Régression linéaire simple

Résumé

Ce chapitre introduit la notion de modèle linéaire par la version la plus élémentaire : expliquer Y par une fonction affine de X. Après avoir expliciter les hypothèses nécessaires et les termes du modèle, les notions d'estimation des paramètres du modèle, de prévision par intervalle de confiance, la signification des tests d'hypothèse sont discutées. Enfin une attention particulière est faite aux outils de diagnostics disponibles : valeurs influentes, et surtout graphe des résidus.

Retour au plan du cours.

1 Introduction

Ce chapitre est une introduction à la modélisation linéaire par le modèle le plus élémentaire, la régression linéaire simple où une variable X est expliquée, modélisée par une fonction affine d'une autre variable y. La finalité d'un tel modèle est multiple et dépend donc du contexte et surtout des questions sous-jacentes. Ce peut-être juste une approche exploratoire ou alors la recherche d'une réponse à une question du type : une variable quantitative X (e.g. la concentration d'une molécule) a-t-elle une influence sur la variable quantitative Y (e.g. une culture bactérienne) ? Ou enfin la recherche d'un modèle de prévision de Y en fonction de X : calibration d'un appareil de mesure d'une concentration à partir d'une mesure optique. Des concepts clefs : modèle, estimations, tests, diagnostics sont introduits et déclinés dans ce contexte élémentaire. Leur emploi et leur signification dépendent des objectifs. Ils se retrouvent dans une présentation plus général du modèle de régression multiple et ce chapitre sert donc d'introduction.

Avant tout travail de modélisation, une approche descriptive ou exploratoire est nécessaire pour dépister au plus tôt des difficultés dans les données : dissymétrie des distributions, valeurs atypiques, liaison non linéaire entre les variables. En fonction des résultats obtenus, une transformation préalable des variables peut s'avérer nécessaire. Dans l'exemple de la figure 1, le choix d'une variable explicative homogène à un volume semble plus judicieux pour estimer le poids d'un arbre.

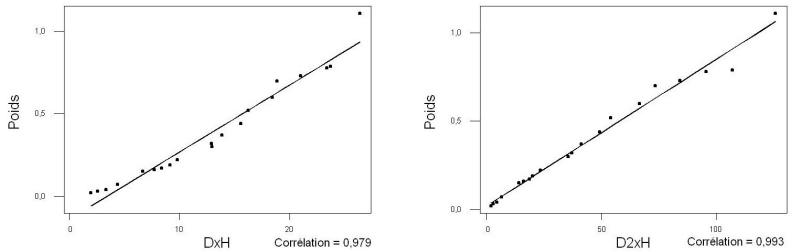

FIGURE 1 – *Exemple de régression du poids d'un arbre en fonction de la variable diamètre × hauteur et diamètre × hauteur au carré*

2 Modèle

On note Y la variable aléatoire réelle à expliquer (variable endogène, dépendante ou réponse) et X la variable explicative ou effet fixe (exogène). Le modèle revient à supposer, qu'en moyenne, $E(Y)$, est une fonction affine de X. L'écriture du modèle suppose implicitement une notion préalable de *causalité* dans le sens où Y dépend de X car le modèle n'est pas symétrique.

$$E(Y) = f(X) = \beta_0 + \beta_1 X \quad \text{ou} \quad Y = \beta_0 + \beta_1 X + \varepsilon$$

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.

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

1. la distribution de l'erreur ε est indépendante de X **ou** X est fixe,
2. l'erreur est centrée et de variance constante (homoscédasticité) :

$$\forall i = 1, \ldots, n \quad E(\varepsilon_i) = 0, \quad \text{Var}(\varepsilon_i) = \sigma^2.$$

3. β_0 et β_1 sont constants, pas de rupture du modèle.

4. Hypothèse complémentaire pour les inférences : $\varepsilon \sim \mathcal{N}(0, \sigma^2)$.

3 Estimation

3.1 Paramètres

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). Les deux approches conduisent aux mêmes estimation tandis que le maximum de vraisemblance induit de meilleure propriétés des estimateurs. Pour une séquence d'observations $\{(x_i, y_i) i = 1 \ldots, n\}$, 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 minimisés par :

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

qui sont les réalisations des estimateurs $\widehat{\beta_0}$ et $\widehat{\beta_1}$. On montre que ces 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 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^2 est estimée par la variation résiduelle :

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

```
Exemple : Analyse de régression : Poids en fonction de D2xH

L'équation de régression est
Poids = 0,0200 + 0,00829 D2xH

Régresseur       Coef       Er-T coef       T        P
Constante     0,01999(1)    0,01365(3)     1,46     0,160
D2xH       0,0082897(2) 0,0002390(4)    34,68     0,000
```

(1)	b_0
(2)	b_1
(3)	écart-type de $\widehat{\beta_0}$: s_{b_0}
(4)	écart-type de $\widehat{\beta_1}$: s_{b_1}

3.2 Qualité d'ajustement

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 :

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

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.

```
Exemple : Analyse de régression : Poids en fonction de D2xH

Analyse de variance
Source            DL         SC          CM          F         P
Régression         1(1)   1,8108(2)   1,8108(5)  1202,89    0,000
Erreur résid      18      0,0271(3)   0,0015(6)
Total             19      1,8379(4)

S = 0,03880(7)    R-carré = 98,5%(8)  R-carré (ajust) = 98,4%
```

(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_ε^2
(7)	s =racine de MSE
(8)	Coefficient de détermination R^2 ou carré du coefficient de corrélation.

4 Inférence

4.1 Loi des paramètres

Les estimateurs $\widehat{\beta_0}$ et $\widehat{\beta_1}$ sont des variables aléatoires réelles de matrice de covariance :

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

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

$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{(n-2)}$$

et donc que les statistiques

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

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 :

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

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.

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.

4.2 Prévision par intervalle de confiance

Connaissant une valeur x_0, on définit deux *intervalles de confiance de prévision* 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 : $\sigma^2 + \text{Var}(\widehat{y_0})$:

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

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 prévision 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), homoscédasticité, soient valides. Éviter les extrapolations.

4.3 Tests d'hypothèse

Les tests précédents prennent une signification particulière avec un objectif "explicatif" ; α désigne le niveau des tests, souvent $\alpha = 5\%$. Comme pour tous les tests usuels de comparaison d'échantillon, les logiciels fournissent les probabilités critiques ou P-valeurs qui, en pratique, sont comparées avec le seuil prédéterminé.

Le test de Fisher s'intéresse à la significativité globale d'un modèle. Dans le cas de la régression simple, seul le paramètre β_1 est concerné :

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

suit une loi de Fisher à $(1, n-2)$ degrés de liberté. L'hypothèse $H_0 : \beta_1 = 0$, est rejetée si $F > f_{1;n-2;1-\alpha/2}$ ou si la P-valeur associée est inférieure à α.

Plus précisément, l'hypothèse $H_0 : \beta_1 = 0$ répond aussi à la question de l'influence de X sur Y. La réponse est négative si H_0 est acceptée : la pente de la droite de régression est nulle, le nuage de point est réparti sans structure linéaire significative. La réponse est positive lorsque le test est significatif et donc l'hypothèse rejetée. Ce paramètre suit une loi de Student et H_0 rejetée lorsque $t_1 = \frac{|b_1|}{s_{b_1}} > t_{n-2;1-\alpha/2}$ ou si la P-valeur associée est inférieure à α. Ce test est strictement équivalent au test de Fisher précédent, il conduit à la même P-valeur.

Enfin, le test de l'hypothèse $H_0 : \beta_0 = 0$ qui signifie : "la droite passe par l'origine", a un intérêt limité à des situations très particulières comme la calibration du "zéro" d'un appareil de mesure. Elle est rejetée si $t_0 = \frac{|b_0|}{s_{b_0}} > t_{n-2;1-\alpha/2}$

5 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.

5.1 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 :

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

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 $\widehat{y}_i$.

5.2 Résidus et PRESS

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). Ce type de résidu conduit à la définition du PRESS (*predicted residual sum of squares*) dit de Allen :

$$\text{PRESS} = \frac{1}{n}\sum_{i=1}^{n} e_{(i)i}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2$$

C'est une estimation sans biais de la qualité de prévision d'un modèle car une même observation n'est pas utilisée, à la fois, pour estimer le modèle et l'erreur de prévision. Le PRESS est très utile pour comparer les qualités prédictives de plusieurs modèles. Ce point important sera développé dans le cas du modèle linéaire multiple : le coefficient R^2 permet de comparer les qualités d'ajustement mais la meilleure prévision n'est pas nécessairement fournie par un modèle de R^2 maximum. Le PRESS encore appelé *leave one out cross validation (loo CV)* est plus pertinent pour atteindre cet objectif. *Remarquer* que dans le cas particulier du modèle linéaire, le PRESS est calculé directement à partir des résidus initiaux et des termes diagonaux h_{ii} de la matrice $\mathbf{H}$. Pour d'autres modèles, le calcul du PRESS nécessite l'estimation, éventuellement coûteuse, de n modèles.

Résidus standardisés : Même si l'hypothèse d'homoscédasticité est vérifiée, ceux-ci n'ont pas la même variance : $E(e_i) = 0$ et $\text{Var}(e_i) = \sigma^2(1-h_{ii})$. 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 Diagnostics

6.1 Distance de Cook

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 prévisions $\widehat{y}_i$, 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.

6.2 Graphe des résidus

Attention : la présentation "pédagogique" des concepts de la régression linéaire ne doit pas faire négliger l'étape de diagnostic des résidus. Concrètement, le graphe des résidus est la première chose à consulter après l'estimation d'un modèle linéaire. L'appréciation de sa forme, même si celle-ci reste "subjective", renseigne précisément sur la validité des hypothèses implicites

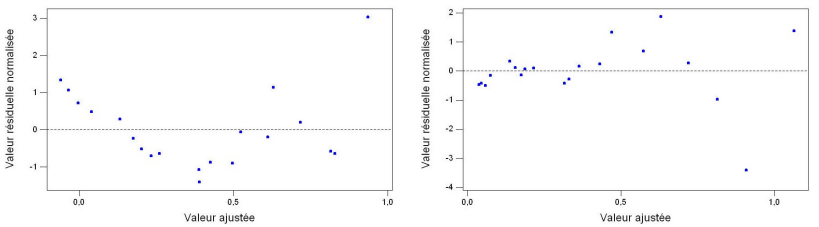

FIGURE 2 – *Les résidus (à gauche) de la régression du poids en fonction du produit (diamètre × hauteur) montre clairement un problème de linéarité. La transformation de la variable diamètre (carré) améliore ce diagnostic mais soulève (à droite) une problème d'hétéroscédasticité*

du modèle dont surtout celle de linéarité et celle d'homoscédasticité. Dans le cas contraire, toutes les décisions issues de tests et les intervalles de confiances n'ont plus de légitimité. Si certaines des hypothèses ne sont pas vérifiées, des mesures s'imposent comme la recherche de transformation des variables.

L'homoscédasticité et la linéarité du modèle sont évalués par un graphique des résidus studentisés ou non : (x_i, t_i) qui doit se disperser "normalement" de part et d'autre de l'axe $y = 0$: symétriquement et sans forme particulière. Des formes d'"entonnoir", ou de "diabolo" du nuage font suspecter une hétéroscédasticité des résidus, celle d'une "banane" indique une possible relation non linéaire entre Y et X.

Même si cette hypothèse est moins sensible, le modèle est robuste surtout en cas de grand échantillon, il est sage de vérifier la normalité des résidus en étudiant leur distribution par exemple par une simple droite de Henri.

Enfin l'auto-corrélation des résidus dans le cas par exemple où la variable explicative est le temps pose également des problèmes. Une modélisation de type série chronologique (ARMA, SARIMA) des résidus serait à tester.

7 Exemples

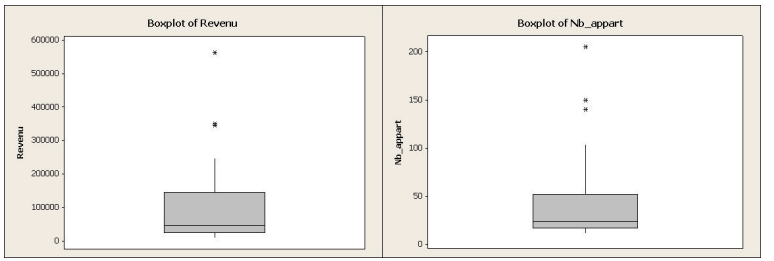

FIGURE 3 – Distribution des variables revenus et nombre d'appartements

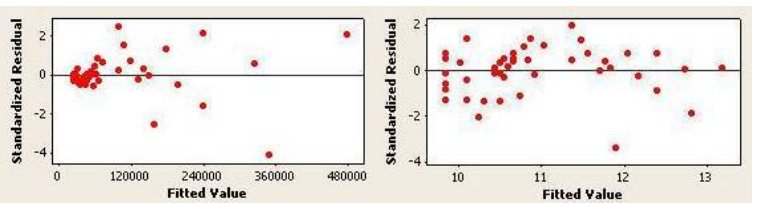

FIGURE 4 – *Les résidus de la régression du revenu sur le nombre d'appartements (à gauche) met nettement en évidence un problème d'hétéroscédasticité ; problème résolu (à droite) par des transformations des variables.*

7.1 Revenu fonction du nombre d'appartements

La variable Y est le revenu d'un immeuble exprimé en fonction de la variable x, nombre d'appartement ; 47 observations sont disponibles. L'erreur naïve consiste à se précipiter sur le premier modèle venu. Les résultats numériques ci-dessous sont satisfaisants, le modèle est significatif avec une qualité correcte d'ajustement (R^2 proche de 0,8).

Mais le graphique des résidus (figure 4) est nettement moins sympathique. Le statisticien amateur est allé trop vite, il a sauté l'étape descriptive des variables. Les diagrammes boites (figure 3)montrent des distributions très dissymétriques, une transformation par la fonction logarithme dégrade certes un peu l'ajustement mais améliore considérable la dispersion des résidus. Attention, le R^2 n'est surtout pas le premier critère à regarder pour comparer des modèles.

```
Regression Analysis: Revenu versus Nb_appart

The regression equation is
Revenu = - 4872 + 2351 Nb_appart

Predictor     Coef  SE Coef      T      P
Constant    -4872    10655  -0,46  0,650
Nb_appart  2350,7    183,8  12,79  0,000

S = 51240,8   R-Sq = 78,4%   R-Sq(adj) = 77,9%

Analysis of Variance

Source          DF           SS           MS       F      P
Regression       1  4,29512E+11  4,29512E+11  163,59  0,000
Residual Error  45  1,18153E+11   2625616822
Total           46  5,47665E+11


Regression Analysis: LRevenu versus LNb_appart

The regression equation is
LRevenu = 6,87 + 1,19 LNb_appart

Predictor      Coef  SE Coef      T      P
Constant     6,8678   0,3332  20,61  0,000
LNb_appart  1,18863  0,09593  12,39  0,000

S = 0,496742   R-Sq = 77,3%   R-Sq(adj) = 76,8%

Analysis of Variance

Source          DF      SS      MS       F      P
Regression       1  37,886  37,886  153,54  0,000
Residual Error  45  11,104   0,247
Total           46  48,990
```

7.2 Étalonnage d'un appareil de mesure

Il s'agit de tester le bon calibrage d'un spectromètre dans le proche infra-rouge (SPIR) évaluant le taux de protéines de variétés de blé. La mesure de référence (TxProtRef) prend plusieurs heures et celle-ci est comparée avec une mesure par le spectromètre (TxprotIR) qui est quasi instantanée. L'opération est répétée sur $n = 26$ échantillons.

```
Regression Analysis: TxprotIR versus TxProtRef

The regression equation is
TxprotIR = 0,16 + 0,981 TxProtRef

Predictor     Coef  SE Coef     T      P
Constant     0,157    1,174  0,13  0,895
TxProtRef  0,9808   0,1046  9,38  0,000
```

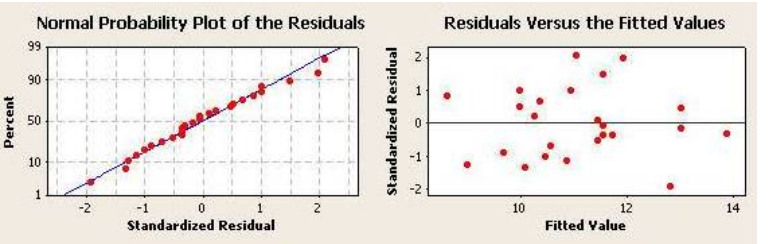

FIGURE 5 – *Droite de Henri et graphe des résidus de l'appareil de spectrométrie.*

```
S = 0,663595   R-Sq = 78,6%   R-Sq(adj) = 77,7%

PRESS = 12,3132   R-Sq(pred) = 75,02%
```