{a}_j$ dans notre feuille Excel (Figure 9.2). Nous les retrouverons souvent sous cette forme dans les sorties des logiciels que nous analyserons au chapitre 15.

$$\begin{cases} \hat{\sigma}_{\hat{a}_0} = 0.63205 \\ \hat{\sigma}_{\hat{a}_1} = 0.00078 \\ \hat{\sigma}_{\hat{a}_2} = 0.01424 \\ \hat{\sigma}_{\hat{a}_3} = 0.00094 \end{cases}$$

9.6.4 Résultats fournis par la fonction DROITEREG

Il est temps de voir un peu ce que nous propose la fonction DROITEREG d'Excel en matière de régression linéaire multiple. Elle sait gérer un nombre d'exogènes supérieur à 1. Il faut simplement que les colonnes des données soient contiguës dans la feuille de calcul. Il ne faut pas intégrer la colonne de constante parmi les exogènes. Une option nous permet de spécifier si nous souhaitons ou pas la constante a_0 dans la régression. Dans la plage de résultats, nous sélectionnons donc $(p + 1 = 4)$ colonnes pour notre régression, et 5 lignes pour qu'Excel puisse intégrer les informationnelles additionnelles permettant d'analyser les résultats.

Dans l'exemple que nous reproduisons ici (Figure 9.3) [3], nous avons inséré la commande *DROITEREG(F3 :F30 ;B3 :D30 ;1 ;1)*. Le première paramètre correspond à la colonne de valeurs de Y ; le second au(x) colonne(s) de X ; le troisième paramètre indique que nous réalisons une régression avec constante (0 si nous souhaitons une régression sans constante) ; et le dernier indique que l'on souhaite obtenir des informations additionnelles en plus des coefficients estimés (0 dans le cas contraire).

	a^	sigma^(a^)
constante	1.70205	0.63205
cylindree	0.00049	0.00078
puissance	0.01825	0.01424
poids	0.00423	0.00094

n	28
p	3

ddl	24

SCR	13.5807

(sigma^)²eps	0.56586
(sigma^)eps	0.75224

DROITEREG

	poids	puissance	cylindree	constante
a^	0.00423	0.01825	0.00049	1.70205
sigma^(a^)	0.00094	0.01424	0.00078	0.63205
R²	0.89911	0.75224	#N/A	#N/A
F	71.2965	24	#N/A	#N/A
SCE	121.0318	13.5807	#N/A	#N/A
		SCR		

Fig. 9.3. Comparaison estimation manuelle et DROITEREG d'Excel - Consommation des véhicules

Mettons en parallèle les résultats de DROITEREG avec ceux calculés manuellement avec les fonctions matricielles d'Excel (Figure 9.3) :

3. `reg_multiple_consommation_automobiles.xlsx` - "variance erreur (droitereg)"

- Sur la première ligne, nous observons les coefficients estimés $\hat{a}$. La constante est toujours en dernière position à droite. En revanche, les coefficients associés aux variables sont dans l'ordre inverse des colonnes des données. Bon, on ne voit pas trop où est la logique. Il faudra s'en souvenir tout simplement. Dans notre tableau de valeurs (Figure 9.2), nous avons de gauche à droite $(cylindree, puissance, poids)$. Dans le tableau fourni par DROITEREG, nous avons de gauche à droite les coefficients associés à $(poids, puissance, cylindree)$.
- Mis à part cette petite incongruité, nous constatons que les coefficients sont les bons, ($\hat{a}_0 = 1.70205, \hat{a}_{cylindree} = 0.00049, \hat{a}_{puissance} = 0.01825, \hat{a}_{poids} = 0.00423$).
- Sur la seconde ligne, nous avons les écart-types estimés des coefficients. En prenant en compte le décalage, nous constatons que les valeurs coïncident avec l'estimation à l'aide des fonctions matricielles d'Excel.
- Dans la case $(3, 2)$, nous avons l'estimation de l'écart-type de l'erreur $\hat{\sigma}_\varepsilon = 0.75224$.
- Dans la case $(4, 2)$, nous observons les degrés de liberté de la régression, $n - p - 1 = 28 - 3 - 1$.
- Enfin, dans la case $(5, 2)$, nous observons la $SCR = 13.5807$.

D'autres informations sont fournies, nous les détaillerons par la suite.

10

Tests de significativité

10.1 Tableau d'analyse de variance et coefficient de détermination

10.1.1 Tableau d'analyse de variance et coefficient de détermination

La décomposition de la variabilité de Y (SCT) en variabilité expliquée par le modèle (SCE) et variabilité résiduelle (SCR) reste valable. Nous pouvons construire une nouvelle version du tableau d'analyse de variance qui tient compte des nouvelles valeurs des degrés de liberté puisque nous estimons $(p+1)$ paramètres maintenant.

Source de variation	Somme des carrés	Degrés de liberté	Carrés moyens
Expliquée	$SCE = \sum_i (\hat{y}_i - \bar{y})^2$	p	$CME = \frac{SCE}{p}$
Résiduelle	$SCR = \sum_i (y_i - \hat{y}_i)^2$	$n-p-1$	$CMR = \frac{SCR}{n-p-1}$
Totale	$SCT = \sum_i (y_i - \bar{y})^2$	$n-1$	-

Tableau 10.1. Tableau d'analyse de variance pour la régression multiple

La part de variance de Y expliquée par le modèle est toujours traduit par le coefficient de détermination

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} \tag{10.1}$$

Bien évidemment $(0 \leq R^2 \leq 1)$, plus il tend vers 1 meilleur sera le modèle. Lorsqu'il est proche de 0, cela veut dire que les exogènes X_j n'expliquent en rien les valeurs prises par Y. Nous retiendrons cette idée dans le test de significativité globale du modèle.

10.1.2 R^2 corrigé ou ajusté

Le R^2 est un indicateur de qualité, mais il présente un défaut ennuyeux : plus nous augmentons le nombre de variables explicatives, même non pertinentes, n'ayant aucun rapport avec le problème que l'on cherche à résoudre, plus grande sera sa valeur, mécaniquement.

A l'extrême, si nous multiplions le nombre d'explicatives jusqu'à ce que $(p+1)$ soit égal à n, nous obtiendrions un $R^2 = 1$.

Teneur en oxyde de carbone des cigarettes. Voyons un petit exemple pour montrer l'inconvénient du R^2 dans la comparaison des modèles. Nous souhaitons expliquer la teneur en oxyde de carbone (CO) des cigarettes à partir de leur composition en goudron (TAR), en nicotine (NICOTINE) et leur poids (WEIGHT). Nous disposons de $n = 24$ observations. Nous réalisons la régression à l'aide de DROITEREG, nous obtenons le R^2 dans la case $(3,1)$ du tableau de résultats : $R^2 = 0.93498$ (Figure 10.1) [1]. Le degré de liberté est $ddl = 24 - 3 - 1 = 20$.

TAR (mg)	NICOTINE (mg)	WEIGHT (g)	ALEA	CO (mg)
14.1	0.86	0.9853	0.26780	13.6
16	1.06	1.0938	0.35783	16.6
8	0.67	0.928	0.12693	10.2
4.1	0.4	0.9462	0.22795	5.4
15	1.04	0.8885	0.10896	15
8.8	0.76	1.0267	0.03880	9
12.4	0.95	0.9225	0.39589	12.3
16.6	1.12	0.9372	0.27504	16.3
14.9	1.02	0.8858	0.85236	15.4
13.7	1.01	0.9643	0.16240	13
15.1	0.9	0.9316	0.67295	14.4
7.8	0.57	0.9705	0.67509	10
11.4	0.78	1.124	0.84745	10.2
9	0.74	0.8517	0.54966	9.5
1	0.13	0.7851	0.43216	1.5
17	1.26	0.9186	0.97986	18.5
12.8	1.08	1.0395	0.39638	12.6
15.8	0.96	0.9573	0.43537	17.5
4.5	0.42	0.9106	0.55342	4.9
14.5	1.01	1.007	0.65460	15.9
7.3	0.61	0.9806	0.51565	8.5
8.6	0.69	0.9693	0.50194	10.6
15.2	1.02	0.9496	0.72085	13.9
12	0.82	1.1184	0.81714	14.9

DROITEREG (TAR, NICOTINE, WEIGHT) - 1

	weight	nicotine	tar	constante
	2.07934	0.51847	0.88758	-0.55170
	3.17842	3.25233	0.19548	2.97128
R^2	0.93498	1.15983	#N/A	#N/A
	95.85850	20	#N/A	#N/A
	386.85	26.90	#N/A	#N/A

DROITEREG (TAR, NICOTINE, WEIGHT, ALEA) - 2

	alea	weight	nicotine	tar	constante
	0.81653	1.87048	0.93450	0.85569	-0.72260
	0.96657	3.21095	3.31268	0.20048	2.99961
R^2	0.93733	1.16822	#N/A	#N/A	#N/A
	71.04289	19	#N/A	#N/A	#N/A
	387.82	25.93	#N/A	#N/A	#N/A

R^2 ajusté (1)	0.92522
R^2 ajusté (2)	0.92414

Fig. 10.1. Comparaison de modèles imbriqués via le R^2 et R^2-ajusté - Données cigarettes

Ajoutons la colonne ALEA dans le tableau de données. Elle a été générée aléatoirement avec la fonction ALEA() d'Excel [loi uniforme $U(0,1)$]. Nous effectuons de nouveau la régression en intégrant ALEA parmi les explicatives. Le degré de liberté est diminué, il est passé à $ddl = 19$, témoin que la variable supplémentaire a bien été prise en compte. Malgré que la variable n'ait aucun rapport avec le problème que nous traitons, nous découvrons que le R^2 a été augmenté, passant à $R^2 = 0.9373$. Diable, ALEA permettrait donc d'expliquer la teneur en carbone des cigarettes ?

Clairement le R^2 en tant que tel n'est pas un bon outil pour évaluer le rôle de variables supplémentaires lors de la comparaison de modèles imbriqués. En augmentant le nombre d'explicatives, nous augmentons de manière mécanique la valeur du R^2 mais, dans le même temps, nous diminuons le degré de liberté. Il faudrait donc intégrer cette dernière notion pour contrecarrer l'évolution du R^2. C'est exactement ce que fait le R^2-ajusté (ou R^2-corrigé).

Le R^2-ajusté est défini de la manière suivante :

1. `cigarettes-regressionmultiple.xls` - "R2 ajusté"

$$\bar{R}^2 = 1 - \frac{CMR}{CMT} = 1 - \frac{SCR/(n-p-1)}{SCT/(n-1)} \tag{10.2}$$

Il s'agit donc d'un R^2 corrigé par les degrés de liberté, il peut s'exprimer en fonction du R^2 d'ailleurs :

$$\bar{R}^2 = 1 - \frac{n-1}{n-p-1}(1-R^2) \tag{10.3}$$

Attention, la lecture en termes de part de variance expliquée n'est plus possible dans ce cas. De même, le $\bar{R}^2$ peut prendre des valeurs négatives. Il ne faut pas s'en offusquer.

Le R^2-ajusté en tant que tel n'est pas d'une grande utilité. Son principal avantage est qu'il permet de comparer des modèles imbriqués. Si nous prenons notre exemple des cigarettes (Figure 10.1), nous constatons que le R^2-ajusté du second modèle est plus faible avec $\bar{R}^2_2 = 0.92414 < \bar{R}^2_1 = 0.92522$, indiquant clairement que l'adjonction de ALEA parmi les exogènes n'amène pas d'information *pertinente* supplémentaire dans l'explication de Y.

Remarque 4 (Comparaison des R^2). La comparaison directe des R^2 (bruts) n'est pas une bonne idée pour évaluer la pertinence de variables supplémentaires dans la régression disions-nous. C'est certain. En revanche, nous pouvons tourner le problème d'une autre manière en posant la question : "est-ce que l'introduction de nouvelles exogènes induit une augmentation *significative* du R^2 ? L'affaire devint intéressante dans ce cas, car nous nous situons dans un schéma de test d'hypothèses. Au résultat est associé un niveau de crédibilité traduit par le risque du test. Nous exploiterons cette idée plus loin dans ce fascicule pour tester la significativité d'un groupe de variables (section 10.4).

10.1.3 Coefficient de corrélation linéaire multiple

A l'instar de la régression linéaire simple, le coefficient de corrélation linéaire multiple est égal à la racine carrée du coefficient de détermination :

$$R = \sqrt{R^2}$$

En revanche, à la différence de la régression simple, il ne correspond plus à la corrélation entre l'endogène et l'exogène, tout simplement parce que nous avons plusieurs exogènes dans notre équation.

Dans le cas de la régression linéaire multiple, on montre que le coefficient de corrélation linéaire multiple correspond à la corrélation entre les valeurs observées et les valeurs prédites de l'endogène (Tenenhaus, page 117) c.-à-d.

$$r_{y,\hat{y}} = R \tag{10.4}$$

Cela suggère d'ailleurs de construire le graphique nuage de points confrontant y_i et $\hat{y}_i$ pour évaluer la qualité de la régression. Si le modèle est parfait, les points seraient alignés sur la première bissectrice.

10.1.4 Application aux données "Consommation des véhicules"

Reprenons notre fichier "Consommation de véhicules". Nous exploitons les sorties de la fonction DROITEREG ($SCE = 121.0318$, $SCR = 13.5807$) pour reconstituer le tableau d'analyse de variance. Nous en déduisons le $R^2 = 1 - \frac{SCR}{SCE+SCR} = 1 - \frac{13.5807}{121.0318+13.5807} = 0.89911$ déjà fourni par Excel en réalité (Figure 10.2) [2].

X			Y	Y^
cylindree	puissance	poids	consommation	
846	32	650	5.7	5.4523
993	39	790	5.8	6.2447
899	29	730	6.1	5.7621
1390	44	955	6.5	7.2296
1195	33	895	6.8	6.6789
658	32	740	6.8	5.7402
1331	55	1010	7.1	7.6339
1597	74	1080	7.4	8.4079
1761	74	1100	9	8.5734
2165	101	1500	11.7	10.9571
1983	85	1075	9.5	8.7780
1984	85	1155	9.5	9.1168
1998	89	1140	8.8	9.1333
1580	65	1080	9.3	8.2353
1390	54	1110	8.6	8.0676
1396	66	1140	7.7	8.4164
2435	106	1370	10.8	10.6319
1242	55	940	6.6	7.2939
2972	107	1400	11.7	11.0420
2958	150	1550	11.9	12.4542
2497	122	1330	10.8	10.7853
1998	66	1300	7.6	9.3902
2496	125	1670	11.3	12.2774
1998	89	1560	10.8	10.9094
1997	92	1240	9.2	9.6105
1984	85	1635	11.6	11.1467
2438	97	1800	12.8	12.2875
2473	125	1570	12.7	11.8432

DROITEREG

	poids	puissance	cylindree	constante
a^	0.00423	0.01825	0.00049	1.70205
sigma^(a^)	0.00094	0.01424	0.00078	0.63205
R^2	0.89911	0.75224	#N/A	#N/A
F	71.2965	24	#N/A	#N/A
SCE	121.0318	13.5807	#N/A	#N/A
		SCR		

Tableau d'analyse de variance

Source	SC	ddl	CM
Expliquée	121.0318	3	40.3439
Résiduelle	13.5807	24	0.5659
Totale	134.6125	27	

R^2	0.89911
R^2-ajusté	0.88650

R	0.94822

r(y,y^)	0.94822

Fig. 10.2. Tableau d'analyse de variance, R^2, $\bar{R}^2$ et R - Consommation des véhicules

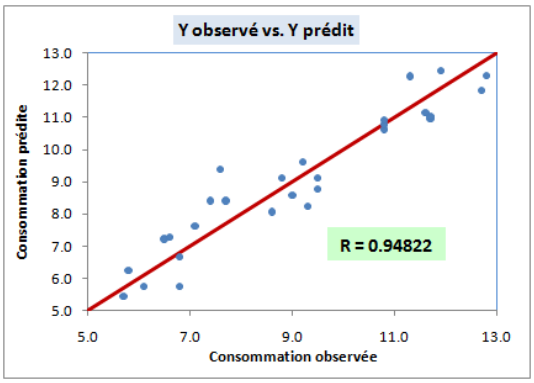


Fig. 10.3. Y observé et Y calculé - Coefficient de corrélation linéaire multiple - Consommation des véhicules

2. `reg_multiple_consommation_automobiles.xlsx` - "anova et R2"

Nous calculons les ratios supplémentaires :

- Le R^2-ajusté, $\bar{R}^2 = 1 - \frac{n-1}{n-p-1}(1-R^2) = 1 - \frac{27}{24}(1-0.89911) = 0.88650$.
- Le coefficient de corrélation linéaire multiple, $R = \sqrt{R^2} = \sqrt{0.89911} = 0.94822$
- Nous vérifions aisément qu'il est égal au coefficient de corrélation linéaire entre l'endogène observée Y et l'endogène prédite par le modèle $\hat{Y}$, $r_{y,\hat{y}} = 0.94822$.

Un R^2 assez élevé laisse à penser que le modèle est plutôt bon. En construisant le graphique croisant Y et $\hat{Y}$, nous constatons effectivement que les points sont plutôt bien alignés sur la première bissectrice (Figure 10.3).

10.2 Test de significativité globale de la régression

10.2.1 Formulation

Le test de significativité globale consiste à vérifier si le modèle, pris dans sa globalité, est pertinent. L'hypothèse nulle correspond à la situation où aucune des exogènes n'emmène de l'information utile dans l'explication de Y c.-à-d. le modèle ne sert à rien. Le test s'écrit :

$$\begin{cases} H_0 : a_1 = a_2 = \cdots = a_p = 0 \\ H_1 : \exists j \,/\, a_j \neq 0 \end{cases}$$

Remarque 5 (Le cas de la constante). **Attention, seuls les coefficients associés aux variables X_j sont inclus dans le test**. En effet, c'est bien l'influence des exogènes sur l'endogène que l'on cherche à établir. Si H_0 est vrai, on sait que a_0 est égal à la moyenne des Y. Sauf cas particulier des variables centrées, la moyenne des Y est non nulle. Inclure a_0 dans le test fausserait les résultats.

Remarque 6 (Tester la significativité du R^2). Un autre manière d'exprimer le test consiste à poser la question : est-ce que le R^2 est significativement supérieur à 0 ? Très prisée des anglo-saxons (cf. quelques références dans la section 3.1), on retrouve très rarement cette formulation dans les ouvrages francophones. Qu'importe. L'essentiel est de bien comprendre que l'on cherche à établir le pouvoir explicatif des X_j, pris dans leur globalité, sur Y.

10.2.2 Statistique de test et région critique

La statistique de test est extraite du tableau d'analyse de variance, elle s'écrit

$$F = \frac{CME}{CMR} = \frac{SCE/p}{SCR/(n-p-1)} \tag{10.5}$$

Nous pouvons aussi l'exprimer à partir du coefficient de détermination

$$F = \frac{R^2/p}{(1-R^2)/(n-p-1)} \tag{10.6}$$

Sous H_0, F suit une loi de Fisher $\mathcal{F}(p, n-p-1)$. Au risque α, la région critique (rejet de H_0) du test correspond aux valeurs exceptionnellement grandes de F :

$$R.C. : F > F_{1-\alpha}(p, n-p-1) \tag{10.7}$$

Application aux données "Consommation de véhicules. Revenons aux résultats de notre régression sur les véhicules (Figure 10.2). A partir du R^2, nous obtenons :

$$F = \frac{R^2/p}{(1-R^2)/(n-p-1)} = \frac{0.89911/3}{(1-0.89911)/(24)} = 71.2965$$

En passant par le tableau d'analyse de variance, nous aurions

$$F = \frac{SCE/p}{SCR/(n-p-1)} = \frac{121.0318/3}{13.5807/(24)} = \frac{40.3439}{0.5659} = 71.2965$$

On constate par ailleurs que la valeur de F est directement fournie par DROITEREG (Figure 10.2).

Nous la comparons avec le quantile d'ordre 0.95 pour un test à 5%, à savoir [3] $F_{0.95}(3, 24) = 3.00879$. Nous constatons que nous sommes dans la région critique. Au risque 5%, nous concluons que le modèle est globalement significatif : la cylindrée, la puissance et poids, pris dans leur globalité, emmènent de l'information pertinente sur la consommation.

En passant par le calcul de la probabilité critique, nous aurions obtenu [4] $\alpha' = 4.26 \times 10^{-12}$, largement inférieure à $\alpha = 5\%$. La conclusion est cohérente.

10.3 Test de significativité d'un coefficient

10.3.1 Définition du test

Après avoir établi la significativité globale de la régression, nous devons évaluer la pertinence des variables prises individuellement. La démarche est analogue à celle définie pour la régression simple (section 3.2.3). Toujours parce que $\varepsilon_i \equiv \mathcal{N}(0, \sigma_\varepsilon)$, on montre que

$$\frac{\hat{a}_j - a}{\hat{\sigma}_{\hat{a}_j}} \equiv \mathcal{T}(n-p-1) \tag{10.8}$$

A partir de là, nous pouvons définir les tests de conformité à un standard, les intervalles de confiance et, ce qui nous intéresse dans cette section, les tests de significativité.

Le test consiste à opposer :

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

3. INVERSE.LOI.F(0.05 ;3 ;24) dans Excel.
4. LOI.F(71.2965 ;3 ;24) dans Excel.

Le retrait de la variable X_j de la régression est possible si l'hypothèse nulle est avérée. *Par rapport aux autres variables*, la contribution de X_j dans l'explication de Y n'est pas significative. Méfiance néanmoins, des problèmes de colinéarité peuvent parfois perturber les résultats. Nous en reparlerons lors du traitement du fichier "Consommation de véhicules".

La statistique de test s'écrit :

$$t_{\hat{a}_j} = \frac{\hat{a}_j}{\hat{\sigma}_{\hat{a}_j}} \tag{10.9}$$

Et la région critique pour un risque α, le test étant bilatéral :

$$R.C. : |t_{\hat{a}_j}| > t_{1-\frac{\alpha}{2}}(n-p-1) \tag{10.10}$$

10.3.2 Tests pour la régression "Consommation des véhicules"

Voyons ce qu'il en est concernant notre régression "Consommation des véhicules". DROITEREG nous fournit à la fois $\hat{a}_j$ et $\hat{\sigma}_{\hat{a}_j}$. Nous sommes armés pour définir les tests de significativité (Figure 10.4) [5].

DROITEREG				
	poids	puissance	cylindree	constante
a^	0.00423	0.01825	0.00049	1.70205
sigma^(a^)	0.00094	0.01424	0.00078	0.63205
R^2	0.89911	0.75224	#N/A	#N/A
F	71.2965	24	#N/A	#N/A
SCE	121.0318	13.5807	#N/A	#N/A
	SCR			

Test de significativité des coefficients				
t-calculé	4.51838	1.28161	0.63304	2.69289
ddl	24	24	24	24
t-théorique	2.06390	2.06390	2.06390	2.06390
p-value	0.00014	0.21222	0.53269	0.01271

Fig. 10.4. Tests de significativité des coefficients - Consommation des véhicules

Nous n'avons pas intégré la constante dans la procédure. En effet, comme nous l'avons souligné dans la régression simple, remettre en cause a_0 modifie la nature de la régression. Pour chaque variable, nous avons calculé la statistique de test :

$$\begin{cases} t_{\hat{a}_1} = \frac{0.00049}{0.00078} = 0.63304 \\ t_{\hat{a}_2} = \frac{0.01825}{0.01424} = 1.28161 \\ t_{\hat{a}_3} = \frac{0.00423}{0.00094} = 4.51838 \end{cases}$$

Pour un risque $\alpha = 5\%$, le seuil critique [6] est égal à $t_{0.975}(24) = 2.06390$. Nous constatons que seul le coefficient a_3 associé à ($X3$ - *Poids*) est significatif, puisque $|t_{\hat{a}_3}| = 4.51838 > t_{0.975} = 2.06390$.

5. `reg_multiple_consommation_automobiles.xlsx` - "test.signif"
6. LOI.STUDENT.INVERSE(0.05 ;24) dans Excel.

Ni *cylindrée*, ni *puissance* en revanche ne semblent pertinentes. Pris individuellement, il ne semblent pas contribuer significativement dans l'explication de la consommation. C'est étrange si l'on connaît un peu les automobiles. Nous reviendrons sur cet aspect dans la section suivante (section 10.4).

Une autre manière de parvenir aux mêmes conclusions est de calculer la probabilité critique[7], nous les affichons dans notre feuille Excel (Figure 10.4) :

$$\begin{cases} \alpha'_{\hat{a}_1} = 0.53269 \\ \alpha'_{\hat{a}_2} = 0.21222 \\ \alpha'_{\hat{a}_3} = 0.00014 \end{cases}$$

10.3.3 Tests pour la régression "Cigarettes" incluant la variable ALEA

Pour montrer l'intérêt du R^2-ajusté, nous avions décrit l'exemple d'une régression où l'on cherchait à expliquer la quantité d'oxyde de carbone ingérée par les personnes fumant des cigarettes (Figure 10.1). L'adjonction d'une variable ALEA générée aléatoirement parmi les exogènes provoquait une baisse du $\bar{R}^2$, indiquant sa non pertinence dans la régression. Voyons si le test de significativité permet d'établir le même résultat.

ALEA est la 4-ème variable de la régression (Figure 10.5) [8], nous avons $\hat{a}_4 = 0.81653$ et $\hat{\sigma}_{\hat{a}_4} = 0.96657$. Nous formons la statistique de test

$$t_{\hat{a}_4} = \frac{0.81653}{0.96657} = 0.84477$$

TAR (mg)	NICOTINE (mg)	WEIGHT (g)	ALEA	CO (mg)
14.1	0.86	0.9853	0.26780	13.6
16	1.06	1.0938	0.35783	16.6
8	0.67	0.928	0.12693	10.2
4.1	0.4	0.9462	0.22795	5.4
15	1.04	0.8885	0.10896	15
8.8	0.76	1.0267	0.03880	9
12.4	0.95	0.9225	0.39589	12.3
16.6	1.12	0.9372	0.27504	16.3
14.9	1.02	0.8858	0.85236	15.4
13.7	1.01	0.9643	0.16240	13
15.1	0.9	0.9316	0.67295	14.4
7.8	0.57	0.9705	0.67509	10
11.4	0.78	1.124	0.84745	10.2
9	0.74	0.8517	0.54966	9.5
1	0.13	0.7851	0.43216	1.5
17	1.26	0.9186	0.97986	18.5
12.8	1.08	1.0395	0.39638	12.6
15.8	0.96	0.9573	0.43537	17.5
4.5	0.42	0.9106	0.55342	4.9
14.5	1.01	1.007	0.65460	15.9
7.3	0.61	0.9806	0.51565	8.5
8.6	0.69	0.9693	0.50194	10.6
15.2	1.02	0.9496	0.72085	13.9
12	0.82	1.1184	0.81714	14.9

DROITEREG (TAR, NICOTINE, WEIGHT, ALEA) - 2

	alea	weight	nicotine	tar	constante
a^	0.81653	1.87048	0.93450	0.85569	-0.72260
sigma^(a^)	0.96657	3.21095	3.31268	0.20048	2.99961
	0.93733	1.16822	#N/A	#N/A	#N/A
	71.04289	19	#N/A	#N/A	#N/A
	387.82	25.93	#N/A	#N/A	#N/A
t calculé	0.84477	0.58253	0.28210	4.26811	
p-value	0.40875	0.56706	0.78092	0.00042	

Fig. 10.5. Tests de significativité du coefficient de ALEA - Cigarettes

Nous en déduisons la probabilité critique $\alpha'_{\hat{a}_4} = 0.40875$. Définitivement, la variable ALEA n'est absolument pas pertinente dans la régression.

7. LOI.STUDENT(ABS(t-calculé) ;24 ;2) dans Excel. Le dernier paramètre correspond à un test bilatéral.
8. `cigarettes-regressionmultiple.xls` - "tests.coefs.avec.alea"

On constate par ailleurs que ni le poids (weight) ni la nicotine ne semblent peser non plus dans l'explication de CO.

10.4 Test de significativité d'un bloc de coefficients

10.4.1 Principe du test

Dans notre exemple des "Consommation des véhicules" (section 10.3.2), nous avions constaté que la cylindrée et la puissance n'étaient pas individuellement significatifs à 5%. Est-ce que cela veut dire que nous pouvons retirer directement les deux variables de la régression ?

Clairement non. Nous ne pouvons nous baser sur les tests individuels pour supprimer en bloc des exogènes du modèle. En effet, les coefficients correspondent à des contributions partielles, tenant compte de l'impact des autres variables. Si ces dernières sont corrélées, elles se gênent mutuellement dans la régression, partageant leur influence au point que, individuellement, elles ne semblent pas intéressantes.

Pour évaluer la contribution de q variables *prises simultanément*, nous introduisons un nouveau type de test. L'hypothèse nulle du test s'écrit (sans restreindre la généralité du propos, nous ne testons pas forcément les q premiers coefficients) :

$$H_0 : a_1 = a_2 = \cdots = a_q = 0$$

Pour résoudre ce problème, nous confrontons deux régressions : celle sous hypothèse nulle, avec $(p-q)$ variables explicatives, nous obtenons un premier coefficient de détermination R_0^2 ; et celle avec les p variables, nous obtenons R_1^2. Les deux modèles sont imbriqués et, forcément, $R_1^2 \geq R_0^2$. Nous posons alors la question suivante : est-ce que l'adjonction des q exogènes supplémentaires dans la régression induit une augmentation *significative* du R^2 au risque α.

Formons la statistique de test F (Jaccard et Turrisi, page 12 ; Hardy, page 24) :

$$F = \frac{(R_1^2 - R_0^2)/q}{(1 - R_1^2)/(n - p - 1)} \tag{10.11}$$

Sous H_0, elle suit une loi de Fisher à $(q, n - p - 1)$ degrés de liberté.

Un autre manière de voir les choses est de considérer que l'on oppose le modèle incluant la totalité des variables avec la régression sous la contrainte H_0.

10.4.2 Tester la nullité simultanée des coefficients de "cylindrée" et "puissance"

Testons donc la nullité simultanée des coefficients de cylindrée et puissance dans la régression "Consommation de véhicules" (Figure 10.6) [9].

9. `reg_multiple_consommation_automobiles.xlsx` - "test.signif.cyl.puissance"

X			Y
cylindree	puissance	poids	consommation
846	32	650	5.7
993	39	790	5.8
899	29	730	6.1
1390	44	955	6.5
1195	33	895	6.8
658	32	740	6.8
1331	55	1010	7.1
1597	74	1080	7.4
1761	74	1100	9
2165	101	1500	11.7
1983	85	1075	9.5
1984	85	1155	9.5
1998	89	1140	8.8
1580	65	1080	9.3
1390	54	1110	8.6
1396	66	1140	7.7
2435	106	1370	10.8
1242	55	940	6.6
2972	107	1400	11.7
2958	150	1550	11.9
2497	122	1330	10.8
1998	66	1300	7.6
2496	125	1670	11.3
1998	89	1560	10.8
1997	92	1240	9.2
1984	85	1635	11.6
2438	97	1800	12.8
2473	125	1570	12.7

DROITEREG

	poids	constante
	0.00669386	1.05269123
	0.00053388	0.65925237
$R^2(1)$	0.85808	0.85718963
	157.202573	26
	115.508374	19.1041258

DROITEREG

	poids	puissance	cylindree	constante
	0.00423	0.01825	0.00049	1.70205
	0.00094	0.01424	0.00078	0.63205
$R^2(3)$	0.89911	0.75224	#N/A	#N/A
	71.2965	24	#N/A	#N/A
	121.0318	13.5807	#N/A	#N/A

F	4.88057
ddl1	2
ddl2	24
p-value	0.01665

r(cyl,puis)	0.94755

Fig. 10.6. Significativité simultanée des coefficients de cylindrée et puissance - Consommation des véhicules

Dans un premier temps, nous réalisons la régression avec la seule variable poids. Nous obtenons $R_0^2 = 0.85808$. Dans un deuxième temps, nous construisons le modèle incluant toutes les variables c.-à-d. au modèle précédent, nous adjoignons les $(q = 2)$ variables cylindrée et puissance que nous souhaitons éprouver. Nous obtenons $R_1^2 = 0.89911$. Le coefficient de détermination est plus élevée, il ne peut pas en être autrement. Mais est-ce qu'il est significativement plus grand ?

Nous formons la statistique de test

$$F = \frac{(R_1^2 - R_0^2)/q}{(1 - R_1^2)/(n - p - 1)} = \frac{(0.89911 - 0.85808)/2}{(1 - 0.89911)/(28 - 3 - 1)} = 4.88057$$

Avec la loi de Fisher à $(2, 24)$ degrés de liberté, nous obtenons une p-value de 0.01665. Contrairement au test individuel où ils n'étaient pas significatifs à 5%, les coefficients pris en bloc le deviennent pour le même niveau de risque. Cette apparente contradiction n'en est pas une, elle s'explique simplement par la forte corrélation entre les deux variables, en effet $r_{cylindree,puissance} = 0.94755$. Les deux variables se neutralisent dans la régression. Clairement, opérer une sélection de variables serait appropriée ici. Vient alors une question cruciale : laquelle retenir ? Nous étudions en détail cette question dans le fascicule consacré à la pratique de la régression [13] (chapitre 3).

10.4.3 Tester la nullité de 3 coefficients dans la régression "Cigarettes"

Lors du traitement du fichier "Cigarettes" (section 10.3.3), nous avons montré que le coefficient associé à ALEA n'était pas significatif à 5%. Dans le même temps, nous avons constaté qu'il en était de même concernant les coefficients de NICOTINE et WEIGHT. Dans cette section, nous allons tester la nullité simultanée des $q = 3$ coefficients.

Nous opposons "$H_0 : a_{alea} = a_{nicotine} = a_{weight} = 0$" à "$H_1$: un de ces coefficients est non nul".

TAR (mg)	NICOTINE (mg)	WEIGHT (g)	ALEA	CO (mg)
14.1	0.86	0.9853	0.26780	13.6
16	1.06	1.0938	0.35783	16.6
8	0.67	0.928	0.12693	10.2
4.1	0.4	0.9462	0.22795	5.4
15	1.04	0.8885	0.10896	15
8.8	0.76	1.0267	0.03880	9
12.4	0.95	0.9225	0.39589	12.3
16.6	1.12	0.9372	0.27504	16.3
14.9	1.02	0.8858	0.85236	15.4
13.7	1.01	0.9643	0.16240	13
15.1	0.9	0.9316	0.67295	14.4
7.8	0.57	0.9705	0.67509	10
11.4	0.78	1.124	0.84745	10.2
9	0.74	0.8517	0.54966	9.5
1	0.13	0.7851	0.43216	1.5
17	1.26	0.9186	0.97986	18.5
12.8	1.08	1.0395	0.39638	12.6
15.8	0.96	0.9573	0.43537	17.5
4.5	0.42	0.9106	0.55342	4.9
14.5	1.01	1.007	0.65460	15.9
7.3	0.61	0.9806	0.51565	8.5
8.6	0.69	0.9693	0.50194	10.6
15.2	1.02	0.9496	0.72085	13.9
12	0.82	1.1184	0.81714	14.9

DROITEREG (TAR, NICOTINE, WEIGHT, ALEA)

	alea	weight	nicotine	tar	constante
a^	0.81653	1.87048	0.93450	0.85569	-0.72260
sigma^(a^)	0.96657	3.21095	3.31268	0.20048	2.99961
R²	0.93733	1.16822	#N/A	#N/A	#N/A
	71.04289	19	#N/A	#N/A	#N/A
	387.82	25.93	#N/A	#N/A	#N/A

t calculé	0.84477	0.58253	0.28210	4.26811
p-value	0.40875	0.56706	0.78092	0.00042

DROITEREG (TAR)

	tar	const
a^	0.92813	1.41285
sigma^(a^)	0.05283	0.64822
R²	0.93346	1.11865
	308.63769	22
	386.21948	27.53011

F	0.39082
ddl1	3
ddl2	19
p-value	0.76096

Fig. 10.7. Tests de significativité simultanée de $q = 3$ coefficients - Cigarettes

Pour ce faire, nous réalisons les deux régressions (Figure 10.7) [10] : la première avec la totalité ($p = 4$) des variables, nous obtenons $R_1^2 = 0.93733$ avec un degré de liberté de $(n - p - 1 = 24 - 4 - 1 = 19)$; la seconde avec TAR seulement, le coefficient de détermination diminue et passe à $R_0^2 = 0.93346$, avec un degré de liberté $n - (p - q) - 1 = 24 - (4 - 3) - 1 = 22$. Formons la statistique de test :

$$F = \frac{(R_1^2 - R_0^2)/q}{(1 - R_1^2)/(n - p - 1)} = \frac{(0.93733 - 0.93346)/3}{(1 - 0.93733)/(19)} = 0.39082$$

Avec une loi de Fisher à $(3, 19)$ degrés de liberté, nous obtenons un p-value $= 0.76096$, largement supérieure à $\alpha = 5\%$. Clairement, nous pouvons retirer le bloc de variables (alea, nicotine et weight) de la régression, elles n'emmènent rien *par rapport à TAR* pour expliquer CO.

10.4.4 Exprimer la statistique de test avec les SCR

Notons que la statistique de test peut s'écrire sous la forme d'une confrontation entre les erreurs résiduelles. Si SCR_0 est la somme des carrés des résidus sous la contrainte H_0 (q coefficients sont nuls,

10. `cigarettes-regressionmultiple.xls` - "tests.bloc.coefs"

la régression comporte $p - q$ variables) et SCR_1 celle de la régression incluant toutes les p variables, forcément $(SCR_0 \geq SCR_1)$ [11], alors :

$$F = \frac{(SCR_0 - SCR_1)/q}{SCR_1/(n-p-1)} \tag{10.12}$$

La valeur obtenue est identique à celle basée sur les coefficients de détermination (équation 10.11).

Voyons notre exemple de la nullité de cylindrée et puissance dans la régression "Consommation des véhicules" (Figure 10.6). Nous y lisons les valeurs adéquates :

- $SCR_0 = 19.1041$
- $SCR_1 = 13.5807$
- Nous en déduisons

$$F = \frac{(SCR_0 - SCR_1)/q}{SCR_1/(n-p-1)} = \frac{(19.1041 - 13.5807)/2}{13.5807/(28-3-1)} = 4.88057$$

Les valeurs de F sont exactement les mêmes.

11. L'erreur résiduelle de la régression non contrainte est toujours plus faible que celle de la régression contrainte. Attention, si on se base sur le coefficient de détermination, la relation est inversée c.-à-d. nous avons forcément $(R_1^2 \geq R_0^2)$. En effet, $R^2 = 1 - \frac{SCR}{SCT}$; et SCT - basé uniquement sur les valeurs de Y - est toujours constant quelle que soit le modèle étudié.

11

Généralisation de l'étude des coefficients

Concernant l'inférence sur les coefficients, nous pouvons aller plus loin que les simples tests de significativité. Dans ce chapitre, nous décrivons la panoplie des outils que l'on pourrait mettre en oeuvre pour les étudier. Nous verrons ainsi que tous les tests exposés dans ce fascicule peuvent s'écrire sous une forme générique unique, le test de combinaisons linéaires des coefficients.

11.1 Inférence sur les coefficients

11.1.1 Intervalle de confiance

La distribution de $\hat{a}_j$ telle que nous l'avons décrite précédemment (Équation 10.8) est valable quel que soit le voisinage. Nous pouvons définir facilement un intervalle de confiance des coefficients au niveau de confiance $(1-\alpha)$ avec

$$\hat{a}_j \pm t_{1-\frac{\alpha}{2}} \times \hat{\sigma}_{\hat{a}_j} \tag{11.1}$$

	DROITEREG			
	poids	puissance	cylindree	constante
a^	0.00423	0.01825	0.00049	1.70205
sigma^(a^)	0.00094	0.01424	0.00078	0.63205
R^2	0.89911	0.75224	#N/A	#N/A
F	71.2965	24	#N/A	#N/A
SCE	121.0318	13.5807	#N/A	#N/A
t_(1-alpha/2)	2.06390	2.06390	2.06390	2.06390
b.basse	0.00230	-0.01114	-0.00112	0.39756
b.haute	0.00616	0.04764	0.00210	3.00654

Fig. 11.1. Intervalle de confiance des coefficients - Consommation des véhicules

Nous reprenons notre fichier des Consommations de véhicules. Nous souhaitons construire les intervalles de variation des coefficients au niveau de confiance 95% (Figure 11.1)[1]. Nous utilisons le quantile

1. `reg_multiple_consommation_automobiles.xlsx` - "intv.conf.coefs"

$t_{0.975}(24) = 2.06390$ de la loi de Student à $(n - p - 1 = 24)$ degrés de liberté. Avec les $\hat{a}_j$ et $\hat{\sigma}_{\hat{a}_j}$, nous formons les bornes basses et bornes hautes. Pour la variables poids, nous obtenons :

$$bb(a_{poids}) = 0.00423 - 2.06390 \times 0.00094 = 0.00230$$
$$bh(a_{poids}) = 0.00423 + 2.06390 \times 0.00094 = 0.00616$$

Les résultats sont cohérents avec le test de significativité. A savoir, le coefficient est significatif au risque α si l'intervalle de confiance au niveau $(1-\alpha)$ ne contient pas la valeur 0. C'est le cas du coefficient de poids, pas pour ceux de puissance et cylindrée.

11.1.2 Test de conformité à un standard

Nous pouvons également mettre en place des tests de conformité à un standard pour répondre à des problèmes très concrets.

Concernant la consommation des véhicules, un expert du domaine peut nous poser la question suivante par exemple : est-ce que l'on peut montrer que, toutes choses égales par ailleurs, l'augmentation du poids de 400 kg des véhicules induit une augmentation de la consommation supérieure à 1 litre/100 km ?

Pour répondre à cela, nous opposons :

$$\begin{cases} H_0 : a_{poids} = \frac{1}{400} = 0.0025 \\ H_1 : a_{poids} > \frac{1}{400} \end{cases}$$

Nous formons la statistique de test

$$t_{a_{(poids}>0.0025)} = \frac{\hat{a}_{poids} - 0.0025}{\hat{\sigma}_{\hat{a}_{poids}}}$$

Au risque α, la région critique s'écrit, le test étant unilatéral :

$$R.C. : t_{(a_{poids}>0.0025)} > t_{1-\alpha}(n - p - 1)$$

Sur nos données (Figure 11.2) [2], cela donne

$$t_{a_{(poids}>0.0025)} = \frac{0.00423 - 0.0025}{0.00094} = 1.84722$$

A comparer avec $t_{0.95}(24) = 1.71088$. Puisque nous sommes dans la région critique au risque 5%, nous pouvons dire qu'une augmentation du poids des véhicules de 400 kg, *à puissance et cylindrée égale*, induit une augmentation de la consommation supérieure à 1 L / 100 km.

2. `reg_multiple_consommation_automobiles.xlsx` - "test.poids.conformité"

DROITEREG				
	poids	puissance	cylindree	constante
a^	0.00423	0.01825	0.00049	1.70205
sigma^(a^)	0.00094	0.01424	0.00078	0.63205
R^2	0.89911	0.75224	#N/A	#N/A
F	71.2965	24	#N/A	#N/A
SCE	121.0318	13.5807	#N/A	#N/A
		SCR		

	poids
t-calculé	1.84722
ddl	24
t-théorique	1.71088
p-value	0.03854

Fig. 11.2. Test de conformité à un standard du coefficient de "poids" - Consommation des véhicules

11.2 Test de conformité pour un bloc de coefficients

11.2.1 Principe du test pour un groupe de coefficient

Nous pouvons généraliser le test de conformité à un groupe de q coefficients (Bourbonnais, page 60 ; Giraud et Chaix, pages 102 à 105). Les hypothèses s'écrivent (en tout généralité, on teste q coefficients, pas nécessairement les q premières, nous adoptons cette écriture pour simplifier les notations) :

$$\begin{cases} H_0 : \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_q \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_q \end{pmatrix} \Leftrightarrow a_{(q)} = c_{(q)} \\ H_1 : \exists j \,/\, a_j \neq c_j \end{cases} \tag{11.2}$$

Les c_j représentent les standards auxquels nous comparons nos coefficients.

Attention, nous ne pouvons absolument réduire ce test à une succession de tests individuels. Il est tentant d'utiliser des règles du type "si on accepte H_0 pour tous les tests pris individuellement, alors on accepte H_0 pour l'égalité simultanée" ou bien "si on rejette H_0 au moins une fois sur un des tests individuels, alors on rejette H_0 pour le test simultané". Ces formulations sont erronées tout simplement parce qu'elles ne tiennent pas compte de l'interaction entre les variables, traduite numériquement par les covariances des coefficients. Ces dernières interviennent dans la construction de la statistique de test. Elle s'écrit :

$$F = \frac{1}{q} \left[\hat{a}_{(q)} - c_{(q)}\right]' \hat{\Omega}^{-1}_{\hat{a}_{(q)}} \left[\hat{a}_{(q)} - c_{(q)}\right] \tag{11.3}$$

$\hat{a}_{(q)}$ représente le sous-vecteur des coefficients estimés mis à contribution dans le test ; $\hat{\Omega}_{\hat{a}_{(q)}}$ est la matrice de variance covariance réduite aux coefficients testés.

Sous H_0, la quantité F suit une loi de Fisher $\mathcal{F}(q, n-p-1)$.

11.2.2 Reconsidérer le test de significativité d'un bloc de coefficients

Le test de significativité est un cas particulier du test de conformité. Pour illustrer cette technique, nous allons reprendre notre exemple de nullité simultanée des coefficients de cylindrée et puissance dans la régression "Consommation des véhicules" (section 10.4.2).

Nous avons $q = 2$ coefficients dans la procédure. L'hypothèse nulle s'écrit :

$$\left\{ H_0 : \begin{pmatrix} a_{puissance} \\ a_{cylindree} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right.$$

MVCV(a^)	constante	cylindree	puissance	poids
constante	3.9949E-01	-8.3229E-05	3.1611E-03	-3.9632E-04
cylindree	-8.3229E-05	6.0783E-07	-8.9923E-06	-2.6529E-07
puissance	3.1611E-03	-8.9923E-06	2.0279E-04	-2.2162E-06
poids	-3.9632E-04	-2.6529E-07	-2.2162E-06	8.7595E-07

MVCV[a(2)]	cylindree	puissance
cylindree	6.0783E-07	-8.9923E-06
puissance	-8.9923E-06	2.0279E-04

MVCV[a(2)]^(-1)	cylindree	puissance
cylindree	4782997.0660	212097.2404
puissance	212097.2404	14336.5614

	a^	c	diff
cylindree	0.000494	0	0.000494
puissance	0.018251	0	0.018251

F	4.88057
F_0.95(2,24)	3.40283
p-value	0.01665

H0 : a(cylindrée) = 0 ET a(puissance) = 0
Conclusion Rejet H0

Fig. 11.3. Test de conformité à un standard d'un bloc de coefficients - Consommation des véhicules

Nous avons élaboré une feuille de calcul Excel pour former la statistique de test (Figure 11.3) [3]. Nous distinguons :

- A partir de la matrice de variance covariance des coefficients $\hat{\Omega}_{\hat{a}}$,
- Nous extrayons la sous-matrice correspondant aux coefficients de cylindrée et puissance

$$\hat{\Omega}_{\hat{a}_{(2)}} = \begin{pmatrix} 6.0783 \times 10^{-7} & -8.9923 \times 10^{-6} \\ -8.9923 \times 10^{-6} & 2.0279 \times 10^{-4} \end{pmatrix}$$

- Que nous inversons

$$\hat{\Omega}^{-1}_{\hat{a}_{(2)}} = \begin{pmatrix} 4782997.0660 & 212097.2404 \\ 212097.2404 & 14336.5614 \end{pmatrix}$$

3. `reg_multiple_consommation_automobiles.xlsx` - "test.conformité.cyl.puissance"

– Nous formons la statistique de test en confrontant les coefficients estimés avec les standards :

$$F = \frac{1}{2}\left(0.000494 - 0\,;\, 0.018251 - 0\right)\hat{\Omega}^{-1}_{\hat{a}_{(q)}}\begin{pmatrix}0.000494 - 0\\ 0.018251 - 0\end{pmatrix} = 4.88057$$

– Le seuil critique est $F_{0.95}(2;24) = 3.40283$. Nous sommes dans la région critique. Au risque 5%, nous rejetons l'hypothèse nulle d'égalité des coefficients (la p-value est $\alpha' = 0.01665$).

Les coefficients, qui étaient égaux à 0 pris individuellement (acceptation de H_0), deviennent non nuls lorsque nous les traitons en bloc (rejet de H_0). Tout simplement parce que nous avons pris en compte leur covariance dans la procédure.

Notons un résultat intéressant, cette procédure est totalement équivalente au test de significativité basé sur la comparaison des coefficients de détermination R^2 mis en oeuvre sur les mêmes données (section 10.4). La valeur de la statistique de test est exactement la même.

11.2.3 Test de conformité pour plusieurs coefficients - Données "Cigarettes"

Bien évidemment, la procédure peut aller au delà du test de significativité. Reprenons l'exemple des données "Cigarettes". Les exogènes sont dans l'ordre TAR (X1), NICOTINE (X2), WEIGHT (X3), ALEA(X4), nous souhaitons mettre en place le test suivant :

$$\begin{cases} H_0 : \begin{pmatrix}a_1\\ a_2\\ a_4\end{pmatrix} = \begin{pmatrix}1\\ 1\\ 0\end{pmatrix} \\ H_1 : \begin{pmatrix}a_1\\ a_2\\ a_4\end{pmatrix} \neq \begin{pmatrix}1\\ 1\\ 0\end{pmatrix} \end{cases} \tag{11.4}$$

Par rapport à l'exemple précédent, l'originalité est qu'il s'agit ici d'un test de conformité quelconque ; la difficulté réside dans le fait que les coefficients analysés se rapportent à des colonnes non adjacentes du tableau de données. Il faudra faire très attention lors de l'extraction des valeurs dans la matrice de variance covariance des coefficients.

Les calculs sont détaillés dans une feuille Excel (Figure 11.4) [4] :

– Nous avons exécuté la fonction DROITEREG pour obtenir les coefficients. Ils sont dans l'ordre inverse des colonnes de données dans le tableau de résultats. Pour éviter les confusions, énumérons-les

4. `cigarettes-regressionmultiple.xls` - "tests.conformite.coefs"

const	TAR (mg)	NICOTINE (mg)	WEIGHT (g)	ALEA	CO (mg)
1	14.1	0.86	0.9853	0.26780	13.6
1	16	1.06	1.0938	0.35783	16.6
1	8	0.67	0.928	0.12893	10.2
1	4.1	0.4	0.9462	0.22795	5.4
1	15	1.04	0.8885	0.10896	15
1	8.8	0.76	1.0267	0.03880	9
1	12.4	0.95	0.9225	0.39589	12.3
1	16.6	1.12	0.9372	0.27504	16.3
1	14.9	1.02	0.8858	0.85236	15.4
1	13.7	1.01	0.9643	0.16240	13
1	15.1	0.9	0.9316	0.67295	14.4
1	7.8	0.57	0.9705	0.67509	10
1	11.4	0.78	1.124	0.84745	10.2
1	9	0.74	0.8517	0.54966	9.5
1	1	0.13	0.7851	0.43216	1.5
1	17	1.26	0.9186	0.97986	18.5
1	12.8	1.08	1.0396	0.39638	12.6
1	15.8	0.96	0.9573	0.43537	17.5
1	4.5	0.42	0.9106	0.55342	4.9
1	14.5	1.01	1.007	0.65460	15.9
1	7.3	0.61	0.9806	0.51565	8.5
1	8.6	0.69	0.9693	0.50194	10.6
1	15.2	1.02	0.9496	0.72085	13.9
1	12	0.82	1.1184	0.81714	14.9

DROITEREG

	a4	a3	a2	a1	a0
	alea	weight	nicotine	tar	constante
a^	0.81653	1.87048	0.93450	0.85569	-0.72260
	0.96657	3.21095	3.31268	0.20048	2.99961
	0.93733	1.16822	#N/A	#N/A	#N/A
	71.04289	19	#N/A	#N/A	#N/A
	387.82	25.93	#N/A	#N/A	#N/A

(X'X)

	const	tar	nicotine	weight	alea
const	24.0000	275.6000	19.8800	23.0921	11.5665
tar	275.6000	3613.1600	254.1770	267.4617	137.7298
nicotine	19.8800	254.1770	18.0896	19.2668	9.8011
weight	23.0921	267.4617	19.2668	22.3637	11.1847
alea	11.5665	137.7298	9.8011	11.1847	7.1289

(X'X)^-1

	const	tar	nicotine	weight	alea
const	6.5930	0.0685	-1.0121	-6.6833	-0.1433
tar	0.0685	0.0295	-0.4656	-0.0084	-0.0267
nicotine	-1.0121	-0.4656	8.0410	-0.4882	0.3488
weight	-6.6833	-0.0084	-0.4882	7.5547	-0.1751
alea	-0.1433	-0.0267	0.3488	-0.1751	0.6846

MVCV(a^)

	const	tar	nicotine	weight	alea
const	8.9977	0.0935	-1.3812	-9.1209	-0.1955
tar	0.0935	0.0402	-0.6355	-0.0115	-0.0365
nicotine	-1.3812	-0.6355	10.9738	-0.6663	0.4760
weight	-9.1209	-0.0115	-0.6663	10.3102	-0.2390
alea	-0.1955	-0.0365	0.4760	-0.2390	0.9343

	a^	ref	diff
tar	0.85569	1	**-0.14431**
nicotine	0.93450	1	**-0.06550**
alea	0.81653	0	**0.81653**

MVCV(tar, nicotine,alea)

	tar	nicotine	alea
tar	0.0402	-0.6355	-0.0365
nicotine	-0.6355	10.9738	0.4760
alea	-0.0365	0.4760	0.9343

INV(MVCV)

	tar	nicotine	alea
tar	302.1196	17.3665	2.9529
nicotine	17.3665	1.0914	0.1223
alea	2.9529	0.1223	1.1234

F	2.22172
ddl1	3
ddl2	19
p-value	0.11880

Fig. 11.4. Test de conformité à un standard d'un bloc de coefficients - Cigarettes

$$\hat{a}_0 = -0.72260$$

$$\hat{a}_1 = 0.85569$$

$$\hat{a}_2 = 0.93450$$

$$\hat{a}_3 = 1.87048$$

$$\hat{a}_4 = 0.81653$$

- Nous observons également l'estimation de l'écart-type de l'erreur, $\hat{\sigma}_\varepsilon = 1.16822$ (en violet dans le tableau DROITEREG).
- Nous calculons successivement $(X'X)$ et $(X'X')^{-1}$ pour obtenir la matrice de variance covariance des coefficients $\hat{\Omega}_{\hat{a}} = \hat{\sigma}^2_\varepsilon \times (X'X)^{-1}$.
- Sur la diagonale de cette matrice, nous avons les variances. On remarque par exemple pour la variable TAR que $\sqrt{\hat{\sigma}^2_{\hat{a}_1}} = \sqrt{0.0402} = 0.20048$, la valeur correspond à l'écart type fourni par DROITEREG (2-ème ligne du tableau).
- Les coefficients mis à contribution dans le test sont a_1, a_2 et a_4. Nous devons piocher les valeurs adéquates des variances et covariances dans $\hat{\Omega}_{\hat{a}}$ (cellules en fond vert) pour former la matrice réduite $\hat{\Omega}_{\hat{a}_{(q)}}$

$$\hat{\Omega}_{\hat{a}_{(q)}} = \begin{pmatrix} 0.0402 & -0.6355 & -0.0365 \\ -0.6355 & 10.9738 & 0.4760 \\ -0.0365 & 0.4760 & 0.9343 \end{pmatrix}$$

- Que nous inversons

$$\hat{\Omega}^{-1}_{\hat{a}_{(q)}} = \begin{pmatrix} 302.1196 & 17.3665 & 2.9529 \\ 17.3665 & 1.0914 & 0.1223 \\ 2.9529 & 0.1223 & 1.1234 \end{pmatrix}$$

– Reste à former la statistique de test :

$$F = \frac{1}{3}\left(0.85569-1;\, 0.93450-1;\, 0.81653-0\right)\hat{\Omega}^{-1}_{\hat{a}_{(q)}}\begin{pmatrix} 0.85569-1 \\ 0.93450-1 \\ 0.81653-0 \end{pmatrix} = 2.22172$$

– Avec la loi de Fisher $\mathcal{F}(q=3, n-p-1=19)$, nous obtenons un probabilité critique de $\alpha' = 0.11880$.

– Au risque 5%, nous ne pouvons pas rejeter l'hypothèse nulle. Les données ne la contredisent pas.

11.2.4 Cas particulier : lorsque $q = 1$

Lorsque $q = 1$, nous retrouvons le test de conformité d'un coefficient (bilatéral) tel que nous l'avons décrit ci dessus (section 11.1.2). En effet, dans ce cas, pour un coefficient quelconque $\hat{a}_j$, l'inverse $(\hat{\sigma}^2_{\hat{a}_j})^{-1}$ devient $\frac{1}{\hat{\sigma}^2_{\hat{a}_j}}$, et nous avons :

$$F = \left(\frac{\hat{a}_j - c_j}{\hat{\sigma}_{\hat{a}_j}}\right)^2$$

C'est le carré de la statistique du test de conformité à un standard d'un coefficient de la régression multiple décrite dans la section 11.1.2.

11.3 Test de contraintes linéaires sur les coefficients

La formulation du test de combinaisons linéaires des coefficients permet de couvrir tous les tests exposés dans ce fascicule. C'est déjà intéressant en soi. Il est toujours plaisant intellectuellement de produire une procédure suffisamment globale qui permette de résoudre tous les problèmes possibles (Bourbonnais, page 69 ; Johnston et DiNardo, page 96). Mais au delà de la curiosité scientifique, nous constatons que cette écriture permet d'introduire de nouveaux tests : les tests de comparaisons de coefficients.

11.3.1 Formulation du test de combinaison linéaire

Le test d'hypothèses s'écrit

$$\begin{cases} H_0 : Ra = r \\ H_1 : Ra \neq r \end{cases} \tag{11.5}$$

Où a est le vecteur des coefficients, de dimension $(p+1, 1)$; R est la matrice décrivant les contraintes linéaires de dimension $(q, p+1)$, q désignant le nombre de contraintes ; r est le vecteur des valeurs de référence, de dimensions $(q, 1)$.

Nous utilisons la statistique :

$$F = \frac{\frac{1}{q}(R\hat{a} - r)'[R(X'X)^{-1}R']^{-1}(R\hat{a} - r)}{SCR/(n-p-1)} \tag{11.6}$$

Sous H_0, elle suit une loi de Fisher $(q, n-p-1)$ degrés de liberté.

11.3.2 Écriture de la matrice M pour les tests de conformité

Tout les autres tests peuvent s'écrire avec cette formulation disions-nous. Voyons ce qu'il en est pour les différentes situations que nous avons analysées dans ce document. Nous considérons que la constante (a_0) est en première position dans la matrice $(X'X)^{-1}$, puis nous avons dans l'ordre : cylindrée (a_1), puissance (a_2), poids (a_3).

Tester la significativité du coefficient a_3 c.-à-d. $H_0 : a_3 = 0$

Ici, $q = 1$, $R = \begin{pmatrix} 0 & 0 & 0 & 1 \end{pmatrix}$ et $r = (0)$.

Tester la significativité globale de la régression

L'hypothèse nulle correspond à la nullité simultanée des coefficients associées aux variables ($H_0 : a_1 = a_2 = a_3 = 0$). Nous avons $q = 3$ contraintes, avec

$$R = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, r = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

Tester la nullité des coefficients de cylindrée (a_1) et puissance (a_2)

Dans ce cas, nous avons $q = 2$ contraintes, avec

$$R = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, r = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

11.3.3 Aller plus loin avec les tests portant sur des contraintes linéaires

Pourquoi faire simple quand on peut faire compliqué, n'est-ce pas ? En réalité, le principal intérêt de cette nouvelle formulation est qu'elle ouvre la porte à toute une série de tests qui vont au delà du test de conformité, notamment les tests de comparaison de coefficients ou les test d'égalité de combinaisons linéaires de coefficients à un standard.

Comparaison de coefficients - Consommation des véhicules

Pour rendre la lecture plus simple, notre exemple ne porte que sur $q = 1$ contrainte linéaire. Mais que cela ne nous trompe pas, le passage à $q > 1$ contraintes ne pose aucun problème.

Nous retrouvons notre fichier "Consommation des véhicules" pour illustrer la technique[5].

La puissance spécifique est une notion largement utilisée pour qualifier les moteurs. Il s'agit du nombre de chevaux développés par unité de cylindrée, le plus souvent en litres, soit 1000 cm^3. Les véhicules sportifs développent plus de 100ch/L. Mais plus trivialement, sur les voitures courantes, elle tourne autour de 40ch/L (à peu près la moyenne constatée sur notre fichier).

Nous souhaitons savoir si, dans la régression, les coefficients conservent le même rapport dans leur impact sur la consommation c.-à-d. nous souhaitons tester :

$$\begin{cases} H_0 : 1000 \times a_{cylindree} = 40 \times a_{puissance} \\ H_1 : 1000 \times a_{cylindree} \neq 40 \times a_{puissance} \end{cases}$$

Pour être en adéquation avec la formulation matricielle, nous la ré-écrivons :

$$\begin{cases} H_0 : 0 \times a_0 + 1000 \times a_1 + (-40) \times a_2 + 0 \times a_3 = 0 \\ H_1 : 0 \times a_0 + 1000 \times a_1 + (-40) \times a_2 + 0 \times a_3 \neq 0 \end{cases}$$

On peut la ré-écrire sous la forme de contraintes linéaires sur les coefficients de la régression. Nous avons $q = 1$ dans notre exemple. Nous en déduisons les matrices :

$$R = \begin{pmatrix} 0; 1000; -40; 0 \end{pmatrix}, r = \begin{pmatrix} 0 \end{pmatrix}$$

Réalisons les calculs à l'aide d'une feuille Excel (Figure 11.5)[6] :

- Pour rappel, nous avons les coefficients

$$\hat{a} = \begin{pmatrix} 1.70205 \\ 0.00049 \\ 0.01825 \\ 0.00423 \end{pmatrix}$$

- La matrice $(X'X)^{-1}$ a déjà été obtenue par ailleurs ; il en est de même pour la somme des carrés des résidus $SCR = 13.58067$ et le degré de liberté $n - p - 1 = 24$.
- Nous formons le vecteur $(R\hat{a} - r)$. Comme nous n'avons qu'une seule $(q = 1)$ contrainte, le résultat est un scalaire

$$R\hat{a} - r = \begin{pmatrix} 0; 1000; -40; 0 \end{pmatrix} \times \begin{pmatrix} 1.70205 \\ 0.00049 \\ 0.01825 \\ 0.00423 \end{pmatrix} - \begin{pmatrix} 0 \end{pmatrix} = \begin{pmatrix} -0.23648 \end{pmatrix}$$

5. Cet exemple est décrit sur notre site de tutoriels, `http://tutoriels-data-mining.blogspot.com/2011/02/regression-lineaire-lecture-des.html`

6. `reg_multiple_consommation_automobiles.xlsx` - "test.comb.lineaire"

(X'X)^-1	constante	cylindrée	puissance	poids
constante	7.060E-01	-1.471E-04	5.586E-03	-7.004E-04
cylindrée	-1.471E-04	1.074E-06	-1.589E-05	-4.688E-07
puissance	5.586E-03	-1.589E-05	3.584E-04	-3.916E-06
poids	-7.004E-04	-4.688E-07	-3.916E-06	1.548E-06

a^	
constante	1.70205
cylindrée	0.00049
puissance	0.01825
poids	0.00423

SCR	13.58067
ddl	24

H0 : 1000 x a(cylindrée) = 40 x a(puissance)

	constante	cylindrée	puissance	poids
R	0	1000	-40	0

r	0

Ra^-r	-0.23648

R(X'X)^(-1)R'	2.91886

0.34260

F-Numérateur	0.01916
F-Dénominateur	0.56586

F	0.03386

p-value	0.85555

F_0.95(1,24)	4.25968

Conclusion	Accepter H0

Fig. 11.5. Test de comparaison de coefficients - Consommation de véhicules

- La quantité $[R(X'X)^{-1}R']$ est également à un scalaire, il est égal à 2.91886. Son inverse est égal à $[R(X'X)^{-1}R']^{-1} = \frac{1}{2.91886} = 0.34260$.
- Nous formons la statistique F (*Remarque : la transposée d'un scalaire est le scalaire lui-même*) :

$$F = \frac{\frac{1}{1}(-0.23648)'(1/2.91886)(-0.23648)}{13.58067/24} = \frac{0.01916}{0.56586} = 0.03386$$

- Le seuil critique au risque $\alpha = 5\%$ est $F_{0.95}(1, 24) = 4.25968$.
- Nous sommes dans la région d'acceptation de H0. Au regard des résultats, l'hypothèse nulle ne peut pas être rejetée.
- La probabilité critique (*p-value*) du test est égale à $\alpha' = 0.85555$.

11.3.4 Régression sous contraintes - Estimation des coefficients

Dans la régression sous-contraintes (régression restreinte), nous introduisons des impératifs - sous forme de combinaisons linéaires de coefficients - sur les paramètres estimés lors du processus de minimisation de la somme des carrés des résidus.

Cela peut survenir par exemple consécutivement aux tests de contraintes linéaires tels que nous les avons étudiés dans les sections précédentes. Après avoir accepté l'hypothèse nulle, nous souhaitons que les coefficients estimés par les MCO reflètent les conditions émises.

Il s'agit donc d'un optimisation sous q contraintes linéaires. A résoudre directement, ça paraît très compliqué. Fort heureusement, il est possible de dériver les nouveaux coefficients des résultats de la régression sans contraintes. Soit $\hat{a}$ le vecteur des coefficients estimés obtenus avec la procédure habituelle. Si nous souhaitons introduire q contraintes linéaires sous la forme $Ra = r$ dans la régression [R est une matrice $(q, p+1)$ et r un vecteur $(q, 1)$], à l'instar de l'hypothèse nulle du test décrit ci-dessus, l'estimateur sous contrainte $\tilde{a}$ s'écrit alors (Johnston et DiNardo, page 102) :

$$\tilde{a} = \hat{a} + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - R\hat{a}) \tag{11.7}$$

Clairement, l'expression n'est pas très simple. Mais on constate néanmoins qu'elle s'appuie uniquement sur les résultats produits par la régression sans contrainte c.-à-d. par la méthode des MCO classique proposée par n'importe quel logiciel de statistique[7].

En ce qui concerne les performances, la somme des carrés des résidus, forcément plus élevée ici puisque nous introduisons des contraintes dans l'optimisation, peut être déduite de la SCR de la régression usuelle (Johnston et DiNardo, page 103) :

$$SCR_{\tilde{a}} = SCR_{\hat{a}} + (\tilde{a} - \hat{a})'(X'X)(\tilde{a} - \hat{a}) \tag{11.8}$$

$SCR_{\tilde{a}}$ est la SCR de la régression sous contrainte, $SCR_{\hat{a}}$ est la SCR de la régression usuelle, et $SCR_{\tilde{a}} \geq SCR_{\hat{a}}$.

Consommation des véhicules

Dans la régression précédente (Figure 11.5), nous avons constaté que l'hypothèse nulle (H_0 : $1000 \times a_{cylindree} = 40 \times a_{puissance}$) n'était pas démentie par les données. Nous souhaitons donc introduire explicitement cette contrainte dans l'estimation des paramètres du modèle. Il n'est pas nécessaire de relancer les opérations, nous pouvons nous appuyer sur les résultats des calculs précédents. Nous complétons la feuille Excel (Figure 11.6)[8] :

- Nous avons $R = (0; 1000; -40; 0)$ et $r = (0)$.
- A partir des coefficients estimés $\hat{a}$, nous calculons $r - R\hat{a} = 0.23648$.
- Vu précédemment, $R(X'X)^{-1}R' = 2.91886$ est un scalaire, son inverse est donc $[R(X'X)^{-1}R']^{-1} = 1/2.91886 = 0.34260$.
- Le produit matriciel

7. Les manipulations telles que nous les décrirons sous Excel paraissent fastidieuses. Je le concède. Mais écrire les mêmes formules sous R, pour peu que l'on connaisse un peu les opérations matricielles, est un jeu d'enfant.

8. `reg_multiple_consommation_automobiles.xlsx` - "reg.sous.contraintes"

(X'X)^-1	constante	cylindrée	puissance	poids
constante	7.060E-01	-1.471E-04	5.586E-03	-7.004E-04
cylindrée	-1.471E-04	1.074E-06	-1.589E-05	-4.688E-07
puissance	5.586E-03	-1.589E-05	3.584E-04	-3.916E-06
poids	-7.004E-04	-4.688E-07	-3.916E-06	1.548E-06

a^

constante	**1.70205**
cylindrée	**0.00049**
puissance	**0.01825**
poids	**0.00423**

SCR (a^)	13.58067

	constante	cylindrée	puissance	poids
R	0	1000	-40	0
r	0			

r - Ra^	0.23648
R(X'X)^(-1)R'	2.91886
[R(X'X)^(-1)R']^(-1)	0.34260

R'[R(X'X)^(-1)R']^(-1)(r-Ra^)	(X'X)^(-1)R'[R(X'X)^(-1)R']^(-1)(r-Ra^)
0.00000	-0.03002
81.01861	0.00014
-3.24074	-0.00245
0.00000	-0.00003

a~

constante	**1.67203**
cylindrée	**0.00063**
puissance	**0.01580**
poids	**0.00420**

SCR(a^) - SCR (a~)
0.01916

SCR (a~)	13.59983

1000 x a~(cylindree)	**0.63207**
40 x a~(puissance)	**0.63207**

Fig. 11.6. Régression sous contrainte $1000 \times a_{cylindree} = 40 \times a_{puissance}$ - Consommation de véhicules

$$R'[R(X'X)^{-1}R']^{-1}(r - R\hat{a}) = \begin{pmatrix} 0.0000 \\ 81.01861 \\ -3.24074 \\ 0.00000 \end{pmatrix}$$

fourni une matrice de taille $(p+1, q)$, comme $q = 1$ dans notre exemple, nous avons un vecteur colonne.

– Pré-multipliée par $(X'X)^{-1}$, nous avons toujours un vecteur

$$(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - R\hat{a}) = \begin{pmatrix} -0.03002 \\ 0.00014 \\ -0.00245 \\ -0.00003 \end{pmatrix}$$

– Il ne reste plus qu'à corriger l'estimation des MCO hors contrainte

$$\tilde{a} = \hat{a} + (X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}(r - R\hat{a}) = \begin{pmatrix} 1.70205 \\ 0.00049 \\ 0.01825 \\ 0.00423 \end{pmatrix} + \begin{pmatrix} -0.03002 \\ 0.00014 \\ -0.00245 \\ -0.00003 \end{pmatrix} = \begin{pmatrix} 1.67203 \\ 0.00063 \\ 0.01580 \\ 0.00420 \end{pmatrix}$$

Nous avons les nouveaux coefficients avec, notamment, $\tilde{a}_{cylindree} = 0.00063$ et $\tilde{a}_{puissance} = 0.01580$. Et nous vérifions aisément que $1000 \times \tilde{a}_{cylindree} = 40 \times \tilde{a}_{puissance}$. C'est assez épatant je trouve ! Dans le même temps, les autres paramètres (coefficient de poids, constante) ont été légèrement modifiés.

Concernant la somme des carrés des résidus, nous récupérons $SCR_{\hat{a}} = 13.58067$ auquel nous additionnons la quantité définie dans l'équation 11.8 :

$$SCR_{\tilde{a}} = SCR_{\hat{a}} + (\tilde{a} - \hat{a})'(X'X)(\tilde{a} - \hat{a}) = 13.58067 + 0.01916 = 13.59983$$

Ici également, il n'est nullement besoin d'accéder au tableau de données. Nous exploitons pleinement les résultats de la régression sans contraintes. On constate que ($SCR_{\tilde{a}} > SCR_{\hat{a}}$).

11.3.5 Test de contraintes linéaires via la confrontation des régressions

A la lumière de ces nouveau développements concernant la régression sous contraintes, nous pouvons éclairer sous un angle nouveau le test de contraintes linéaires sur les coefficients (section 11.3.1). Il s'agit de confronter les résultats de deux modèles, l'une construite sous l'hypothèse H_0, la régression sous contrainte, l'autre normalement, en dehors de toute contrainte c.-à-d. hors H_0.

Dès lors, l'hypothèse nulle n'est justifiée que si la somme des carrés des résidus n'augmente pas de manière significative, si l'introduction de la contrainte définie par H_0 n'entraîne pas une trop forte dégradation en termes de SCR tout simplement.

La seconde formulation de la statistique du test de q contraintes linéaires sur les paramètres de la régression devient (Bourbonnais, page 70 ; Johnston et DiNardo, page 103) :

$$F = \frac{(SCR_{\tilde{a}} - SCR_{\hat{a}})/q}{SCR_{\hat{a}}/(n - p - 1)} \tag{11.9}$$

Elle suit une loi de Fisher $\mathcal{F}(q, n - p - 1)$ sous l'hypothèse nulle. La région critique correspond aux grandes valeurs de F.

En reprenant notre exemple "Consommation des véhicules" (Figure 11.6), nous avons

$$F = \frac{(SCR_{\tilde{a}} - SCR_{\hat{a}})/q}{SCR_{\hat{a}}/(n - p - 1)} = \frac{(13.599983 - 13.58067)/1}{13.58067/24} = 0.03386$$

La valeur de la statistique est exactement la même que celle obtenue avec la première formulation du test sur les contraintes linéaires (Figure 11.5).

12

Prédiction ponctuelle et par intervalle

12.1 Prédiction ponctuelle

Comme pour la régression linéaire simple, il s'agit, pour un nouvel individu $i*$, de fournir une prédiction de la valeur de l'endogène $\hat{y}_{i*}$ à partir de sa description c.-à-d. les valeurs prises par les exogènes $(x_{i*,1}, \ldots, x_{i*,p})$.

La prédiction ponctuelle est obtenue en appliquant les coefficient estimés sur la description de l'individu à traiter

$$\begin{aligned}\hat{y}_{i*} &= \hat{y}(x_{i*}) \\ &= \hat{a}_0 + \hat{a}_1 \times x_{i*,1} + \cdots + \hat{a}_p \times x_{i*,p}\end{aligned}$$

L'expression est plus facile à manipuler en utilisant la notation matricielle :

$$y_{i*} = X_{i*} \times \hat{a} \tag{12.1}$$

Où X_{i*} est un vecteur ligne de dimension $(1, p+1)$: $X_{i*} = (1\,;\, x_{i*,1}\,;\, \cdots\,;\, x_{i*,p})$. La première valeur 1 permet de prendre en compte la constante $\hat{a}_0$. Le résultat est bien un scalaire puisque $\hat{a}$ est de dimension $(p+1, 1)$.

On montre aisément que **la prédiction ponctuelle est sans biais**. Pour ce faire, intéressons nous à l'erreur de prédiction $\hat{\varepsilon}_{i*}$:

$$\begin{aligned}\hat{\varepsilon}_{i*} &= \hat{y}_{i*} - y_{i*} \\ &= X_{i*}\hat{a} - (X_{i*}a + \varepsilon_{i*}) \\ &= X_{i*}(\hat{a} - a) + \varepsilon_{i*}\end{aligned}$$

Et

$$E(\hat{\varepsilon_{i*}}) = X_{i*} \times E(\hat{a} - a) + E(\varepsilon_{i*}) = 0$$

L'espérance de l'erreur de prévision est nulle parce que les estimateurs sont sans biais $[E(\hat{a}) = a]$ et l'espérance de l'erreur est nulle $[E(\varepsilon_{i*}) = 0]$ par hypothèse.

Par conséquent, la prédiction ponctuelle est sans biais :

$$E(\hat{y}_{i*}) = y_{i*}$$

12.2 Intervalle de prédiction

Pour construire l'intervalle de prédiction (la fourchette), nous devons connaître la variance estimée de l'erreur de prédiction et la distribution de cette dernière. L'esprit de l'approche a déjà été développée lors de la présentation de la régression simple. Nous donnons directement les résultats ici (pour plus de détails, voir Bourbonnais, pages 77 et 78; Giraud et Chaix, pages 72 et 73; Johnston et DiNardo, pages 105 à 107).

Concernant la variance estimée de l'erreur de prédiction, nous avons :

$$\hat{\sigma}^2_{\hat{\varepsilon}_{i*}} = \hat{\sigma}^2_{\varepsilon} \left[1 + X_{i*}(X'X)^{-1}X'_{i*}\right] \tag{12.2}$$

La variance sera d'autant plus grande que la régression est de mauvaise qualité ($\hat{\sigma}^2_{\varepsilon}$ est élevé) et que l'on est loin du barycentre du nuage de points ($h_{i*} = X_{i*}(X'X)^{-1}X'_{i*}$ – le levier – est élevé). L'analogie avec la régression simple est totale.

Le ratio erreur/écart-type est distribué selon une loi de Student à $(n - p - 1)$ degrés de liberté :

$$\frac{\hat{\varepsilon}_{i*}}{\hat{\sigma}_{\hat{\varepsilon}_{i*}}} = \frac{\hat{y}_{i*} - y_{i*}}{\hat{\sigma}_{\hat{\varepsilon}_{i*}}} \equiv \mathcal{T}(n - p - 1) \tag{12.3}$$

On en déduit l'intervalle de confiance au niveau de confiance $(1 - \alpha)$:

$$\hat{y}_{i*} \pm t_{1-\frac{\alpha}{2}} \times \hat{\sigma}_{\hat{\varepsilon}_{i*}} \tag{12.4}$$

12.3 Prédiction pour le modèle "Consommation de véhicules"

Nous souhaitons prédire la consommation d'un véhicule présentant les caractéristiques suivantes : cylindrée = 1984 cm^3, puissance = 85 ch et poids = 1155 kg (Figure 12.1) [1].

Nous obtenons la prédiction en appliquant les coefficients estimés du modèle sur cette description :

$$\hat{y}_{i*} = X_{i*} \times \hat{a} = \Big(1;\, 1984;\, 85;\, 1155\Big) \times \begin{pmatrix} 1.70205 \\ 0.00049 \\ 0.01825 \\ 0.00423 \end{pmatrix} = 9.12$$

1. `reg_multiple_consommation_automobiles.xlsx` - "prediction"

const	eng.size	horsepower	weight
1	1984	85	1155

a^

constante	1.70205
cylindrée	0.00049
puissance	0.01825
poids	0.00423

p.ponctuelle
9.12

(X'X)^-1	constante	cylindrée	puissance	poids
constante	7.060E-01	-1.471E-04	5.586E-03	-7.004E-04
cylindrée	-1.471E-04	1.074E-06	-1.589E-05	-4.688E-07
puissance	5.586E-03	-1.589E-05	3.584E-04	-3.916E-06
poids	-7.004E-04	-4.688E-07	-3.916E-06	1.548E-06

Levier	0.05910
sigma^2(err)	0.56586
sigma^2(err^)	0.59931
t_0.95 (24)	2.06390
lower.limit	**7.52**
upper.limit	**10.71**

Fig. 12.1. Prédiction ponctuelle et fourchette de prédiction - Consommation de véhicules

Calculons le levier de l'observation :

$$h_{i*} = X_{i*}(X'X)^{-1}X'_{i*} = \left(1;\, 1984;\, 85;\, 1155\right)(X'X)^{-1}\begin{pmatrix}1\\1984\\85\\1155\end{pmatrix} = 0.05910$$

Avec l'estimation de variance de l'erreur ($\hat{\sigma}^2_\varepsilon = 0.56586$) fournie par DROITEREG, nous produisons l'estimation de la variance de l'erreur de prévision :

$$\hat{\sigma}^2_{\hat{\varepsilon}_{i*}} = \hat{\sigma}^2_\varepsilon\left[1 + X_{i*}(X'X)^{-1}X'_{i*}\right] = 0.56586 \times [1 + 0.05910] = 0.59931$$

Pour un niveau de confiance de 95%, le quantile de la loi de Student à (24) degrés de liberté est $t_{0.975}(24) = 2.06390$, nous calculons finalement les bornes basses et hautes de la fourchette de prédiction :

$$b.b. = \hat{y}_{i*} - t_{1-\frac{\alpha}{2}} \times \hat{\sigma}_{\hat{\varepsilon}_{i*}} = 9.12 - 2.06390 \times \sqrt{0.59931} = 7.52$$
$$b.h. = \hat{y}_{i*} + t_{1-\frac{\alpha}{2}} \times \hat{\sigma}_{\hat{\varepsilon}_{i*}} = 9.12 + 2.06390 \times \sqrt{0.59931} = 10.71$$

13

Interprétation des coefficients

13.1 Coefficient brut et partiel

Le charme de la régression tient en grande partie à ses capacités opérationnelles. A partir des valeurs des exogènes, elle peut fournir une prédiction et une fourchette de prédiction de la valeur de l'exogène. Mais il tient beaucoup également aux possibilités d'interprétations qu'elle propose. On parle alors **d'analyse structurelle**. En effet, la régression cherche à établir l'existence d'une relation entre Y et les X mais, en plus, elle quantifie à travers les coefficients du modèle l'importance des associations : dans quelle mesure les exogènes influent sur les valeurs (ou les variations de valeurs) de l'endogène.

L'interprétation cherche à mettre à jour les causalités entre les variables. Elle ne peut être basée uniquement sur des critères numériques. L'expertise du domaine joue un rôle important. Revenons à notre exemple fétiche de "Consommation des véhicules". On peut comprendre que le poids ait une influence sur la consommation. En s'attachant à diminuer la première, on espère diminuer également la gloutonnerie des automobiles. En revanche, la relation inverse paraît incongrue. Manipuler la consommation, par exemple en prenant un gicleur de carburateur de plus grosse section (ouh là là, ça devient technique là, je me demande qui connaît encore les carburateurs de nos jours – `http://fr.wikipedia.org/wiki/Carburateur`, un beau weber double ou quadruple corps avec les bruits d'aspiration qui vont bien, ahhh...) ne va pas modifier le poids de la voiture. C'est d'ailleurs la raison pour laquelle je m'attache à prendre des exemples simples dans ce support. Il ne s'agit surtout pas de se lancer dans des interprétations plus ou moins heureuses (foireuses) dans des domaines que je maîtrise mal (ex. médecine, écologie, etc.).

Dans cette section, nous nous attacherons à lire les coefficients fournis par la modélisation, tout d'abord dans une régression simple, on parle de coefficients bruts, puis dans la régression multiple, on parle de coefficients partiels.

13.1.1 Coefficient brut

On cherche à expliquer la consommation à partir du poids (Figure 13.1 ; Régression simple) [1]. Nous obtenons le modèle :

1. `reg_multiple_consommation_automobiles.xlsx` - "coef.interprétation"

$$consommation = 0.00669 \times poids + 1.06269$$

La pente de la régression est (largement) significative à 5% avec un t de Student à $t_{poids} = 0.00669/0.00053 = 12.53805$.

cylindree	poids	consommation
846	650	5.7
993	790	5.8
899	730	6.1
1390	955	6.5
1195	895	6.8
658	740	6.8
1331	1010	7.1
1597	1080	7.4
1761	1100	9
2165	1500	11.7
1983	1075	9.5
1984	1155	9.5
1998	1140	8.8
1580	1080	9.3
1390	1110	8.6
1396	1140	7.7
2435	1370	10.8
1242	940	6.6
2972	1400	11.7
2958	1550	11.9
2497	1330	10.8
1998	1300	7.6
2496	1670	11.3
1998	1560	10.8
1997	1240	9.2
1984	1635	11.6
2438	1800	12.8
2473	1570	12.7

Régression simple

	poids	constante
a^	0.00669	1.06269
sigma^(a^)	0.00053	0.65925
t(a^)	12.53805	-

Régression multiple

	poids	cylindree	constante
a^	0.00443	0.00130	1.41755
sigma^(a^)	0.00093	0.00046	0.59935
t(a^)	4.73780	2.81339	-

Fig. 13.1. Lecture du coefficient de "Poids" - Consommation de véhicules

Nous pouvons lire le coefficient de la manière suivante : une augmentation du poids d'un véhicule de 1 kg entraîne une consommation supplémentaire de 0.00669 litres au km. On mesure l'association brute, en dehors de toute considération des autres variables qui pourraient influencer la consommation.

13.1.2 Coefficients partiels

Réalisons maintenant la même régression en introduisant la variable cylindrée (Figure 13.1 ; Régression multiple) [2]. Le coefficient de poids a été modifié :

$$consommation = 0.00443 \times poids + 0.00130 \times cylindree + 1.41755 \tag{13.1}$$

Les deux variables sont significatives à 5%.

2. `reg_multiple_consommation_automobiles.xlsx` - "coef.interprétation"

La modification du coefficient de poids s'explique par le fait que la variable poids est liée à cylindrée. Le coefficient de corrélation $r_{poids,cylindree} = 0.8616$ le montre bien. Lorsque la cylindrée augmente, le poids varie également, dans le même sens : $\hat{a}_{poids}$ en tient compte.

Le nouveau coefficient se lit de la manière suivante : ***à cylindrée égale***, lorsque le poids augmente de 1 kg, la consommation s'accroîtra de 0.00443 litres au km. C'est le fameux *"toutes choses égales par ailleurs"* dont on nous rabâche les oreilles en économétrie. On parle alors de coefficient partiel. Nous avons neutralisé l'impact de la cylindrée sur le poids dans la détermination de l'influence de ce dernier sur la consommation. Ces notions sont à rapprocher du concept de corrélation partielle et semi-partielle que nous étudions en détail dans notre fascicule consacré à l'étude des dépendances entre variables quantitatives [12] (Partie II - Chapitres 4 et 5).

Régression sur résidus

Tentons une petite expérience pour décrypter ce phénomène. Nous allons retrancher la partie de poids expliquée par la cylindrée en calculant le résidu de la régression ($poids = a_1 \times cylindree + a_0$). Puis, nous introduisons ce résidu comme variable explicative dans la régression simple expliquant la consommation $consommation = b_1 \times residu + b_0$). Si notre explication tient la route, la pente b_1 devrait correspondre au coefficient partiel 0.00443.

Nous avons monté une nouvelle feuille Excel (Figure 13.2) [3]. Dans un premier temps, nous régressons poids sur cylindrée. Nous obtenons le modèle :

$$poids = 0.42686 \times cylindree + 424.74778$$

Il est significatif avec un coefficient de détermination $R^2 = 0.74228$. Nous calculons les résidus en déduisant du poids observé le poids prédit par le modèle

$$residus(poids/cylindree) = poids - (0.42686 \times cylindree + 424.74778)$$

Le résidu représente la fraction de poids qui n'est pas expliquée par la cylindrée. Nous l'introduisons comme variable explicative dans la régression expliquant la consommation :

$$consommation = 0.00443 \times residus + 9.07500$$

$b_1 = 0.00443$ représente l'impact du poids sur la consommation en dehors de (*en contrôlant, en neutralisant*) l'influence de la cylindrée et, oh miracle, nous retrouvons le coefficient partiel de la régression multiple (Équation 13.1).

13.2 Comparer l'impact des variables - Les coefficients standardisés

Revenons à la régression multiple expliquant la consommation à partir du poids et de la cylindrée (Figure 13.1 ; Régression multiple). Nous avons

3. `reg_multiple_consommation_automobiles.xlsx` - "coef.interprétation"

cylindree	poids	consommation	poids^	résidus (poids/cyl)
846	650	5.7	785.9	-135.9
993	790	5.8	848.6	-58.6
899	730	6.1	808.5	-78.5
1390	955	6.5	1018.1	-63.1
1195	895	6.8	934.8	-39.8
658	740	6.8	705.6	34.4
1331	1010	7.1	992.9	17.1
1597	1080	7.4	1106.4	-26.4
1761	1100	9	1176.4	-76.4
2165	1500	11.7	1348.9	151.1
1983	1075	9.5	1271.2	-196.2
1984	1155	9.5	1271.6	-116.6
1998	1140	8.8	1277.6	-137.6
1580	1080	9.3	1099.2	-19.2
1390	1110	8.6	1018.1	91.9
1396	1140	7.7	1020.6	119.4
2435	1370	10.8	1464.1	-94.1
1242	940	6.6	954.9	-14.9
2972	1400	11.7	1693.4	-293.4
2958	1550	11.9	1687.4	-137.4
2497	1330	10.8	1490.6	-160.6
1998	1300	7.6	1277.6	22.4
2496	1670	11.3	1490.2	179.8
1998	1560	10.8	1277.6	282.4
1997	1240	9.2	1277.2	-37.2
1984	1635	11.6	1271.6	363.4
2438	1800	12.8	1465.4	334.6
2473	1570	12.7	1480.4	89.6

poids = f(cylindrée)

a_1	a_0
0.42686	424.74778

consommation = f(résidus)

b_1	b_0
0.00443	9.07500

Fig. 13.2. Régression sur le résidu de poids / cylindrée - Consommation de véhicules

$$consommation = 0.000443 \times poids + 0.00130 \times cylindree + 1.41755$$

Les coefficients indiquent l'impact des exogènes en contrôlant les autres variables. Il reste néanmoins une question clé : quelle est la variable qui a le plus d'influence sur la consommation, le poids ou la cylindrée ?

La tentation est grande de comparer les coefficients puisqu'ils mesurent l'impact des variables. Ce n'est pas une bonne idée tout simplement parce que les variables sont exprimées dans des unités différentes. Les variations d'une unité de poids et d'une unité de cylindrée ne représentent pas la même chose, elles ne sont pas opposables.

Pour les rendre comparables, nous devons standardiser les coefficients et raisonner en termes d'écarts-type. Nous obtiendrions une lecture du type : lorsque le poids (la cylindrée) varie de 1 écart-type, la consommation varie de m écarts-type.

Régression sur données centrées et réduites. Une technique simple permettant d'obtenir ces coefficients consiste à centrer et réduire toutes les variables (exogènes et endogène) et à lancer la régression sur les données transformées. Nous avons réalisé cette opération sur notre fichier. Pour la variable consommation (Y) par exemple, nous avons utilisé :

$$\bar{y} = \frac{1}{n}\sum_i y_i = 9.0750$$

$$\sigma_y = \sqrt{\frac{1}{n}\sum_i (y_i - \bar{y})^2} = 2.1926$$

$$\overset{cr}{y_i} = \frac{y_i - \bar{y}}{\sigma_y}$$

Nous obtenons de nouveaux coefficients avec la régression sur le tableau de données centrées et réduites (Figure 13.3 - Régression sur données centrées et réduites) [4] :

$$\overset{cr}{consommation} = 0.61281 \times \overset{cr}{poids} + 0.36390 \times \overset{cr}{cylindree} + 0.00000$$

Données originelles

cylindree	poids	consommation
846	650	5.7
993	790	5.8
899	730	6.1
1390	955	6.5
1195	895	6.8
658	740	6.8
1331	1010	7.1
1597	1080	7.4
1761	1100	9
2165	1500	11.7
1983	1075	9.5
1984	1155	9.5
1998	1140	8.8
1580	1080	9.3
1390	1110	8.6
1396	1140	7.7
2435	1370	10.8
1242	940	6.6
2972	1400	11.7
2958	1550	11.9
2497	1330	10.8
1998	1300	7.6
2496	1670	11.3
1998	1560	10.8
1997	1240	9.2
1984	1635	11.6
2438	1800	12.8
2473	1570	12.7
moyenne	1809.0714	1196.9643	9.0750
écart-type	612.4232	303.4249	2.1926

Données centrées-réduites

cylindree	poids	consommation
-1.5726	-1.8026	-1.5393
-1.3325	-1.3412	-1.4936
-1.4860	-1.5390	-1.3568
-0.6843	-0.7974	-1.1744
-1.0027	-0.9952	-1.0376
-1.8795	-1.5060	-1.0376
-0.7806	-0.6162	-0.9007
-0.3463	-0.3855	-0.7639
-0.0785	-0.3196	-0.0342
0.5812	0.9987	1.1972
0.2840	-0.4020	0.1938
0.2856	-0.1383	0.1938
0.3085	-0.1877	-0.1254
-0.3740	-0.3855	0.1026
-0.6843	-0.2866	-0.2166
-0.6745	-0.1877	-0.6271
1.0221	0.5703	0.7867
-0.9259	-0.8469	-1.1288
1.8989	0.6691	1.1972
1.8760	1.1635	1.2884
1.1233	0.4384	0.7867
0.3085	0.3396	-0.6727
1.1217	1.5590	1.0148
0.3085	1.1965	0.7867
0.3069	0.1418	0.0570
0.2856	1.4436	1.1516
1.0270	1.9874	1.6989
1.0841	1.2294	1.6533

Rég. Sur centrées-réduites

poids	cylindrée	constante
0.61281	0.36390	0.00000

Rég. Sur données originelles

poids	cylindrée	constante
0.00443	0.00130	1.41755

Coef. Corrigés par les écarts-type

poids	cylindrée	constante
0.61281	0.36390	-

Fig. 13.3. Coefficients standardisés - Consommation de véhicules

Les variables étant centrées, la constante est nulle. Nous pouvons lire les résultats en termes d'écarts-type et comparer les coefficients. Lorsque le poids (resp. la cylindrée) augmente de 1 écart-type, la consommation augmente de 0.61281 fois (resp. 0.36390) son écart-type. Maintenant, nous pouvons dire que le poids pèse comparativement plus sur la consommation que la cylindrée.

Ces coefficients standardisés sont souvent directement fournis par les logiciels de statistique pour indiquer l'importance relative des variables (*Standardized coefficients* - *Beta weight* pour SPSS – `http://faculty.chass.ncsu.edu/garson/PA765/regress.htm#bcoeff`).

4. `reg_multiple_consommation_automobiles.xlsx` - "coef.comparaison"

Correction des coefficients de la régression usuelle. Si nous avons la réponse à la question, la procédure est un peu lourde quand même. Elle devient contraignante si nous avons à manipuler un fichier volumineux. Et, en réalité, elle n'est pas nécessaire. Il est possible de corriger les coefficients de la régression sur les données originelles avec les écarts-type des variables pour obtenir les coefficients standardisés. Pour la variable X_j, dont le coefficient estimé est initialement $\hat{a}_j$, nous obtenons le coefficient standardisé $\hat{\beta}_j$ avec

$$\hat{\beta}_j = \hat{a}_j \times \frac{\sigma_{x_j}}{\sigma_y} \tag{13.2}$$

C'est ainsi que pour la variables poids, nous retrouvons (Figure 13.3 - Coefficients corrigés par les écarts-type) :

$$\hat{\beta}_{poids} = \hat{a}_{poids} \times \frac{\sigma_{poids}}{\sigma_{consommation}} = 0.00443 \times \frac{303.4249}{2.1926} = 0.61281$$

le coefficient obtenu sur les données centrées et réduites. Il en de même pour la variable cylindrée.

13.3 Contribution au R^2 des variables dans la régression

Les coefficients standardisés $\hat{\beta}_j$ permettent de comparer l'impact des variables explicatives dans la régression, ils permettent également de calculer leurs contributions.

En effet, il existe une relation entre le coefficient de détermination R^2 et les coefficients standardisés [5] :

$$R^2 = \sum_j \hat{\beta}_j \times r_{y,x_j} \tag{13.3}$$

Où r_{y,x_j} est le coefficient de corrélation linéaire entre l'endogène Y et l'explicative X_j.

La formule étant additive, nous pouvons interpréter la quantité...

$$CRT_j = \hat{\beta}_j \times r_{y,x_j} \tag{13.4}$$

... comme la *contribution* au R^2 de la variable exogène X_j.

Exemple *Traitement du fichier "Consommation des véhicules".* Tout d'abord, nous réalisons la régression avec les variables originelles. Nous obtenons un $R^2 = 0.89221$ (Figure 13.4). Puis, à partir des coefficients bruts et des écarts-type estimés, nous calculons les coefficients standardisés en utilisant la formule (13.2). Enfin, nous estimons les corrélations entre l'endogène et les exogènes, nous avons $r_{conso,poids} = 0.56766$ et $r_{conso,cylindree} = 0.32455$. Nous pouvons dès lors calculer les contributions à l'aide de l'équation (13.4). Nous résumons cela dans le tableau suivant.

Variables	Poids	Cylindrée
Coefficients standardisés	0.61281	0.36390
Corrélation avec *poids*	0.92633	0.89187
Contributions	0.56766	0.32455

5. Daniel Borcard, "Régression et corrélations multiples et partielles", Université de Montréal, Département de Sciences Biologiques, 2002, `http://biol09.biol.umontreal.ca/borcardd/r2partiel.pdf`

On vérifiera facilement l'égalité :

$$R^2 = 0.56766 + 0.32455 = 0.89221$$

Données originelles		
cylindree	poids	consommation
846	650	5.7
993	790	5.8
899	730	6.1
1390	955	6.5
1195	895	6.8
658	740	6.8
1331	1010	7.1
1597	1080	7.4
1761	1100	9
2165	1500	11.7
1983	1075	9.5
1984	1155	9.5
1998	1140	8.8
1580	1080	9.3
1390	1110	8.6
1396	1140	7.7
2435	1370	10.8
1242	940	6.6
2972	1400	11.7
2958	1550	11.9
2497	1330	10.8
1998	1300	7.6
2496	1670	11.3
1998	1560	10.8
1997	1240	9.2
1984	1635	11.6
2438	1800	12.8
2473	1570	12.7
moyenne	1809.0714	1196.9643	9.0750
écart-type	612.4232	303.4249	2.1926

Rég. Sur données originelles		
poids	cylindrée	constante
0.00443	0.00130	1.41755
0.00093	0.00046	0.59935
R^2 0.89221	0.76184	#N/A
103.46435	25	#N/A
120.10238	14.51012	#N/A

Coefficients standardisés		
poids	cylindrée	constante
0.61281	0.36390	-

Corrélation avec "consommation"	
poids	cylindrée
0.92633	0.89187

Contributions au R^2	
poids	cylindrée
0.56766	0.32455

Somme des contributions
0.89221

Fig. 13.4. Contributions au R^2 des variables - Consommation de véhicules

Remarque : *Prudence autour de la notion de "contribution".* Il faut être très prudent quant à la notion de "contribution" telle qu'elle est définie ici. En effet, les variables exogènes sont plus ou moins liées entres elles. Une fraction de l'influence des autres variables pèse en réalité dans le calcul de CTR_j via l'estimation du coefficient $\hat{\beta}_j$. On notera d'ailleurs que les quantités $\hat{\beta}_j$ et r_{y,x_j} peuvent être de signes opposés. Cela peut arriver lorsqu'il y a très forte colinéarité entre les variables par exemple ([13], chapitre 3). On aboutirait alors à une contribution négative au R^2 de X_j. L'interprétation devient très hasardeuse dans ce cas. On peut difficilement dire qu'une variable retire de l'information dans une régression.

De fait, l'idée de la contribution d'une variable X_j à l'explication de Y n'est réellement rigoureuse que lorsque les exogènes sont 2 à 2 orthogonales. Dans ce cas, et uniquement dans ce cas, la contribution d'une variable X_j est égale au R^2 de la régression simple de Y sur X_j. Et la lecture sous forme de *fraction*

de variance expliquée de la contribution - via le rapport $(\frac{CTR_j}{R^2})$ - pour exprimer le gain consécutif à l'introduction de X_j dans la régression multiple devient justifiée.

Il reste que la relation (13.3) mérite d'être connue. Elle est très peu citée dans la littérature.

13.4 Traitement des variables exogènes qualitatives

Nous nous contentons de donner les principaux repères dans cette section, lorsqu'une des variables explicative est binaire. Pour une étude détaillée des exogènes qualitatives, nous renvoyons le lecteur à notre fascicule "Pratique de la régression linéaire multiple - Diagnostic et Sélection de variables" ([13], chapitre 4).

13.4.1 Explicative binaire dans la régression simple

Comparaison de moyennes

Nous souhaitons mettre en lumière les différences entre les salaires (Y, en euros) selon le genre (X, variable "sexe") [6] : les hommes sont codés 0 et les femmes 1.

Une approche très simple consiste à réaliser un test de comparaison de moyennes [7]. Nous confrontons :

$$\begin{cases} H_0 : \mu_{y/1} = \mu_{y/0} \\ H_1 : \mu_{y/1} \neq \mu_{y/0} \end{cases}$$

Où $\mu_{y/1}$ (resp. $\mu_{y/0}$) est la moyenne des salaires chez les femmes (resp. chez les hommes).

Nous disposons de $n = 40$ observations. A l'aide du tableau croisé dynamique d'Excel (Figure 13.5) [8], nous calculons les moyennes, les écarts-type et les effectifs conditionnels.

Sexe	Moyennes	Ecarts-type	Nombre
Homme (0)	$\bar{y}_0 = \frac{1}{n_0}\sum_{i:x_i=0} y_i = 3110.800$	$s_0 = \sqrt{\frac{1}{n_0-1}\sum_{i:x_i=0}(y_i - \bar{y}_0)^2} = 1517.327$	$n_0 = 20$
Femme (1)	$\bar{y}_1 = 1947.250$	$s_1 = 1021.592$	$n_1 = 20$

Nous calculons l'écart entre les salaires, la statistique de test sera basée sur cet indicateur

$$D = \bar{y}_1 - \bar{y}_0 = 1947.250 - 3110.800 = -1163.550$$

Pour obtenir la variance de D, nous devons passer dans un premier temps par l'estimation de la variance commune aux deux groupes, la variance intra-classes. Nous faisons donc l'hypothèse que les

6. Les données proviennent du site `http://www.cabannes.net/`

7. Rakotomalala R., *Comparaison de populations - Tests paramétriques*, Chapitre 1 : Comparaison de 2 moyennes - Cas des variances égales, `http://eric.univ-lyon2.fr/~ricco/cours/cours/Comp_Pop_Tests_Parametriques.pdf`

8. `regression-salaire-sexe.xlsx` - "comp.moyenne"

SAL	SEXE
2249	0
1374	0
3749	0
3749	0
2750	0
6243	0
5099	0
6250	0
3494	0
4000	0
875	0
1249	0
2500	0
3749	0
1457	0
2275	0
3249	0
1968	0
2500	0
3437	0
2916	1
2221	1
1374	1
2702	1
1350	1
1874	1
2500	1
3571	1
2437	1
837	1
850	1
1406	1
2500	1
1525	1
1042	1
1199	1
1312	1
1237	1
4843	1
1249	1

	Données		
SEXE	Moyenne de SAL	Écartype de SAL	Nombre de SAL
0	3110.800	1517.327	20
1	1947.250	1021.592	20

D	-1163.550

n_0	20
n_1	20

s²	1672965.70921
sigma^(D)	409.01903

s	1293.43176

t calculé	-2.84473
ddl	38

t théorique	2.02439
p-value	0.00712

Fig. 13.5. Comparaisons des moyennes - Salaires

variances sont identiques dans les groupes [9].

$$s^2 = \frac{(n_0 - 1) \times s_0^2 + (n_1 - 1) \times s_1^2}{n_0 + n_1 - 2} = \frac{19 \times 1517.327 + 19 \times 1021.592}{20 + 20 - 2} = 1672965.70921$$

Enfin

$$\hat{\sigma}_D = s \times \sqrt{\frac{1}{n_0} + \frac{1}{n_1}} = 1293.43176 \times \sqrt{\frac{1}{20} + \frac{1}{20}} = 409.01903 \tag{13.5}$$

La statistique de test s'écrit :

$$t_{calc} = \frac{D}{\hat{\sigma}_D} = \frac{-1163.550}{409.01903} = -2.84473$$

Sous H_0, elle suit une loi de Student à $(n_0 + n_1 - 2 = n - 2 = 38)$ degrés de liberté. La région critique au risque α correspond à

9. Lorsque les effectifs sont équilibrés comme c'est le cas ici, cette approche est très robuste. Même si les variances sont sensiblement différentes, la procédure tient parfaitement la route.

$$R.C. : |t_{calc}| > t_{1-\frac{\alpha}{2}}(n_0 + n_1 - 2)$$

Dans notre exemple, au risque 5%, $t_{0.975}(38) = 2.02439$. Nous sommes dans la région critique. Nous rejetons l'hypothèse nulle. Les salaires sont différents selon le sexe de la personne.

Traitement avec la régression - Explicative binaire dans la régression multiple

Peut-on obtenir les mêmes résultats via la régression ? La réponse est oui. Tout dépend du codage adopté. Dans le cas d'une explicative binaire, il n'y a pas trop à se poser de questions : une de modalité est codée 1 (les femmes), l'autre 0 (les hommes). Lorsqu'il s'agit d'une variable nominale à plus de 2 modalités ou d'une variable ordinale, le codage pèse sur la lecture des résultats [13] (chapitre 4).

Dans notre exemple, nous réalisons la régression

$$salaire = a \times sexe + b$$

DROITEREG			
	a	b	
coef^	-1163.550	3110.800	
sigma^(coef^)	409.019035	289.220133	
R²	0.17557098	1293.43176	
F	8.09250659	38	sigma^(err)
SCE	13538486	63572697	SCR
t-calculé	-2.84473		
p-value	0.00712		

Fig. 13.6. Régression simple - $salaire = a \times sexe + b$ - Salaires

Nous obtenons les coefficients (Figure 13.6) [10] :

$$\hat{b} = 3110.800 = \hat{y}_0$$
$$\hat{a} = -1165.550 = \hat{y}_1 - \hat{y}_0 = D$$

On constate que la constante correspond à la moyenne conditionnelle du salaire pour la modalité de référence de sexe (celle qui est codée 0 c.-à-d. les hommes). Et la pente correspond au différentiel entre les salaires.

De fait, tester la significativité de la pente dans la régression revient à tester la significativité de l'écart entre les salaires. La statistique de test ($t_{\hat{a}} = -2.84473$) prend exactement la même valeur, la conclusion est la même bien évidemment. Notons cependant une information importante, dans la régression on fait implicitement l'hypothèse que la variance de Y est la même dans les sous-populations. Hypothèse d'homoscédasticité que nous émettions explicitement dans la comparaison des moyennes.

10. `regression-salaire-sexe.xlsx` - "comp.moyenne"

13.4.2 Coefficient partiel avec une explicative binaire

Un expert vient vous dire que tout ça est bien beau mais on sait par ailleurs que les hommes font plus d'études que les femmes. Comparer les salaires en se basant sur le sexe uniquement fausse les résultats et laisse croire des choses qui n'existent pas. Il en veut pour preuve que dans notre fichier, en intégrant la variable ETUDES, les hommes ont effectué en moyenne 13.5 années d'études, et les femmes 12.2 seulement.

En toute rigueur, il faudrait effectuer l'extraction d'un sous-échantillon chez les hommes, faire de même chez les femmes, et s'arranger que les deux sous-échantillons présentent une moyenne d'années d'étude identique. Ou encore pondérer les individus de manière à ce que les moyennes conditionnelles d'ETUDES soient identiques. Enfin, une autre piste serait d'effectuer un appariement c.-à-d. créer un fichier où chaque ligne confronte des personnes de sexe opposé mais ayant effectué un nombre d'années d'études identique.

Tout cela induit des manipulations de fichier plus ou moins hasardeuses. Il y a une solution plus simple. S'appuyer sur le fait que la régression produit des coefficients partiels. Nous réalisons donc la régression (Figure 13.7) [11]

$$salaire = a_2 \times etudes + a_1 \times sexe + a_0$$

L'écart de salaires selon le sexe est moindre $\hat{a}_2 = -881.44020$ (contre -1165.550 pour le coefficient brut). Cela veut dire qu'à *années d'études égales*, les femmes reçoivent en moyenne un salaire inférieur de 881 euros par rapport à celui des hommes. Et l'écart reste significatif à 5% avec un t-calculé de $t_{\hat{a}_2} = -2.22922$ et une p-value de $\alpha' = 0.03195$.

Ainsi, même si les hommes et les femmes ont un niveau d'études identique, ces dernières ont tendance à obtenir un salaire moins élevé. A partir de la régression, nous arrivons à répondre précisément à la question posée. Monsieur l'expert peut rentrer chez lui.

Les férus de statistique n'auront pas manqué de voir dans cet exemple une illustration simplifiée d'une **analyse de covariance** (ANCOVA) [12]), technique où l'on cherche à étudier l'impact d'une variable catégorielle sur une variable dépendante quantitative, en contrôlant l'effet d'une tierce variable sur cette dernière.

Je détaille la régression sur exogènes qualitatives dans l'ouvrage consacré à la pratique de la régression ([13], chapitre 4). Les différents types de codages et les interprétations y afférentes sont analysés.

11. `regression-salaire-sexe.xlsx` - "reg.multiple"
12. `http://pages.usherbrooke.ca/spss/pages/statistiques-inferentielles/analyse-de-covariance.php` ; `http://faculty.chass.ncsu.edu/garson/PA765/anova.htm`

SAL	SEXE	ETUDES
2249	0	10
1374	0	12
3749	0	12
3749	0	14
2750	0	8
6243	0	17
5099	0	17
6250	0	14
3494	0	14
4000	0	12
875	0	12
1249	0	14
2500	0	16
3749	0	16
1457	0	13
2275	0	13
3249	0	12
1968	0	16
2500	0	14
3437	0	14
2916	1	12
2221	1	8
1374	1	12
2702	1	12
1350	1	16
1874	1	12
2500	1	10
3571	1	14
2437	1	15
837	1	12
850	1	8
1406	1	14
2500	1	16
1525	1	10
1042	1	12
1199	1	12
1312	1	12
1237	1	9
4843	1	16
1249	1	12

Moyenne de E	
SEXE	Total
0	13.5
1	12.2
Total général	12.85

	DROITEREG		
	ETUDES	SEXE	CONSTANTE
coef^	217.00754	-881.44020	181.19821
sigma^(coef^)	82.60886	395.40231	1147.22271
R^2	0.30516	1203.37046	#N/A
F	8.12494	37	#N/A
SCE	23531466	53579717	#N/A

t-calculé	-2.22922
p-value	0.03195

Fig. 13.7. Régression simple - $salaire = a_2 \times etudes + a_1 \times sexe + a_0$ - Salaires

14

Étude de cas : Analyse du taux de chômage en France

Récapitulons les différents thèmes abordés dans ce document en réalisant une étude de cas. Nous souhaitons comprendre les tenants et aboutissants du taux de chômage en France métropolitaine à la fin de l'année 2008. Le sujet et les données proviennent du site de Mme Aurélie Bonein (`http://aurelie.bonein.free.fr/`), nous reprenons le second thème de son cours d'économétrie (`http://aurelie.bonein.free.fr/telechargement/Econometrie/2010-2011/TD2_sujet.pdf`).

Pour expliquer le taux de chômage (Y), nous disposons de $p = 5$ variables explicatives :

- X_1 le nombre de faillites d'entreprises par région au cours de l'année 2008 ;
- X_2 le nombre d'établissements de construction par région en 2008 ;
- X_3 le nombre de commerces par région en 2008 ;
- X_4 le nombre d'établissement de services par région en 2008 ;
- X_5 le nombre d'industries agro-alimentaire par région en 2008.

Le fichier comporte $n = 22$ observations (régions). Nous reproduisons ici le contenu du fichier (Figure 14.1). Attention, la précision de l'affichage a été limitée à 4 décimales. En réalité, les données en comportent beaucoup plus.

14.1 Lecture des résultats de la régression

Nous avons lancé la fonction DROITEREG sur ces données (Figure 14.2) [1]. Nous en avons déduit les informations importantes pour la compréhension des résultats :

- Le tableau d'analyse de variance permet de porter un jugement sur la qualité globale de la régression. Les SCE et SCR sont directement fournis par Excel, nous avons calculé $SCT = SCE + SCR = 28.5832 + 13.8800 = 40.7332$ et les carrés moyens

$$CME = \frac{SCE}{p} = \frac{26.8532}{5} = 5.3706,\ CMR = \frac{SCR}{n-p-1} = \frac{13.8800}{16} = 0.8675$$

- Nous pouvons en déduire le R^2 et le R^2-ajusté

1. `analysetauxdechomage.xlsx` - "analyse"

	Y	X1	X2	X3	X4	X5
Région	Chômage	Nb faillites en	Construction	Commerce	Services	IIAA
Alsace	8.4	4.4080	4.4526	11.4346	30.3529	0.9766
Aquitaine	8.7	5.4625	7.6495	13.7757	35.4038	1.3584
Auvergne	8.4	3.6644	6.7569	12.3236	32.6443	1.5891
Bourgogne	8.5	3.7866	6.1018	11.9393	30.1655	1.3072
Bretagne	7.7	3.8187	6.2947	10.8079	30.2383	1.4307
Centre	8.4	3.6643	5.8452	10.4330	27.3298	1.1395
Champagne-Ardenne	10	3.4160	5.0540	10.7609	28.0360	1.5187
Corse	8.3	8.3344	12.0563	19.3477	51.2517	2.1954
Franche-Comté	9.7	3.5263	5.6411	10.7990	29.2295	1.3443
Île-de-France	7.8	6.6301	5.7058	14.9091	46.2009	0.6347
Languedoc-Roussillon	12.4	6.9784	8.6708	15.0351	38.2325	1.5428
Limousin	7.7	3.6367	7.0653	11.7401	31.8218	1.4463
Lorraine	9.9	3.4998	4.8387	9.9311	26.1562	1.0561
Midi-Pyrénées	9	5.2993	8.0155	13.1663	36.3999	1.4303
Nord-Pas-de-Calais	12.8	3.2611	2.2576	9.5878	24.4222	0.9028
Basse-Normandie	9	3.5284	6.0062	11.8154	30.3903	1.4361
Haute-Normandie	10.2	3.4077	4.6496	10.0198	25.9107	0.9901
Pays de la Loire	8.2	3.8615	5.7309	10.4515	28.9145	1.1551
Picardie	10.8	3.1442	4.6068	9.0968	24.6421	0.9168
Poitou-Charentes	8.9	4.0384	6.8451	12.4550	30.1779	1.4802
Provence-Alpes-Côte d'Azur	10.3	7.5831	8.4880	17.3224	44.0305	1.3396
Rhône-Alpes	8.6	5.4735	6.9763	13.1199	37.5726	1.2438

Fig. 14.1. Analyse du taux de chômage - par région - en France (2008) - Données

	X5	X4	X3	X2	X1	constante
a^	2.7246	-0.3975	0.1766	-0.8975	2.1239	12.5732
sigma^(a^)	1.6054	0.1240	0.4847	0.3433	0.5792	1.7078
R^2	0.6592	0.9314	#N/A	#N/A	#N/A	#N/A
F	6.1909	16	#N/A	#N/A	#N/A	#N/A
SCE	26.8532	13.8800	#N/A	#N/A	#N/A	#N/A
		SCR				

Tableau d'analyse de variance

Source	SC	DDL	CM
Expliquée	26.8532	5	5.3706
Résiduelle	13.8800	16	0.8675
Totale	40.7332	21	

R^2	0.6592
R^2-ajusté	0.5528

Test de significativité globale

F	6.1909
ddl1	5
ddl2	16
p-value	0.002238911

Test de significativité des variables

	X5	X4	X3	X2	X1
t-calculé	1.6971	-3.2062	0.3644	-2.6145	3.6669
p-value	0.1090	0.0055	0.7204	0.0188	0.0021

Fig. 14.2. Analyse du taux de chômage - par région - en France (2008) - Régression

$$R^2 = \frac{SCE}{SCT} = \frac{29.8532}{40.7332} = 0.6592$$
$$\bar{R}^2 = 1 - \frac{n-1}{n-p-1}(1-R^2) = 1 - \frac{22-1}{22-5-1}(1-0.6592) = 0.5528$$

– Pour tester globalement le modèle, nous utilisons la statistique F

$$F = \frac{CME}{CMR} = \frac{5.3706}{0.8675} = 6.1909$$

– Distribuée selon une loi de Fisher $\mathcal{F}(5,16)$ sou H_0, nous obtenons la probabilité critique $\alpha' = 0.00224$. Au risque 5%, le modèle est globalement significatif.

– Voyons le rôle de chacune des variables maintenant, nous construisons un tableau intermédiaire

Variable	X_5	X_4	X_3	X_2	X_1
$\hat{a}_j$	2.7246	-0.3975	0.1766	-0.8975	2.1239
$\hat{\sigma}_{\hat{a}_j}$	1.6054	0.1240	0.4847	0.3433	0.5792
$t_{\hat{a}_j} = \frac{\hat{a}_j}{\hat{\sigma}_{\hat{a}_j}}$	1.6971	-3.2062	0.3644	-2.6145	3.6669
p-value	0.1090	0.0055	0.7204	0.0188	0.0021
Signif. à 5%	non	oui	non	oui	oui

– Les variables qui influent sur le taux de chômage à 5%, toutes choses égales par ailleurs (en contrôlant l'effet des autres variables) sont : X_1, le nombre d'entreprises en faillites dans la région ; X_2, le nombre d'établissements de construction ; X_4, le nombre d'établissement de service.

14.2 Tester simultanément les coefficients de (X_2, X_3, X_5)

En se basant sur ses connaissances en économie, un expert vient expliquer que seules les variables X_1 et X_4 influent réellement sur le taux de chômage. Il nous demande de vérifier la nullité simultanée des coefficients des variables (X_2, X_3, X_5) à 5%.

Nous sommes un peu étonné quant à ces affirmations. Certes, X_3 et X_5 pris individuellement ne sont pas pertinentes. En revanche, X_2 l'est, l'enlever de la régression semble intuitivement un peu hasardeux.

Laissons de côté l'intuition et réalisons les calculs. Nous construisons le modèle avec uniquement les variables X_1 et X_4 (Figure 14.3) [2] : le coefficient de détermination R^2 est égal à 0.5053. Il était de 0.6592 avec la totalité des $(p = 5)$ variables. Est-ce que cette dégradation est significative ?

	Y	X1	X4
Région	Chômage	Nb faillites	Services
Alsace	8.4	4.4080	30.3529
Aquitaine	8.7	5.4625	35.4038
Auvergne	8.4	3.6644	32.6443
Bourgogne	8.5	3.7866	30.1655
Bretagne	7.7	3.8187	30.2383
Centre	8.4	3.6643	27.3298
Champagne-Ardenne	10	3.4160	28.0360
Corse	8.3	8.3344	51.2517
Franche-Comté	9.7	3.5263	29.2295
Île-de-France	7.8	6.6301	46.2009
Languedoc-Roussillon	12.4	6.9784	38.2325
Limousin	7.7	3.6367	31.8218
Lorraine	9.9	3.4998	26.1562
Midi-Pyrénées	9	5.2993	36.3999
Nord-Pas-de-Calais	12.8	3.2611	24.4222
Basse-Normandie	9	3.5284	30.3903
Haute-Normandie	10.2	3.4077	25.9107
Pays de la Loire	8.2	3.8615	28.9145
Picardie	10.8	3.1442	24.6421
Poitou-Charentes	8.9	4.0384	30.1779
Provence-Alpes-Côte d'Azur	10.3	7.5831	44.0305
Rhône-Alpes	8.6	5.4735	37.5726

Droitereg

	X4	X1	constante
	-0.4319	1.8680	14.8566
	0.0981	0.4530	1.4340
R²	0.5053	1.0298	#N/A
	9.7036	19	#N/A
	20.5826	20.1506	#N/A

R² (5)	0.6592
R² (2)	0.5053

Siginificativité (X2,X3,X5)

F	2.4095
ddl1	3
ddl2	16
p-value	0.1050

Test des coefficients restants (X1,X4)

	X4	X1
t-calculé	-4.4032	4.1236
p-value	0.0003	0.0006

Fig. 14.3. Taux de chômage en France (2008) - Test de significativité des coefficients de (X_2, X_3, X_5)

Nous calculons la statistique de test

$$F = \frac{(R_1^2 - R_0^2)/q}{(1 - R_1^2)/(n - p - 1)} = \frac{(0.6592 - 0.5053)/3}{(1 - 0.9592)/(22 - 5 - 1)} = 2.4095$$

2. `analysetauxdechomage.xlsx` - "test - X5.X3.X2"

Avec la distribution $\mathcal{F}(3, 16)$, nous avons une p-value de $\alpha' = 0.1050$. Effectivement, l'expert avait raison, l'hypothèse selon laquelle les coefficients de (X_2, X_3, X_5) sont simultanément nuls n'est pas démentie par les données.

Dans le modèle réduit à 2 variables (X_1, X_4), nous constatons que tous deux sont individuellement significatifs. Un nombre de faillites accru entraîne une augmentation du chômage ; lorsque le nombre d'entreprises de services est élevé, le chômage est moindre. Oui, tout ça tombe sous le sens. On se demande parfois pourquoi on s'enquiquine avec des techniques compliquées pour sortir des évidences pareilles....

C'est tout le charme de la modélisation. Beaucoup d'appelés (on tente, on tente, on triture les données comme on peut), mais peu d'élus (trouver des modèles réellement intéressants, opérationnels, reste rare).

14.3 Prédiction ponctuelle et par intervalle

Les données qui ont servi à la construction du modèle proviennent de la France métropolitaine. Nous souhaitons l'appliquer aux DOM-TOM c.-à-d. à partir de leur description $x = (1, X_1 = 3.45, X_2 = 4.01, X_3 = 11.2, X_4 = 28, X_5 = 2.54)$ (la première valeur **1** représente la constante), proposer une prédiction ponctuelle et par intervalle de son taux de chômage.

En toute rigueur, il serait plus approprié de recourir au modèle simplifié, avec les deux explicatives (X_1, X_4), puisque nous avons montré dans la section précédente que (X_2, X_3, X_5) n'étaient pas pertinentes dans l'explication de Y. Mais, pour être raccord avec le corrigé proposé en ligne sur notre site de référence [3], nous utiliserons le modèle complet avec $p = 5$ exogènes.

La prédiction est très simple à obtenir. Il suffit d'appliquer les coefficients estimés du modèle sur la description de la nouvelle observation à traiter

$$\hat{y} = x.\hat{a} = \Big(1, 3.45, 4.01, 11.2, 28, 2.54\Big) . \begin{pmatrix} 12.5732 \\ 2.1239 \\ -0.8975 \\ 0.1766 \\ -0.3975 \\ 2.7246 \end{pmatrix} = 14.07$$

Plus compliquées à chiffrer sont les bornes de l'intervalle de prédiction (Figure 14.4) [4].

- Il nous faut au préalable calculer la matrice $(X'X)^{-1}$. Ce que nous faisons dans la feuille Excel.
- Puis calculer le levier pour les DOM-TOM

$$h = x(X'X)^{-1}x' = 6.4385$$

3. `http://aurelie.bonein.free.fr/telechargement/Econometrie/2010-2011/Exercice1_corrige.xlsx`
4. `analysetauxdechomage.xlsx` - "prédiction"

– Nous calculons alors l'écart type de l'erreur de prédiction, en nous servant de l'estimation de l'erreur du modèle fournie par Droitereg $\hat{\sigma}_\varepsilon = 0.9314$,

$$\hat{\sigma}_{\hat{\varepsilon}} = \hat{\sigma}_\varepsilon \sqrt{1+h} = 0.93124 \times \sqrt{1+6.4385} = 2.5403$$

– Avec le quantile d'ordre $1-\alpha/2$ de la loi de Student $\mathcal{T}(n-p-1=16)$, nous établissons les bornes basses et hautes de l'intervalle au niveau de confiance $(1-\alpha) = 95\%$

$$b.b. = \hat{y} - t_{0.975} \times \hat{\sigma}_{\hat{\varepsilon}} = 14.07 - 2.1199 \times 2.5403 = 8.6849$$

$$b.h. = \hat{y} + t_{0.975} \times \hat{\sigma}_{\hat{\varepsilon}} = 14.07 + 2.1199 \times 2.5403 = 19.4551$$

	constante	X1	X2	X3	X4	X5
a^	12.5732	2.1239	-0.8975	0.1766	-0.3975	2.7246
DOM TOM	1	3.45	4.01	11.2	28	2.54

Y^	14.0700

X					
1	4.4080	4.4526	11.4346	30.3529	0.9766
1	5.4625	7.6495	13.7757	35.4038	1.3584
1	3.6644	6.7569	12.3236	32.6443	1.5891
1	3.7866	6.1018	11.9393	30.1655	1.3072
1	3.8187	6.2947	10.8079	30.2383	1.4307
1	3.6643	5.8452	10.4330	27.3298	1.1395
1	3.4160	5.0540	10.7609	28.0360	1.5187
1	8.3344	12.0563	19.3477	51.2517	2.1954
1	3.5263	5.6411	10.7990	29.2295	1.3443
1	6.6301	5.7058	14.9091	46.2009	0.6347
1	6.9784	8.6708	15.0351	38.2325	1.5428
1	3.6367	7.0653	11.7401	31.8218	1.4463
1	3.4998	4.8387	9.9311	26.1562	1.0561
1	5.2993	8.0155	13.1663	36.3999	1.4303
1	3.2611	2.2576	9.5878	24.4222	0.9028
1	3.5284	6.0062	11.8154	30.3903	1.4361
1	3.4077	4.6496	10.0198	25.9107	0.9901
1	3.8615	5.7309	10.4515	28.9145	1.1551
1	3.1442	4.6068	9.0968	24.6421	0.9168
1	4.0384	6.8451	12.4550	30.1779	1.4802
1	7.5831	8.4880	17.3224	44.0305	1.3396
1	5.4735	6.9763	13.1199	37.5726	1.2438

Levier	6.4385
sigma^(epsilon)	0.9314
ddl	16
sigma^(eps^)	2.5403
t-théorique	2.1199
b.basse	8.6849
b.haute	19.4551

(X'X)					
22.0000	100.4235	139.7086	270.2721	719.5237	28.4343
100.4235	508.0899	687.0238	1312.1386	3501.6399	133.5432
139.7086	687.0238	967.4812	1806.7822	4803.2442	190.9269
270.2721	1312.1386	1806.7822	3456.7276	9205.1134	358.6286
719.5237	3501.6399	4803.2442	9205.1134	24592.4369	949.5666
28.4343	133.5432	190.9269	358.6286	949.5666	38.9120

(X'X)^(-1)					
3.3622	0.5604	0.2815	-0.3807	-0.0635	-0.7032
0.5604	0.3867	-0.0733	-0.2120	-0.0011	0.6035
0.2815	-0.0733	0.1358	0.0184	-0.0122	-0.4923
-0.3807	-0.2120	0.0184	0.2708	-0.0452	-0.4766
-0.0635	-0.0011	-0.0122	-0.0452	0.0177	0.0945
-0.7032	0.6035	-0.4923	-0.4766	0.0945	2.9710

Fig. 14.4. Taux de chômage en France (2008) - Prédiction et intervalle de prédiction pour les DOM-TOM

15

La régression linéaire avec les logiciels de statistique

Dans ce chapitre, nous décrirons la mise en oeuvre de la régression linéaire multiple à l'aide de quelques logiciels connus (ou non) des praticiens de l'économétrie. Nous mettrons l'accent sur la lecture des résultats. Pour faciliter les comparaisons, nous utiliserons le seul et unique fichier "conso-vehicules.xls" correspondant au problème de "Consommation des véhicules" maintes fois analysé dans ce fascicule.

Pour les outils que je connais bien (Tanagra, Regress et R principalement), nous creuserons un peu plus en abordant des sujets qui sont par ailleurs détaillés dans notre second fascicule relatif à la régression [13] (ex. sélection de variables, détection des points atypiques, etc.).

Un petit aparté avant de commencer. "Bon sang ne saurait mentir" a-t-on l'habitude de dire. A travers le choix des logiciels que j'ai choisi de mettre en avant dans ce fascicule, tout le monde aura bien compris quelle est ma véritable culture. D'autres auraient plutôt choisi de parler de *EViews, Gauss, Rats (que j'ai beaucoup utilisé naguère), Stata, TSP,* etc. Ils auraient très bien fait également. Comme j'ai l'habitude de le dire : qu'importe le logiciel, le plus important est que nous sachions quoi faire avec l'outil, puis comment exploiter efficacement les résultats. C'est justement pour dégager les étudiants du logiciel que je m'évertue à détailler tous les calculs à l'aide d'un tableur.

15.1 Tanagra

15.1.1 Régression linéaire multiple avec Tanagra

Tanagra est un logiciel gratuit de Data Mining (`http://eric.univ-lyon2.fr/~ricco/tanagra/`, version 1.4.38). Il comporte un onglet dédié à l'analyse de régression. On y retrouve des outils pour la régression linéaire telle qu'elle est décrite dans ce document. Les outils associés sont également proposés.

De nombreux tutoriels décrivent l'importation d'un fichier Excel dans Tanagra [1], nous ne reviendrons pas là-dessus. Une fois les données importées et le problème spécifié à l'aide de l'outil DEFINE STATUS (consommation en TARGET, les autres variables en INPUT), nous introduisons la régression linéaire

1. `http://tutoriels-data-mining.blogspot.com/`

multiple à l'aide du composant MULTIPLE LINEAR REGRESSION. Détaillons les résultats affichés par Tanagra (Figure 15.1) :

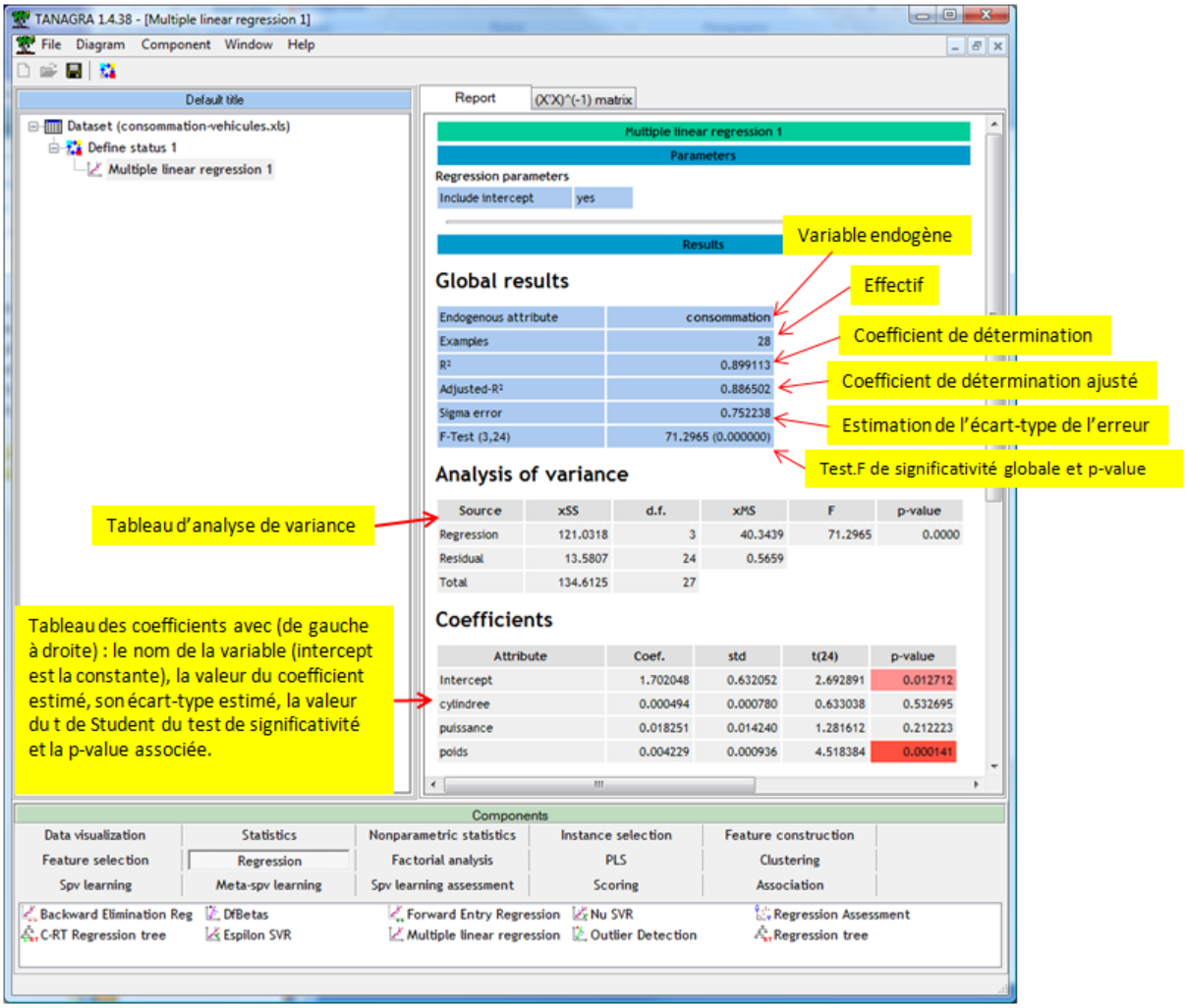


Fig. 15.1. Résultats de la régression avec Tanagra - Consommation des véhicules

- Un premier tableau "Global Results" décrit les résultats globaux (*tiens donc* !) permettant d'expertiser rapidement la qualité de l'ajustement. Nous y apprenons, entres autres, que le coefficient de détermination $R^2 = 0.899113$. Le modèle explique près de 90% de la variance de consommation.
- Le second tableau "Analysis of variance" correspond au tableau d'analyse de variance. Tanagra y adjoint le statistique F du test de significativité globale de la régression et la p-value associée. Dans notre exemple, le modèle est très fortement significatif.
- Le troisième tableau correspond aux coefficients associés aux variables explicatives du modèle. "Intercept" est la constante. Parmi les exogènes, seul poids semble significatif. Mais nous avons vu par ailleurs que "cylindrée" et "puissance" se gênent dans la régression parce que fortement corrélées.

Residuals analysis

Att. name	Full statistics		Histogram			
	Statistics		Values	Count	Percent	Histogram
Err_Pred_imreg_1	Average	0.0000	x_<_-1.5047	1	3.57%	
	Median	0.1446	-1.5047_=<_x_<_-1.2192	0	0.00%	
	Std dev. [Coef of variation]	0.7092 [-99999.0000]	-1.2192_=<_x_<_-0.9337	2	7.14%	
	MAD [MAD/STDDEV]	0.5930 [0.8361]	-0.9337_=<_x_<_-0.6482	3	10.71%	
	Min * Max [Full range]	-1.79 * 1.06 [2.85]	-0.6482_=<_x_<_-0.3627	4	14.29%	
	1st * 3rd quartile [Range]	-0.54 * 0.52 [1.07]	-0.3627_=<_x_<_-0.0772	2	7.14%	
	Skewness (std-dev)	-0.5180 (0.4405)	-0.0772_=<_x_<_0.2083	3	10.71%	
	Kurtosis (std-dev)	-0.1893 (0.8583)	0.2083_=<_x_<_0.4937	5	17.86%	
			0.4937_=<_x_<_0.7792	5	17.86%	
			x>=_0.7792	3	10.71%	

Fig. 15.2. Description succincte des résidus dans Tanagra - Consommation des véhicules

Un dernier tableau dans la partie basse de la fenêtre donne un aperçu des caractéristiques des résidus, si importants dans la régression (Figure 15.2). Nous avons un histogramme de fréquences et quelques caractéristiques numériques. On sait par exemple que si le rapport MAD/STDDEV (écart absolu moyen / écart type) s'écarte résolument de 0.8 ($\sqrt{\frac{2}{\pi}}$ pour être précis [2]), l'hypothèse de normalité des résidus est mise à mal. Dans notre cas, nous avons $MAD/STDEV = 0.8361$. Nous détaillons l'analyse des résidus dans le chapitre 1 du second fascicule de cours [13].

Multiple linear regression 1

Report | (X'X)^(-1) matrix

(X'X)^(-1)	cylindree	puissance	poids	intercept
cylindree	1.0741665E-6	-1.589138E-5	-4.6883004E-7	-0.00014708385
puissance	-1.589138E-5	0.00035836594	-3.9164528E-6	0.0055863437
poids	-4.6883004E-7	-3.9164528E-6	1.547989E-6	-0.00070037619
intercept	-0.00014708385	0.0055863437	-0.00070037619	0.70598604

Fig. 15.3. La matrice $(X'X)^{-1}$ dans Tanagra - Consommation des véhicules

Dans le second onglet de la fenêtre d'affichage (Figure 15.3), nous disposons de la matrice $(X'X)^{-1}$ qui ouvre la porte à toute une batterie de tests statistiques (ex. tests de conformité simultanée, combinaison linéaire de variables, levier pour les intervalles de prévision...). Il est facile d'en copier les valeurs dans un tableur. Un tutoriel en détaille l'usage (`http://tutoriels-data-mining.blogspot.com/2011/02/regression-lineaire-lecture-des.html`).

2. Rakotomalala R., *Tests de normalité - Techniques empiriques et tests statistiques*, `http://eric.univ-lyon2.fr/~ricco/cours/cours/Test_Normalite.pdf`

15.1.2 Autres outils liés à la régression dans Tanagra

Sélection de variables

Tanagra intègre 2 composants de sélection de variables (version 1.4.38) : l'un implémente la procédure *forward*, l'autre la procédure *backward* [13] (chapitre 3). Dans la copie d'écran ci-jointe, nous avons mis en oeuvre la sélection *backward* sur nos données. La variable "cylindrée" a été éliminée, puis le processus a été stoppé car toutes les variables restants étaient significatives au niveau de signification choisi par l'utilisateur (Figure 15.4).

Détection des points atypiques et influents

Tanagra intègre toute une panoplie d'outils de détection de points atypiques et influents dans la régression. Les formules et les interprétations sont longuement décrites dans le chapitre 2 de notre second fascicule [13].

Dans notre exemple, nous l'appliquons sur la régression portant sur les variables sélectionnées (puissance et poids). Nous avons d'une part les valeurs des indicateurs et les valeurs de coupures, les observations suspectes sont mis en évidence (Figure 15.5), d'autre part un récapitulatif permet d'établir un diagnostic rapidement (Figure 15.6).

Enfin, le composant DFBETAS permet d'identifier le coefficient du modèle sur lequel agit inconsidérément une observation par trop influente (Figure 15.7).

Le diagramme de traitement

Comme la très grande majorité des logiciels de Data Mining, Tanagra retrace les opérations menées sur les données à l'aide d'un diagramme. Nous pouvons le sauvegarder pour des traitements ultérieurs. Soit parce que le fichier a été mis à jour, soit tout simplement parce que nous souhaitons compléter notre étude.

Concernant les analyses décrites dans cette section, nous avons réalisé (Figure 15.8) : une importation des données (Dataset), spécifié l'endogène et les exogènes (Define Status), mené une première analyse de régression (Multiple linear regression), effectué une sélection de variables *backward*, opéré une première détection des points atypiques et influents (Outlier Detection), puis une second analyse approfondie permettant de déterminer sur quels coefficients agissent ces points (Dfbetas).

15.1.3 Tutoriels Tanagra

Tanagra est un logiciel, mais c'est aussi et surtout plus de 150 tutoriels en français (à peu près 130 en anglais) dédiés à la pratique du Data Mining[3]. Plusieurs d'entre eux ont trait à la régression (`http://tutoriels-data-mining.blogspot.com/search/label/Régression`). Nous citerons entres autres :

3. A ce jour, Mai 2011.

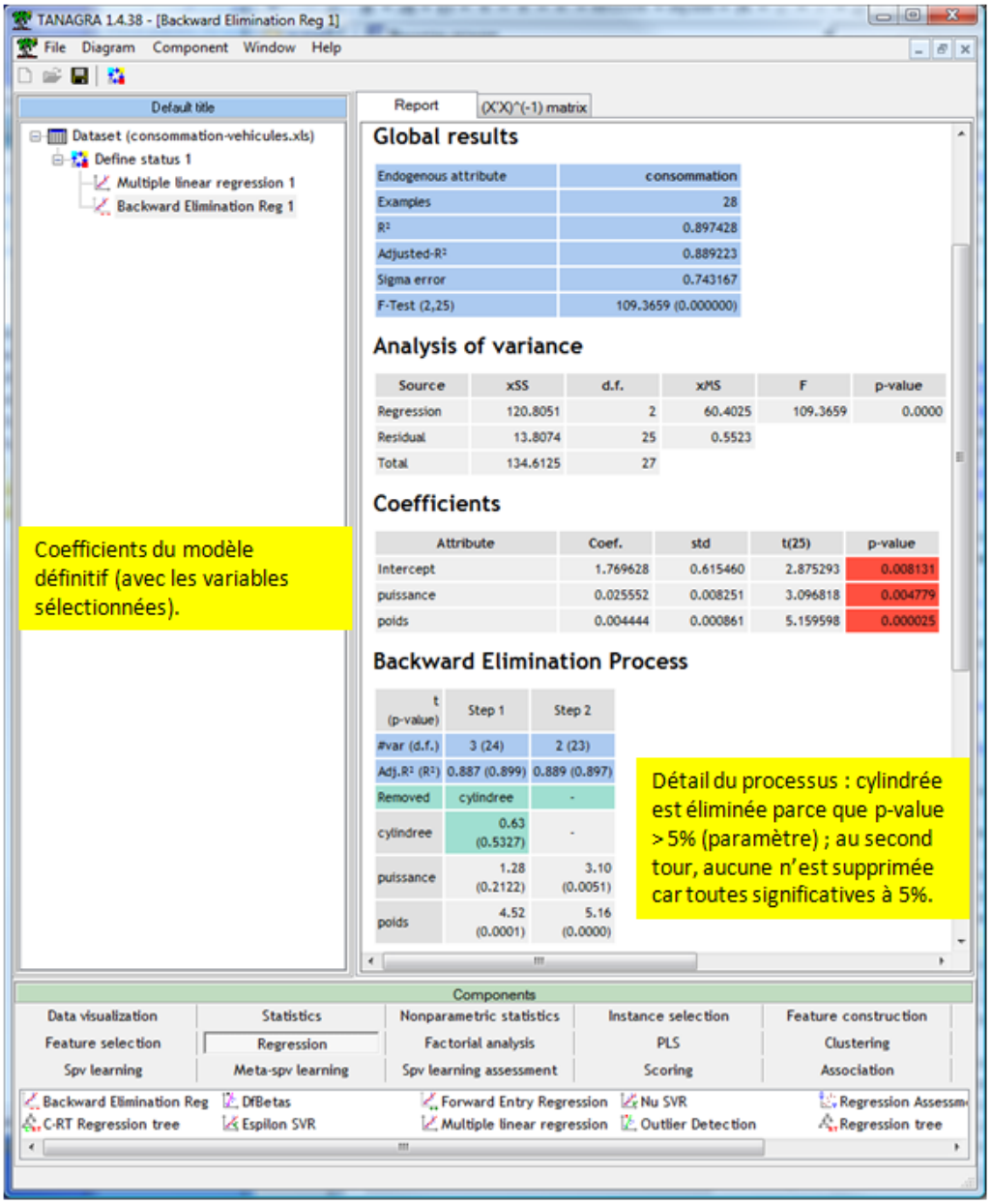
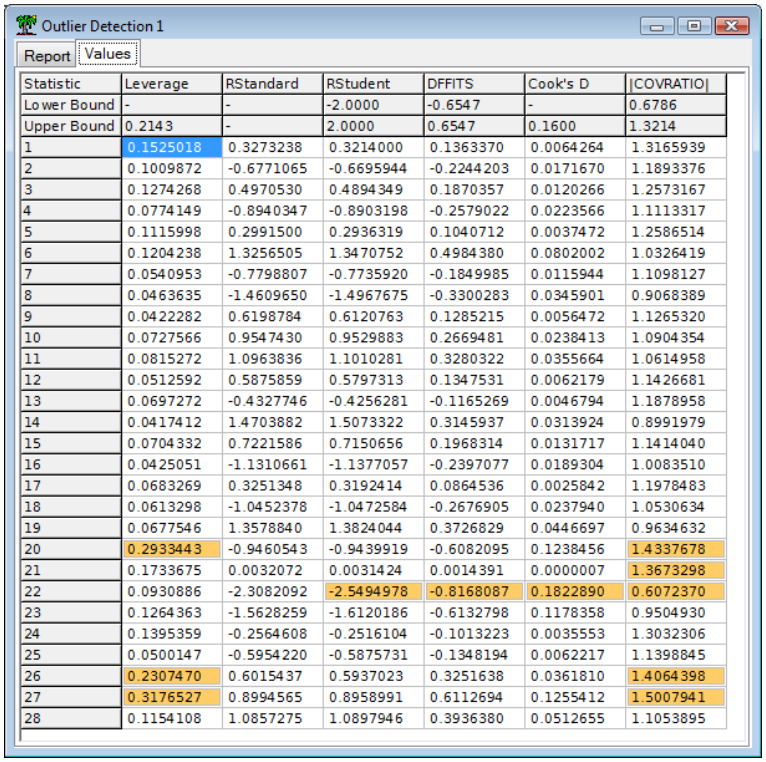


Fig. 15.4. Utilisation du composant "Backward Elimination Reg" dans Tanagra - Consommation des véhicules

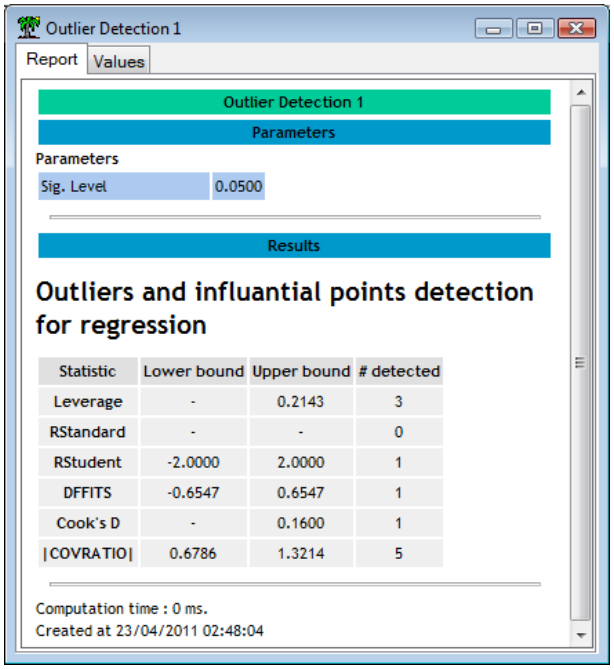
– **Régression linéaire - Lecture des résultats** (`http://tutoriels-data-mining.blogspot.com/2011/02/regression-lineaire-lecture-des.html`). Ce document retrace les principales formules utilisés dans la régression. Il met en parallèle les sorties de Tanagra, mais aussi ceux de R. On peut le voir comme une version très abrégée de ce support de cours. Un accent particulier est mis sur l'utilisation de la matrice $(X'X)^{-1}$ dans différentes opérations subséquentes (test de conformité, test de combinaison linéaires, intervalle de prédiction). L'analyse complète menée avec le tandem Tanagra + Excel est entièrement reproduite à l'aide du logiciel R.


Outlier Detection 1

Report | Values

Statistic	Leverage	RStandard	RStudent	DFFITS	Cook's D	|COVRATIO|
Lower Bound	-	-	-2.0000	-0.6547	-	0.6786
Upper Bound	0.2143	-	2.0000	0.6547	0.1600	1.3214
1	0.1525018	0.3273238	0.3214000	0.1363370	0.0064264	1.3165939
2	0.1009872	-0.6771065	-0.6695944	-0.2244203	0.0171670	1.1893376
3	0.1274268	0.4970530	0.4894349	0.1870357	0.0120266	1.2573167
4	0.0774149	-0.8940347	-0.8903198	-0.2579022	0.0223566	1.1113317
5	0.1115998	0.2991500	0.2936319	0.1040712	0.0037472	1.2586514
6	0.1204238	1.3256505	1.3470752	0.4984380	0.0802002	1.0326419
7	0.0540953	-0.7798807	-0.7735920	-0.1849985	0.0115944	1.1098127
8	0.0463635	-1.4609650	-1.4967675	-0.3300283	0.0345901	0.9068389
9	0.0422282	0.6198784	0.6120763	0.1285215	0.0056472	1.1265320
10	0.0727566	0.9547430	0.9529883	0.2669481	0.0238413	1.0904354
11	0.0815272	1.0963836	1.1010281	0.3280322	0.0355664	1.0614958
12	0.0512592	0.5875859	0.5797313	0.1347531	0.0062179	1.1426681
13	0.0697272	-0.4327746	-0.4256281	-0.1165269	0.0046794	1.1878958
14	0.0417412	1.4703882	1.5073322	0.3145937	0.0313924	0.8991979
15	0.0704332	0.7221586	0.7150656	0.1968314	0.0131717	1.1414040
16	0.0425051	-1.1310661	-1.1377057	-0.2397077	0.0189304	1.0083510
17	0.0683269	0.3251348	0.3192414	0.0864536	0.0025842	1.1978483
18	0.0613298	-1.0452378	-1.0472584	-0.2676905	0.0237940	1.0530634
19	0.0677546	1.3578840	1.3824044	0.3726829	0.0446697	0.9634632
20	0.2933443	-0.9460543	-0.9439919	-0.6082095	0.1238456	1.4337678
21	0.1733675	0.0032072	0.0031424	0.0014391	0.0000007	1.3673298
22	0.0930886	-2.3082092	-2.5494978	-0.8168087	0.1822890	0.6072370
23	0.1264363	-1.5628259	-1.6120186	-0.6132798	0.1178358	0.9504930
24	0.1395359	-0.2564608	-0.2516104	-0.1013223	0.0035553	1.3032306
25	0.0500147	-0.5954220	-0.5875731	-0.1348194	0.0062217	1.1398845
26	0.2307470	0.6015437	0.5937023	0.3251638	0.0361810	1.4064398
27	0.3176527	0.8994565	0.8958991	0.6112694	0.1255412	1.5007941
28	0.1154108	1.0857275	1.0897946	0.3936380	0.0512655	1.1053895

Fig. 15.5. Détection des points atypiques - Indicateurs, bornes basses et hautes - Consommation des véhicules


Outlier Detection 1

Report | Values

Outlier Detection 1

Parameters

Parameters

Sig. Level	0.0500

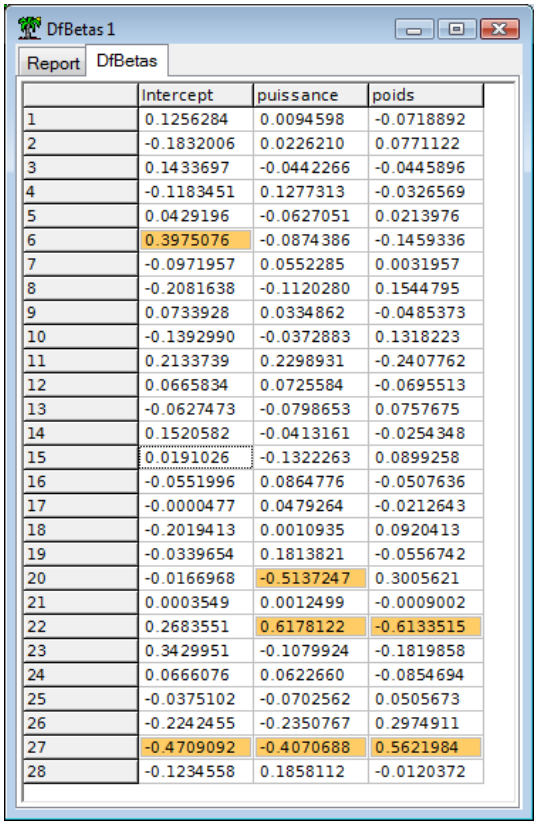
Results

Outliers and influantial points detection for regression

Statistic	Lower bound	Upper bound	# detected
Leverage	-	0.2143	3
RStandard	-	-	0
RStudent	-2.0000	2.0000	1
DFFITS	-0.6547	0.6547	1
Cook's D	-	0.1600	1
|COVRATIO|	0.6786	1.3214	5

Computation time : 0 ms.
Created at 23/04/2011 02:48:04

Fig. 15.6. Détection des points atypiques - Bilan - Consommation des véhicules



DfBetas 1

Report | DfBetas

	Intercept	puissance	poids
1	0.1256284	0.0094598	-0.0718892
2	-0.1832006	0.0226210	0.0771122
3	0.1433697	-0.0442266	-0.0445896
4	-0.1183451	0.1277313	-0.0326569
5	0.0429196	-0.0627051	0.0213976
6	0.3975076	-0.0874386	-0.1459336
7	-0.0971957	0.0552285	0.0031957
8	-0.2081638	-0.1120280	0.1544795
9	0.0733928	0.0334862	-0.0485373
10	-0.1392990	-0.0372883	0.1318223
11	0.2133739	0.2298931	-0.2407762
12	0.0665834	0.0725584	-0.0695513
13	-0.0627473	-0.0798653	0.0757675
14	0.1520582	-0.0413161	-0.0254348
15	0.0191026	-0.1322263	0.0899258
16	-0.0551996	0.0864776	-0.0507636
17	-0.0000477	0.0479264	-0.0212643
18	-0.2019413	0.0010935	0.0920413
19	-0.0339654	0.1813821	-0.0556742
20	-0.0166968	-0.5137247	0.3005621
21	0.0003549	0.0012499	-0.0009002
22	0.2683551	0.6178122	-0.6133515
23	0.3429951	-0.1079924	-0.1819858
24	0.0666076	0.0622660	-0.0854694
25	-0.0375102	-0.0702562	0.0505673
26	-0.2242455	-0.2350767	0.2974911
27	-0.4709092	-0.4070688	0.5621984
28	-0.1234558	0.1858112	-0.0120372

Fig. 15.7. Détection des points atypiques - DFBETAS - Consommation des véhicules

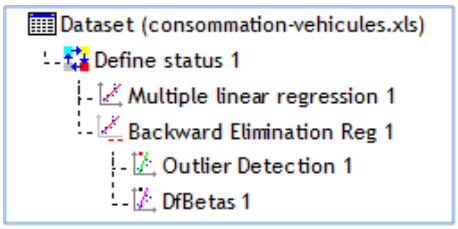


Fig. 15.8. Diagramme de traitements Tanagra - Consommation des véhicules

- **Points aberrants et influents dans la régression** (`http://tutoriels-data-mining.blogspot.com/2008/04/points-aberrants-et-influents-dans-la.html`). Ce tutoriel détaille la mise en oeuvre des outils de détection des points atypiques et influents dans Tanagra. Tous les résultats sont comparés avec ceux de R et SAS.
- **Colinéarité et régression** (http ://tutoriels-data-mining.blogspot.com/2008/04/ colinarit-et-rgression.html). Dans un premier temps, il décrit les outils à utiliser pour détecter rapidement les problèmes de colinéarité (redondance des explicatives) dans la régression. Dans un deuxième temps, différentes solutions pour combattre la colinéarité sont étudiées. Tout d'abord une première solution basée sur la sélection de variables est proposée. Puis, par la suite, d'autres pistes sont

explorées : la régression sur les axes d'une ACP (analyse en composante principale) et la régression PLS (partial least squares). Enfin, nous comparons les coefficients des différents modèles obtenus.

- **Diagnostic de la régression avec R** (`http://tutoriels-data-mining.blogspot.com/2009/05/diagnostic-de-la-regression-avec-r.html`). Sous forme de "slides", il montre les principales commandes de R pour le diagnostic de la régression : graphique des résidus, repérage des points atypiques, détection et traitement de la colinéarité.
- D'autres tutoriels décrivant les autres techniques de régression peuvent nous intéresser également : les arbres de régression (`http://tutoriels-data-mining.blogspot.com/2008/04/arbres-de-rgression.html`, les support vector regression (SVR - `http://tutoriels-data-mining.blogspot.com/2009/04/support-vector-regression.html`), ...

15.2 REGRESS

Le logiciel REGRESS est un logiciel très simplifié de régression linéaire multiple que j'ai développé il y a fort longtemps. Je l'ai mis à jour à l'occasion de l'écriture de ce document. Mon idée est de le mettre en totale adéquation avec les formules présentées dans mes fascicules consacrés à la régression.

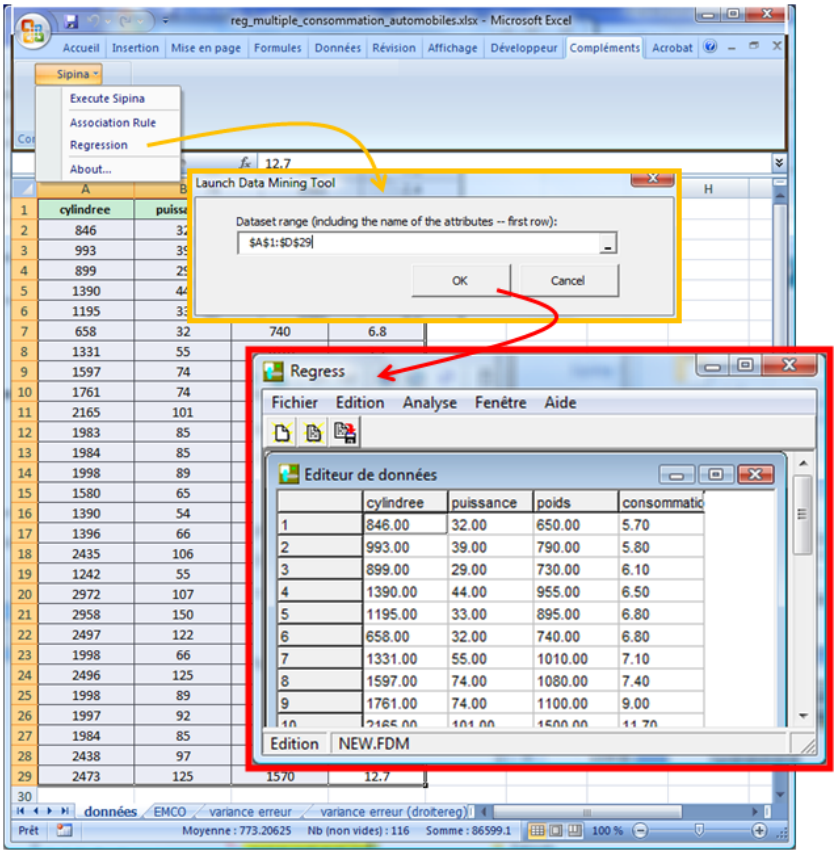


Fig. 15.9. Envoi des données d'Excel vers REGRESS via la macro complémentaire SIPINA.XLA

REGRESS est distribué de deux manières. Il peut être chargé et installé individuellement (`http://eric.univ-lyon2.fr/~ricco/regress.html`). Il peut être installé en même temps que la distribution SIPINA (`http://eric.univ-lyon2.fr/~ricco/sipina.html`). Cette seconde solution est préférable. En effet, il bénéficie dans ce cas d'une intégration privilégiée dans Excel via la macro complémentaire SIPINA.XLA. Tout comme TANAGRA ou SIPINA, il est dès lors possible de charger ses données dans le tableur Excel, de procéder à toutes les opérations de préparation et de transformations possibles et imaginables [4], puis de les envoyer à REGRESS pour la modélisation (Figure 15.9).

REGRESS est exclusivement piloté par menu. En cela, il se rapproche de OPEN STAT (`http://www.statpages.org/miller/openstat/`), un excellent logiciel gratuit et source libre, très complet, que j'utilise souvent pour vérifier mes calculs dans le domaine de la statistique [5].

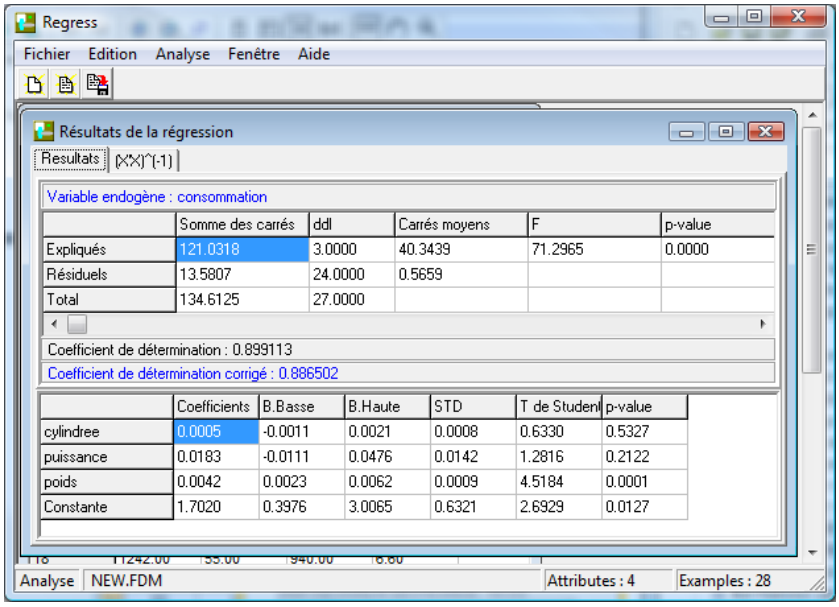


Fig. 15.10. Fenêtre de résultats de REGRESS

Après avoir spécifié l'endogène et les exogènes dans la boîte de dialogue de paramétrage, les principaux résultats apparaissent dans une fenêtre dédiée (Figure 15.10). Nous observons successivement : le tableau d'analyse de variance avec la statistique F du test de significativité globale ; le R^2 et le R^2-ajusté ; la grille des coefficients, avec notamment leurs intervalles de confiance à 95% (paramétrable).

4. Excel est très largement utilisé dans ce contexte - `http://www.kdnuggets.com/polls/2010/data-mining-analytics-tools.html`

5. Et qui est très complet concernant la régression linéaire multiple. Hélas, je ne peux pas présenter tous les outils existants dans ce fascicule. J'ai du faire des choix. Sur le site de OPEN STAT, vous trouverez plusieurs tutoriels, rédigés et sous forme d'animation vidéo. C'est vraiment du travail de très très grande qualité.

La mise en oeuvre de REGRESS et l'accès aux résultats sont décrits dans un tutoriel accessible en ligne (`http://tutoriels-data-mining.blogspot.com/2011/05/regress-dans-la-distribution-sipina.html`).

15.3 Le logiciel R

R est un logiciel extraordinaire (`http://www.r-project.org/`). Je ne lui vois qu'un seul défaut : il faut faire l'apprentissage de son langage de commande (de programmation) pour en tirer profit. Pour les personnes réfractaires à l'informatique, la barrière (psychologique) peut paraître insurmontable. Mais une fois cet écueil passé, on constate rapidement les immenses possibilités de l'outil.

Comme pour Tanagra, plusieurs tutoriels relatifs à la pratique de la régression avec R sont disponibles sur notre site web – `http://tutoriels-data-mining.blogspot.com/`. Mais, de toute manière, vous trouverez de très nombreux documents gratuits et de qualité sur internet via Google. Citons, entres autres, le fameux (parce précurseur) tutoriel de Julian J. Faraway, *Practical Regression and Anova using R*, 2002 ; `http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf`.

Et n'allez surtout pas acheter les livres qui prétendent présenter la régression *et* sa mise en oeuvre avec R dans un chapitre de quelques pages, noyées au milieu de tout un tas de techniques statistiques, décrites également de manière expéditive[6]. Ca ne vous servira pas à grand chose. Mis à part constater que le label **R** fait vendre. Curieuse destinée pour un logiciel gratuit.

15.3.1 La procédure *lm()*

La procédure *lm*() lance la régression dans R (version 2.12.0). Les sorties paraissent éminemment laconiques, voire lapidaires, dans un premier temps. Seuls les coefficients sont affichés (Figure 15.11).

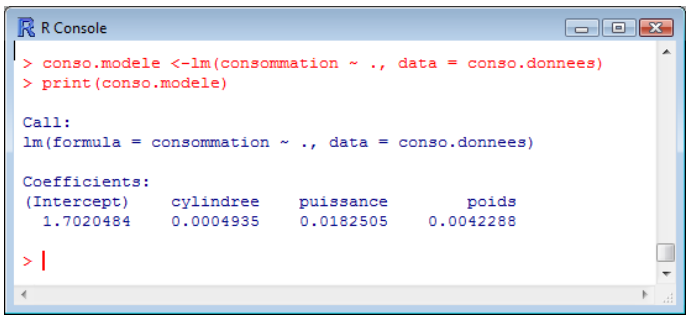


Fig. 15.11. La commande *lm*() de R - Consommation des véhicules

6. Et ils sont nombreux, surtout en anglais. J'en ai moi-même acheté. Honte à moi. A part caler mon étagère, je ne vois pas très bien à quoi ils peuvent servir.

Il ne faut pas s'arrêter à ce premier résultat. Si on connaît un peu R, on sait que des champs sont associés à la grande majorité des objets de R. Pour en obtenir la liste, nous utilisons la commande *attributes*(.). On se rend compte alors qu'on peut avoir accès aux coefficients sous forme de tableau ($coefficients), ça peut toujours être intéressant pour des manipulations ultérieures ; mais nous avons également accès à d'autres informations comme les résidus ($*residuals*) (Figure 15.12).

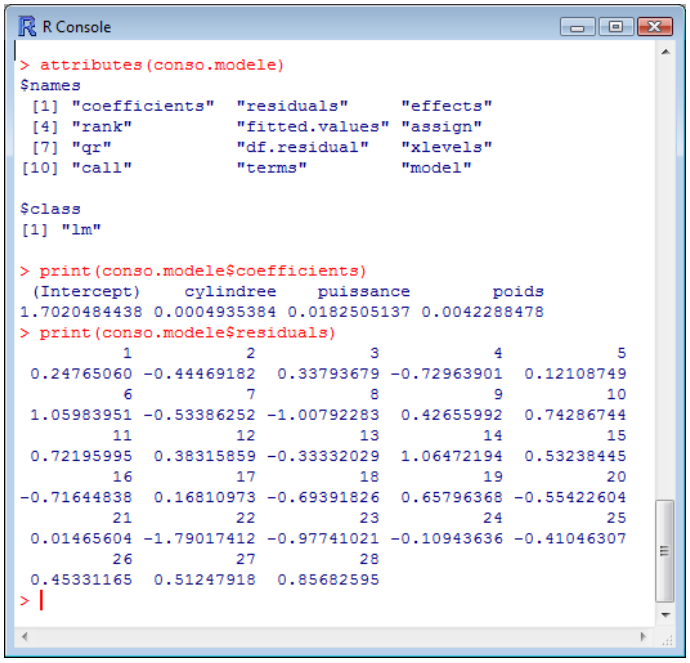


Fig. 15.12. Accès aux champs de l'objet *lm*() de R - Consommation des véhicules

15.3.2 L'objet *summary* de *lm()*

Les résultats détaillés viennent avec l'objet *summary* de *lm*. Nous obtenons le tableau de coefficients accompagnée cette fois du test de significativité individuelle. Un rapport sur le test de significativité globale est également proposé (Figure 15.13).

Comme toujours dans R, nous avons accès aux champs de l'objet. Dans notre copie d'écran, nous affichons l'estimation de l'écart-type de l'erreur et la fameuse matrice $(X'X)^{-1}$ (Figure 15.14).

A partir de là, toutes les post-traitements possibles et imaginables sont réalisables pour peu que l'on sache transcrire les bonnes commandes.

15.3.3 Sélection de variables avec *stepAIC*

Concernant la sélection de variables, la littérature met souvent en avant la commande *stepAIC* du package MASS. La procédure consiste à trouver la combinaison de variable qui minimise le critère AIC

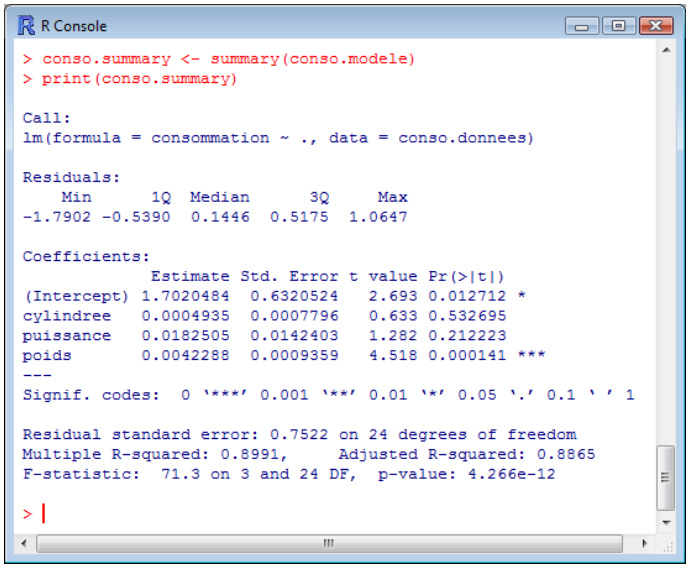


Fig. 15.13. Sorties de l'objet *summary* de *lm*() - Consommation des véhicules

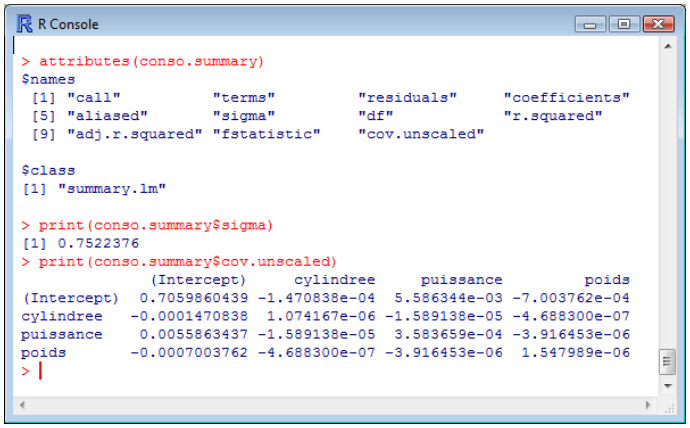


Fig. 15.14. Accès aux champs de *summary* de *lm*() - Consommation des véhicules

(Akaike) ou, c'est paramétrable, le critère BIC de Schwartz. Les stratégies usuelles de recherche (*forward*, *backward*, *stepwise* - bidirectionnelle) sont proposés.

Pour notre part, nous avons réalisé une sélection *bacwkard* avec pour point de départ la régression sur la totalité des variables, et en demandant à ce que le détail des opérations soit affiché. A la sortie, nous obtenons un modèle avec les variables poids et puissance (Figure 15.15).

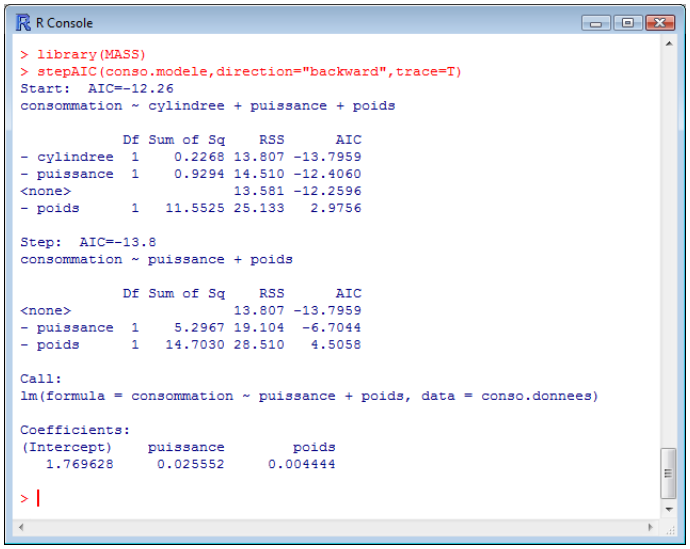


Fig. 15.15. Sélection de variables avec la commande *stepAIC* - Consommation des véhicules

15.4 Régression avec les tableurs

15.4.1 DROITEREG sous Open Office Calc

J'utilise beaucoup Excel tout simplement parce que c'est l'outil dont je dispose pour mes cours à l'Université Lyon 2. En réalité, le terme "tableur" est plus approprié. Dans cette optique, j'aurais tout aussi bien pu utiliser le tableur CALC de la suite bureautique *gratuite* OPEN OFFICE (`http://fr.openoffice.org/`) pour l'élaboration de ce document.

Ainsi, outre les fonctions de calculs standards et les opérations matricielles, Calc propose également la fonction DROITEREG, avec exactement le même mode opératoire. Cela n'est absolument pas étonnant. Il sait importer sans pertes (à ma connaissance) les fichiers au format XLSX de Excel 2007 et 2010. Les données et les formules sont préservées.

Par curiosité, j'ai inséré la fonction Droitereg de Calc sur les données "Consommation de véhicules" (cf. l'expression dans la barre de formules), et j'ai copié (collage spécial valeurs) en dessous les valeurs proposées par Excel. Tout doute, s'il y en avait un, est absolument levé quant aux capacités de calcul de Calc en matière de régression (Figure 15.16) [7].

15.4.2 Add-on pour Open Office Calc

Il est possible d'enrichir les fonctionnalité de Calc en intégrant des "greffons" (*add-on* en anglais). Le plus souvent, il s'agit de macro complémentaires qui installent de nouveaux menus dans Open Office. Ils permettent de faire le lien avec des logiciels externes. Ainsi, toute la gestion des données, opérations

7. `reg_multiple_consommation_automobiles.ods` - "droitereg - comparaison"

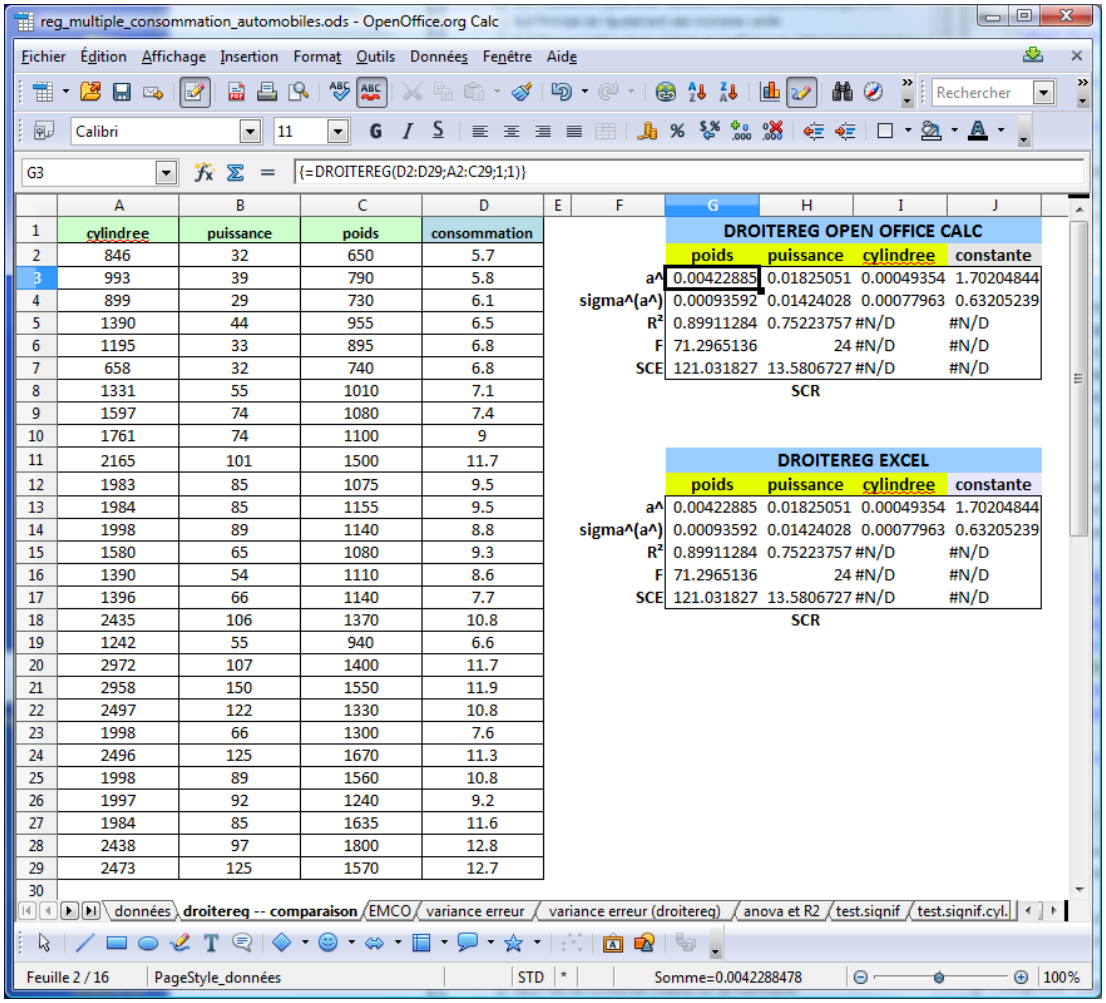


Fig. 15.16. DROITEREG sous Open Office Calc - Comparaison avec Excel

souvent fastidieuses, est dévolue au tableur. Les calculs scientifiques sont en revanche réalisés à l'aide des logiciels spécialisés. Chaque outil oeuvre dans l'espace qui lui est le plus favorable.

Parmi les innombrables add-ons disponibles, nous citerons volontiers, parce que faisant référence à des logiciels gratuits que tout le monde peut charger et installer, ceux de Tanagra [8] et de R [9].

8. http://tutoriels-data-mining.blogspot.com/2008/03/connexion-open-office-calc.html
9. http://wiki.services.openoffice.org/wiki/R_and_Calc

15.4.3 L'utilitaire d'analyse du tableur Excel

Il est également possible d'intégrer des "greffons" dans Excel. Tanagra en propose (*tanagra.xla* – pour Excel 2003 et versions antérieures [10] ; pour Excel 2007 et plus récentes [11]). Je ne doute absolument pas qu'il ne puisse y avoir de solutions analogues pour R (*il suffit de chercher un peu*).

Dans cette section, j'ai choisi de mettre en avant "l'utilitaire d'analyse" parce qu'elle fait partie de la distribution standard d'Excel. Aucune installation additionnelle n'est requise. Parmi les techniques statistiques proposées se trouve la régression linéaire. Par rapport à DROITEREG, ses sorties sont plus riches, d'où l'intérêt de les décrire de manière détaillée.

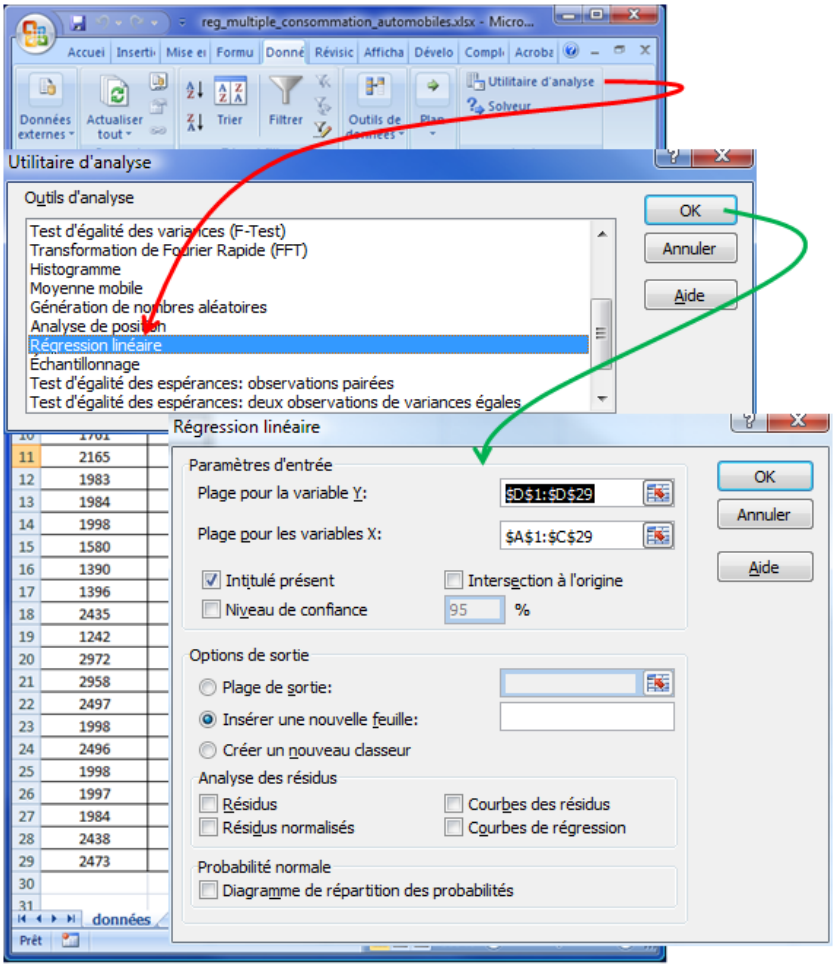


Fig. 15.17. Utilitaire d'analyse - Excel - Paramétrage

Dans Excel 2007, l'utilitaire d'analyse est accessible dans l'onglet "Données". Nous sélectionnons la régression linéaire. La boîte de paramétrage apparaît (Figure 15.17) :

10. http://tutoriels-data-mining.blogspot.com/2008/03/importation-fichier-xls-excel-macro.html
11. http://tutoriels-data-mining.blogspot.com/2010/08/ladd-in-tanagra-pour-excel-2007-et-2010.html

- Nous spécifions les plages de valeurs pour l'endogène et les exogènes. Nous pouvons sélectionner les étiquettes de colonnes, il faut simplement préciser à Excel que la première ligne correspond aux noms des variables dans ce cas.
- Les résultats sont insérés dans une nouvelle feuille de calcul.
- Nous pouvons, si nous le souhaitons, obtenir des indications détaillées concernant les résidus.

Les résultats sont affichés dans une nouvelle feuille, conformément à notre paramétrage. Par rapport à DROITEREG, les sorties sont mieux organisées, elles intègrent de surcroît les ratios intermédiaires permettant de tester la significativité globale de la régression (tableau d'analyse de variance, test de Fisher) et la significativité de chaque coefficient (t calculé, probabilité critique). Les intervalles de confiance des coefficients sont également fournies. Je ne sais pas du tout en revanche pourquoi les colonnes associées sont dupliquées (Figure 15.18) [12].

RAPPORT DÉTAILLÉ

Statistiques de la régression		
Coefficient de détermination multiple	0.948215609	**R**
Coefficient de détermination R^2	0.899112841	**R^2**
Coefficient de détermination R^2	0.886501946	**R^2-ajusté**
Erreur-type	0.752237571	**sigma^(epsilon)**
Observations	28	**n**

ANALYSE DE VARIANCE

	Degré de liberté	*Somme des carrés*	*Moyenne des carrés*	*F*	*aleur critique de F*
Régression	3	121.03183	40.34394	71.29651	0.00000
Résidus	24	13.58067	0.56586		
Total	27	134.6125			

	Coefficients	*Erreur-type*	*Statistique t*	*Probabilité*	*Limite inférieure pour seuil de confiance = 95%*	*Limite supérieure pour seuil de confiance = 95%*	*Limite inférieure pour seuil de confiance = 95.0%*	*Limite supérieure pour seuil de confiance = 95.0%*
Constante	1.70205	0.63205	2.69289	0.01271	0.39756	3.00654	0.39756	3.00654
cylindree	0.00049	0.00078	0.63304	0.53269	-0.00112	0.00210	-0.00112	0.00210
puissance	0.01825	0.01424	1.28161	0.21222	-0.01114	0.04764	-0.01114	0.04764
poids	0.00423	0.00094	4.51838	0.00014	0.00230	0.00616	0.00230	0.00616
	a^	sigma^(a^)	t - calculé	p-value	Intervalle de confiance à 95%			

Fig. 15.18. Utilitaire d'analyse - Excel - Sorties

15.5 SAS

SAS est un logiciel connu des statisticiens, bien en place depuis de très nombreuses années déjà. Il doit faire face à une concurrence de plus en plus accrue aujourd'hui. Beaucoup de praticiens se posent la question du passage à d'autres logiciels libres (ou non) de qualité (KDnuggets Poll, *Switching from SAS to WPS, R...*, `http://www.kdnuggets.com/polls/2010/switching-from-sas-to-wps.html`).

12. `reg_multiple_consommation_automobiles.xlsx` - "utilitaire d'analyse"

Je me contenterai d'une description assez succincte dans ce fascicule (SAS version 9.2). Pour le lecteur désireux d'en apprendre plus sur la pratique de la régression sous SAS, je conseille l'excellent tutoriel de Confais et Leguen (2005) [4] paru dans la non moins excellente revue gratuite en ligne MODULAD (`http://www-roc.inria.fr/axis/modulad/`).

La régression sur les données "Consommation des véhicules" a été réalisée à l'aide des commandes suivantes

```
proc reg data = ucidata.consovehicules;
model consommation = poids puissance cylindree;
run;
```

Nous obtenons les sorties standards de la régression, à savoir : le tableau d'analyse de variance et les ratios associés (test F de significativité globale et R^2), le tableau des coefficients et le tests de significativité individuels (Figure 15.19). Les résultats sont bien évidemment les mêmes que ceux des autres logiciels.

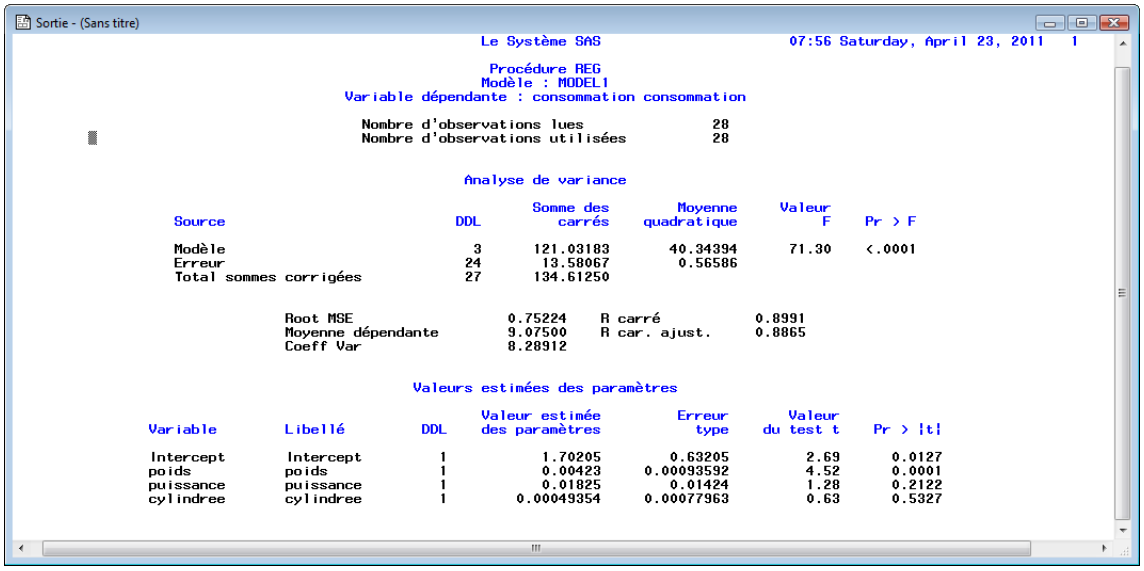


Fig. 15.19. Régression avec la PROC REG de SAS - Consommation des véhicules

15.6 SPAD

SPAD (version 7.3) est un logiciel de traitement statistique qui a fait les beaux jours de l'analyse de données "à la française". Depuis quelques années, il étend ses compétences en investissant, entres autres, les domaines de la modélisation et du data mining.

Nous avons construit une filière pour réalisé la régression linéaire multiple (Figure 15.20). Le composant dédié "Régression Anova" encapsule plusieurs techniques connexes : la régression, l'analyse de

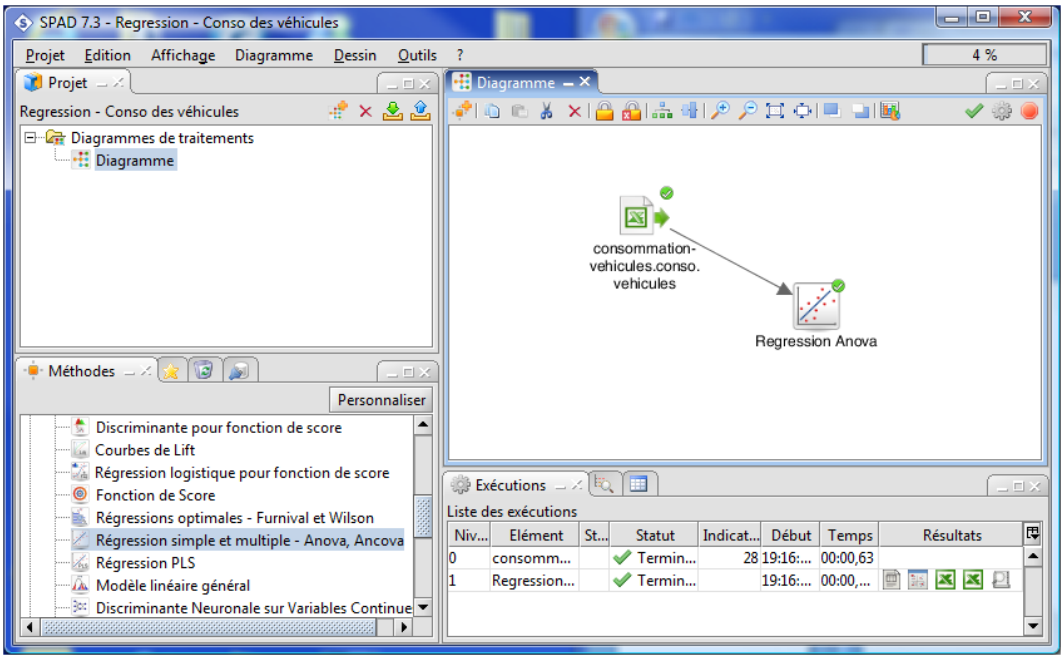


Fig. 15.20. La "filière" SPAD pour la Régression - Consommation des véhicules

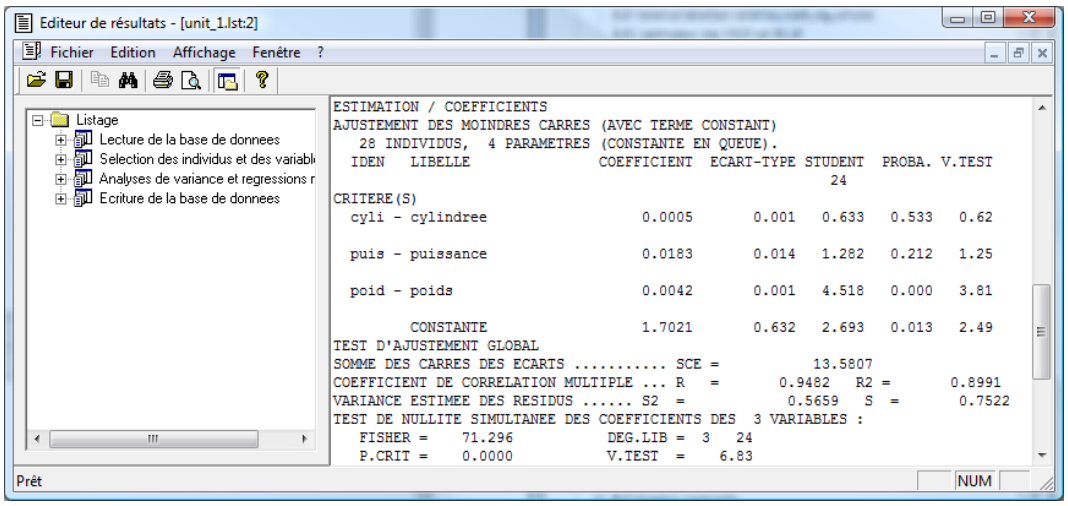


Fig. 15.21. Résultats de SPAD pour la Régression - Consommation des véhicules

variance (anova) et l'analyse de covariance. Dans notre étude, la variable à expliquer est quantitative, les facteurs simples également, nous opérons bien une analyse de régression.

Les résultats peuvent être visualisés de différentes manières. Pour ma part, je préfère l'éditeur de résultats car il permet d'obtenir directement une vision globale : tous les éléments importants tiennent sur une seule page (Figure 15.21). L'autre option est de transférer les résultats dans le tableur Excel, la présentation est certainement meilleure, mais le test de significativité globale et la grille des coefficients sont sur deux feuilles différentes. Tout dépend des souhaits de l'utilisateur en définitive.

15.7 SPSS

Nous lançons la régression linéaire standard (Analyse / Régression / Linéaire...) dans SPSS version 12.0. Dans le fenêtre de rapport sont affichés : le tableau indiquant la qualité globale du modèle (R^2, $\hat{\sigma}_\varepsilon$) ; le tableau d'analyse de variance et le test F d'évaluation globale du modèle ; la grille des paramètres de la régression avec les coefficients standardisés et les tests individuels de significativité (Figure 15.22).

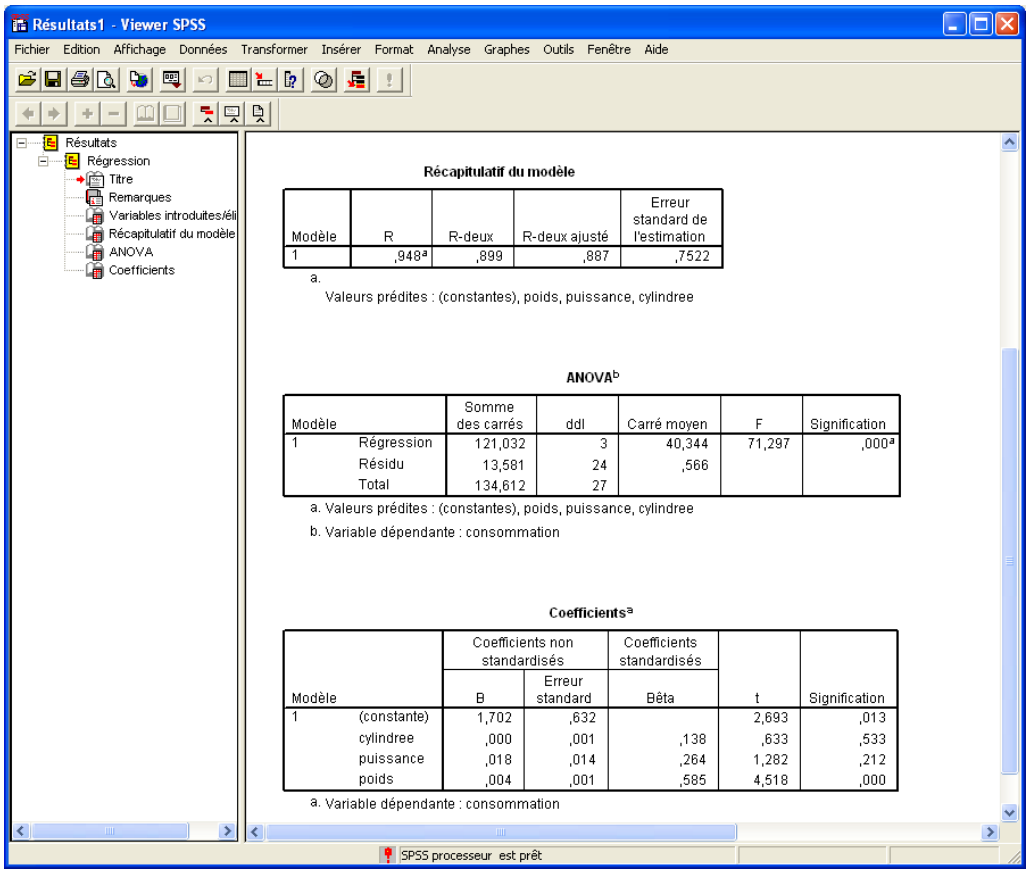


Fig. 15.22. Rapport relatif à la Régression Linéaire sous SPSS - Consommation des véhicules

15.8 STATISTICA

Ma version de STATISTICA est plutôt ancienne (version 5.5). Mais bon, la régression telle que nous l'abordons n'ayant pas connu de bouleversements théoriques forts ces dernières années (enfin j'imagine), nous pouvons considérer que les sorties restent d'actualité.

Les données ont été importées, nous lançons la régression en spécifiant la variable dépendante (endogène) et les variables indépendantes (exogènes). Nous obtenons un bilan global de la régression dans une première fenêtre (Figure 15.23). Nous y trouvons le coefficient de détermination R^2, la valeur de la statistique F, l'écart type estimé de l'erreur, etc.

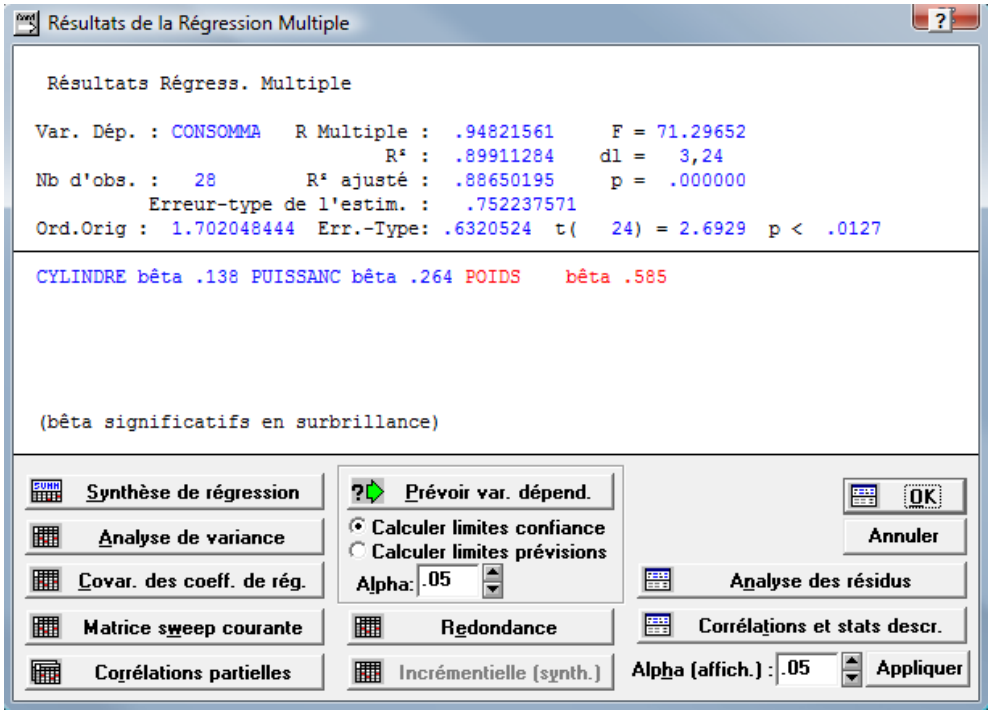
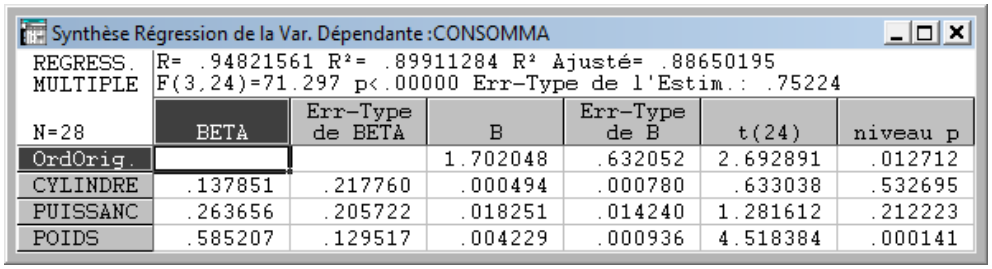


Fig. 15.23. Régression linéaire avec STATISTICA - Premiers résultats - Consommation des véhicules



	BETA	Err-Type de BETA	B	Err-Type de B	t(24)	niveau p
OrdOrig.			1.702048	.632052	2.692891	.012712
CYLINDRE	.137851	.217760	.000494	.000780	.633038	.532695
PUISSANC	.263656	.205722	.018251	.014240	1.281612	.212223
POIDS	.585207	.129517	.004229	.000936	4.518384	.000141

Fig. 15.24. Grille des coefficients estimés sous STATISTICA - Consommation des véhicules

Cette fenêtre nous permet d'accéder à d'autres résultats.Si nous cliquons sur le bouton "Synthèse de régression" par exemple, nous obtenons la grille des coefficients avec les tests de significativité individuels. Notons que STATISTICA propose directement les coefficients standardisés (BETA) (Figure 15.24).

D'autres analyses sont possibles bien évidemment. Si nous actionnons le bouton "Analyse des résidus", nous accédons à un panneau de commande particulièrement complet permettant de scruter en détail les caractéristiques des résidus de la régression (Figure 15.25). Nous pouvons obtenir, entres autres, la "Droite de Henry" (graphique Q-Q Plot ; [13], chapitre 1) permettant de vérifier la compatibilité de la distribution observée des résidus avec l'hypothèse gaussienne (Figure 15.26).

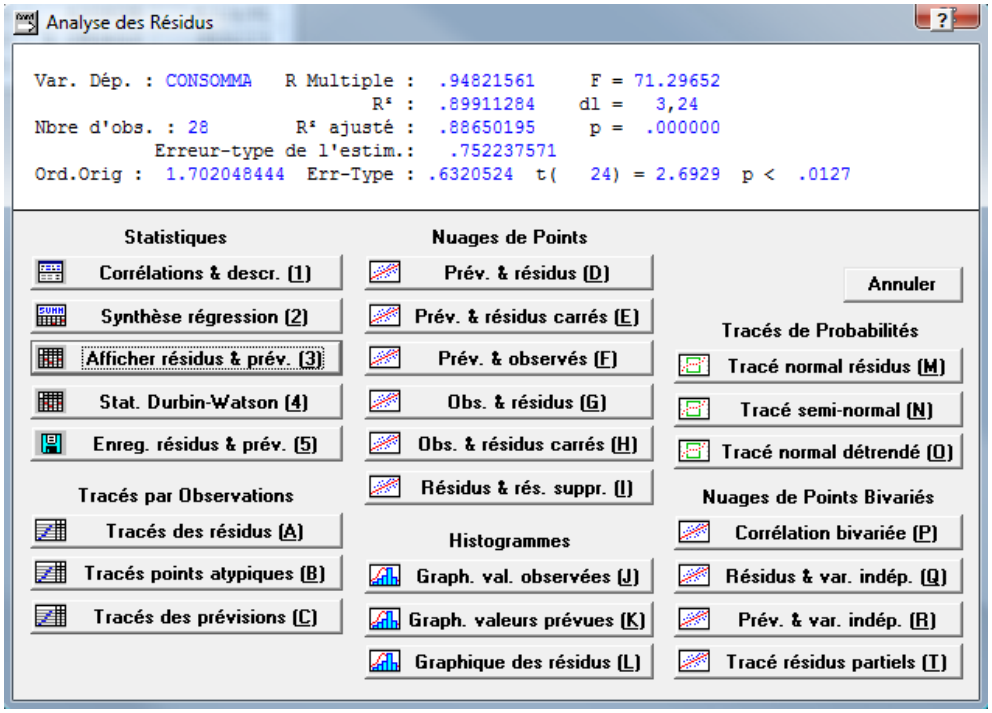


Fig. 15.25. Panneau de commande de l'analyse des résidus sous STATISTICA - Consommation des véhicules

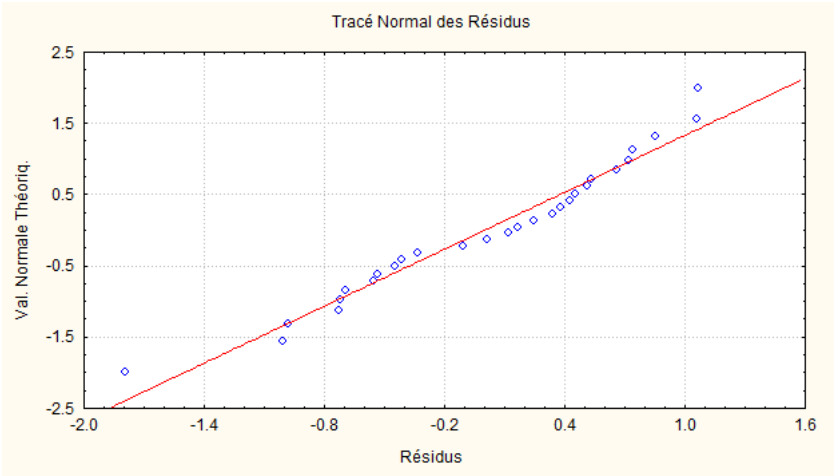


Fig. 15.26. Droite de Henry (Q-Q Plot) sous STATISTICA - Consommation des véhicules

15.9 A propos des logiciels

Sur des calculs reposant sur des algorithmes déterministes et maîtrisés (opérations matricielles), les logiciels fournissent des résultats identiques. Le contraire aurait été inquiétant. Après, privilégier tel ou tel outil dépend essentiellement d'autres considérations comme la possibilité d'initier des calculs supplémentaires simplement (tests statistiques additionnels...), les facilités en termes de manipulation de données

(*data management*), l'accès au logiciel, etc. Il dépend aussi, soyons honnête, de la culture ambiante dans lequel évolue le statisticien.

Je me garderai bien donc de conseiller un logiciel. Le choix appartient pleinement à l'utilisateur. Et c'est très bien ainsi.

A

Gestion des versions

Ce document n'est pas figé. Il est appelé à évoluer dans le temps. Dans cette annexe, nous détaillerons au fur et à mesure son évolution. Le numéro de version est indiquée sur la couverture. En bas de page, nous avons la date et l'heure de la compilation. Toute modification un tant soit peu importante (rajout de section, réorganisation) induit un nouveau numéro de version. Un simple erratum en revanche n'est pas explicitement indiqué (coquilles, fautes d'orthographes), il faut se référer à la date de compilation dans ce cas.

1. **Version 1.0 -** Première version de ce fascicule, terminée et diffusée au mois de mai 2011. Elle comporte 15 chapitres.
2. **Version 1.1 -** Rajout de la section consacrée à la contribution des variables dans la régression via la décomposition du R^2 (section 13.3).

B

Fichiers de données et de calculs

Plusieurs exemples illustrent les sujets traités dans ce document. L'énorme avantage de la distribution par le web est que nous pouvons diffuser les fichiers de données avec les calculs associés.

Tous les fichiers sont au format Excel. Vous avez du le remarquer, chaque copie d'écran est accompagnée en bas de page d'une double référence : le nom du fichier (.xlsx - Excel format 2007) et le nom de la feuille. Vous pouvez ainsi étudier dans le détail la séquence de calculs réalisée pour obtenir les résultats décrits dans le document.

Ces fichiers sont regroupés dans une archive (`http://eric.univ-lyon2.fr/~ricco/cours/cours/econometrie_regression_fichiers.zip`). Nous les listons ici avec les principaux thèmes qui y sont abordés :

1. `regression_simple_rendements_agricoles.xlsx`. Source : Bourbonnais, page 12. Thèmes : régression linéaire simple, intervalle de confiance de la droite de régression, décomposition de la variance, test de significativité globale, test de significativité de la pente, intervalle de confiance de la pente, résultats de droitereg, prédiction ponctuelle, intervalle de prédiction.
2. `conso_poids_vehicules_reg_simple.xlsx`. Thème : étude de cas, consommation de carburant vs. poids.
3. `equipementmagnetoscope.xlsx`. Source : Bourbonnais, page 160. Thèmes : modèle logistique, estimation des coefficients, estimation par balayage de y_{max}.
4. `regression_sans_constante.xlsx`. Thème : régression sans constante, sur données centrées et non-centrées.
5. `comparaisondesregressions.xls`. Thème : comparaison des régressions.
6. `reg_multiple_consommation_automobiles.xlsx`. Thèmes : régression linéaire multiple et sujets associés (en version Open Office Calc : `reg_multiple_consommation_automobiles.ods`).
7. `cigarettes-regressionmultiple.xls`. Thèmes : régression linéaire multiple et sujets associés.
8. `regression-salaire-sexe.xlsx`. Source : `http://www.cabannes.org/exemples_pour_excel.htm`. Thème : régression sur exogène qualitative (binaire).

9. `analysetauxdechomage.xlsx`. Source : `http://aurelie.bonein.free.fr/`. Thème : étude de cas, régression linéaire multiple.

Littérature

1. Aïvazian Z., *Étude statistique des dépendances*, Éditions Mir, 1978.
2. Bourbonnais, R., *Econométrie. Manuel et exercices corrigés*, Dunod, 2è édition, 1998.
3. Bressoux P., *Modélisation statistique appliquées aux sciences sociales*, De Boeck, 2008.
4. Confais J., Le Guen M., *Premier pas en régression linéaire avec SAS®*, Revue Modulad n^o35, pages 220 à 363, 2006.
5. Dagnelie P., *Statistique théorique et appliquées - Inférence Statistique à une et deux dimensions*, vol.2, de Boeck, 2006.
6. Dodge, Y, Rousson, V., *Analyse de régression appliquée*, Dunod, 2è édition, 2004.
7. Giraud, R., Chaix, N., *Econométrie*, Presses Universitaires de France (PUF), 1989.
8. Hardy M., *Regression with Dummy Variables*, Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-093, Newbury Park, CA : Sage, 1993.
9. Jacquard J., Turrisi R., *Interaction effects in multiple regression*, (2^{nd} ed). Sage University Papers Series on Quantitative Applications in the Social Sciences, 07-072, Thousands Oaks, CA : Sage, 2003.
10. Johnston, J., DiNardo, J., *Méthodes Econométriques*, Economica, 4è édition, 1999.
11. Labrousse, C., *Introduction à l'économétrie. Maîtrise d'économétrie*, Dunod, 1983.
12. Rakotomalala R., *Analyse de corrélation - Étude des dépendances - Variables quantitatives*, `http://eric.univ-lyon2.fr/~ricco/cours/cours/Analyse_de_Correlation.pdf`.
13. Rakotomalala R., *Pratique de la régression linéaire multiple - Diagnostic et sélection de variables*, `http://eric.univ-lyon2.fr/~ricco/cours/cours/La_regression_dans_la_pratique.pdf`.
14. Rakotomalala, R., *Pratique de la régression logistique - Régression Logistique Binaire et Polytomique*, `http://eric.univ-lyon2.fr/~ricco/cours/cours/pratique_regression_logistique.pdf`.
15. Saporta, G., *Probabilités, Analyse des données et Statistique*, Technip, 2ème édition, 2006.
16. Scherrer B., *Biostatistique*, Volume 1, Gaëtan Morin Editeur, 2007.
17. Tenenhaus, M., *Statistique - Méthodes pour décrire, expliquer et prévoir*, Dunod, 2007.