Régression linéaire simple

Résumé

Introductions au modèle linéaire et modèle linéaire général. Retour au plan du cours.

1 Introduction

Ce chapitre élémentaire permet d'introduire simplement certains concepts clefs : modèle, estimations, tests, diagnostics, qui seront ensuite déclinés dans des cadres plus généraux. Il vient en complément d'un cours traditionnel de Statistique de niveau bac+3 sur l'estimation et les tests.

2 Modèle

On note Y la variable aléatoire réelle à expliquer et X la variable explicative (déterministe) ou effet fixe ou facteur contrôlé. Le modèle revient à supposer, qu'en moyenne, $E(Y)$, est une fonction affine de X.

$$E(Y) = f(X) = \beta_0 + \beta_1 X.$$

Remarque : Nous supposerons pour simplifier que X est déterministe. Dans le cas contraire, X aléatoire, le modèle s'écrit alors conditionnellement aux observations de X : $E(Y|X = x) = \beta_0 + \beta_1 x$ et conduit aux mêmes estimations.

Pour une séquence d'observations aléatoires identiquement distribuées $\{(y_i, x_i)i = 1, \ldots, n\}$ ($n > 2$, et les x_i non tous égaux), le modèle s'écrit avec les observations :

$$y_i = \beta_0 + \beta_1 x_i + u_i \quad i = 1, \ldots, n$$

ou sous la forme matricielle :

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} u_1 \\ \vdots \\ u_n \end{bmatrix},$$
$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

où le vecteur $\mathbf{u}$ contient les erreurs.

Les *hypothèses* relatives à ce modèle sont les suivantes :

1. la distribution de l'erreur $\mathbf{u}$ est indépendante de X **ou** X est fixe,
2. l'erreur est centrée et de variance constante (homoscédasticité) :
$$\forall i = 1, \ldots, n \quad E(u_i) = 0, \quad \text{Var}(u_i) = \sigma_u^2.$$
3. β_0 et β_1 sont constants, pas de rupture du modèle.
4. Hypothèse complémentaire pour les inférences : $\mathbf{u} \sim \mathcal{N}(0, \sigma_u^2 \mathbf{I}_p)$.

3 Estimation

L'estimation des paramètres $\beta_0, \beta_1, \sigma^2$ est obtenue en maximisant la vraisemblance, sous l'hypothèse que les erreurs sont gaussiennes, ou encore par minimisation de la somme des carrés des écarts entre observations et modèle (moindres carrés). Pour un jeu de données $\{(x_i, y_i)i = 1 \ldots, n\}$, le critère des moindres carrés s'écrit :

$$\min_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

On pose :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i, \qquad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i,$$
$$s_x^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2, \qquad s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2,$$
$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \qquad r = \frac{s_{xy}}{s_x s_y};$$

Les moindres carrés sont obtenus par :

$$\begin{aligned} b_1 &= \frac{s_{xy}}{s_x^2}, \\ b_0 &= \bar{y} - b_1\bar{x}. \end{aligned}$$

On montre que ce sont des estimateurs sans biais et de variance minimum parmi les estimateurs fonctions linéaires des y_i (resp. parmi tous les estimateurs dans le cas gaussien). À chaque valeur de X correspond la valeur *estimée* (ou prédite, ajustée) de Y :

$$\widehat{y_i} = b_0 + b_1 x_i,$$

les *résidus* calculés ou estimés sont :

$$e_i = y_i - \widehat{y_i}.$$

La variance σ_u^2 est estimée par la variation résiduelle :

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2.$$

3.1 Inférence

Les estimateurs b_0 et b_1 sont des variables aléatoires réelles de matrice de covariance :

$$\sigma_u^2 \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2} & -\frac{\bar{x}}{(n-1)s_x^2} \\ -\frac{\bar{x}}{(n-1)s_x^2} & \frac{1}{(n-1)s_x^2} \end{bmatrix}$$

qui est estimée en remplaçant σ_u^2 par son estimation s^2. Sous l'hypothèse que les résidus sont gaussiens, on montre que

$$\frac{(n-2)s^2}{\sigma_u^2} \sim \chi^2_{(n-2)}$$

et donc que les statistiques

$$(b_0-\beta_0) \Big/ s\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)^{1/2} \quad \text{et} \quad (b_1-\beta_1) \Big/ s\left(\frac{1}{(n-1)s_x^2}\right)^{1/2}$$

suivent des lois de Student à $(n-2)$ degrés de liberté. Ceci permet de tester l'hypothèse de nullité d'un de ces paramètres ainsi que de construire les intervalles de confiance :

$$\begin{aligned} b_0 &\pm t_{\alpha/2;(n-2)}\, s\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)^{1/2}, \\ b_1 &\pm t_{\alpha/2;(n-2)}\, s\left(\frac{1}{(n-1)s_x^2}\right)^{1/2}. \end{aligned}$$

Attention : une inférence conjointe sur β_0 et β_1 ne peut être obtenue en considérant séparément les intervalles de confiance. La région de confiance est en effet une ellipse d'équation :

$$n(b_0-\beta_0)^2 + 2(b_0-\beta_0)(b_1-\beta_1)\sum_{i=1}^{n} x_i + (b_1-\beta_1)^2 \sum_{i=1}^{n} x_i^2 = 2s^2\mathcal{F}_{\alpha;2,(n-2)}$$

qui est inclue dans le rectangle défini par les intervalles. Un grande part des valeurs du couple (β_0, β_1) est donc exclue de la région de confiance et ce d'autant plus que b_0 et b_1 sont corrélés.

4 Qualité d'ajustement, prédiction

Il est d'usage de décomposer les sommes de carrés des écarts à la moyenne sous la forme ci-dessous ; les notations sont celles de la plupart des logiciels :

Total sum of squares	SST	$= (n-1)s_y^2$,
Regression sum of squares	SSR	$= (n-1)\frac{s_{xy}^2}{s_x^2}$,
Error sum of squares	SSE	$= (n-2)s^2$,

et on vérifie : SST = SSR + SSE.

On appelle *coefficient de détermination* la quantité

$$R^2 = r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = 1 - \frac{n-2}{n-1}\frac{s^2}{s_y^2} = \frac{\text{SSR}}{\text{SST}}$$

qui exprime le rapport entre la variance expliquée par le modèle et la variance totale.

Sous l'hypothèse : $\beta_1 = 0$, la statistique

$$(n-2)\frac{R^2}{1-R^2} = (n-2)\frac{\text{SSR}}{\text{SSE}}$$

suit une distribution de Fisher $\mathcal{F}_{1,(n-2)}$. Cette statistique est le carré de la statistique de Student correspondant à la même hypothèse.

Connaissant une valeur x_0, on définit deux *intervalles de confiance de prédiction* à partir de la valeur prédite $\widehat{y_0} = b_0 + b_1 x_0$. Le premier encadre $E(Y)$ sachant $X = x_0$; le deuxième, qui encadre $\widehat{y_0}$ est plus grand car il tient compte de la variance totale : $\sigma_u^2 + \text{Var}(\widehat{y_0})$:

$$\widehat{y_0} \pm t_{\alpha/2;(n-2)} s \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2} \right)^{1/2},$$

$$\widehat{y_0} \pm t_{\alpha/2;(n-2)} s \left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2} \right)^{1/2}.$$

Les logiciels proposent également une *bande de confiance* entre deux arcs d'hyperboles pour la droite de régression. À chaque point (b_0, b_1) de l'ellipse de confiance de (β_0, β_1) correspond une droite d'équation $\widehat{y} = b_0 + b_1 x$. Toutes ces droites sont comprises entre les bornes :

$$\widehat{y} \pm s\sqrt{\mathcal{F}_{1,(n-2)}} \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{(n-1)s_x^2} \right)^{1/2}.$$

Ceci signifie que cette bande recouvre la "vraie" ligne avec une probabilité $1 - \alpha$. Elle est plus grande que celle associée aux intervalles de confiance des $E(Y)$.

Attention : la prédiction par intervalle n'est justifiée que pour des observations appartenant à la population échantillonnée et à condition que les hypothèses : linéarité, erreurs i.i.d., (normalité), soient valides. Éviter les extrapolations.

5 Nuage de points, transformations

Toute tentative de modélisation nécessite une étude descriptive préalable afin de s'assurer, au moins graphiquement, de la validité des hypothèses considérées. Ceci passe

1. par une étude uni-variée de chaque distribution pour détecter des dissymétries ou encore des valeurs atypiques (outliers) : boîtes à moustaches, histogrammes, estimation non-paramétrique de la densité,
2. puis par une représentation du nuage de points dans le repère (X, Y) et une régression non-paramétrique afin de déceler une éventuelle liaison non-linéaire entre les variables. *Attention*, même si elle est forte, une liaison non-linéaire, par exemple de type quadratique entre X et Y, peut conduire néanmoins a un coefficient de corrélation linéaire très faible.

Dans les deux cas, en cas de problèmes, le remède consiste souvent à rechercher des transformations des variables permettant de rendre les distributions symétriques, de "banaliser" les points atypiques et de rendre linéaire la relation. La qualité de l'estimation d'une distribution par un histogramme dépend beaucoup du découpage en classe. Malheureusement, plutôt que de fournir des classes d'effectifs égaux et donc de mieux répartir l'imprécision, les logiciels utilisent des classes d'amplitudes égales et tracent donc des histogrammes parfois peu représentatifs. Ces 20 dernières années, à la suite du développement des moyens de calcul, sont apparues des méthodes d'estimation dites *fonctionnelles* ou *non-paramétriques* qui proposent d'estimer la distribution d'une variable ou la relation entre deux variables par une fonction construite point par point (noyaux) ou dans une base de fonctions *splines*. Ces estimations sont simples à calculer (pour l'ordinateur) mais nécessitent le choix d'un paramètre dit de *lissage*. Les démonstrations du caractère optimal de ces estimations fonctionnelles, liée à l'optimalité du choix de la valeur du paramètre de lissage, font appel à des outils théoriques plus sophistiquées sortant du cadre de ce cours (Eubank 1988, Silverman 1986).

Nous résumons ci-dessous les techniques non-paramétriques, simples et efficaces dans ce genre de situation, trop rarement enseignées dans un cours de statistique descriptive, mais déjà présentes dans certains logiciels (SAS/INSIGHT).

5.1 Estimation de la densité

L'estimation de la densité par la méthode du noyau se met sous la forme générale :

$$\widehat{g}_\lambda(x) = \frac{1}{n\lambda} \sum_{i=1}^{n} K\left(\frac{x - x_i}{\lambda} \right)$$

où λ est le paramètre de lissage optimisée par une procédure automatique qui minimise une approximation de l'erreur quadratique moyenne intégrée (MISE : norme dans l'espace L^2) ; K est une fonction symétrique, positive, concave, appelée *noyau* dont la forme précise importe peu. C'est souvent la fonction densité de la loi gaussienne :

$$K(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$$

qui possède de bonnes propriétés de régularité. Le principe consiste simplement à associer à chaque observation un "élément de densité" de la forme du noyau K et à sommer tous ces éléments. Un histogramme est une version particulière d'estimation dans laquelle l'"élément de densité" est un "petit rectangle" dans la classe de l'observation.

5.2 Régression non-paramétrique

On considère un modèle de régression de la forme

$$y_i = f(x_i) + \varepsilon_i$$

où les erreurs sont centrées et la fonction f est supposée régulière : existence de dérivées jusqu'à un certain ordre. Dans ce contexte, de nombreux estimateurs de f ont été proposés. Ils conduisent souvent à des résultats assez voisins, le point le plus sensible étant le choix de λ.

Spline

Le *lissage spline* élémentaire consiste à rechercher, dans l'espace des fonctions continûment différentiables et avec une dérivée seconde de carré intégrable, le minimum d'un critère combinant ajustement des observations et régularité de la solution :

$$\widehat{f_\lambda} = \arg\min_f \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_{-\infty}^{+\infty} (f''(x))^2 dx.$$

On montre que l'on obtient une fonction polynômiale (de degré 3) par morceaux. La valeur optimale du paramètre de lissage est fixée par validation croisée généralisée (GCV).

Noyau

La régression non-paramétrique par la méthode du noyau consiste à calculer une moyenne pondérée autour de chaque observation. La pondération est fixée par une fonction K du même type que celle utilisée pour l'estimation de la densité.

$$\widehat{f_\lambda}(x) = \sum_{i=1}^{n} \frac{K\left(\frac{x-x_i}{\lambda}\right) x_i}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{\lambda}\right)}.$$

Loess

L'estimateur précédent est susceptible de biais même dans le cas simple de points alignés. Une adaptation propose de calculer, plutôt qu'une moyenne locale pondérée, une régression linéaire ou même quadratique locale. On parle alors de lisseur polynômial local.

5.2.1 transformations

Dans le cas où des problèmes (distribution, non-linéarité) ont été identifiés, l'étape suivante consiste à rechercher des transformations élémentaires (logarithme, puissance) des variables susceptibles de les résoudre. Ceci amène à étudier les modèles des exemples suivants :

$$\begin{aligned}
Y &= \beta_0 + \beta_1 \ln X \\
\ln Y &= \beta_0 + \beta_1 X \quad \text{ou} \quad Y = ab^X \quad \text{avec} \quad \beta_0 = \ln a \text{ et } \beta_1 = \ln b \\
\ln Y &= \beta_0 + \beta_1 \ln X \quad \text{ou} \quad Y = aX^{\beta_1} \quad \text{avec} \quad \beta_0 = \ln a \\
Y &= \beta_0 + \beta_1 (1/X) \\
Y &= \beta_0 + \beta_1 X^{1/2} \\
Y &= \beta_0 + \beta_1 X^2 \quad \text{ou, plus généralement,} \\
Y &= \beta_0 + \beta_1 X^\alpha \\
&\dots
\end{aligned}$$

6 Influence

Le critère des moindres carrés, comme la vraisemblance appliquée à une distribution gaussienne douteuse, est très sensible à des observations atypiques,

hors "norme" (outliers) c'est-à-dire qui présentent des valeurs trop singulières. L'étude descriptive initiale permet sans doute déjà d'en repérer mais c'est insuffisant. Un diagnostic doit être établi dans le cadre spécifique du modèle recherché afin d'identifier les observations *influentes* c'est-à-dire celles dont une faible variation du couple (x_i, y_i) induisent une modification importante des caractéristiques du modèle.

Ces observations repérées, il n'y a pas de remède universel : supprimer un valeur aberrante, corriger une erreur de mesure, construire une estimation robuste (en norme L_1), ne rien faire..., cela dépend du contexte et doit être négocié avec le commanditaire de l'étude.

6.1 Effet levier

Une première indication est donnée par l'éloignement de x_i par rapport à la moyenne $\bar{x}$. En effet, écrivons les prédicteurs $\widehat{y_i}$ comme combinaisons linéaires des observations (cf. exo 3) :

$$\widehat{y_i} = b_0 + b_1 x_i = \sum_{j=1}^{n} h_{ij} y_j \quad \text{avec} \quad h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{j=1}^{n}(x_j - \bar{x})^2} \; ;$$

en notant $\mathbf{H}$ la matrice (hat matrix) des h_{ij} ceci s'exprime encore matriciellement :

$$\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}.$$

Les éléments diagonaux h_{ii} de cette matrice mesurent ainsi l'impact ou l'importance du rôle que joue y_i dans l'estimation de $\widehat{y_i}$.

6.2 Résidus

Différents types de résidus sont définis afin d'affiner leurs propriétés.

Résidus : $e_i = y_i - \widehat{y_i}$

Résidus$_i$: $e_{(i)i} = y_i - \widehat{y_{(i)i}} = \frac{e_i}{1-h_{ii}}$
où $\widehat{y_{(i)i}}$ est la prévision de y_i calculée sans la $i$ème observation (x_i, y_i). On note

$$\text{PRESS} = \sum_{i=1}^{n} e_{(i)i}^2 \quad \text{(predicted residual sum of squares)}$$

la somme des carrés de ces résidus.

Résidus standardisés : Même si l'hypothèse d'homoscédasticité est vérifiée, ceux-ci n'ont pas la même variance : $E(e_i) = 0$ et $\text{Var}(e_i) = \sigma_u^2(1-h_{ii})$. Il est donc d'usage d'en calculer des versions *standardisées* afin de les rendre comparables :

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}.$$

Résidus studentisés : La standardisation ("interne") dépend de e_i dans le calcul de s estimation de $\text{Var}(e_i)$. Une estimation non biaisée de cette variance est basée sur

$$s_{(i)}^2 = \left[(n-2)s^2 - \frac{e_i^2}{1-h_{ii}}\right]/(n-3)$$

qui ne tient pas compte de la $i$ème observation. On définit alors les résidus *studentisés* par :

$$t_i = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}.$$

Sous hypothèse de normalité, on montre que ces résidus suivent une loi de Student à $(n-3)$ degrés de liberté.

Il est ainsi possible de construire un test afin tester la présence d'*une* observation atypique ou de plusieurs en utilisant l'inégalité de Bonferroni. Plus concrètement, en pratique, les résidus studentisés sont comparés aux bornes ± 2.

6.3 Diagnostics

Les deux critères précédents contribuent à déceler des observations potentiellement influentes par leur éloignement à $\bar{x}$ ou la taille des résidus. Ces informations sont synthétisées dans des critères évaluant directement l'influence d'une observation sur certains paramètres : les prédictions $\widehat{y_i}$, les paramètres b_0, b_1, le déterminant de la matrice de covariance des estimateurs. Tous ces indicateurs proposent de comparer un paramètre estimé sans la $i$ème observation et ce même paramètre estimé avec toutes les observations.

Le plus couramment utilisé est la distance de Cook :

$$D_i = \frac{\sum_{j=1}^{n}(\widehat{y_{(i)j}} - \widehat{y_j})^2}{2s^2} = \frac{h_{ii}}{2(1-h_{ii})} r_i^2 \qquad \text{pour} \quad i = 1, \ldots, n$$

qui mesure donc l'influence d'une observation sur l'ensemble des prévisions en prenant en compte effet levier et importance des résidus.

La stratégie de détection consiste le plus souvent à repérer les points atypiques en comparant les distances de Cook avec la valeur 1 puis à expliquer cette influence en considérant, pour ces observations, leur résidu ainsi que leur effet levier.

7 Graphe des résidus

Le nuage des points (x_i, y_i) assorti d'un lissage permet de détecter une éventuelle relation non-linéaire entre les variables. D'autres hypothèses doivent être validées :

- l'homoscédasticité par un graphique des résidus studentisés ou non : (x_i, t_i) afin de repérer des formes suspectes de ce nuage qui devrait se répartir uniformément de part et d'autre de l'axe des abscisses,
- éventuellement la normalité des résidus en étudiant leur distribution,
- l'autocorrélation des résidus dans le cas, par exemple, où la variable explicative est le temps.

Une transformation des variables ou une modélisation spécifique à une série chronologique (SARIMA) permet, dans les situations favorables, de résoudre les difficultés évoquées.

8 Exemple

Pour 47 immeubles d'appartements locatifs d'une grande ville américaine, les données (Jobson, 1991) fournissent le "revenu net" en fonction du "nombre d'appartements". Les tableaux ci-dessous sont des extraits des résultats fournis par la procédure `reg` du module SAS/STAT. Cette procédure génère beaucoup d'autres résultats comme les matrices $\mathbf{X'X}$ (crossproducts), $\mathbf{X'DX}$ (model crossproducts) et son inverse, matrices des variances et corrélations des estimateurs.

```
proc reg data=sasuser.suitinco all;
model revenu=nbappart /dw Influence cli clm;
output out=hubout h=lev p=pred r=res student=resstu ;
run;
```

```
Descriptive Statistics
Variables     Sum         Mean      Uncorrected SS    Variance    Std Deviation
INTERCEP       47            1              47            0              0
NBAPPART     1942    41.319148936         157970    1689.7437558   41.106492866
REVENU    4336086    92257.148936    947699637616    11905754472    109113.49354
Correlation : 0.8856
```

```
Analysis of Variance
                          Sum of          Mean
Source          DF       Squares        Square          F Value      Prob>F
                (1)
Model            1 429511948724 (2) 429511948724 (5)   163.585 (7)  0.0001 (8)
Error           45 118152756990 (3)  2625616822  (6)
C Total         46 547664705714 (4)

   Root MSE   51240.77304 (9)    R-square       0.7843   (12)
   Dep Mean   92257.14894 (10)   Adj R-sq       0.7795
   C.V.          55.54125 (11)
```

(1)	degrés de liberté de la loi de Fisher du test global ($H_0 : \beta_1 = 0$)
(2)	SSR
(3)	SSE ou déviance
(4)	SST=SSE+SSR
(5)	SSR/DF
(6)	s^2 =MSE=SSE/DF est l'estimation de σ_u^2
(7)	Statistique F du test de Fisher du modèle global
(8)	$P(f_{p;n-p-1} > F)$; H_0 est rejetée au niveau α si $P < \alpha$
(9)	s =racine de MSE
(10)	moyenne empirique de la variable à expliquée
(11)	Coefficient de variation $100\times$ (9)/(10) sans échelle ni dimension
(12)	Coefficient de détermination R^2 ou carré du coefficient de corrélation.

```
Parameter Estimates
                        Parameter       Standard     T for H0:
Variable   DF            Estimate          Error    Parameter=0    Prob > |T|
                             (1)            (2)          (3)           (4)
INTERCEP    1     -4872.015285    10655.276212       -0.457        0.6497
NBAPPART    1      2350.705828     183.79188506       12.790        0.0001
```

(1)	estimations des paramètres (b_j)
(2)	écarts-types de ces estimations (s_{b_j})
(3)	statistique T du test de Student de $H_0 : b_j = 0$ $((b_j - 0)/s_{b_j})$
(4)	$P(t_{n-p-1} > T)$; H_0 est rejetée au niveau α si $P < \alpha$

Connaissant les fractiles de la loi de Student : $t_{0,975;45} = 2,015$, on construit facilement des intervalles de confiance des estimateurs, ici au niveau 5% : $[b_j - t_{0,975;n-2}s_{b_j}; b_j + t_{0,975;n-2}s_{b_j}]$.

```
   Dep Var Predict Std Err Lower95 Upper95 Lower95 Upper95          Std Err Student
Obs REVENU Value   Predict   Mean    Mean   Predict Predict  Resid. Resid.  Resid.
      (1)   (2)      (3)     (4)     (5)      (6)     (7)     (8)    (9)     (10)
```

```
 1  119202 131469  8078.5  115198  147740  26989.9 235948 -12266.9 50599.9 -0.242
...
23  345608 239601  13732.  211943  267260   132755 346448   106007 49366.3  2.147
24  350633 324227  19616.  284717  363736   213718 434735  26406.2 47337.2  0.558
25  226375 98559.  7490.4  83472.  113646  -5742.0 202860   127816 50690.3  2.522
26  247203 178483  10065.  158210  198756  73306.5 283660  68720.0 50242.4  1.368
27  28519. 157327  9041.4  139116  175537  52528.2 262125  -128808 50436.7  -2.55
28  154278 347734  21327.  304779  390689   235947 459521  -193456 46591.4  -4.15
29  157332 140872  8385.2  123983  157760  36294.8 245449  16460.3 50550.0  0.326
30  171305 197289  11104.  174924  219653  91689.0 302888 -25983.7 50023.1  -0.52
...
                    Cook's          Hat Diag  Cov            INTERCEP NBAPPART
Obs     -2-1-0 1 2     D  Rstudent      H  Ratio    Dffits Dfbetas  Dfbetas
          (11)       (12)    (13)     (14)    (15)     (15)    (15)       (15)
 1 |       |       | 0.001 -0.2399  0.0249  1.0698 -0.0383 -0.0145 -0.0145
...
23 |       |****   | 0.178  2.2413  0.0718  0.9078  0.6235 -0.1347  0.5230
24 |       |*      | 0.027  0.5535  0.1466  1.2087  0.2294 -0.0898  0.2121
25 |       |***** | 0.069  2.6906  0.0214  0.7881  0.3976  0.2597  0.0262
26 |       |**     | 0.038  1.3815  0.0386  0.9994  0.2768  0.0120  0.1854
27 | *****|       | 0.105 -2.7310  0.0311  0.7893 -0.4896 -0.0876 -0.2755
28 |******|       | 1.806 -5.2275  0.1732  0.4814 -2.3929  1.0090 -2.2411
29 |       |       | 0.001  0.3224  0.0268  1.0697  0.0535  0.0162  0.0242
30 |      *|       | 0.007 -0.5152  0.0470  1.0844 -0.1144  0.0063 -0.0846
...
```

(1)	variable à expliquer y_i
(2)	valeur ajustée $\widehat{y_i}$
(3)	écart-type de cette estimation$s_{\widehat{y_i}}$
(4)et (5)	Intervalle de confiance pour l'estimation de $E(y_i)$
(6) et (7)	Intervalle de confiance pour l'estimation de y_i
(8)	résidus calculés $e_i = y_i - \widehat{y_i}$
(9)	écarts-types de ces estimations
(10)	résidus standardisés (ou studentisés internes) r_i
(11)	repérage graphique des résidus standardisés : $* = 0.5$.
(12)	Distance de Cook
(13)	résidus studentisés (externes) t_i
(14)	Termes diagonaux de la matrice chapeau $\mathbf{H}$
(15)	autres indicateurs d'influence

Les observations 28 et 16 seraient à inspecter avec attention. Certaines, dont la 28, présentent une valeur observée hors de l'intervalle de prédiction.

Le graphique des résidus sont présentés dans la figure 1. Il montre clairement que l'hypothèse d'homoscédasticité n'est pas satisfaite. Une autre modélisation faisant intervenir une transformation des variables serait nécessaire. Ainsi la modélisation du logarithme du revenu en fonction du logarithme du nombre d'appartements représentée par la figure 2 est nettement plus satisfaisante. Une étude descriptive préalable des distributions aurait permis de conduire à ce choix.

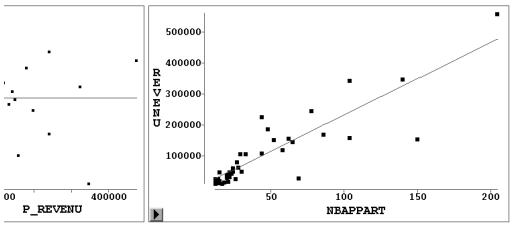

FIGURE 1 – Graphe des résidus et nuage de points de la régression du revenu en fonction du nombre d'appartements.

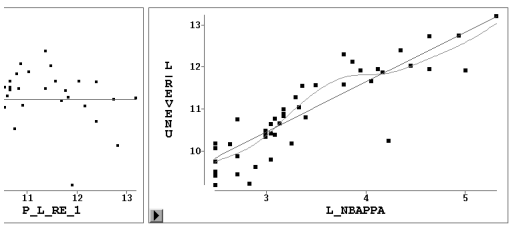

FIGURE 2 – Graphe des résidus et nuage de points de la régression (linéaire et non paramétrique) du logarithme du revenu en fonction du logarithme du nombre d'appartements.