Premiers pas en régression linéaire avec SAS
Monique Le Guen, Josiane Confais

To cite this version:

HAL Id: cel-00357697
https://cel.archives-ouvertes.fr/cel-00357697
Submitted on 31 Jan 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Premiers pas en régression linéaire avec SAS®

Josiane Confais (UPMC-ISUP) –
Monique Le Guen (CNRS-CES-MATISSE-UMR8174)
e-mail : confais@ccr.jussieu.fr
e-mail : monique.leguen@univ-paris1.fr

Résumé
Ce tutoriel1 montre, de façon intuitive et sans formalisme excessif, les principales notions théoriques nécessaires à la compréhension et à l'interprétation des résultats d’analyses de régression linéaire, simple et multiple, produits par la procédure REG de SAS® et par le menu FIT de SAS/INSIGHT2.

Ce tutoriel est issu d’un cours enseigné par les auteurs dans différentes formations : ISUP, Master de Paris 1, formation permanente du CNRS, au CEPE de l’INSEE. Il fait suite à un premier document de travail publié à l’Unité Méthodes statistiques de l’INSEE.

Nous avons ajouté de nombreux graphiques et affichages de SAS/INSIGHT qui par ses possibilités de visualisation et d’interactivité, facilite la compréhension à la fois des données et des techniques. Nous avons également ajouté des liens vers des applets ou d’autres documents accessibles sur internet.

Nous insistons dans ce tutoriel sur l’importance des graphiques exploratoires et sur les limites des résultats obtenus par une régression linéaire, si l’étape de vérification des suppositions n’est pas systématiquement entreprise.

Mots clés : Régression linéaire simple, Régression linéaire multiple, Moindres carrés ordinaires, SAS, Proc REG, SAS/INSIGHT, Graphiques exploratoires, Validation

First steps in linear regression with SAS®

Abstract
This tutorial3 shows in an intuitive way and without excessive formalism, the theoretical notions necessary to understand and interpret simple and multiple regression produced by SAS® PROC REG and by the menu FIT of SAS/INSIGHT4.

This tutorial is based heavily on training courses given by the authors in high-profile institutions like ISUP, Master Degree at Paris 1 University, CNRS and CEPE-INSEE. It follows a first working paper published by UMS-INSEE. Thanks to SAS/INSIGHT interactivity and visualization tools, we created numerous graphs and displays to improve the understanding of data and statistical methods. This tutorial also includes various links towards applets or other documents from the internet.

We insist in this tutorial on the significance of exploratory graphs and on the limits of results obtained by a linear regression if the assumptions are not systematically checked.

Keywords: Simple Linear Regression, Multiple Linear Regression, Ordinary Least Squares, SAS, Proc REG, SAS/INSIGHT, Exploratory Graphics, Validation.

JEL Classification : C01, C52 AMS Classification : 62J05

1 Ce tutoriel est publié dans le n°33 de la revue électronique MODULAD à l’adresse http://www-rocq.inria.fr/axis/modulad/numero-35/Tutoriel-confais-35/confais-35.pdf
2 SAS® et SAS/INSIGHT sont les marques déposées de SAS Institute Inc., Cary, NC, USA
3 This tutorial was published in the electronics review MODULAD, n° 33, http://www-rocq.inria.fr/axis/modulad/numero-35/Tutoriel-confais-35/confais-35.pdf
4 SAS® and SAS/INSIGHT are registered trademarks of SAS Institute Inc., Cary, NC, USA
Sommaire

1. SENSIBILISATION À LA RÉGRESSION LINÉAIRE SIMPLE .......................................................... 5
   1.1. Où se place la régression linéaire ? ................................................................................. 5
   1.2. Ajustement affine ou Régexion Simple............................................................................ 6
       1.2.1. Comment trouver la droite qui passe « au plus près » de tous les points? ............... 8
       1.2.2. Méthode d’estimation des paramètres β0 et β1 .................................................. 9
       1.2.3. Effet d’un point observation sur la droite de régression ...................................... 11
       1.2.4. Décomposition de l’écart entre Y et la moyenne de Y ........................................ 11
       1.2.5. Analyse de la variance ......................................................................................... 12
       1.2.6. Statistiques liées au paramètre β0 ......................................................................... 13
              Calcul de l’intervalle de confiance pour β0 ............................................................. 31
              Test portant sur le paramètre β0 ............................................................................ 32
              Test portant sur le paramètre β1 ........................................................................... 33
       1.3. 2.1. Théorie de la régression pondérée........................................................................... 34
              Exemple d’estimation des paramètres avec Proc REG .......................................... 34
       1.3. 2.2. Effet d’un point observation sur la droite de régression ...................................... 11
       1.3. 2.3. Test portant sur le paramètre β0 ......................................................................... 13
              Calcul de l’intervalle de confiance pour β0 ............................................................. 31
              Test portant sur le paramètre β0 ............................................................................ 32
       1.3. 2.4. Test portant sur le paramètre β1 ......................................................................... 13
              Calcul de l’intervalle de confiance pour β1 ............................................................. 32
              Test portant sur le paramètre β1 ............................................................................ 33

1.4. Confiance à accorder aux résultats ................................................................................... 27
   1.4.1. Test de la signification globale de la régression ..................................................... 27
       1.4.2. Statistiques liées au paramètre β1 ......................................................................... 28
              Calcul de la variance de β1 .................................................................................. 29
              Test portant sur le paramètre β1 ........................................................................... 30
       1.4.3. Statistiques liées au paramètre β0 ......................................................................... 31
              Test portant sur le paramètre β0 ........................................................................... 32
              Calcul de l’intervalle de confiance pour β0 ............................................................. 33
              Exemple d’estimation des paramètres avec Proc REG .......................................... 34

2. LA RÉGRESSION LINÉAIRE MULTIPLE .................................................................................. 41
   2.1. Le critère des moindres carrés ....................................................................................... 41
   2.2. Formalisation de la régression linéaire multiple ............................................................ 42
       2.3. Exemples de régression linéaire multiple avec Proc REG ........................................ 44
           2.3.1. Présentation des données .................................................................................. 44
           2.3.2. Régression linéaire multiple avec Proc REG sans options ................................ 45
       2.4. TYPE I SS et TYPE II SS de Proc REG .................................................................... 48
           2.4.1. Définition de TYPE I SS et TYPE II SS .......................................................... 48
           2.4.2. Interprétations conjointes de TYPE I SS et TYPE II SS ............................ 51
           2.4.3. Options SS1 et SS2 de l’instruction model de Proc REG .................................. 51
           2.4.4. Tester la nullité de r paramètres pour tester un sous modèle ......................... 53
       2.5. Ce qu’il faut retenir des ‘SS’ ......................................................................................... 56
       2.6. Les résidus ................................................................................................................... 57

3. QUAND LES RÉSULTATS D’UNE RÉGRESSION NE SONT PAS FORCÉMENT PERTINENTS ....... 59
   3.1. Exemples en régression simple ..................................................................................... 59
       3.1.1. Une même valeur pour des situations différentes .............................................. 59
       3.1.2. Pondérations et régression linéaire par morceaux .............................................. 61
              Théorie de la régression pondérée ........................................................................ 64
       3.1.3. Transformation des données ................................................................................ 64
       3.1.4. Méthode non paramétrique du LOWESS ........................................................... 68
   3.2. Exemples en régression multiple .................................................................................. 70
       3.2.1. Y « expliquée » par la corrélation entre deux régresseurs ...................................... 70
       3.2.2. Instabilité des coefficients de la régression, en cas de multicollinéarité ............... 72

©WP CES 2007.47 - Premiers pas en régression linéaire avec SAS- Confaïs Le Guen - 2
Exemple sur données réelles ................................................................. 72
Exemple sur données avec modèle théorique connu et régresseurs corrélés .......... 74
3.3. Conditions d'utilisation de la régression, les diagnostics ........................................ 76
  3.3.1. Modèle Inadapté ........................................................................... 77
  3.3.2. L'influence de certaines données, les données atypiques -Outliers- ................. 77
  3.3.3. Corrélation et colinéarité entre les régresseurs ........................................ 78
4. VALIDATION D'UNE RÉGRESSION ................................................................. 79
  4.1. Introduction.......................................................................................... 79
  4.1.1. Modèle et notations......................................................................... 79
  4.1.2. Problèmes à étudier........................................................................ 80
  4.2. Vérification des suppositions de base sur les erreurs ........................................ 80
  4.2.1. Espérance nulle............................................................................... 80
  4.2.2. Indépendance.................................................................................. 80
      Cas particulier où les observations sont apparentées (cas des chroniques) : .... 81
  4.2.3. Égalité des variances (homoscédasticité)........................................ 82
  4.2.4. Normalité des erreurs....................................................................... 84
  4.2.5. Exemple......................................................................................... 84
  4.4.3. Condition index et variance proportion ........................................... 90
  4.4.2. Variance Inflation Factor............................................................... 90
  4.3.10. Tableau récapitulatif...................................................................... 93
  4.3.9. Comment obtenir les mesures d'influence dans SAS.......................... 92
  4.3.8. Coefficient global PRESS.............................................................. 92
  4.3.7. Influence sur la valeur ajustée: DFFITS ........................................ 92
  4.3.6. Précision des estimateurs : COVRATIO ........................................ 91
  4.3.5. Influence sur chacun des coefficients : DFBETAS ......................... 91
  4.3.4. Mesure globale de l'influence sur le vecteur des coefficients: Distance de COOK 90
  4.3.3. Résidus studentisés externes ......................................................... 90
  4.3.2. Résidus studentisés internes ......................................................... 90
  4.3.1. Hat matrice et leverages .............................................................. 88
  4.2.3. Égalité des variances (homoscédasticité)........................................ 82
  4.2.2. Indépendance.................................................................................. 80
  4.1.1. Modèle et notations......................................................................... 79
  3.3.3. Corrélation et colinéarité entre les régresseurs ........................................ 78
  3.3.2. L'influence de certaines données, les données atypiques -Outliers- ................. 77
  3.3.1. Modèle Inadapté ........................................................................... 77
4.3. Influence d'observations............................................................................... 88
  4.3.1. Hat matrice et leverages .............................................................. 88
  4.3.2. Résidus studentisés internes ......................................................... 90
  4.3.3. Résidus studentisés externes ......................................................... 90
  4.3.4. Mesure globale de l'influence sur le vecteur des coefficients: Distance de COOK 90
  4.3.5. Influence sur chacun des coefficients : DFBETAS ......................... 91
  4.3.6. Précision des estimateurs : COVRATIO ........................................ 91
  4.3.7. Influence sur la valeur ajustée: DFFITS ........................................ 92
  4.3.8. Coefficient global PRESS.............................................................. 92
  4.3.9. Comment obtenir les mesures d'influence dans SAS.......................... 92

Dans PROC REG.......................................................... 92
Dans SAS/INSIGHT ................................................. 93
  4.3.10. Tableau récapitulatif................................................................ 93
  4.3.11. Exemple..................................................................................... 95
4.4. Colinéarité des régresseurs................................................................. 99
  4.4.1. Méthodes basées sur l'étude de la matrice XX .................................. 100
      Etude de la matrice de corrélation des régresseurs .............................. 101
  4.4.2. Variance Inflation Factor............................................................... 101
  4.4.3. Condition index et variance proportion ........................................ 102
      Les indices de colinéarité ................................................................. 103
  4.4.4. Remèdes en cas de multi-colinéarité ............................................ 104
  4.4.5. Exemple...................................................................................... 105
  4.5. Choix des régresseurs.......................................................................... 106
  4.5.1. Utilisation des sommes de carrés ................................................ 107
      Rappel sur les somme de carrés apportés par un régresseur ................. 107
      Tests des apports à SS_{modèle} d’une variable .................................... 108
      Exemple d’élimination progressive ..................................................... 109
  4.5.2. Différentes méthodes basées sur les sommes de carrés .................... 111
      Méthode FORWARD (ascendante) ..................................................... 111
      Méthode BACKWARD (descendante) ............................................. 112
      Méthode STEPWISE (progressive) .................................................. 112
      Exemples de sélection STEPWISE ............................................... 113
  4.5.3. Amélioration de R^2 ..................................................................... 115
      Maximum R^2 Improvement (MAXR) ............................................ 115
      Minimum R^2 Improvement (MINR) ................................................ 116
  4.5.4. Autres méthodes basées sur R^2 : RSQUARE et ADJRSQ .................. 116
  4.5.5. Coefficient CP de Mallows........................................................... 116
      Sélection suivant le coefficient CP .................................................... 117
      Utilisation du coefficient CP dans une sélection de régresseurs .......... 117
  4.5.6. Critères AIC et BIC ..................................................................... 117
  4.5.7. Exemple de sélection RSQUARE ................................................ 118

CONCLUSION............................................................................................................ 120
ANNEXES .................................................................................................................. 122
ANNEXE 1..................................................................................................................... 123
SYNTAXE SIMPLIFIÉE DE LA PROCÉDURE REG DE SAS................................................................. 123
PROC REG options ;...................................................................................................................... 123
MODEL dépendante = régresseurs / options ;................................................................................ 124
Instructions BY FREQ ID WEIGHT ;............................................................................................. 125
REWEIGHT expression / WEIGHT = valeur ;.................................................................................... 125
TEST equation(s) ;.......................................................................................................................... 125
RESTRICT equation(s);..................................................................................................................... 125
OUTPUT OUT = nomtab mot_clef = nom_var ; ................................................................................... 126
PLOT Yvar1*Xvar1='s' Yvar2*Xvar2='s' / options ;.............................................................................. 126
PRINT mots-clefs; ............................................................................................................................ 126
Options RIDGE et PCOMIT des instructions PROC REG ou MODEL........................................... 127

ANNEXE 2.......................................................................................................................................... 128
MODE D'EMPLOI TRÈS SUCCINCT DE SAS/INSIGHT................................................................. 128
Le lancement de SAS/INSIGHT ..................................................................................................... 128
Rôle statistique des variables dans SAS/INSIGHT....................................................................... 129
Menu principal de SAS/INSIGHT .................................................................................................. 130
Graphiques standard en SAS/INSIGHT .......................................................................................... 130
Les Analyses Statistiques avec SAS/INSIGHT .............................................................................. 132
Exemple de Régression linéaire sur la Table SAS : Chenille (processionnaire du pin du §2.3.1.).... 132
Impression et Sauvegarde ............................................................................................................... 133
Pour imprimer ................................................................................................................................. 133
Pour sauvegarder les résultats graphiques ou tableaux dans un fichier .......................................... 133
Pour insérer un fichier externe .bmp dans Word ........................................................................... 134
Pour plus d’information sur les graphiques.................................................................................. 135

ANNEXE 3.......................................................................................................................................... 136
STATISTIQUES RELATIVES À L’ANALYSE DE LA VARIANCE ................................................ 136
Statistiques sur les paramètres ....................................................................................................... 137

ANNEXE 4.......................................................................................................................................... 138
RELATIONS ENTRE LA LOI NORMALE ET LES STATISTIQUES DE LOIS ................................. 138

ANNEXE 5.......................................................................................................................................... 139
CONSTRUCTION D’UN QQ-PLOT..................................................................................................... 139
Principe de la droite de Henry ......................................................................................................... 139
1. Soit une variable X dont on veut vérifier l’adéquation à une loi normale (m,\(\sigma\)).................. 139
2. Si X suit une loi normale (m,\(\sigma\)) alors : ................................................................................. 139
3. En pratique, on ordonne les valeurs \(x_i\) : on note \(x_{(i)}\) les valeurs ordonnées. ......................... 139
Généralisation............................................................................................................................... 140
QQ-PLOT AVEC SAS...................................................................................................................... 140

BIBLIOGRAPHIE............................................................................................................................ 141
1. Sensibilisation à la régression linéaire simple

Cette sensibilisation à la régression présente de manière détaillée la logique et les calculs permettant la compréhension de la régression simple.
On montre tout d'abord la démarche algébrique qui conduit à un ajustement affine, puis par un détour obligé à l'Histoire, on « glisse » vers la modélisation en s'appuyant sur la Statistique.

1.1. Où se place la régression linéaire ?

La régression linéaire se classe parmi les méthodes d’analyses multivariées qui traitent des données quantitatives.

C'est une méthode d'investigation sur données d'observations, ou d'expérimentations, où l'objectif principal est de rechercher une liaison linéaire entre une variable Y quantitative et une ou plusieurs variables X également quantitatives.

C’est la méthode la plus utilisée pour deux raisons majeures :

- c’est une méthode ancienne,
- c’est l'outil de base de la plupart des modélisations plus sophistiquées comme la régression logistique, le modèle linéaire généralisé, les méthodes de traitement des séries temporelles, et surtout des modèles économétriques, etc.

A l'aide du tableau 1.1, on peut repérer les méthodes les plus courantes d'analyses statistiques et les procédures SAS utiles pour rechercher des liaisons, selon le type (nominal, ordinal, intervalle, ratio) des variables Y et X. Le lecteur peu familiarisé avec la terminologie des variables SAS pourra voir sur le site de MODULAD, le tutoriel5 « La Proc FREQ de SAS, Tests d’indépendance et d’association », de J. CONFAIS, Y. GRELET, M. LE GUEN.

### Tableau 1.1 Procédures SAS adaptées selon le type des variables
(nominal, ordinal, intervalle, ratio)

<table>
<thead>
<tr>
<th>X intervalle/ratio</th>
<th>X ordinaire/nominal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Régression linéaire</td>
<td>Analyse de la variance</td>
</tr>
<tr>
<td>PROC REG</td>
<td>PROC ANOVA</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Y intervalle/ratio</th>
<th>Y ordinaire/nominal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Si Y est ordinaire ou à 2 modalités</td>
<td>Analyses de tableaux de contingence</td>
</tr>
<tr>
<td>Régression logistique</td>
<td>PROC FREQ</td>
</tr>
<tr>
<td>PROC LOGISTIC</td>
<td>Régression logistique</td>
</tr>
<tr>
<td></td>
<td>PROC LOGISTIC</td>
</tr>
</tbody>
</table>

Modèles linéaires généralisés
⇔ PROC GLM

Traitements des variables catégorielles
⇔ PROC CATMOD

Pour la régression linéaire la procédure REG est la plus complète. Cependant le module SAS/INSIGHT, qui est à la fois un tableur, un grapheur et un analyseur, est particulièrement adapté pour étudier des données dans une problématique de régression linéaire couplée à une analyse exploratoire des données.

Dans les exemples nous utiliserons l’une ou l’autre de ces possibilités. En annexe 2, on trouvera un mode d’emploi très succinct de SAS/INSIGHT.

### 1.2. Ajustement affine ou Régression Simple

**Exemple**
Soient les 2 mesures de poids (variable X) et taille (variable Y) relevées sur un échantillon de 20 objets.
Tableau 1.2 Données Taille et Poids

<table>
<thead>
<tr>
<th>identifiant</th>
<th>poids (X)</th>
<th>taille (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>46</td>
<td>152</td>
</tr>
<tr>
<td>2</td>
<td>78</td>
<td>158</td>
</tr>
<tr>
<td>3</td>
<td>85</td>
<td>160</td>
</tr>
<tr>
<td>4</td>
<td>85</td>
<td>162</td>
</tr>
<tr>
<td>5</td>
<td>85</td>
<td>158</td>
</tr>
<tr>
<td>6</td>
<td>85</td>
<td>159</td>
</tr>
<tr>
<td>7</td>
<td>95</td>
<td>165</td>
</tr>
<tr>
<td>8</td>
<td>95</td>
<td>165</td>
</tr>
<tr>
<td>9</td>
<td>100</td>
<td>166</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>159</td>
</tr>
<tr>
<td>11</td>
<td>100</td>
<td>166</td>
</tr>
<tr>
<td>12</td>
<td>103</td>
<td>168</td>
</tr>
<tr>
<td>13</td>
<td>105</td>
<td>163</td>
</tr>
<tr>
<td>14</td>
<td>105</td>
<td>164</td>
</tr>
<tr>
<td>15</td>
<td>115</td>
<td>168</td>
</tr>
<tr>
<td>16</td>
<td>115</td>
<td>166</td>
</tr>
<tr>
<td>17</td>
<td>115</td>
<td>162</td>
</tr>
<tr>
<td>18</td>
<td>130</td>
<td>165</td>
</tr>
<tr>
<td>19</td>
<td>135</td>
<td>167</td>
</tr>
<tr>
<td>20</td>
<td>150</td>
<td>172</td>
</tr>
</tbody>
</table>

Le graphique du nuage de points, d’abscisse le poids et d’ordonnée la taille montre qu’il existe une relation linéaire entre ces deux variables. Lorsque le poids augmente, la taille a tendance à croître également.

Figure 1.1 Taille*Poids

Les points du nuage sont approximativement alignés sur une droite (y=ax+b) à une erreur près.

Taille = $\beta_0 + \beta_1$ Poids + erreur

La variable Taille (Y) est appelée la variable “réponse”, ou selon les domaines disciplinaires, variable à expliquer, ou encore variable dépendante. La variable Poids (X) est la variable “régresseur”, encore appelée variable explicative, ou variable indépendante.

$\beta_0$ est l’ordonnée à l’origine.
\( \beta_1 \) est la pente de la droite d’ajustement.

**Note :** Dans ce document nous n’utiliserons que les termes « réponse » et « régresseurs », pour éviter toutes confusions sémantiques très dommageables lors des interprétations des résultats, et particulièrement lors de la communication des résultats à un tiers.

Par exemple, la variable dite expliquée n’est pas forcément expliquée par les variables dénommées explicatives. Quand aux variables dites indépendantes, elles sont, dans le cas de données réelles, rarement indépendantes.

### 1.2.1. Comment trouver la droite qui passe « au plus près » de tous les points?

Pour trouver la droite qui passe « au plus près » de tous les points il faut se donner un **critère d’ajustement.**

![Figure 1.2 Projection des points M₁...M₄ sur la droite.](image)

La somme des carrés des écarts (SCE) des points observés \( M_i \) à la droite solution soit minimum.

La droite solution sera appelée **droite de régression de Y sur X.**

Le critère est le « **critère des Moindres Carrés Ordinaires** » (MCO, *Ordinary Least Squares* en anglais), appelé aussi par les statisticiens « **critère de Norme L²** ». Les écarts sont calculés en projetant les points M parallèlement à l’axe des Y. On pourrait aussi projeter les points M parallèlement à l’axe des X, on aurait alors une autre droite solution (régression de X sur Y). Dans ces deux régressions Y et X ne jouent pas le même rôle.
On pourrait aussi projeter les points M perpendiculairement à la droite solution. Y et X joueraient dans ce cas le même rôle. C’est la situation que l’on rencontre dans une Analyse en Composantes Principales\(^6\), illustrée dans la figure 1.3.

![Figure 1.3 Trois projections possibles du point \((X_i, Y_i)\)](image)

### 1.2.2. Méthode d’estimation des paramètres \(\beta_0\) et \(\beta_1\)

La Somme des Carrés des Ecarts (SCE) est donnée par :

\[
S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2
\]

La valeur de cette fonction \(S\) est minimum lorsque les dérivées de \(S\) par rapport à \(\beta_0\) et \(\beta_1\) s’annulent. La solution est obtenue en résolvant le système :

\[
\frac{\partial S}{\partial \beta_0} = 0 \quad \text{et} \quad \frac{\partial S}{\partial \beta_1} = 0
\]

Les dérivées par rapport à \(\beta_0\) et \(\beta_1\) sont :

\[
\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)
\]

\[
\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i)
\]

Ces dérivées s’annulent pour deux valeurs \(b_0\) et \(b_1\) solutions des 2 équations à 2 inconnues :

---

\(^6\) On pourrait encore prendre comme critère la somme des valeurs absolues des écarts des points observés à la droite, ce serait alors un critère de norme L\(^1\), et pourquoi pas prendre un exposant non entier appartenant à l’intervalle [1,2], ce serait une norme L\(^p\).

©WP CES 2007.47 - Premiers pas en régression linéaire avec SAS- Confais Le Guen - 9
équation 1 : \[ \sum_{i=1}^{n} (Y_i - b_0 - b_1X_i) = 0 \]

équation 2 : \[ \sum_{i=1}^{n} X_i(Y_i - b_0 - b_1X_i) = 0 \]

Ce système de 2 équations à 2 inconnues déterminent les **équations normales**.

Développons ces 2 équations normales :

- l'équation 1 donne : \[ \sum Y_i - nb_0 - b_1 \sum X_i = 0 \]
  et en divisant par \( n \)
  \[ \bar{Y} = b_0 + b_1 \bar{X} \]

On remarque que la droite solution passe par le centre de gravité du nuage

\[ (\bar{X}, \bar{Y}) = \left( \frac{\sum X_i}{n}, \frac{\sum Y_i}{n} \right) \]

- L'équation 2 donne
  \[ \sum Y_iX_i - b_0 \sum X_i^2 - b_1 \sum X_i^2 = 0 \]
  dans laquelle on remplace \( b_0 \)

  \[ \sum Y_iX_i - (\bar{Y} - b_1 \bar{X}) \sum X_i - b_1 \sum X_i^2 = 0 \]

Solution :

\[ b_1 = \frac{\sum X_iY_i - (\sum X_i \sum Y_i)/n}{\sum X_i^2 - (\sum X_i)^2 / n} \]

en divisant numérateur et dénominateur par \( n \) on retrouve les expressions de la covariance et de la variance empiriques :

\[ \sum (X_i - \bar{X})(Y_i - \bar{Y}) \]

\[ \sum (X_i - \bar{X})^2 \]

**formule n° 1**

Les points qui sont sur la droite de régression ont pour ordonnée : \( \hat{Y} = b_0 + b_1X \)

Le coefficient \( b_1 \) dépend au numérateur de la covariance entre \( X \) et \( Y \), et de la variance de \( X \) pour le dénominateur.

**Terminologie**

\( \hat{Y} \) est l'**estimation** de \( Y \) obtenue à partir de l’équation de régression.

\( \hat{Y} \) *se prononce* \( Y \ chapeau.\)
b₀ et b₁ sont les estimateurs des moindres carrés des paramètres inconnus β₀ et β₁. On appelle estimations les valeurs particulières (solutions) prises par les estimateurs b₀ et b₁. 

Dans la suite du document on ne fera pas de différence de notations entre les estimateurs b₀ ou b₁ et leurs estimations.

1.2.3. Effet d’un point observation sur la droite de régression

Avec cet applet java http://www.stat.sc.edu/~west/javahtml/Regression.html on peut voir l’effet de levier (leverage) sur le calcul de la droite de régression en ajoutant un point -rouge- par un simple clic de souris. Ici le point rouge est un point influent dans la liaison (X,Y). Plus le point est éloigné de la tendance plus son levier sera grand. Il peut aussi exister des points atypiques -Outliers- seulement en direction des X, ou dans la direction des Y (voir le chapitre 4).

1.2.4. Décomposition de l’écart entre Yi et la moyenne de Y

En un point d’observation (Xᵢ, Yᵢ) on décompose l’écart entre Yi et la moyenne des Y en ajoutant puis retranchant Yi la valeur estimée de Y par la droite de régression. Cette procédure fait apparaître une somme de deux écarts :

\[(Yᵢ - \bar{Y}) = (Yᵢ - \hat{Yᵢ} + \hat{Yᵢ} - \bar{Y})\]
\[(Yᵢ - \bar{Y}) = (Yᵢ - \hat{Yᵢ}) + (\hat{Yᵢ} - \bar{Y})\]

Ainsi l’écart total \((Yᵢ - \bar{Y})\) peut être vu comme la somme de deux écarts :

- un écart entre \(Yᵢ\) observé et \(\hat{Yᵢ}\) la valeur estimée par le modèle
- un écart entre \(\hat{Yᵢ}\) la valeur estimée par le modèle et la moyenne \(\bar{Y}\).

Le graphique suivant montre l’explication géométrique de cette décomposition. Cet artifice de décomposition aura un intérêt fondamental dans l’analyse de la variance abordée au paragraphe suivant.
Figure 1.4 Décomposition des différents écarts

Ecart total \( (Y_i - \bar{Y}) \) = écart dû au modèle \( (Y_i - \hat{Y}_i) \) + écart résiduel \( (\hat{Y}_i - Y) \)

1.2.5. Analyse de la variance

Ce que le modèle explique et ce qu’il n’explique pas

A partir de l’équation de la droite de régression (modèle retenu), on peut pour tout point \( i \) d’abscisse \( X_i \) calculer son estimation (ordonnée) \( \hat{Y}_i \)

\[
\hat{Y}_i = b_0 + b_1 X_i \quad \text{avec} \quad b_0 = \bar{Y} - b_1 \bar{X}
\]

ce qui donne :

\[
\hat{Y}_i = \bar{Y} + b_1 (X_i - \bar{X})
\]

ou encore

\[
\hat{Y}_i - \bar{Y} = b_1 (X_i - \bar{X}) \quad \text{formule n° 2}
\]

En un point \( i \) l’écart ou résidu est :

\[
Y_i - \hat{Y}_i = (Y_i - \bar{Y}) - (\hat{Y}_i - \bar{Y})
\]

On élève les deux membres au carré et on somme sur les observations \( i \) :

\[
\sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \bar{Y})^2 + \sum (\hat{Y}_i - \bar{Y})^2 - 2 \sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})
\]

En utilisant la formule n°2 :

\[
\sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \bar{Y})^2 + \sum (\hat{Y}_i - \bar{Y})^2 - 2 \sum (Y_i - \bar{Y}) \cdot b_1 (X_i - \bar{X})
\]

En utilisant une transformation de la formule n°1 :

\[
b_1 \sum (X_i - \bar{X})^2 = \sum (X_i - \bar{X})(Y_i - \bar{Y})
\]

on obtient
\[ \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \bar{Y})^2 + \sum (\hat{Y}_i - \bar{Y})^2 - 2 \cdot b_1 \sum (X_i - \bar{X})^2 \]

En utilisant la formule n° 2 :

\[ \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - \bar{Y})^2 + \sum (\hat{Y}_i - \bar{Y})^2 - 2 \sum (\hat{Y}_i - \bar{Y})^2 \]

On aboutit enfin à l'égalité fondamentale :

\[ \sum (Y_i - \hat{Y}_i)^2 = \sum \hat{Y}_i - \bar{Y})^2 + \sum (Y_i - Y_i)^2 \]

La SCE (Somme des Carrés des Écartsw) totale est égale à la somme des carrés des écarts dus au modèle augmentée de la somme des carrés des écarts dus aux erreurs

\[ \text{SCE totale} = \text{SCE modèle} + \text{SCE erreur}. \]

Cette formule montre que :

Les variations de Y autour de sa moyenne, c'est-à-dire SCE Totale (SS Total pour Sum of Squares en anglais) peuvent être expliquées par :

- le modèle grâce à SCE Modèle (SS Model en anglais) ;
- et ce qui ne peut être expliqué par le modèle, est contenu dans SCE Erreur (SS Error en anglais).

L'erreur est aussi appelée le « résidu ».

**Standard de présentation de l'Analyse de la Variance**

On a l'habitude de représenter l'analyse de la variance sous forme d'un tableau, faisant apparaître les 3 sources de variation : le total en 3ème ligne qui se décompose en la partie modèle et la partie erreur.

A chaque source de variation (Total, Modèle, Erreur) correspond un nombre de degrés de liberté (ddl) respectivement égal à n-1, p, n-p-1,

- n : nombre d'observations
- p : nombre de variables régresseurs (la variable X₀, constante égale à 1, correspondant au paramètre β₀, n'est pas comprise).

Nous présentons le tableau général de l'analyse de variance pour p régresseurs. Pour la régression simple, p=1 (une seule variable régresseur).
**Tableau 1.3 Analyse de variance (version anglaise)**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODEL</td>
<td>p</td>
<td>$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$</td>
<td>$\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 / p$</td>
</tr>
<tr>
<td>ERROR</td>
<td>n-p-1</td>
<td>$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$</td>
<td>$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 / (n - p - 1)$</td>
</tr>
<tr>
<td>TOTAL</td>
<td>n-1</td>
<td>$\sum_{i=1}^{n}(Y_i - \bar{Y})^2$</td>
<td></td>
</tr>
</tbody>
</table>

**Abréviations:**

DF : Degrees of Freedom se traduit par degrés de liberté (ddl).
Ils vérifient : DF<sub>total</sub>=DF<sub>model</sub> + DF<sub>erreur</sub>

SS : Sum of Squares se traduit par Somme des Carrés des Ecarts (SCE)

MS : Mean Square, est le rapport SS/DF, relatif soit au modèle soit à l'erreur

MSE : Mean Square Error = $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 / (n - p - 1)$ représente le carré de l'écart moyen résiduel.

Tous ces « indicateurs » SS, MS, MSE, vont jouer un rôle important dans l'appréciation du modèle calculé à partir des observations.

**Figure 1.5 Décomposition des SS –Sums of Squares**

La figure 1.5 montre les liens entre SS<sub>total</sub>, SS<sub>model</sub> et SS<sub>error</sub> lorsque l'on somme les carrés des écarts sur tous les points $i$. 

©WP CES 2007.47 - Premiers pas en régression linéaire avec SAS- Confais Le Guen - 14
Il est remarquable que la formule de décomposition de l’écart total en un point \(i\), vu au § 1.2.4.

\[
(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)
\]

prendra la même forme pour la somme des carrés.

\[
\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2
\]

**Comment apprécier globalement la régression**

Les deux quantités SCE totale (SS total) et SCE modèle (SS model) sont des sommes de carrés donc toujours positives ou nulles et telles que

\[
\text{SCE Modèle} \leq \text{SCE Totale}
\]

Le rapport \(\frac{\text{SCE Modèle}}{\text{SCE Totale}}\) est donc compris entre 0 et 1.

On appelle ce rapport **le coefficient de détermination**

\[
R^2 = \frac{\text{SCE Modèle}}{\text{SCE Totale}} = \frac{SS model}{SS Total}
\]

**Cas particulier :**
Si tous les points \(Y_i\) observés sont alignés sur la droite de régression, le modèle est parfaitement adapté et SCE Erreur = 0,

Dans ce cas: \(\frac{\text{SCE Modèle}}{\text{SCE Totale}} = 1\)

**Interprétation de \(R^2\)**

\(R^2\) qui varie entre 0 et 1, mesure la proportion de variation totale de \(Y\) autour de la moyenne expliquée par la régression, c’est-à-dire prise en compte par le modèle. Plus \(R^2\) se rapproche de la valeur 1, meilleure est l’adéquation du modèle aux données.

Un \(R^2\) faible signifie que le modèle a un faible pouvoir explicatif.

On démontre que \(R^2\) représente aussi le carré du **coefficient de corrélation linéaire** entre \(Y\) et \(Y\) estimé:

\[
R^2 = \text{Corr}^2(Y, \hat{Y})
\]

Dans le cas de la régression simple, \(R\) est aussi la valeur absolue du coefficient de corrélation linéaire entre \(Y\) et \(X\).

\[
R = |\text{Corr} (Y, X)|
\]
Lien entre coefficient de corrélation de 2 variables et le cosinus de leur angle

Soient 2 vecteurs $X_1$ et $X_2$ définis dans un espace $\mathbb{R}^n$ (espace des n observations), le coefficient de corrélation entre $X_1$ et $X_2$ est aussi le cosinus de l’angle $\theta$ entre ces 2 vecteurs.

En utilisant les conventions de notation, le produit scalaire de 2 vecteurs $X_1$ et $X_2$ se note $<X_1, X_2> = \|X_1\| \|X_2\| \cos(\theta)$

On a :

$$\cos(X_1, X_2) = \frac{<X_1, X_2>}{(\langle X_1, X_1 \rangle <X_2, X_2 \rangle)^{1/2}} = \frac{1}{n} \sum_{i=1}^{n} \frac{(X_{1i} - \overline{X})(X_{2i} - \overline{X})}{s_1 * s_2} = \text{Corrélation (}X_1, X_2\text{)}$$

$s_1 * s_2$ étant le produit des écarts-type des 2 vecteurs.

L’interprétation d’un coefficient de corrélation comme un cosinus est une propriété importante.

Comme le remarque TOMASSONE (1992), les variables $X$ n’étant pas des variables aléatoires, il est plus correct de parler de « cosinus » des angles formés par les vecteurs associés, en réservant le terme « coefficient de régression » pour sa similitude avec l’estimation de ce coefficient à partir d’un échantillon.

**Exemple : Régression de la Taille en fonction du Poids**


**Programme SAS**

```
Proc gplot data=libreg.tailpoid;
  plot Y*X;
  title 'Graphique taille en fonction du Poids ';
Proc corr data=libreg.tailpoid;
  title 'Corrélation ';
  var Y X;
Proc REG data=libreg.tailpoid;
  title 'Régression de la Taille en fonction du Poids ';
  model Y=x;
run;
```
Le coefficient de corrélation CORR entre Y et X vaut 0.83771.

Sortie standard de la Proc REG sans options

Dans la sortie de Proc REG on obtient d'abord le tableau d'analyse de la variance, puis les estimations des paramètres.
Lecture de l'Analyse de la Variance

\[ SS\ Model = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = 280.52918 \]

\[ SS\ Error = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = 119.22082 \]

\[ SS\ Total = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 399.75 \]

\[ Mean\ Square\ Model = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{p} = 280.52918 \]

\[ Mean\ Square\ Error = \frac{\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2}{(n - p - 1)} = 6.62338 \]

\[ Root\ MSE = \sqrt{MS\ ERROR} = 2.57359 \]

\[ Dependant\ Mean = \bar{Y} = 163.25 \]

\[ R-Square = \frac{SS\ Model}{SS\ Total} = 0.7018 = CORR(X,Y)^2 = (0.83771)^2. \]

Autres indicateurs

- CV = 1.57647  C'est le Coefficient de Variation = \( \frac{Root\ MSE \times 100}{Dep\ Mean} \)

Le CV est un indicateur sans dimension -exprimé en %- permettant de comparer l'écart moyen résiduel à la moyenne de la variable dépendante \( Y \). Ce pourcentage est plutôt utilisé pour comparer 2 modèles (donc 2 CV) portant sur le même jeu de données.

- Le coefficient \( R^2 \) ajusté, Adj R-sq

Le \( R^2 \) ajusté (utilisé en régression multiple) tient compte du nombre de paramètres du modèle.

\[ R^{2}\ ajusté = 1 - \left( \frac{n - \text{intercept}}{n - p} \right) \left( 1 - R^2 \right) \]

Avec \( \text{Intercept}=0 \), si il n'y a pas de constante \( b_0 \) à l'origine\(^7\) sinon \( \text{Intercept}=1 \).

Le reproche fait au coefficient de détermination est qu'il peut approcher la valeur 1, interprété comme un ajustement parfait, si on ajoute suffisamment de variables régresseurs.

Le \( R^2 \) ajusté tient compte du rapport \( p/n \) entre le nombre de paramètres du modèle et le nombre d'observations.

Selon certains auteurs ce coefficient permet de comparer des modèles de régression sur différents ensembles de données, mais il ne fait pas l'unanimité.

Attention : Adj R-sq peut prendre des valeurs inférieures à zéro !

\(^7\) S'il n'y a pas de constante \( b_0 \) à l'origine, les statistiques relatives à l'analyse de la variance n'ont pas la même interprétation.
Lecture du tableau des paramètres
Intercept = $b_0 = 145.98$ donne la valeur de la constante à l'origine. On peut remarquer que dans cet exemple, cette valeur n’a pas de signification dans le monde physique. On ne peut concevoir qu’à un poids de valeur nulle corresponde une taille de 145.98.

La pente de la droite (coefficient de $X$) = $b_1 = 0.1703$.

On l'interprète comme augmentation de la taille lorsque le poids augmente de une unité.

Equation de la droite : Taille = $145.98 + 0.1703 \times $Poids

Là encore il faut se préserver de toute interprétation causale. Peut-on agir et augmenter le poids en espérant faire augmenter la taille ?

Nous verrons les autres indicateurs dans la suite du chapitre.

Pour mieux comprendre la technique de la régression, voyons certaines représentations géométriques.

1.2.6. Représentations géométriques

**Régression simple de $Y$ sur $X$**

Afin d'avoir une idée géométrique de la régression prenons un exemple avec $n=3$ observations $(y_1,x_1)$, $(y_2,x_2)$ et $(y_3,x_3)$.

Le vecteur réponse $Y = (y_1,y_2,y_3)$, et le vecteur régresseur $X = (x_1,x_2,x_3)$ peuvent se représenter dans l'espace à 3 dimensions des observations. On nomme 1,2,3 les axes de ce repère.

Dans l'espace des observations représenté figure 1.6, la droite $\Delta$ des constantes a pour vecteur directeur $(1,1,1)$.

---

8 La figure 1.6 est une synthèse des graphiques de DRAPER & SMITH (1966) pp112-113 et SAPORTA (2006) p208
Figure 1.6 Régression de Y sur (X, Δ) dans l’espace des 3 observations.

L’interprétation géométrique de la régression est la suivante :

Régresser Y sur (Δ et X) consiste à projeter orthogonalement Y sur le plan (Δ, X) ce qui donne le point P.
Si d’autre part on projette Y sur la droite Δ, on obtient le point Q.
Par le théorème des 3 droites perpendiculaires Q est aussi la projection orthogonale de P sur Δ.

Dans le triangle YQP, rectangle en P, on peut appliquer le théorème de Pythagore :

\[ YQ^2 = YP^2 + PQ^2 \]

\[ \sum_i (Y_i - \bar{Y})^2 = \sum_i (Y_i - \hat{Y}_i)^2 + \sum_i (\hat{Y}_i - \bar{Y})^2 \]

La longueur YQ² représente la somme des carrés corrigée SCE Totale (SS Total).
La longueur YP² représente la somme des carrés non expliquée par le régression SCE Erreur (SS error) .
La longueur PQ² représente la somme des carrés expliquée par la régression soit SCE Modèle (SS model).

C’est l’équation fondamentale de l’analyse de la variance vue précédemment :

SS Total = SS Model + SS Error
Le coefficient de détermination $R^2$ est le rapport $\frac{PQ^2}{YQ^2}$.

$R^2$ représente donc le carré du cosinus de l'angle $(YQ, QP)$, c'est à dire l'angle entre $Y$ et $\hat{Y}$.

Plus l'angle entre $Y$ et $\hat{Y}$ est faible, meilleur est le pouvoir explicatif du modèle.

Et maintenant il suffit de généraliser mentalement à l'ordre n cette représentation à 3 dimensions.

**Remarque en régression multiple**

Si au lieu d'avoir une seule variable régresseur $X$, on avait plusieurs variables $X_1, \ldots, X_p$, alors le plan de projection $(X, \Delta)$ serait remplacé par l'hyperplan formé par les vecteurs $X_1, \ldots, X_p, \Delta$.

Régresser $Y$ sur les $p$ variables régresseurs consisterait à projeter orthogonalement $Y$ sur l'hyperplan déterminé par $X_1, \ldots, X_p, \Delta$.

**Distribution en un point fixé de $X$**

Jusqu'ici, on ne s'est appuyé que sur des calculs algébriques et sur des notions de géométrie, sans faire appel à des notions de statistique. On ne cherchait que la droite d'ajustement sur l'échantillon.

Aucune supposition n'a été nécessaire dans toutes les démonstrations.

Si maintenant, on souhaite utiliser les résultats obtenus à partir des observations, vues comme un échantillon, pour inférer sur la population, il faut faire appel à des notions de probabilité, et de statistique puisque dans les relevés de données (exemple : Poids et Taille) à notre disposition on n'a qu'un échantillon de valeurs et non toute la population.

Sur la figure 1.7, on remarque que pour une même valeur du Poids, par exemple 85, il y a plusieurs valeurs possibles de la Taille (158, 159, 160 et 162).
Il n'y a pas de valeur unique associée à une valeur $X_i$ mais une distribution de valeurs.

Pour chaque valeur du poids ($X$) existe une distribution théorique des tailles ($Y$). Les valeurs de centrage sont les espérances des tailles de la population correspondant à chaque poids $X_i$. L'espérance (moyenne théorique $\mu_i$) de chaque distribution de $Y$, est appelée « statistiquement parlant » l'espérance de $Y_i$ sachant $X_i$ que l'on note $E(Y_i/X_i)$.

L'hypothèse de la régression linéaire est que les $\mu_i$ sont alignés sur la vraie droite de régression qui est inconnue.

**Remarque** : pour simplifier l'écriture on note $E(Y_i)$ au lieu de $E(Y_i/X_i)$, soit :

$$\mu_i = E(Y_i) = \beta_0 + \beta_1 X_i$$

**Représentation de $X$ fixé et $Y$ aléatoire**

![Diagramme de régression](image.png)

**Figure 1.8 Distributions de $Y$ pour $X$ fixé**

Pour un même poids $X_1$ fixé on a une distribution de taille $Y_1$, dont on a observé une réalisation $y_1$, ou plusieurs.

Par exemple sur le graphique Taille*Poids de la figure 1.7, on remarque que pour $X=46$ on a une seule valeur observée $Y=152$, tandis que pour $X=85$ on observe plusieurs valeurs de $Y$ (158, 159, 160 et 162).

Chaque $Y_i$ est une variable aléatoire qui a une distribution de probabilité de $Y_i$ sachant $X_j$ notée $p(Y_i|X_j)$. Des hypothèses sur la régularité de ces distributions devront être faites :
• les distributions, pour tous les points $X_i$, sont supposées normales
• les espérances des distributions sont centrées sur la droite de régression
• les variances de chaque $Y_i$ conditionnellement à $X_i$ sont toutes égales.

De plus les variables aléatoires $Y_i$ ne doivent pas être reliées entre elles, elles sont supposées indépendantes.
Ces suppositions se résument ainsi :

Les variables aléatoires $Y_i$ sont indépendantes,

d'espérance et de variance :

$$E(Y_i) = \beta_0 + \beta_1 X$$
$$\text{variance}(Y_i) = \sigma^2$$

Il faut avoir à l'esprit que $E(Y_i)$ est une espérance conditionnelle. De même lorsque l'on parle de variance de $Y$, c'est sous-entendu, variance conditionnellement à $X$.

« Vraie » droite de régression et droite estimée par la régression

La figure 1.8 montrant les distributions de $Y$ pour $X$ fixé est une illustration du modèle de régression linéaire.
Toujours en supposant que le modèle linéaire postulé est le véritable modèle, on obtiendrait le vraie droite de régression $E(Y) = \beta_0 + \beta_1 X$, si on avait à notre disposition toute la population.
Comme on n'a qu'un échantillon d'observations, on n'a qu'une estimation $\hat{Y} = b_0 + b_1 X$ ou droite estimée par la régression.

A propos des erreurs

L'erreur théorique $\varepsilon_i$ représente l'écart entre $Y_i$ observé et l'espérance $E(Y_i)$ non observable. On notera que $\varepsilon_i$ non plus n'est pas observable. Ce qui est observable c'est l'erreur $e_i$ correspondant à l'écart entre $Y_i$ observé et $\hat{Y}_i$, son estimation par le modèle.

Le résidu observé $e_i$ est une estimation de l'erreur inobservable $\varepsilon_i$.

1.3. Glissement fonctionnel de la méthode des Moindres Carrés Ordinaires à la Régression.

De la théorie des erreurs en astronomie à l'étude des moyennes en sciences sociales, un siècle les sépare.
1.3.1. De l'Astronomie...

Historiquement la méthode des moindres carrés a d'abord été développée par LEGENDRE en 1805, pour répondre à une question posée par les astronautes et les spécialistes de la géodésie comme le rapporte DESROSIÈRES (1993):

"Comment combiner des observations effectuées dans des conditions différentes, afin d'obtenir les meilleures estimations possibles de plusieurs grandeurs astronomiques ou terrestres liées entre elles par une relation linéaire?".

Ces grandeurs sont mesurées par des instruments imparfaits, et par des observateurs qui ne sont pas tous identiques. Il y a des erreurs de mesures dans les observations.

De là provient le vocabulaire : observation, écart, erreur ou résidu.
Vous pouvez trouver sur internet une traduction anglaise de ce premier article scientifique de Legendre sur les moindres carrés (Least Squares) http://www.stat.ucla.edu/history/legendre.pdf.

1.3.2. … Aux Sciences Sociales

En s'appuyant sur :

1. Le théorème central limite (LAPLACE 1810) montrant que même si la distribution de probabilité des erreurs ne suit pas une loi normale, celle de la moyenne tend vers une loi normale, quand le nombre des observations s'accroît indéfiniment,

2. La synthèse opérée par Laplace et Gauss vers 1810 entre :
   - comment combiner au mieux des observations imparfaites ?
     Réponse : en utilisant le milieu (la moyenne),
   - comment estimer le degré de confiance que mérite une estimation ?
     Réponse : en terme de probabilité,

Galton inventeur de la "régression" et PEARSON inventeur de la "corrélation" appliquèrent l'ajustement des moindres carrés à des données sociales dans les années 1880.

Nous reproduisons ci-après le graphique de GALTON, révélateur d'une « Reversion », et accessible sur internet :
http://www.stat.ucla.edu/history/regression.gif.

1.3.3. Galton Diagram Regression

En 1885 GALTON réalisa le tableau qui croise la taille de 928 enfants (devenus adultes) nés de 203 parents, en fonction de la taille moyenne de leurs parents (la taille de la mère étant préalablement multipliée par un coefficient 1.08).

---

9 F GALTON, Regression towards mediocrity in hereditary stature", Journal of the Anthropological Institute 15 (1886), 246-263.
En présentant ce tableau sous forme d’un graphique, GALTON remarqua que l’on pouvait voir des ellipses de densités. Si les parents sont plus grands que la moyenne, les enfants seront également plus grands que la moyenne mais avec une taille plus proche de la moyenne que celle de leurs parents.

Si les parents sont de petites tailles, leurs enfants seront également plus petits que la moyenne, mais avec une taille plus proche de la moyenne que celle de leurs parents.

D'où le terme de « régression ».

Ce n'est que vers les années 1930 que le formalisme de la méthode des moindres carrés associé à une interprétation probabiliste est devenu la « Régression » (ARMATTE (1995)).
Le glissement des méthodes d’analyse, des erreurs en Astronomie vers des estimations de moyennes en Sciences Sociales, a conduit à appeler erreur ou perturbation ou encore aléa, l’écart de Y par rapport à sa moyenne. Le modèle s’écrit :

\[ Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \]

où les erreurs \( \varepsilon_i \) sont des aléas indépendants d’espérance =0 de variance \( \sigma^2 \)

1.3.4. Formalisation des Suppositions

L’ensemble des suppositions nécessaires pour élaborer les tests statistiques se résume ainsi :

- l’erreur \( \varepsilon_i \) est une variable aléatoire d’espérance nulle et de variance constante \( \sigma^2 \).

\[ \mathbb{E}(\varepsilon_i) = 0 \text{ et } \text{Var}(\varepsilon_i) = \sigma^2 \]

- l’erreur \( \varepsilon_i \) est non corrélée à \( \varepsilon_j \).

\[ \text{Cov}(\varepsilon_i, \varepsilon_j) = 0 \text{ pour } i \neq j \]

- les erreurs \( \varepsilon_i \) sont normalement distribuées

\[ \varepsilon_i \sim N(0, \sigma^2) \]

On résume souvent ces 3 suppositions par l’expression "i i d selon une loi normale" qui signifie

Indépendantes et Identiquement Distribuées selon une loi normale

Il faut de plus que les variables aléatoires Y conditionnellement à X soient indépendantes, et que les régresseurs \( X_j \) soient non aléatoires et non corrélés.

Lorsque ces suppositions sont vérifiées, l’estimateur MCO est non biaisé et efficace (de variance minimum). En anglais on utilise l’acronyme BLUE (BEST Linear Unbiased Estimator)

Nous verrons au paragraphe suivant, comment interviennent ces suppositions sur les aléas dans les raisonnements statistiques.
1.4. Confiance à accorder aux résultats

Pour inférer de l'échantillon à la population dont sont issues les observations, la logique statistique nous conduit à effectuer des tests d'hypothèses, et à déterminer des intervalles de confiance autour des valeurs estimées.

Successivement on va chercher à :

- **tester** la signification globale de la régression,
- **tester** l'hypothèse nulle $\beta_1=0$ et à **calculer** l'intervalle de confiance de $\beta_1$,
- **tester** l'hypothèse nulle $\beta_0=0$ et à **calculer** l'intervalle de confiance de $\beta_0$,
- **calculer** la précision de l'estimation de $Y$ pour la moyenne et pour une observation individuelle.

1.4.1. Test de la signification globale de la régression

Ce test a surtout un intérêt dans le cadre de la régression multiple, c'est à dire avec $p$ régresseurs. En anticipant sur le chapitre 2, qui présente la régression multiple, on généralise le modèle de régression à un régresseur au cas d'un modèle à $p$ régresseurs $X_1, X_2, \ldots, X_p$ :

$$\mu_i = E(Y_i) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Ce test permet de connaître l'apport global de l'ensemble des variables $X_1, \ldots, X_p$ à la détermination de $Y$.

On veut tester l'hypothèse nulle:

$H_0$: $\beta_1 = \cdots = \beta_p = 0$ contre

$H_a$: Il existe au moins un $\beta_j$ parmi $\beta_1, \ldots, \beta_p$ non égal à 0.

On calcule la statistique de test $F = \frac{MS_{model}}{MS_{error}}$ avec $MS_{model} = \frac{SS_{model}}{p}$ et $MS_{error} = \frac{SS_{error}}{n-p-1}$ représentant respectivement une somme de carrés des écarts moyens respectivement pour le modèle et pour l'erreur.

Si $H_0$ est vraie et sous réserve des suppositions suivantes, ce rapport $F$ est une valeur observée d'une variable qui suit une loi de Fisher-Snedecor à $p$ et $n-p-1$ degrés de liberté.

Si les $\varepsilon_i$ sont indépendants et suivent une loi normale de même variance $\varepsilon_i \approx N(0, \sigma^2)$.
Alors la statistique $F$ suit une loi de Fisher-Snedecor

$$F = \frac{MS \text{ model}}{MS \text{ error}} \approx F(p, n - p - 1)$$

**Règle de décision**

Si $F$ observé $\geq F_{1-\alpha}(p, n - p - 1)$

Alors $H_0: \beta_1 = \cdots = \beta_p = 0$ doit être rejetée au niveau $\alpha$

où $F_{1-\alpha}(p, n - p - 1)$ représente le quantile d'ordre $(1 - \alpha)$ de la loi de Fisher-Snedecor à $(p)$ et $(n-p-1)$ degrés de liberté.

**Note:** Dans SAS, la fonction de répartition inverse pour une loi de Fisher-Snedecor est donnée par la fonction FINV.

**Instruction SAS**

$$F = \text{FINV}(1 - \alpha, p, n - p - 1)$$

Avec $n$ nombre d'observations et $p$ nombre de régresseurs (non compris la constante).

Pour éviter de raisonner sur $F$, SAS fournit la *p-value* associée au $F$ observé. La *p-value* est le niveau de significativité du test de Fisher-Snedecor, c'est-à-dire la probabilité de dépasser le $F$ observé si l'hypothèse nulle est vraie. On compare la *p-value* au risque $\alpha$ choisi (par exemple $\alpha=0.05$).

**Raisonnement sur la p-value**

Si $p-value \leq \alpha$

Alors on rejette l'hypothèse nulle $\beta_1 = \cdots = \beta_p = 0$

**Interprétation**

On dit que la régression est significative au niveau $\alpha$.

Le modèle retenu améliore la prévision de $Y$ par rapport à la simple moyenne des $Y$.

Pour la régression simple, ce test porte uniquement sur le paramètre $\beta_1$.

Ce test fournit un moyen d'apprécier la régression dans son ensemble, ce qui ne signifie pas que chacun des coefficients de la régression soit significativement différent de 0.

**1.4.2. Statistiques liées au paramètre $\beta_1$**

Pour s'assurer de la significativité du paramètre $\beta_1$, on va dans une première étape calculer la variance de $b_1$, puis en deuxième étape tester l'hypothèse nulle $\beta_1 = 0$, en troisième étape on pourra alors déterminer un intervalle de confiance pour $\beta_1$ autour de $b_1$. 

©WP CES 2007.47 - Premiers pas en régression linéaire avec SAS- Confais Le Guen - 28
Calcul de la variance de \( b_1 \)
On a vu que \( b_1 \) est le rapport de la covariance entre \( X \) et \( Y \) divisé par la variance de \( X \):

\[
b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}
\]

On développe le numérateur

\[
b_1 = \frac{\sum (X_i - \bar{X})Y_i - \bar{Y}\sum (X_i - \bar{X})}{\sum (X_i - \bar{X})^2}
\]

Comme le 2\(\text{ème} \) terme du numérateur est nul, par définition de la moyenne \( \sum (X_i - \bar{X}) = 0 \), il ne reste que le 1\(\text{er} \) terme.

\[
b_1 = \frac{((X_1 - \bar{X})Y_1 + \cdots + (X_n - \bar{X})Y_n)}{\sum (X_i - \bar{X})^2}
\]

On ne peut calculer la variance de \( b_1 \) que si on fait des suppositions sur les \( X_i \) et sur les liaisons entre les \( Y_i \).

**Suppositions** pour calculer la variance de \( b_1 \)

Si les \( X_i \) sont non aléatoires
Si les \( Y_i \) sont non corrélés et de même variance \( \sigma^2 \)
Et comme par construction \( \text{Cov}(\bar{Y}, b_1) = 0 \)
Alors :

\[
\text{Var}(b_1) = a_1^2 \text{Var}(Y_1) + \cdots + a_n^2 \text{Var}(Y_n)
\]

avec \( a_i = \frac{(X_i - \bar{X})}{\sum (X_i - \bar{X})^2} \) assimilés à des constantes

Ce qui permet d’aboutir à :

\[
\text{Var}(b_1) = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}
\]

\( \sigma^2 \) représente la variance inconnue de \( Y \). Il faut de nouveau faire une supposition.

**Supposition**

Si le modèle postulé est le modèle correct
Alors \( \sigma^2 \) peut être estimé par les erreurs entre les \( Y_i \) observés et \( \hat{Y}_i \)
Mean Square Error:\[ \text{MSE} = s^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2} \]

Note : Pour la régression multiple : \[ \text{MSE} = s^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-p-1} \]

Compte tenu de toutes ces suppositions, l’estimateur de l’écart-type de \( b_1 \) devient :

\[ s(b_1) = \frac{s}{\sqrt{\sum(X_i - \bar{X})^2}} \]

Remarques:
- La variance de \( b_1 \) est inversement proportionnelle à la dispersion des \( X_i \) autour de la moyenne.
  Donc, si on veut améliorer la précision de \( b_1 \) il faut, si possible, augmenter la variance empirique des \( X_i \).
- La variance de \( b_1 \) est inversement proportionnelle à \( (n-2) \), \( n \) étant la taille de l’échantillon.
  Donc, si on veut améliorer la précision de \( b_1 \) il faut augmenter la taille de l’échantillon.

**Test portant sur le paramètre \( \beta_1 \)**

On s’intéresse au test de l’hypothèse nulle:

\( H_0 : \text{paramètre } \beta_1 = 0 \) contre \( H_a : \text{paramètre } \beta_1 \neq 0 \)

On calcule la statistique de test \[ T_{\text{observé}} = \frac{b_1}{s(b_1)} \]

Si \( \beta_1 = 0 \) la statistique \( T_{\text{observé}} \) suit une loi de Student, **sous l’hypothèse** que les erreurs soient indépendantes et identiquement distribuées selon la loi Normale.

**Suppositions**

- Si \( \beta_1 = 0 \)
- Si \( \varepsilon_i \approx N(0, \sigma^2) \)
  Alors \( T \text{ observé suit une loi de Student à } n-1 \) degrés de liberté

**Raisonnement**

On compare la \( p\text{-value} \) associée à \( T \text{ observé} \), au risque \( \alpha \) choisi (par ex:\( \alpha = 0.05 \)).

\[ \text{Si } p\text{-value} \leq \alpha \]
\[ \text{Alors on rejette l’hypothèse } \beta_1 = 0 \]

Conclusion : \( \beta_1 \) est significativement différent de zéro au niveau \( \alpha \).
Calcul de l'intervalle de confiance de \( \beta_1 \)

On peut calculer un intervalle de confiance (IC de niveau 1-\( \alpha \)) autour de \( b_1 \), ce qui permet de statuer sur le paramètre \( \beta_1 \) :

\[
|I_C_{1-\alpha}(\beta_1)| = \left[ b_1 - t_{1-\alpha/2} \cdot s(b_1); b_1 + t_{1-\alpha/2} \cdot s(b_1) \right]
\]

où \( t_{1-\alpha/2} \) représente le quantile d'ordre 1-\( \alpha/2 \) de la loi de Student à (n-2) degrés de liberté.

Note
- Dans SAS, la fonction de répartition inverse pour une loi de Student est donnée par la fonction TINV.
  Instruction SAS \( \Rightarrow T = \text{TINV}(1 - \alpha/2, n - 2) \) avec \( n \) nombre d'observations
- Dans le cas de la régression multiple avec \( p \)=nombre de régresseurs, la formule précédente devient:
  Instruction SAS \( \Rightarrow T = \text{TINV}(1 - \alpha/2, n - p - 1) \)

En pratique : si \( \alpha = 5\% \) et si \( n \) est assez grand (n>30), pour approcher la loi de Student par la loi Normale,

 alors

\[
|I_C_{0.95}(\beta_1)| = \left[ b_1 - 1.96 \cdot s(b_1); b_1 + 1.96 \cdot s(b_1) \right]
\]

Interprétation
Si la valeur 0 est dans l'intervalle de confiance de \( \beta_1 \), alors l'introduction de la variable \( X \) dans le modèle n'apporte aucun pouvoir explicatif sur \( Y \).

1.4.3. Statistiques liées au paramètre \( \beta_0 \)

La première étape consiste à calculer la variance de \( b_0 \), puis en deuxième étape à tester l'hypothèse nulle \( \beta_0=0 \), en troisième étape on pourra alors déterminer un intervalle de confiance pour \( \beta_0 \).

Calcul de la variance de \( b_0 \)

On a vu que \( b_0 \) vaut :

\[
b_0 = \bar{Y} - b_1 \bar{X}
\]

la variance vaut:

\[
\text{Var}(b_0) = \text{Var}(\bar{Y} - b_1 \bar{X})
\]

Raisonnement pour calculer la variance de \( b_0 \)

Pour pouvoir calculer la variance il faut faire des suppositions sur les termes de cette expression.
On suppose que les \( X_i \) sont non aléatoires.
Seuls la moyenne des $Y_i$ et le coefficient $b_1$ sont des variables aléatoires. On peut montrer de plus que la covariance entre $\bar{Y}$ et le coefficient $b_1$ est nulle\(^{10}\).

**Suppositions** pour calculer la variance de $b_0$

- Si les $X_i$ sont non aléatoires
- Si les $Y_i$ sont non corrélés et de même variance $\sigma^2$
- Et comme par construction $\text{Cov}(\bar{Y}, b_1) = 0$

Alors :

$$\text{Var}(b_0) = \text{Var}(\bar{Y}) + \frac{\sigma^2}{n} \sum (X_i - \bar{X})^2 = \frac{\sigma^2}{n} \sum X_i^2$$

$\sigma^2$ représente la variance inconnue de $Y$. Il faut de nouveau faire une supposition.

**Supposition**

- Si le modèle postulé est le modèle correct
Alors $\sigma^2$ peut être estimé par les erreurs entre les $Y$ observés et $\hat{Y}$

$$s^2 = \frac{\sum (Y_i - \hat{Y}_i)^2}{n - 2} = \text{MSE}$$

L'estimateur de la variance de $b_0$ devient :

$$s^2(b_0) = \frac{s^2 \sum X_i^2}{n \sum (X_i - \bar{X})^2}$$

**Remarque**:
La variance de $b_0$ est proportionnelle à la somme des carrés des $X_i$.
Si le plan d'expérience est tel que les valeurs des $X_i$ sont très grandes, la variance de $b_0$ sera très grande, et l'estimation de $b_0$ n'aura aucune signification.

**Test portant sur le paramètre $\beta_0$**

Test de l'hypothèse nulle $H_0 : \beta_0 = 0$ contre $H_a : \beta_0 \neq 0$

On calcule la statistique de test

$$T_{\text{observé}} = \frac{b_0}{s(b_0)}$$

\(^{10}\) voir démonstration dans NETER, WASSERMAN, KUTNER pp75-77.
Si $\beta_0=0$ la statistique $T_{\text{observé}}$ suit une loi de Student à $n-2$ degrés de liberté, **sous l'hypothèse** que les erreurs sont indépendantes et identiquement distribuées selon la loi Normale.

**Supposition**

$$\text{Si } \varepsilon_i \approx N(0,\sigma^2)$$
$$\text{Alors } T_{\text{observé}} \text{ suit une loi de Student}$$

**Raisonnement**

On compare la *p-value* associée à $T_{\text{observé}}$, c'est-à-dire la probabilité de dépasser le $T_{\text{observé}}$ en valeur absolue, au risque $\alpha$ choisi (par exemple $\alpha=0.05$).

$$\text{Si } p\text{-value} \leq \alpha$$
$$\text{Alors on rejette l'hypothèse } \beta_0=0$$

**Conclusion**  $\beta_0$ est significativement différent de zéro au niveau $\alpha$

**Calcul de l'intervalle de confiance de $\beta_0$**

On peut assigner un intervalle de confiance autour de $b_0$, ce qui permet de statuer sur le paramètre $\beta_0$:

$$[C_{1-\alpha}(\beta_0)=[b_0 - t_{1-\alpha/2} \cdot s(b_0); b_0 + t_{1-\alpha/2} \cdot s(b_0)]$$

où $t_{1-\alpha/2}$ représente le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-2$ degrés de liberté.

**Note**

- Dans SAS, la fonction de répartition inverse pour une loi de Student est donnée par la fonction TINV.
  Instruction SAS ⇔ $T = \text{TINV}(1-\alpha/2,n-2)$ avec $n=$ nombre d'observations

- Dans le cas de la régression multiple avec $p=$nombre de régresseurs, la formule précédente devient:
  Instruction SAS ⇔ $T = \text{TINV}(1-\alpha/2,n-p)$

En pratique si on choisit le risque $\alpha=5\%$ et si $n$ est assez grand ($n>30$) pour approcher la loi de Student par la loi Normale, alors l'intervalle de confiance de $\beta_0$ à 95% est donné par :

$$[C_{0.95}(\beta_0)=[b_0 - 1.96 \cdot s(b_0); b_0 + 1.96 \cdot s(b_0)]$$
**Interprétation**
Si la valeur 0 est dans l'intervalle de confiance de $\beta_0$, alors la droite de régression passe par l'origine.

**Exemple d’estimation des paramètres avec Proc REG**
Sur l’exemple de la Taille en fonction du Poids

**Programme SAS**

```sas
Proc REG data=libreg.tailpoid outest=TableSortie;
title 'Régression de la Taille en fonction du Poids';
model y=x ;
proc Print;title "Table de l'option outest";
run;
```

**Sortie de Proc REG**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>280.52918</td>
<td>280.52918</td>
<td>42.35</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>18</td>
<td>119.22082</td>
<td>6.62338</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>19</td>
<td>393.75000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Parameter Estimates**

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|-------------------|----------------|---------|------|---|
| Intercept| 1  | 145.30394         | 2.1384         | 53.79   | <.0001 |
| X        | 1  | 0.17030           | 0.02917        | 5.61    | <.0001 |

**Interprétation du test de la signification globale de la régression**

La statistique $F = \frac{MS \text{ model}}{MS \text{ error}} = \frac{280.529}{6.62} = 42.35$ indique que globalement le modèle avec le régresseur Poids améliore la prévision de la Taille, par rapport à la moyenne seule dans le modèle.
Interprétation des estimations des paramètres

L'estimateur de $\beta_0$ a pour valeur 145.98994. Son écart type vaut 2.71384.

La statistique de Test $t \text{ value} = \frac{145.98994}{2.71384} = 53.79$ et sa $p \text{ value}$ associée est bien inférieure au seuil 0.05.
On rejette l'hypothèse que $\beta_0 = 0$ avec une grande confiance.

Même raisonnement pour l'estimateur de $\beta_1$ qui a pour valeur 0.17030.

**Note**
Dans le cas de la régression simple la statistique de test de l'estimateur de $\beta_1$ et lié à $F: F = (t \text{ value})^2$

Dans la table en sortie par l'option `outest=` SAS enregistre RMSE et les valeurs des paramètres.
SAS n'imprime pas en standard les intervalles de confiance des paramètres mais on peut les récupérer dans cette table en sortie, en utilisant l’option `outest=` et le mot clé `Tableout`.

**Programme SAS**

```sas
Proc REG data=libreg.tailpoid outest=TableSortie Tableout;
title 'Régression de la Taille en fonction du Poids ';
model y=x  ;
proc PRINT data=TableSortie;
title "Table produite par l'option outest avec le mot clé Tableout";
run;
```

<table>
<thead>
<tr>
<th>Ob</th>
<th><em>MODEL</em></th>
<th><em>TYPE</em></th>
<th><em>DEPVAR</em></th>
<th><em>RMSE</em></th>
<th>Intercept</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MODEL1</td>
<td>PARS</td>
<td>Y</td>
<td>2.57359</td>
<td>145.980</td>
<td>0.17030</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>MODEL1</td>
<td>STDErr</td>
<td>Y</td>
<td>2.57359</td>
<td>2.714</td>
<td>0.02617</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>MODEL1</td>
<td>T</td>
<td>Y</td>
<td>2.57359</td>
<td>53.795</td>
<td>6.50000</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>MODEL1</td>
<td>PVALUE</td>
<td>Y</td>
<td>2.57359</td>
<td>0.000</td>
<td>0.00000</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>MODEL1</td>
<td>U50</td>
<td>Y</td>
<td>2.57359</td>
<td>140.290</td>
<td>0.11502</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>MODEL1</td>
<td>U55B</td>
<td>Y</td>
<td>2.57359</td>
<td>151.091</td>
<td>0.22520</td>
<td></td>
</tr>
</tbody>
</table>

Les lignes L95B et U95B donnent les intervalles de confiance à 95% des paramètres.

**1.4.4. Précision sur l'estimation de Y**

On a vu que pour chaque valeur $X$, fixée, la *vraie droite de régression* était le lieu de l'espérance (i.e. la *valeur moyenne*) de $Y$ et que les $Y$ devaient théoriquement se distribuer selon une loi normale centrée sur cette droite avec une variance théorique $\sigma^2$. 

©WP CES 2007.47 - Premiers pas en régression linéaire avec SAS- Confais Le Guen - 35
Pour évaluer la précision sur l'estimation de Y on aura deux optiques à considérer, soit on s'intéressera à l'intervalle de confiance autour de l'estimation de la droite de régression, soit on s'intéresse à l'intervalle de prévision de Y en fonction de X.

**Intervalle de confiance autour de l'estimation de la droite de régression**

Soit $X_k$ représentant un niveau particulier de X pour lequel nous voulons estimer la valeur moyenne de Y. $X_k$ peut être une valeur connue dans l'échantillon, ou une autre valeur de la variable régresseur non repérée dans l'échantillon. La réponse moyenne quand $X=X_k$ est notée $E(Y_k)$. L'estimateur de $E(Y_k)$ est noté $\hat{Y}_k$.

Il faut voir la distribution d'échantillonnage de $\hat{Y}_k$, comme la distribution que l'on obtiendrait si on effectuait des mesures répétées en $X_k$.

♦ **Calcul de l'erreur-type sur $\hat{Y}_k$**

On a vu que l'estimation de $E(Y_k)$ est donnée par :

$$\hat{Y}_k = \bar{Y} + b_1(X - \bar{X}_k)$$

Plaçons-nous en un point $X_k$ et calculons la variance de $\hat{Y}_k$ :

$$\text{Var}(\hat{Y}_k) = \text{Var}(\bar{Y}) + (X_k - \bar{X})^2 \text{Var}(b_1)$$

Pour pouvoir calculer la variance il faut faire des suppositions sur les termes de cette expression.

Comme précédemment on suppose que les $X_i$ sont non aléatoires.

Seuls la moyenne des $Y_i$ et le coefficient $b_1$ sont des variables aléatoires. On peut montrer de plus que la covariance entre $\bar{Y}$ et le coefficient $b_1$ est nulle.

**Suppositions** pour calculer la variance de $\hat{Y}_k$

Si les $X_i$ sont non aléatoires

Si les $Y_i$ sont non corrélés et de même variance $\sigma^2$

Et comme par construction $\text{Cov}(\bar{Y}, b_1) = 0$

Alors :

$$\text{Var}(\hat{Y}_k) = \text{Var}(\bar{Y}) + (X_k - \bar{X})^2 \text{Var}(b_1) = \frac{\sigma^2}{n} + (X_k - \bar{X})^2 \frac{\sigma^2}{\sum (X_i - \bar{X})^2}$$

Comme précédemment, on ne connaît pas la variance théorique $\sigma^2$ de Y. Il faut l'estimer.

---

11 voir démonstration dans NETER, WASSERMAN, KUTNER pp75-77.
Supposition

Si le modèle postulé est le modèle correct
Alors $\sigma^2$ peut être estimé par les erreurs entre les
$Y$ observés et $\hat{Y}$

$$s^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-2} = \text{MSE}$$

L'estimateur de l'erreur-type de $\hat{Y}_k$ devient :

$$s(\hat{Y}_k) = s \left[ \frac{1}{n} + \frac{(X_k - \bar{X})^2}{\sqrt{\sum(X_i - \bar{X})^2}} \right]^{1/2}$$

♦ Calcul de l'intervalle de confiance de $\hat{Y}_k$

On montre que pour un modèle de régression la statistique $\frac{\hat{Y}_k - E(Y)}{s(\hat{Y}_k)}$ suit une
distribution de Student à (n-2) degrés de liberté.
La vraie valeur moyenne $\mu_k$ de $Y$ pour un $X_k$ a une probabilité égale à $(1-\alpha)$
d'appartenir à l'intervalle de confiance :

$$[\hat{Y}_k - t_{1-\alpha/2}.s(\hat{Y}_k), \hat{Y}_k + t_{1-\alpha/2}.s(\hat{Y}_k)]$$

L'intervalle de confiance de $\hat{Y}_k$ se matérialise par deux lignes courbes, des
hyperboles, comme le montre la figure 1.9.

\[Figure 1.9 \quad \text{Intervalle de confiance à 95% de la moyenne des Tailles selon les valeurs des Poids}\]

A propos de la largeur de l'intervalle de confiance, on peut faire les remarques suivantes :
• La largeur varie en fonction de \((X_k - \bar{X})\).
• La largeur est minimum au point \(X_k = \bar{X}\).
  C'est dire que la précision est la **meilleure**, au centre de gravité du nuage des points.
• La largeur croît lorsqu'on s'éloigne du centre de gravité. La précision est la plus **mauvaise** aux extrémités du nuage de points.

**Intervalle de prévision de \(Y\) sachant \(X\)**

Ici on s'intéresse à la prévision d'une *nouvelle* observation individuelle de \(Y\) pour une valeur \(X_k\), de la variable \(X\) et non pas à la valeur moyenne de \(Y\).

Dans ce cas, la variance de \(Y\) a deux composantes :

1. la variance de la position centrale de la distribution d'échantillonnage de \(\hat{Y}_k\), cf. calcul réalisé au paragraphe précédent
2. la variance \(\sigma^2\) de la distribution de \(Y\) autour de sa position centrale au point \(X = X_k\). Comme précédemment, on estime \(\sigma^2\) par \(s^2\).

Pour une explication visuelle de cette décomposition\(^{12}\) voir la figure 1.10.

![Figure 1.10 Illustration de la prédiction d'une nouvelle observation individuelle de \(Y\)](image)

L'estimateur de l'erreur-type de \(Y\) sachant \(X\) devient :

\[
s^2 + s^2(\hat{Y}_k) = s^2 \left[ 1 + \frac{1}{n} + \frac{(X_k - \bar{X})^2}{\sum(X_i - \bar{X})^2} \right]
\]

L'intervalle de confiance d'une prévision de \(Y\) sachant \(X\) se matérialise là aussi par deux lignes courbes décalées d'une distance "\(s\)" par rapport à l'intervalle de confiance calculé pour la moyenne de \(Y_k\).

\(^{12}\)Source : NETER, WASSERMAN et KUTNER, p82.

©WP CES  2007.47 - Premiers pas en régression linéaire avec SAS- Confais Le Guen -
Les remarques faites précédemment sur l’estimation de la moyenne de $Y_k$ sont les mêmes que celles faites pour une observation individuelle. A savoir, la largeur de l'intervalle de confiance varie en fonction de $(X_k - \bar{X})$, c’est au centre de gravité du nuage de points que la précision est la meilleure, et aux extrémités du nuage de points que cette précision est la plus mauvaise.

**Figure 1.11 Intervalle de confiance à 95% des prévisions individuelles des Tailles**

Sur la figure 1.11 on voit que l'intervalle de confiance des prévisions individuelles est évidemment plus grand que l'intervalle de confiance des moyennes théoriques.

**Attention**
En prévision et dans un cadre temporel, on cherche à estimer aux extrémités de la plage de variation de $X$, or c'est justement là que la précision est la moins bonne!

**Exemple avec les options CLI CLM de la Proc REG**
 Les options CLI (Confidence Limit Individual) et CLM (Confidence Limit Mean) de l'instruction model de Proc REG donnent ces intervalles de confiance. Pour sauvegarder ces valeurs dans une table SAS il faut utiliser l'instruction Output.

**Programme SAS**

```sas
Proc REG data=libreg.tailpoid ;
title 'Régression de la Taille en fonction du Poids ';
model y=x /CLI CLM  ;
Output Out=Table2 Predicted=Pred residual=Residu
   LCL=Borne_Inf_ind UCL=Borne_Sup_Ind
   LCLM=Borne_Inf_Moy UCLM=Borne_Sup_Moy;
proc PRINT data=Table2 ;title "Table produite par l'instruction OUTPUT";
run;
```
Sortie de PROC REG

<table>
<thead>
<tr>
<th>Obs</th>
<th>Variable</th>
<th>Predicted Value</th>
<th>Std Error Mean Predict</th>
<th>95% CL Mean</th>
<th>95% CL Predict</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>45</td>
<td>152.824</td>
<td>150.549</td>
<td>157.096</td>
<td>147.563</td>
<td>160.145</td>
</tr>
<tr>
<td>2</td>
<td>58</td>
<td>159.273</td>
<td>157.510</td>
<td>161.029</td>
<td>152.568</td>
<td>164.501</td>
</tr>
<tr>
<td>3</td>
<td>65</td>
<td>160.466</td>
<td>158.950</td>
<td>161.372</td>
<td>154.853</td>
<td>166.678</td>
</tr>
<tr>
<td>4</td>
<td>85</td>
<td>162.169</td>
<td>159.466</td>
<td>161.737</td>
<td>154.853</td>
<td>166.678</td>
</tr>
<tr>
<td>5</td>
<td>85</td>
<td>162.169</td>
<td>158.938</td>
<td>161.372</td>
<td>154.853</td>
<td>166.678</td>
</tr>
<tr>
<td>6</td>
<td>85</td>
<td>162.169</td>
<td>159.466</td>
<td>161.737</td>
<td>154.853</td>
<td>166.678</td>
</tr>
<tr>
<td>7</td>
<td>85</td>
<td>162.169</td>
<td>159.466</td>
<td>161.737</td>
<td>154.853</td>
<td>166.678</td>
</tr>
<tr>
<td>8</td>
<td>100</td>
<td>165.309</td>
<td>161.029</td>
<td>161.372</td>
<td>154.853</td>
<td>166.678</td>
</tr>
<tr>
<td>9</td>
<td>100</td>
<td>165.309</td>
<td>161.029</td>
<td>161.372</td>
<td>154.853</td>
<td>166.678</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>165.309</td>
<td>161.029</td>
<td>161.372</td>
<td>154.853</td>
<td>166.678</td>
</tr>
</tbody>
</table>

Lecture :

Les options CLM CLI donnent pour chaque observation, les valeurs :

- **Dependant variable : Y**
- **Predicted Value : \( \hat{Y} \)**
- **Std Error mean predict** : erreur-type au point \( X_i \)
- **95% CL Mean** : les 2 colonnes suivantes donnent les bornes inférieure et supérieure de l’intervalle de prédiction à 95% de la moyenne.
- **95% CL Predict** : les 2 colonnes suivantes donnent les bornes inférieure et supérieure de l’intervalle pour une prédiction individuelle.
- **Residual** : résidu

L’instruction Output avec les mots clés LCL UCL LCLM UCLM permettent de récupérer ces statistiques dans une table SAS:
2. La régression linéaire multiple

Dans ce chapitre nous reprenons les concepts de la régression linéaire simple pour les formaliser et les étendre à la régression multiple. Nous présentons les différentes formes de décomposition de sommes de carrés (Sum of Squares) et commentons les résultats obtenus avec la procédure REG.

2.1. Le critère des moindres carrés

Tout comme en régression linéaire simple; la régression linéaire multiple cherche à approximer une relation fonctionnelle trop complexe en général, par une fonction mathématique simple telle qu'une équation de la forme:

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \cdots + \beta_pX_p + \epsilon$$

Reprenons le résumé des concepts de la régression linéaire présenté au chapitre 1. L'équation de régression ou modèle postulé, met en relation:

- Y : variable réponse (à expliquer ou variable dépendante).
- Xj : variables régresseurs (explicatives ou variables indépendantes).

Cette équation est linéaire par rapport aux paramètres (coefficients de régression) $\beta_0, \beta_1, \ldots, \beta_p$. Le modèle est dit linéaire. Ces paramètres sont inconnus, on les estime en minimisant le critère des moindres carrés (MCO ou Ordinary Least Squares).

Le critère des moindres carrés correspond à la minimisation de la somme des carrés des écarts (SC Erreur en français, SS Error en anglais) entre Y observé et Y estimé par l'équation de régression.

Y estimé est noté $\hat{Y}$.

$$\hat{Y}_i = b_0 + b_1X_{i1} + \cdots + b_pX_{ip}$$

avec:

- Y : variable réponse
- Xj : p variables régresseurs, j=1,..,p
- i indice de l'observation courante, i=1,...n
- n le nombre d'observations.

Les valeurs qui minimisent ce critère sont des estimations $b_0,b_1,\ldots,b_p$ des paramètres $\beta_0,\beta_1,\ldots,\beta_p$ inconnus.

Estimation des paramètres du modèle

Dans le cas d'un modèle à p variables régresseurs le critère des moindres carrés s'écrit:
Les valeurs des $\beta$ qui minimisent ce critère seront les solutions $b_0$, $b_1$, ...$b_p$ du système linéaire de $(p+1)$ équations à $(p+1)$ inconnues.

\[ S(b_0, \ldots, b_p) = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_1 - \ldots - \beta_p X_p)^2 \]

Avec

\[ S_{b_j} = \sum_{i=1}^{n} (X_{ji} - \bar{X}_j)(X_{ji} - \bar{X}_j) \quad \text{pour } k,j=1,2,\ldots,p \]

\[ S_{xy} = \sum_{i=1}^{n} (X_{ki} - \bar{X}_k)(Y_i - \bar{Y}) \quad \text{pour } k=1,2,\ldots,p \]

Pour résoudre un tel système linéaire les mathématiciens ont développé le calcul (algèbre) matriciel qui permet une présentation et des traitements compacts de grands tableaux de données. La notation matricielle est donc devenue l’unique moyen d’appréhender la régression multiple. Cependant cette présentation cache bien des difficultés du point de vue des résolutions numériques sur données réelles.

Les estimateurs des moindres carrés estiment les paramètres inconnus $\beta_0, \beta_1, \ldots, \beta_p$ avec une certaine précision. Sous les suppositions que les erreurs sont indépendantes et identiquement distribuées selon une loi normale, les estimateurs MCO sont centrés sur une valeur à laquelle est associé un intervalle de confiance. L’intervalle de confiance dépend de l’adéquation du modèle aux données, adéquation qui dépend des erreurs inconnues $\varepsilon_i$ :

\[ \varepsilon_i = Y_i - E(Y_i) \]

### 2.2. Formalisation de la régression linéaire multiple

En notation matricielle :

- $Y$ est le vecteur colonne des $n$ observations de la variable réponse
- $X(n,p)$ la matrice des observations des $p$ vecteurs $X_i$, chacun de dimension $(n,1)$.

A cette matrice on ajoute en première colonne un vecteur constitué uniquement de 1. Ce vecteur correspond à la constante $X_0$. La matrice $X$ est alors de dimension $(n,p+1)$.

Cette représentation permet de traiter la constante $X_0$ comme une variable explicative.
• $\beta$ est le vecteur colonne des (p+1) coefficients de régression ou paramètres inconnus $\beta_i$.

• $\varepsilon$ représente le vecteur des erreurs.

Le modèle s'écrit:\n\[ Y = X\beta + \varepsilon \]

$Y$ estimé par le modèle de régression s'écrit:\n\[ \hat{Y} = X\hat{\beta} = XB \]

Le vecteur colonne $\hat{\beta}$ (noté aussi B) représente le vecteur des estimateurs $b_i$ des moindres carrés des paramètres inconnus $\beta$.

Les notations matricielles permettent d'écrire simplement le système à résoudre pour trouver les coefficients $b_i$ qui minimisent le critère des moindres carrés:
\[ (X'X)B = (X'Y) \]

$X'$ désignant la matrice transposée de $X$.

Le vecteur $B$ des coefficients solution s'obtient en inversant la matrice $(X'X)$:
\[ B = (X'X)^{-1}(X'Y) \]

La résolution de ce système n'est pas toujours possible. Cette résolution est liée à la possibilité d'inversion de la matrice $(X'X)$.

Supposons que 2 variables $X_i$ et $X_j$ soient corrélées entre elles c'est-à-dire qu'il existe une relation linéaire permettant de passer de $X_i$ à $X_j$ on a alors 2 lignes de la matrice $(X'X)$ qui sont proportionnelles et lorsque l'on veut résoudre le système il ne reste plus que p équations indépendantes et toujours (p+1) inconnues à trouver.

Le système est indéterminé, il existe une infinité de solutions.

Les variances des estimateurs ($b$) sont les éléments diagonaux de la matrice de variance-covariance des $X$ inversée multipliés par la variance des erreurs $\sigma^2$.
\[ \sigma^2(b) = \sigma^2(X'X)^{-1} \]

Comme pour la régression simple $\sigma^2$ est estimé par:\n\[ MSE = \frac{SS error}{n-p-1} \]

**La matrice H**

A partir de l'expression du vecteur B des estimateurs des coefficients on peut calculer l'estimation de Y:

\[
\hat{Y} = XB \\
\hat{Y} = X(X'X)^{-1}X'Y \\
\hat{Y} = HY \\
\text{avec} \\
H = X(X'X)^{-1}X'
\]

Cette matrice H - *H comme Hat matrice* - qui ne comporte que des données relatives aux variables régresseurs va jouer un rôle important, et son usage sera développé chapitre 4.

### 2.3. Exemples de régression linéaire multiple avec Proc REG

#### 2.3.1. Présentation des données

Pour présenter la régression multiple avec quelques options de Proc REG, nous avons repris l'exemple de la chenille processionnaire du pin traité dans l'ouvrage de TOMASSONE & al. Cet exemple est fréquemment analysé dans la littérature française (voir FOUCART, AZAIS-BARBET). On pourra ainsi, avec leurs ouvrages, poursuivre des analyses plus complexes de ces données.

Le fichier de données est composé de 33 placettes où sont plantés des arbres infectés par des nids de chenille « processionnaire du pin », une variable réponse (X11 et sa transformée en Log et dix variables régresseurs potentiels (X1-X10).

« Les expérimentateurs souhaitent connaître l'influence de certaines caractéristiques de peuplements forestiers (variables régresseurs X1-X10) sur le développement de la chenille processionnaire du pin (variable réponse X11 ou son logarithme) ».

| X11 : Nombre de nids de processionnaires par arbre d'une placette. |
| Log = Log(X11), transformation de la variable X11 par son logarithme |
| X1 : Altitude (en mètre) |
| X2 : pente (en degré) |
| X3 : nombre de pins dans une placette de 5 ares |
| X4 : hauteur de l’arbre échantillonné au centre de la placette |
| X5 : diamètre de cet arbre |
**2.3.2. Régression linéaire multiple avec Proc REG sans options**

Nous étudions le modèle linéaire de la variable Log en fonction des 4 régresseurs X1, X2, X4, X5.

**Étape 1**: Graphique de la matrice de diagrammes de dispersion (Scatter Plot avec SAS/INSIGHT)

**Étape 2**: Analyse des corrélations entre les variables

**Étape 3**: Régression multiple

Nous utilisons SAS/INSIGHT qui est beaucoup plus efficace pour obtenir des graphiques exploratoires, (voir en annexe 2 le mode d’emploi succinct de SAS/INSIGHT).

**Programme SAS**

```sas
/* étape 2 */
proc CORR data=libreg.chenilles; title 'Corrélation de X1 X2 X4 X5 avec Log';
var X1 X2 X4 X5 Log;
run;
/*étape 3 */
```

©WP CES 2007.47 - Premiers pas en régression linéaire avec SAS- Confais Le Guen - 45
proc REG data=libreg.chenilles;
title 'Régression de LOG avec X1 X2 X4 X5 sans options';
model Log=X1 X2 X4 X5;
run;

Figure 2.1: Matrice des diagrammes de dispersion des variables croisées 2*2.
Sur la diagonale sont affichées les valeurs min et max pour chaque variable.

Sortie SAS de PROC CORR

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Skew</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>33</td>
<td>1181</td>
<td>129.63628</td>
<td>4306</td>
<td>1075</td>
<td>1575</td>
</tr>
<tr>
<td>X2</td>
<td>33</td>
<td>29.63628</td>
<td>7.03012</td>
<td>958.00000</td>
<td>24.00000</td>
<td>65.00000</td>
</tr>
<tr>
<td>X4</td>
<td>33</td>
<td>4.45152</td>
<td>1.64077</td>
<td>146.90000</td>
<td>2.40000</td>
<td>65.00000</td>
</tr>
<tr>
<td>X5</td>
<td>33</td>
<td>1.25362</td>
<td>4.30552</td>
<td>593.30000</td>
<td>56.00000</td>
<td>21.00000</td>
</tr>
<tr>
<td>Log</td>
<td>33</td>
<td>-0.81328</td>
<td>1.24494</td>
<td>-26.38256</td>
<td>-5.58656</td>
<td>1.09861</td>
</tr>
</tbody>
</table>

Pearson Correlation Coefficients, N = 33
Prob > |r| under HO: |r|=0

<table>
<thead>
<tr>
<th></th>
<th>X1</th>
<th>X2</th>
<th>X4</th>
<th>X5</th>
<th>Log</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>1.00000</td>
<td>0.12052</td>
<td>0.32105</td>
<td>0.28377</td>
<td>-0.53361</td>
</tr>
<tr>
<td></td>
<td>0.5041</td>
<td>0.0585</td>
<td>0.1995</td>
<td>-0.0014</td>
<td></td>
</tr>
<tr>
<td>X2</td>
<td>0.12052</td>
<td>1.00000</td>
<td>0.13509</td>
<td>0.11342</td>
<td>-0.42944</td>
</tr>
<tr>
<td></td>
<td>0.5041</td>
<td>0.0585</td>
<td>0.1995</td>
<td>-0.0014</td>
<td></td>
</tr>
<tr>
<td>X4</td>
<td>0.32105</td>
<td>0.13509</td>
<td>1.00000</td>
<td>0.06455</td>
<td>-0.42525</td>
</tr>
<tr>
<td></td>
<td>0.0685</td>
<td>0.0031</td>
<td>0.0501</td>
<td>0.0136</td>
<td></td>
</tr>
<tr>
<td>X5</td>
<td>0.28377</td>
<td>0.11342</td>
<td>0.06455</td>
<td>1.00000</td>
<td>-0.20094</td>
</tr>
<tr>
<td></td>
<td>0.5041</td>
<td>0.0585</td>
<td>0.1995</td>
<td>-0.0014</td>
<td></td>
</tr>
<tr>
<td>Log</td>
<td>-0.53361</td>
<td>-0.42944</td>
<td>-0.42525</td>
<td>-0.20094</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>0.5041</td>
<td>0.0585</td>
<td>0.1995</td>
<td>-0.0014</td>
<td></td>
</tr>
</tbody>
</table>

Le graphique des diagrammes de dispersion de la figure 2.1, donne une image des liaisons entre toutes les variables X1, X2, X4, X5, Log. On voit d’un coup d’œil que les variables X4 et X5 sont très liées. Le coefficient de corrélation vaut 0.90466. D’autre part la variable réponse Log est liée négativement à tous les régresseurs. La matrice de Scatter Plot est un complément utile à l’analyse de la matrice des coefficients de corrélation. Elle permet aussi de repérer les points atypiques (outliers) en X et en Y.
Sortie SAS de PROC REG sans options

### Tableau de l'Analyse de Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>32.09265</td>
<td>12.88</td>
<td>.0001</td>
</tr>
<tr>
<td>Error</td>
<td>28</td>
<td>17.50330</td>
<td>.52512</td>
<td></td>
</tr>
<tr>
<td>Corrected</td>
<td>32</td>
<td>49.59603</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Root MSE = 0.79065
- R-Square = 0.6471
- Adj R-Sq = 0.5967
- Coef Var = 97.21883

### Estimations des paramètres

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|--------------------|----------------|---------|------|---|
| Intercept| 1  | 7.73214            | 1.48858        | 5.19    | .0001|
| X1       | 1  | -0.00392           | 0.00115        | -3.42   | .0019|
| X2       | 1  | -0.05734           | 0.01939        | -2.96   | .0062|
| X4       | 1  | -1.35614           | 0.31983        | -4.24   | .0002|
| X5       | 1  | 0.28306            | 0.07626        | 3.71    | .0009|

Cette sortie est analogue à celle de la régression simple, on retrouve les mêmes informations explicitées au chapitre 1.

### Lecture du test global dans le tableau de l'Analyse de Variance

- **F value = 12.83 avec p value <0.001** rejet de H0 tous les paramètres ne sont pas tous nuls
- **R² = 0.6471** 64% de la variabilité de Y est expliquée par le modèle.

### Lecture des paramètres

- Intercept $b_0 = 7.73214$ $s(b_0) = 1.48858$
- Coefficient de X1 $b_1 = -0.00392$ $s(b_1) = 0.00115$
- Coefficient de X2 $b_2 = -0.05734$ $s(b_2) = 0.01939$
- Coefficient de X4 $b_4 = -1.35614$ $s(b_4) = 0.31983$
- Coefficient de X5 $b_5 = 0.28306$ $s(b_5) = 0.07626$

Toutes les $p$-value associées aux estimateurs des paramètres sont <0.05, on rejette l’hypothèse de nullité pour chacun des coefficients de la régression. Cependant le coefficient de X5 est positif alors que la corrélation (X5, Log) est négative.

Peut-on alors parler d’un effet positif de cette variable X5 sur la variable réponse Log ?

La corrélation entre X4 et X5 provoque une instabilité des valeurs des coefficients.

### 2.4. TYPE I SS et TYPE II SS de Proc REG

Nous verrons d’abord la définition des statistiques Type I SS et Type II SS relatif à un paramètre puis les tests partiels relatifs à plusieurs paramètres.
2.4.1. Définition de TYPE I SS et TYPE II SS

Reprenons l'équation fondamentale de l'analyse de la variance :

\[
SS \text{ Total} = SS \text{ Model} + SS \text{ Error}
\]

SS total qui représente la somme des carrés des écarts entre Y et sa moyenne est invariant quel que soit le nombre de variables régresseurs p dans le modèle. Lorsqu'on introduit une nouvelle variable régresseur dans un modèle SS Model augmente et donc SS Error diminue de la même quantité.

Pour juger de la contribution d'une variable régresseur à la réduction de SS Error, la Proc REG calcule pour chaque variable du modèle, deux sortes de SS Error.

- **TYPE I SS** : représente la réduction de SS Error liée à la variable lorsqu'elle est introduite séquentiellement dans le modèle.
- **TYPE II SS** : représente la réduction de SS Error liée à la variable lorsqu'elle est introduite la dernière dans le modèle.

**• TYPE I SS**
Soit le modèle complet définit avec les p régresseurs de l'instruction MODEL de Proc REG :

\[
IC_{.95}(\beta_i) = b_i - 1.96 \cdot s(b_i) b_i + 1.96 \cdot s(b_i)
\]

Soit d'autre part le modèle restreint aux \( k \) premiers régresseurs :

\[
Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon
\]

Pour la \( k \)ème variable, TYPE I SS correspond à la différence entre SSerror du modèle à \( (k-1) \) régresseurs et SSerror du modèle à \( k \) régresseurs.

**Attention** : Le TYPE I SS d'une variable dépend de l'ordre de la variable dans l'instruction MODEL de Proc REG.

A TYPE I SS peut être associée une statistique de test, la \( F \) Value ou F de Fisher-Snedecor, et sa \( p \)-value, niveau de significativité du test. Le calcul de la \( F \) value et de sa \( p \) value associée n'est pas réalisé dans Proc REG\(^1\).  

• **F VALUE : statistique de Fisher-Snedecor**

Cette statistique de test vaut :

\[
F \text{ VALUE} = \frac{\text{TYPEI SS}}{\text{MS Error}}
\]

Le numérateur TYPE I SS correspond à la réduction de SS error lorsque l'on passe du modèle à \( (k-1) \) régresseurs -la variable étudiée étant exclue- au modèle à \( k \) régresseurs.

Le dénominateur MS Error correspond au modèle complet à \( p \) régresseurs.

La statistique de test \( F \) value ainsi définie permet de tester l'hypothèse nulle de la \( k \)ème variable.

\(^1\) La statistique de Fisher-Snedecor \( F \) value et son niveau de significativité \( p \)-value pour Type I SS et TYPE II SS sont disponibles dans Proc GLM et dans SAS/INSIGHT.

©WP CES 2007.47 - Premiers pas en régression linéaire avec SAS- Confais Le Guen - 48
Hypothèse à tester
On veut tester si le paramètre $\beta_k = 0$.

H$_0$ : $\beta_k = 0$ contre
H$_a$ : $\beta_k \neq 0$

La statistique de test F value est sous H$_0$ une valeur observée d'une variable F de Fisher-Snedecor à 1 et (n-p-1) degrés de liberté. L'hypothèse nulle doit être rejetée au niveau $\alpha$ lorsque :

$$F_{\text{observé}} \geq F_{1, n-p-1}$$

où $F_{1, n-p-1}$ représente le quantile d'ordre $(1 - \alpha)$ de la loi de Fisher-Snedecor à (1) et (n-p-1) degrés de liberté.

Règle de décision

| Si $F$ value $\geq F_{1, n-p-1}$ Alors H$_0$: $\beta_k = 0$ doit être rejeté au niveau $\alpha$ |

Raisonnement

Au seuil $\alpha$% la valeur maximum atteinte par F sous l'hypothèse nulle H$_0$ : $\beta_k = 0$ est $F_{1, n-p-1}$, si donc $F$ value est supérieure on rejette l'hypothèse nulle. La variable contribue significativement à la réduction de SS Error, lorsqu'elle est entrée en dernier dans le modèle à k régresseurs.

• Prob > F
C'est la p-value associée à $F$ value.
On compare la p-value, au risque $\alpha$ choisi (par exemple $\alpha=0.05$).

Raisonnement sur la p-value

| Si p-value $\leq \alpha$ Alors on rejette l'hypothèse nulle $\beta_k = 0$ |

Interprétation : Si la probabilité (Prob>F) est faible (<0.05) la variable contribue significativement à la réduction de SS Error dans le modèle à k régresseurs.

• TYPE II SS
Soit le modèle complet défini avec p régresseurs dans l'instruction MODEL de Proc REG :

$$\text{IC}_{0.05}(\beta_i) = [b_i - 1.96 \cdot s(b_i); b_i + 1.96 \cdot s(b_i)]$$

Pour la $k$ème variable, TYPE II SS correspond à la différence entre SSerror du modèle à (p-1) régresseurs (le $k$ème régresseur étant exclu) et SSerror du modèle complet à p régresseurs.
A TYPE II SS peut être associée une statistique de test, la *F value* ou F de Fisher-Snedecor, et sa *p-value*, niveau de significativité du test.

**Remarque** : Par construction TYPE I SS et TYPE II SS de la dernière variable du modèle ont la même valeur.

Le calcul de la *F value* et de sa *p value* associée n'est pas réalisé dans Proc REG, il faut faire le calcul à la main, ou utiliser l'option de Proc REG, SELECTION= FORWARD ou BACKWARD qui donne les *F value* de TYPE II à chaque pas.

- **F VALUE** : statistique de Fisher-Snedecor

Cette statistique de test vaut : \[ F_{VALUE} = \frac{\text{TYPE II SS}}{\text{MS Error}} \]

Le numérateur TYPE II SS correspond à la réduction de SS error lorsque l'on passe du modèle à (p-1) régresseurs –le régresseur étudié étant exclu- au modèle complet à p régresseurs.

Le dénominateur MS Error correspond au **modèle complet**, avec les p régresseurs.

La statistique de test *F value* ainsi définie permet de tester l'hypothèse nulle du "k"ème régresseur, lorsqu'il entre en dernier dans le modèle. C'est un F dit **partiel**.

**Hypothèse à tester**

On veut tester si le paramètre \( \beta_k = 0 \).

- \( H_0 : \beta_k = 0 \) contre
- \( H_a : \beta_k \neq 0 \)

La statistique de test *F value* est sous \( H_0 \) une valeur observée d'une variable F de Fisher-Snedecor à 1 et (n-p-1) degrés de liberté. L'hypothèse nulle doit être rejetée au niveau \( \alpha \) lorsque :

\[ \text{Fobservé} \geq F_{1,n-p-1,\alpha} \]

où \( F_{1,n-p-1,\alpha} \) représente le quantile d'ordre \( (1-\alpha) \) de la loi de Fisher-Snedecor à (1) et (n-p-1) degrés de liberté.

**Règle de décision**

\[ \text{Si } F_{value} \geq F_{1,n-p-1,\alpha} \]

Alors \( H_0 : \beta_k = 0 \) doit être rejetée au niveau \( \alpha \).
**Raisonnement**

Au seuil $\alpha$% la valeur maximum atteinte par $F$ sous l'hypothèse nulle $H_0 : \beta_k = 0$ est $F_{1, n-p-1}$, si donc $F$ value est supérieure on rejette l'hypothèse nulle.

Le régresseur contribue significativement à la réduction de SS Error, lorsqu'il est entré en **dernier** dans le modèle.

**Remarque** : $F$ value $= T^2$, avec $T$ représentant la valeur du test de Student associé au paramètre.

**• Prob > F**

C'est la $p$-value associée à $F$ value. On compare la $p$-value, au risque $\alpha$ choisi (par ex : $\alpha=0.05$).

**Raisonnement sur la p-value**

| Si $p$-value $\leq \alpha$  
| Alors on rejette l'hypothèse nulle $\beta_k = 0$ |

**Interprétation** : Si la probabilité (Prob>F) est faible, le régresseur contribue significativement à la réduction de SS Error, même lorsqu'il est entré en **dernier** dans le modèle complet à $p$ régressseurs.

**2.4.2. Interprétations conjointes de TYPE I SS et TYPE II SS**

**•** Lorsque le modèle ne comporte qu'une seule variable régresseur :
  
  TYPE I SS = TYPE II SS = SS modèle

**•** Lorsque le modèle comporte plus d'une variable régresseur :

Le TYPE I SS (lié au F séquentiel) dépend de l'ordre d'apparition des variables régresseurs dans l'instruction MODEL, tandis que TYPE II SS (lié au F partiel) n'en dépend pas.

Si pour un régresseur $X_i$, le F séquentiel et le F partiel sont plus grands que ceux des autres régresseurs, alors $X_i$ a une contribution plus grande puisqu'il réduit plus la variation de SS Error, que la variable soit entrée en séquence dans le modèle ou en dernier.

Si pour un régresseur $X_i$, le F séquentiel est significatif et le F partiel ne l'est plus c'est qu'il y a des colinéarités entre les régresseurs.

L’information apportée par ce régresseur est redondante par rapport à l’information apportée par les précédents régresseurs déjà introduits dans le modèle.

**2.4.3. Options SS1 et SS2 de l'instruction model de Proc REG**

Les options SS1 et SS2 de l'instruction model de PROC REG permettent d’obtenir les statistiques Type I SS et Type II SS.
Programme SAS

```sas
proc REG data=libreg.chenilles;
  title 'Régression de Log avec X1 X2 X4 X5 avec Options SS1 SS2 '; 
  model Log=X1 X2 X4 X5/ SS1 SS2;
run;
```

Sortie de Proc REG de SAS

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>Type I SS</th>
<th>Type II SS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>7.73211</td>
<td>1.48858</td>
<td>5.19</td>
<td>&lt;.001</td>
<td>21.02704</td>
<td>16.85629</td>
</tr>
<tr>
<td>X1</td>
<td>1</td>
<td>-0.005392</td>
<td>0.00115</td>
<td>-3.12</td>
<td>0.001</td>
<td>14.12216</td>
<td>7.30671</td>
</tr>
<tr>
<td>X2</td>
<td>1</td>
<td>-0.057344</td>
<td>0.01939</td>
<td>-2.96</td>
<td>0.006</td>
<td>6.70953</td>
<td>5.46832</td>
</tr>
<tr>
<td>X4</td>
<td>1</td>
<td>-1.35614</td>
<td>0.31983</td>
<td>-4.24</td>
<td>0.002</td>
<td>2.64867</td>
<td>11.23888</td>
</tr>
<tr>
<td>X5</td>
<td>1</td>
<td>0.28306</td>
<td>0.07626</td>
<td>3.71</td>
<td>0.000</td>
<td>8.61230</td>
<td>8.61230</td>
</tr>
</tbody>
</table>

Lecture

**Exemple pour X1 :**
- Type I SS = 14.12216 est la réduction de SS error lorsque la variable X1 est entrée la première dans le modèle (elle est alors la seule variable régressee).
- Type II SS = 7.30671 est la réduction de SS error lorsque la variable X1 est entrée la dernière dans le modèle.

Pour tester si cette réduction est significative il faut faire le calcul à la main, car dans cette sortie, Proc REG ne fournit pas les F value et les proba associées pour TYPE I SS et TYPE II SS :

\[
F \text{ value } = \frac{\text{TYPE I SS}}{\text{MS Error}} = \frac{14.1222}{0.62512} = 22.59
\]

\[
F \text{ value } = \frac{\text{TYPE II SS}}{\text{MS Error}} = \frac{7.30671}{0.62512} = 11.69
\]

On peut vérifier ces F value avec les sorties de SAS/INSIGHT.
Note : SAS/ INSIGHT nomme TYPE III tests ce qu’on a appelé TYPE II SS dans la Proc REG

On retrouve bien les calculs faits à la main (cf. F Stat =22.59 pour la F value de Type I SS et F Stat=11.69 pour la F value de Type II SS).

La variable X4 a un comportement bizarre, lorsqu’elle est entrée en 3ème rang dans le modèle elle est limite au niveau significativité (pvalue =0.0490), alors que son apport est très significatif lorsqu’elle est entrée la dernière (p value =0.0002)

La liaison entre X4 et X5 nous joue des tours !

2.4.4. Tester la nullité de r paramètres pour tester un sous modèle

Ce type d'analyse est d'usage courant en Econométrie. L'idée est de mettre à l'épreuve une approche théorique par une validation empirique. L'intérêt porte non sur l'estimation des paramètres mais sur la « spécification » du modèle. Par spécification on entend la recherche des variables-régresseurs intervenant dans la détermination de la variable « à expliquer » Y.

On veut tester la nullité de r (indices k à q) paramètres parmi les p, c'est à dire l'hypothèse nulle

\[ H_0 : \beta_k = \cdots = \beta_q = 0 \] contre

\[ H_a : \text{il y a parmi les } \beta_k \cdots \beta_q \text{ des coefficients non égaux à 0) (not all } \beta_k \cdots \beta_q \text{ equal to 0).} \]

Le modèle sans les r variables est appelé le modèle restreint par opposition au modèle complet à p variables.

Ici aussi on raisonne sur les réductions de SS error.

On note :

RRSS (Restricted Residual Sum of Squares) = Somme des carrés des résidus du modèle restreint

URSS (Unrestricted Residual Sum of Squares)=Somme des carrés des résidus du modèle complet.
L'hypothèse est testée en évaluant la statistique $F$ dite partielle$^{14}$

$$F = \frac{(RRSS - URSS)/r}{URSS/(n - p - 1)} = \frac{(RRSS - URSS)/r}{MSE}$$

La statistique de test $F$ est sous $H_0$ une valeur observée d'une variable $F$ de Fisher-Snedecor à $r$ et $(n-p-1)$ degrés de liberté. L'hypothèse nulle doit être rejetée au niveau $\alpha$ lorsque :

$$F_{\text{observé}} \geq F_{1-\alpha}(r,n-p-1)$$

où $F_{1-\alpha}(r,n-p-1)$ représente le quantile d'ordre $(1 - \alpha)$ de la loi de Fisher-Snedecor à $(r)$ et $(n-p-1)$ degrés de liberté.

**Règle de décision**

| Si $F_{\text{value}} \geq F_{1-\alpha}(r,n-p-1)$ | Alors $H_0 : \beta_k = \cdots = \beta_q = 0$ doit être rejetée au niveau $\alpha$ |

L'instruction **TEST** de la Proc REG réalise ces tests en fournissant la statistique de Fisher-Snedecor $F_{\text{value}}$ et son niveau de significativité $p-value$ associé, noté $Prob>F$.

**2.4.5. Exemple de test partiel avec PROC REG**

On veut tester si les 2 coefficients de $X_4$ et $X_5$ sont nuls.

$q_0 : \beta_4 = \beta_5 = 0$  

instruction SAS $\Rightarrow$ **test** $X_4=0$, $X_5=0$;

**Programme SAS**

```
proc REG data=libreg.chenilles;
   title "Test de l'Hypothèse nulle X4=0 et X5=0";
   model Log=X1 X2 X4 X5;
   test X4=0, X5=0;
```

- Pour le modèle restreint (sans $X_4$ et $X_5$) on a :

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>20.83168</td>
<td>10.41584</td>
<td>10.86</td>
<td>0.0003</td>
</tr>
<tr>
<td>Error</td>
<td>30</td>
<td>28.76444</td>
<td>0.95881</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>32</td>
<td>43.59603</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$^{14}$ La statistique de test $F$ ainsi calculée est appelée $F_{\text{partiel}}$ quand elle ne porte que sur un sous-ensemble de paramètres, pour la distinguer de la statistique $F$, qui porte sur l'ensemble des paramètres du modèle complet.
RRSS = 28.76434

- Pour le modèle complet :

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>32.09285</td>
<td>8.02316</td>
<td>12.88</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>28</td>
<td>17.50338</td>
<td>0.62512</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>32</td>
<td>49.59609</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

RRSS = 17.50338 avec DF=(n-p-1)=28

d'où

\[
F = \frac{(RRSS - URSS)/r}{URSS/(n-p-1) \times MSE} = \frac{(28.76434 - 17.50338)/2}{17.80338/28} = 5.63048 = 9.01
\]

Sortie SAS pour le test partiel

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Numerator</td>
<td>2</td>
<td>5.63048</td>
<td>9.01</td>
<td>0.0010</td>
</tr>
<tr>
<td>Denominator</td>
<td>28</td>
<td>0.62512</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

On trouve bien la valeur \( F \text{ Value} = 9.01 \) avec un niveau de significativité \( \text{Prob} > F \) de 0.0010.

**Conclusion**
Le niveau de significativité (0.0010) étant bien inférieur à 0.05, on rejette l'hypothèse nulle \( H_0: \beta_4 = \beta_5 = 0 \)

Il existe au moins un effet de \( X_4 \) et/ou de \( X_5 \) sachant \( X_1 \) et \( X_2 \) introduit dans le modèle.
2.5. Ce qu’il faut retenir des 'SS'

**Décomposition des SS : Sum of Squares**

SS Total = SS error + SS model

Lorsqu'on introduit une nouvelle variable dans un modèle :

**SS model** augmente de la même quantité que **SS error** décroît

Donc le coefficient de détermination R-square augmente toujours.

Cependant on n'améliore pas nécessairement la précision de l'estimation de Y.

En effet SS Error décroît mais \( s^2 = \text{MSE} = \frac{\text{SS Error}}{n - p - 1} \) peut croître donc augmenter la largeur de l'intervalle de confiance de Y estimé qui est proportionnel à MSE.

A la limite si le nombre de variables p + 1 (le 1 correspond à la variable constante X₀) est égal au nombre d'observations (n), l'équation de régression passera exactement par tous les points du nuage, l'ajustement sera parfait.

Dans ce cas SS Error vaut 0, et le coefficient de détermination \( R^2 \) vaut 1. Ce n'est plus de la statistique mais de la résolution d'équations !

D'autre part les colinéarités entre les régresseurs rendent les résultats instables. En augmentant le nombre de régresseurs on augmente les risques de colinéarités. Le chapitre 4 traitera de ce problème.

**Modèles "parcimonieux"**

Par sagesse, les statisticiens parlent de modèles « parcimonieux », pour signifier qu'un modèle doit comporter un nombre limité de variables par rapport au nombre d'observations, si on veut que le modèle ait une portée prévisionnelle et/ou explicative.
2.6. Les résidus

En un point d'observation i, l'écart entre Y observé et Y estimé par le modèle est le résidu au point i:

\[ e_i = Y_i - \hat{Y}_i \]

Ces résidus \( e_i \) sont vus comme les erreurs observées des vraies erreurs inconnues \( \varepsilon_i \):

\[ e_i = Y_i - E(Y_i) \]

Nous avons vu que les suppositions faites sur les \( \varepsilon_i \) pour élaborer les tests statistiques se résument ainsi « les erreurs doivent être indépendantes et identiquement distribuées selon une loi normale ».

Si le modèle est approprié aux données, les résidus observés \( e_i \) doivent refléter les propriétés des vraies erreurs inconnues \( \varepsilon_i \). C'est donc par le biais de l'analyse des résidus que l'on cherchera à valider le modèle de régression postulé.

Pour cela on effectuera différents graphiques des résidus en fonction de:

- Y la variable réponse
- Y estimé (\( \hat{Y} \))
- \( X_i \) les variables régresseurs
- la variable temporelle, si l'analyse statistique porte sur des séries chronologiques
- etc.

De même on étudiera la normalité des résidus, leur indépendance. En effet, les résidus contiennent d'une part un aléa d'espérance nulle et de variance \( \sigma^2 \), et d'autre part une information concernant l'inadéquation du modèle aux données (c'est-à-dire l'écart entre le modèle postulé et le modèle correct inconnu). Ce que l'on veut c'est que l'importance de cette deuxième partie soit moindre que celle due à l'aléa.

Pour cela on devra rechercher si dans les résidus il n'existe pas une structure organisée ou un contenu informationnel qui prouverait que le modèle postulé se différencie significativement du modèle correct.

Tous les tests sont faits en supposant que le modèle postulé est le modèle correct, si donc l'analyse des résidus prouve l'inadéquation du modèle postulé, les tests ne sont plus valables, ou sont biaisés.

Des orientations pour l'analyse critique des résidus seront données dans le chapitre 4.
Conclusion

Au cours des chapitres 1 et 2 nous avons présenté la majeure partie des concepts théoriques nécessaires à la compréhension d'un modèle de régression. Les démonstrations ont été limitées à l'essentiel. Le lecteur se reporterà à la bibliographie pour avoir plus de précision et de rigueur mathématique. L'ouvrage de Tomassone & al., en particulier, est vivement conseillé.

Dans le chapitre 3, nous analyserons à partir d'exemples de la littérature, les difficultés de la régression linéaire lorsqu'on étudie des données réelles, qui « prennent un malin plaisir » à ne pas se comporter comme la théorie le suppose.

Pour aider aux diagnostics, de nombreuses options sont disponibles dans la procédure REG, ce que nous verrons au chapitre 4.
3. Quand les résultats d'une régression ne sont pas forcément pertinents

Dans ce chapitre nous montrons sur quelques exemples les difficultés rencontrées dans l'application de la régression linéaire simple et la régression linéaire multiple sur données réelles ou simulées, lorsque les suppositions ne sont pas vérifiées. Nous présentons quelques aides très utiles. Ce n'est qu'un petit survol de la littérature sur le sujet de la robustesse qui nécessite à lui seul plusieurs ouvrages.

Ce chapitre a pour objectif de vous sensibiliser à l'importance des diagnostics proposés dans Proc REG, qui seront vus au chapitre 4.

La majorité des résultats d'une régression sont présentés avec les sorties de SAS/INSIGHT, pour montrer l'apport de l'interactivité, et l'importance des graphiques dans la compréhension des analyses.

3.1. Exemples en régression simple

3.1.1. Une même valeur pour des situations différentes

Cet exemple est de TOMASSONE & al. (1986). Dès 1973 ANSCOMBE, neveu et collaborateur de J.W TUKKEY (1977) avait proposé un exemple similaire dans «Graphs in Statistical Analysis ». Soient les 5 couples de 16 observations (X,Ya),(X,Yb),(X,Yc),(X,Yd),(Xe,Ye) sur lesquels on effectue 5 régressions linéaires simples.

<table>
<thead>
<tr>
<th>OBS</th>
<th>X</th>
<th>Ya</th>
<th>Yb</th>
<th>Yc</th>
<th>Yd</th>
<th>Xe</th>
<th>Ye</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>7</td>
<td>5.535</td>
<td>0.113</td>
<td>7.399</td>
<td>3.864</td>
<td>13.715</td>
<td>5.654</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>4.249</td>
<td>7.426</td>
<td>8.468</td>
<td>7.504</td>
<td>13.715</td>
<td>8.491</td>
</tr>
<tr>
<td>16</td>
<td>23</td>
<td>17.690</td>
<td>13.947</td>
<td>29.431</td>
<td>12.187</td>
<td>33.281</td>
<td>27.435</td>
</tr>
</tbody>
</table>
Les estimations des Y sont les suivantes :

<table>
<thead>
<tr>
<th>OBS</th>
<th>Ya_est</th>
<th>Yb_est</th>
<th>Yc_est</th>
<th>Yd_est</th>
<th>Ye_est</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.18</td>
<td>6.18</td>
<td>6.18</td>
<td>6.18</td>
<td>11.61</td>
</tr>
<tr>
<td>3</td>
<td>7.80</td>
<td>7.80</td>
<td>7.80</td>
<td>7.80</td>
<td>11.61</td>
</tr>
<tr>
<td>4</td>
<td>8.61</td>
<td>8.61</td>
<td>8.61</td>
<td>8.61</td>
<td>11.61</td>
</tr>
<tr>
<td>5</td>
<td>10.22</td>
<td>10.23</td>
<td>10.22</td>
<td>10.22</td>
<td>11.61</td>
</tr>
<tr>
<td>6</td>
<td>11.03</td>
<td>11.03</td>
<td>11.03</td>
<td>11.03</td>
<td>11.61</td>
</tr>
<tr>
<td>7</td>
<td>11.84</td>
<td>11.84</td>
<td>11.84</td>
<td>11.84</td>
<td>11.61</td>
</tr>
<tr>
<td>8</td>
<td>11.84</td>
<td>11.84</td>
<td>11.84</td>
<td>11.84</td>
<td>11.61</td>
</tr>
<tr>
<td>9</td>
<td>12.65</td>
<td>12.65</td>
<td>12.65</td>
<td>12.65</td>
<td>11.61</td>
</tr>
<tr>
<td>10</td>
<td>14.27</td>
<td>14.27</td>
<td>14.27</td>
<td>14.27</td>
<td>11.61</td>
</tr>
<tr>
<td>11</td>
<td>15.07</td>
<td>15.08</td>
<td>15.08</td>
<td>15.08</td>
<td>11.61</td>
</tr>
<tr>
<td>12</td>
<td>15.88</td>
<td>15.89</td>
<td>15.89</td>
<td>15.89</td>
<td>11.61</td>
</tr>
<tr>
<td>13</td>
<td>15.88</td>
<td>15.89</td>
<td>15.89</td>
<td>15.89</td>
<td>11.61</td>
</tr>
<tr>
<td>14</td>
<td>16.69</td>
<td>16.69</td>
<td>16.69</td>
<td>16.69</td>
<td>11.61</td>
</tr>
<tr>
<td>15</td>
<td>17.50</td>
<td>17.50</td>
<td>17.50</td>
<td>17.50</td>
<td>11.61</td>
</tr>
<tr>
<td>16</td>
<td>19.12</td>
<td>19.12</td>
<td>19.12</td>
<td>19.12</td>
<td>27.43</td>
</tr>
</tbody>
</table>

Les ajustements par les 5 droites de régressions sont donnés figure 3.1 ; ils sont identiques, mêmes estimations, mêmes statistiques pour $R^2=0.617$, mêmes coefficients $b_0=0.520$, $b_1=0.809$, mêmes erreurs-types sur les coefficients, et pourtant les situations sont bien différentes.

Analyse des résultats

- Sur le 1er graphique on peut voir que le modèle semble bien adapté.
- Sur le 2ème graphique le modèle linéaire est inadapté, un modèle quadratique de la forme $Y = \beta X^2$ serait préférable.
- Sur le 3ème graphique un point est suspect et entraîne la droite de régression vers le haut.
- Sur le 4ème graphique la variance des erreurs varie. Il y a un phénomène d’hétéroscédasticité (variance de Y sachant X non constante).
- Sur le 5ème graphique le plan expérimental défini par les valeurs de Xe est particulièrement mauvais. Un seul point extrême détermine la droite. Que se passe-t-il entre $X=13.715$ et $X=34.281$ ? La liaison est-elle linéaire ?

Un simple diagramme cartésien (Scatter plot) permet dans chacun des 5 cas de vérifier si les suppositions de la régression linéaire sont respectées, et de porter un diagnostic.
Figure 3.1 : 5 droites de régression

3.1.2. Pondérations et régression linéaire par morceaux

Cet exemple inspiré de J.P. BENZECRI & F.BENZECRI (1989), « Calculs de corrélation entre variables et juxtaposition de tableaux », est totalement artificiel.
L'objectif de cet exemple est double : montrer d'une part l'effet d'une pondération des observations sur les résultats d'une analyse de régression et d'autre part de sensibiliser par des graphiques, à l'usage abusif de la régression lorsque l'hypothèse de linéarité sur tout l'intervalle n'est pas valide.

**Tableau des données**

101 observations (variables X et Y) sont générées par programme de la manière suivante : X varie de -50 à +50 avec un pas de 1 et à chaque pas Y est calculé selon les formules linéaires suivantes :

- Si $X < -10$ alors $Y = X + 20$
- Si $-10 \leq X < 11$ alors $Y = -X$
- Si $X \geq 11$ alors $Y = X - 20$

X est la variable régresseur, Y est la variable réponse. X et Y sont donc *rigoureusement* liées par une fonction linéaire par morceaux.

Les 3 parties ont des pentes (paramètre $\beta_1$) respectives de +1 pour les 40 premières observations, -1 pour les observations 41 à 61, et +1 pour les observations 62 à 101.

On effectue une première régression sans pondération, sur toutes les observations (figure 3.2), puis deux régressions pondérées, en pondérant par 100 les observations de la *partie centrale*, les autres observations ayant une pondération de 1, puis une 3ème régression avec une pondération par 1000.

On utilise l'instruction "Weight " de la Proc Reg.

**Programme SAS de génération des observations et analyse de régression**

L'appel de Proc REG se fait par une macro.

```sas
data ligne;
  do x=-50 to 50;
    p1=1;  p100=1;  pmil=1;
    if x <-10 then  y=x+20;
      else if x < 11 then do; y=-x ;
        p100=100;  pmil =1000;  end;
      else y=x-20;
    output;
  end;
%macro reg(poids=);
  proc reg data=ligne;
    model y=x ;
    %if &poids ^=  %then weight &poids ; %str(;;)
    title "Avec ponderation &poids";
  %mend;
  %reg(poids=p1)
  %reg(poids=p100)
  %reg(poids=p1000)
quit;
```

©WP CES 2007.47 - Premiers pas en régression linéaire avec SAS- Confais Le Guen - 62
a) Observations (X,Y) générées sans aléa
b) Régression sans pondération
c) Régression avec pondération de 100
d) Régression avec pondération de 1000

Figure 3.2 Droite de régression selon les pondérations utilisées (1, 100, 1000)

Analyse des résultats
En pondérant de 3 façons différentes la partie centrale on obtient des F value, des coefficients de détermination, des estimations de β₁ droite de régression- totalement différents et cependant toujours significatif (cf. Tableau 3.1)

Tableau 3.1

<table>
<thead>
<tr>
<th>Pondération</th>
<th>F value</th>
<th>P value</th>
<th>R²</th>
<th>estimation de β₁</th>
<th>T Student</th>
</tr>
</thead>
<tbody>
<tr>
<td>sans pondération</td>
<td>261.433</td>
<td>&lt;0.0001</td>
<td>0.7253</td>
<td>0.0255</td>
<td>16.169</td>
</tr>
<tr>
<td>par 100</td>
<td>11.740</td>
<td>&lt;0.0009</td>
<td>0.1060</td>
<td>-0.2512</td>
<td>-3.427</td>
</tr>
<tr>
<td>par 1000</td>
<td>389.810</td>
<td>&lt;0.0001</td>
<td>0.7985</td>
<td>-0.8580</td>
<td>-19.744</td>
</tr>
</tbody>
</table>

En fonction de la pondération, la droite de régression « tourne » jusqu'à s'adapter à la pente de la partie centrale, comme on le voit dans les figures 3.2 (b, c, d).
Si la modélisation devient correcte pour la partie centrale, il n'en va pas de même pour les deux autres parties.

Si, par contre on réalise 3 études séparées sur les 3 intervalles où l'hypothèse de linéarité entre X et Y est exacte, on obtient des coefficients de détermination R²=1 (corrélation parfaite) et des droites de régression totalement adaptées aux données (SS Error=0).

L'exemple met en lumière les questions que l'on doit se poser avant d'effectuer une régression : l'hypothèse de linéarité est elle plausible sur tout l'intervalle ?.
**Théorie de la régression pondérée**

- **L'instruction **Weight** de Proc REG**

L'instruction *Weight* de Proc REG, minimise la somme des carrés résiduels pondérés :

\[
\sum_1^i w_i (Y_i - \hat{Y}_i)^2
\]

\(w_i\) : valeur de la variable spécifiée dans l'instruction Weight,

\(Y_i\) : valeur observée de la variable à expliquer,

\(\hat{Y}_i\) : valeur prédite pour l'observation \(i\).

Les équations normales utilisées sont dans ce cas :

\[
\beta = (X'WX)^{-1}X'WY
\]

\(W\) : matrice diagonale constituée des poids.

- **Quand utiliser l'instruction Weight ?**

Lorsque l'hypothèse de variance constante des erreurs n'est pas vérifiée (hétéroscédasticité des erreurs), la littérature statistique propose d'utiliser la régression pondérée en prenant comme pondération l'inverse des variances théoriques \(\sigma_i^2\).

\[
w_i = \frac{1}{\sigma_i^2}
\]

En général les variances théoriques des erreurs ne sont pas connues, elles sont estimées.

**Conclusion**

L'artifice de pondération des données doit être utilisé avec discernement. En particulier il faut au préalable contrôler si les observations sont homogènes, c'est-à-dire si d'éventuels groupes aux comportements différents peuvent être différenciés. Une approche par analyse de données ou analyse exploratoire des données peut s'avérer nécessaire. On utilisera avec profit des marqueurs de couleurs (SAS/INSIGHT et sa boîte à outils **Tools**) pour repérer si une variable de groupe ne fait pas apparaître des mélanges de populations.

Il existe des méthodes robustes pour pondérer les observations mais elles sortent du cadre de cet ouvrage. La lecture de l'ouvrage de Rousseeuw et al. (2003) est fortement recommandée.

**3.1.3. Transformation des données**

Lorsque sur un graphique, la liaison entre \(X\) et \(Y\) n'apparaît pas linéaire, on peut essayer de transformer les données, pour tenter de linéariser la liaison.
Exemple : Liaison entre Produit National Brut et taux d'urbanisation
La table SAS Paysniv3 porte sur 173 pays des 5 continents (figure 3.3).

Sur la Figure 3.4 a, sont représentées la variable PNB, Produit national brut en fonction du taux d'urbanisation URBA, et la droite de régression, pour 173 pays. Bien que les résultats statistiques soient significatifs ($R^2=0.4358$, $F=122.06$, $p \text{ value} <0.0001$, $T$ de Student = 11.05, $p \text{ value} <0.0001$) cette analyse n'est pas satisfaisante. La supposition de liaison linéaire n'est pas vérifiée, confirmé par le graphique des résidus (figure 3.3 b).

La variance des erreurs n'est pas constante, elle augmente en fonction de URBA. Il y a un effet d'entonnoir caractéristique de l'hétéroscédasticité des erreurs.

Les couleurs des points observations correspondent aux couleurs des continents de la Figure 3.3. On peut noter grâce aux couleurs que l'Afrique (point rouge) se différencie totalement de l'Europe (point bleu), on a à faire à des mélanges de population. Cependant pour notre démonstration sur les transformations on passera sous silence cette remarque.

Pour dilater les valeurs faibles et en même temps compresser les valeurs élevées du PNB on transforme la variable en son logarithme.
En SAS/INSIGHT il suffit de cliquer sur le nom de la variable PNB sur le graphique et de demander par le menu ➔ Edit # Variables # Log, la transformation de la variable PNB en son logarithme Log (L_PNB). Tous les affichages sont modifiés (figures 3.5 a et b).

*a) Avant transformation*  
*b) Après transformation par le logarithme*

**Figure 3.5 Liaison entre le taux d’urbanisation et le Produit National Brut**

La transformation a eu un double effet (figure 3.5). D’une part elle a symétrisé la distribution du PNB, visible sur les Box-plots, et d’autre part elle a linéarisé la liaison entre Log(PNB) et URBA.

Les indicateurs statistiques de la régression sont nettement améliorés. ($R^2=0.6468$, $F=289.39$, $pvalue<0.0001$, $T$ de Student $= 34,05$, $pvalue <0.0001$).

*a) régression Log(PNB) fonction de URBA*  
*b) Graphique des résidus fonction de URBA*

**Figure 3.6**

Les résidus (figure 3.6 b) se trouvent maintenant bien repartis dans la bande (-2, +2), à l’exception de 2 pays atypiques, Oman et Equatorial Guinea, sur lesquels il faut s’interroger.
Echelle de Tukey – Ladder of Power

Les transformations de variables occupent une place importante dans la littérature. On a vu précédemment, que la transformation Log du PNB, permettait de linéariser la liaison entre Y et X. Pour choisir une transformation appropriée, J.W. Tukey (1977) a proposé ce qu'on appelle maintenant l’échelle de transformation de Tukey\textsuperscript{15}.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>$Y^4$</td>
</tr>
<tr>
<td>3</td>
<td>$Y^3$</td>
</tr>
<tr>
<td>2</td>
<td>$Y^2$</td>
</tr>
<tr>
<td>1</td>
<td>$Y^1$</td>
</tr>
<tr>
<td>1/2</td>
<td>$\sqrt{Y}$</td>
</tr>
<tr>
<td>0</td>
<td>Log($Y$)</td>
</tr>
<tr>
<td>-1/2</td>
<td>$-1/\sqrt{Y}$</td>
</tr>
<tr>
<td>-1</td>
<td>$-1/Y$</td>
</tr>
<tr>
<td>-2</td>
<td>$-1/Y^2$</td>
</tr>
</tbody>
</table>

Selon la forme des courbes définies par les points $(X_i,Y_i)$ on pourra soit monter l’échelle, c’est-à-dire transformer X ou Y en ses puissances ($Y^2$, $Y^3$ etc.) soit descendre l’échelle, en prenant $\sqrt{Y}$, Log($Y$), $-1/\sqrt{Y}$, $-1/Y$ etc.

Astuce de lecture proposée par J. Vanpoucke et E. Horber\textsuperscript{16}

La forme des courbes du tableau 3.2 dessine un arc, et la flèche indique selon son orientation s’il faut monter ou descendre l’échelle.

Ainsi pour PNB en fonction de URBA, les données formaient une courbe d’allure n°3, il faut donc soit monter l’échelle en X, soit descendre l’échelle en Y. Comme la distribution de URBA est symétrique (voir figure 3.5 a), c’est plutôt sur PNB qu’il faut agir (distribution non symétrique, voir le Box plot de PNB).

Cette procédure raisonnée permet d’éviter d’agir complètement par « essais-erreurs », cependant pour trouver le bon choix on n’évite pas de faire quelques essais, grandement facilités par l’interactivité de SAS/INSIGHT.

\textsuperscript{15} Tukey utilise le mot « re-expression » et non le mot « transformation » d’une variable.

Tableau 3.2 Transformations appropriées selon la forme des courbes

<table>
<thead>
<tr>
<th>Forme de Courbes</th>
<th>Description de l’action</th>
<th>Transformation sur X</th>
<th>Transformation sur Y</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image" alt="Descendre l’échelle en X ou Monter l’échelle en Y" /></td>
<td>Descendre l’échelle en X ou Monter l’échelle en Y</td>
<td>$\sqrt{X}$, Log$(X)$, $-1/\sqrt{X}$ etc.</td>
<td>$Y^2, Y^3$ etc.</td>
</tr>
<tr>
<td><img src="image" alt="Descendre l’échelle en X ou Descendre l’échelle en Y" /></td>
<td>Descendre l’échelle en X ou Descendre l’échelle en Y</td>
<td>$\sqrt{X}$, Log$(X)$, $-1/\sqrt{X}$ etc.</td>
<td>$\sqrt{Y}$, Log$(Y)$, $-1/\sqrt{Y}$ etc.</td>
</tr>
<tr>
<td><img src="image" alt="Monter l’échelle en X ou Descendre l’échelle en Y" /></td>
<td>Monter l’échelle en X ou Descendre l’échelle en Y</td>
<td>$X^2$, $X^3$ etc.</td>
<td>$\sqrt{Y}$, Log$(Y)$, $-1/\sqrt{Y}$ etc.</td>
</tr>
<tr>
<td><img src="image" alt="Monter l’échelle en X ou Monter l’échelle en Y" /></td>
<td>Monter l’échelle en X ou Monter l’échelle en Y</td>
<td>$X^2$, $X^3$ etc.</td>
<td>$Y^2$, $Y^3$ etc.</td>
</tr>
</tbody>
</table>

Pour vous familiariser avec les transformations aller voir cet applet sur internet [http://noppa5.pc.helsinki.fi/opetus/sd/sdt0.html](http://noppa5.pc.helsinki.fi/opetus/sd/sdt0.html). Développé par JUHA PURANEN de l'Université de Helsinski, cet applet montre l’effet de différentes transformations sur la liaison entre 2 variables, Y= Distance de freinage et X=Vitesse du véhicule, dont tout conducteur sait que la liaison n’est pas linéaire. Il montre également la technique de lissage de courbes par LOWESS.

3.1.4. Méthode non paramétrique du LOWESS

La méthode non paramétrique du LOWESS - [Locally WEighted SMOOTHing Scatter](http://noppa5.pc.helsinki.fi/opetus/sd/sdt0.html)-de Cleveland (1979, 1993, 1994) permet de lisser une courbe. Cette technique combine une technique de régression linéaire ou polynomiale locale avec la flexibilité de la régression non linéaire. Le lissage obtenu permet à l’œil de repérer des points atypiques, de voir d’éventuelles structures, de détecter des non linéarités etc. Le principe repose sur des régressions locales définies sur des fenêtres glissantes. Chaque point observation est estimé, par une droite éventuellement un polynôme, à partir des points de son voisinage, situés dans une fenêtre. Chaque point du voisinage est pondéré en fonction de sa distance au point estimé. Les paramètres sur lesquels on peut agir sont la largeur de la fenêtre (Bandwith) et la fonction de pondération des points. Le principe est analogue à celui des moyennes mobiles, plus la largeur de la fenêtre est grande plus la série sera lissée. Pour la fonction de poids, on utilise généralement la fonction tri-cube.
SAS/INSIGHT permet d’utiliser cette technique. L’exemple\(^{17}\) porte sur les températures quotidiennes maximum de Melbourne de janvier 1981 à décembre 1990 (Source : Australian Bureau of Meteorology, 3650 observations).

Dans le menu Fit de la régression, on demande une régression simple de la variable *Température* en Y en fonction de la variable *Date* en X. Lorsque l’analyse de régression est affichée, on choisit le menu ➔ *Curves # Loess*

### Figure 3.6 Technique du LOWESS appliquée à des données de températures

La droite de régression est strictement horizontale avec une température moyenne de 21°04 sur les 10 années (figure 3.6).

<table>
<thead>
<tr>
<th>Model Equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>temperature = 21.0418 - 0.0001 date</td>
</tr>
</tbody>
</table>

La technique du LOWESS (parfois dénommée LOESS) permet le lissage de la série. Elle nécessite beaucoup de calculs, ce qui ne pose plus de problèmes avec les ordinateurs actuels si la technique est bien programmée. C’est une des techniques modernes les plus attractives puisqu’elle ne nécessite pas de préciser la forme d’un modèle, elle laisse « parler les données » Cette technique est très utile dans la phase exploratoire des données mais elle a son revers, elle ne fournit pas de fonction analytique comme la régression linéaire.

Tableau 3.3

<table>
<thead>
<tr>
<th>Curve</th>
<th>Type</th>
<th>Weight</th>
<th>Interv</th>
<th>Metho</th>
<th>Alpha</th>
<th>K</th>
<th>DF</th>
<th>R-Square</th>
<th>MSE</th>
<th>MSE(GCV)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Line</td>
<td>Tri-Cu</td>
<td>128</td>
<td>GCV</td>
<td>0.0104</td>
<td>17</td>
<td>123.56</td>
<td>0.5419</td>
<td>17.6074</td>
<td>10.307</td>
<td></td>
</tr>
</tbody>
</table>

En cliquant sur le curseur Alpha (paramètre du LOWESS lié à la largeur de bande - bandwith) on peut agir sur la largeur de la fenêtre et voir l’effet sur le filtrage. Avec un coefficient Alpha de 0.005, on fait apparaître un lissage moins régulier.

**Figure 3.7 après modification du paramètre de filtrage Alpha**

Pour connaître la technique du LOWESS programmée dans SAS, il suffit de cliquer sur un mot clé (par exemple Alpha –voir Tableau 3.3) et de demander l’aide en ligne par le menu ➔ Help # Help on Selection.

Il existe aussi dans SAS une procédure LOESS dans le module SAS/ETS.

### 3.2. Exemples en régression multiple

#### 3.2.1. Y « expliquée » par la corrélation entre deux régresseurs

Cet exemple, montre l’influence sur le coefficient de détermination lorsque les régresseurs sont corrélés. Les données sont issues du cours de Georges Monette de York University.

```sas
Data HWH;
input Weight Height Health;
cards;
68 94 120
137 114 60
94 104 123
121 107 94
100 118 104
93 91 117
76 123 139
102 73 100
122 112 91
89 78 91
```

©WP CES 2007.47 - Premiers pas en régression linéaire avec SAS- Confais Le Guen - 70
• Pour le premier modèle, régression simple de HEALTH en fonction de WEIGHT, le coefficient de détermination vaut **0.0738**. **Aucune liaison entre santé et poids.**

• Pour le deuxième modèle, régression simple de HEALTH en fonction de HEIGHT, le coefficient de détermination vaut **0.0367**. **Aucune liaison entre santé et taille.**

• Pour le troisième modèle du tableau 3.4, donnant les résultats de la régression multiple de HEALTH en fonction de WEIGHT et HEIGHT, le coefficient de détermination vaut **0.6551**. La statistique F et les T de student sont tous significatifs. Si on s’arrête à ces seules indications le modèle explique 65 % de la variation de Health, voir le tableau 3.4.

### Tableau 3.4 Régression de Health en fonction de WEIGHT et HEIGHT

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>4295.26012</td>
<td>2147.63006</td>
<td>11.29</td>
<td>0.0017</td>
</tr>
<tr>
<td>Error</td>
<td>14</td>
<td>6555.59355</td>
<td>468.25325</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>16</td>
<td>10840.85355</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Root MSE | 15.22072 |
| Dependent Mean | 109.92233 |
| Adj R-Sq | 0.5976 |
| Ccfl Var | 12.46932 |

### Parameter Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>101.08215</td>
<td>11.15291</td>
<td>8.75</td>
<td>0.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Height</td>
<td>1</td>
<td>-0.77008</td>
<td>0.138369</td>
<td>-4.64</td>
<td>0.0006</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight</td>
<td>1</td>
<td>0.82603</td>
<td>0.04369</td>
<td>18.95</td>
<td>0.0007</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Cette fausse explication de la santé par le Poids et le Taille est due à la corrélation entre les régresseurs (0.8357). Ce qui est visible sur la matrice de corrélation et sur la matrice de scatter plot.
3.2.2. Instabilité des coefficients de la régression, en cas de multicolinéarité

**Exemple sur données réelles**

Cet exemple est de D. LADIRAY. On dispose de 44 observations et on cherche à expliquer le taux d'urbanisation, variable URBA, en fonction de 10 variables régresseurs, POP87 à ESPER.

**Tableau 3.5 Tableau des données**

<table>
<thead>
<tr>
<th>OBS</th>
<th>URBA</th>
<th>POP87</th>
<th>NAT</th>
<th>MORT</th>
<th>ACCR</th>
<th>DOUB</th>
<th>FERTI</th>
<th>MORTI</th>
<th>AGE15</th>
<th>AGE65</th>
<th>ESPER</th>
<th>CCR_CAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>81</td>
<td>0.4</td>
<td>32</td>
<td>5</td>
<td>2.8</td>
<td>25</td>
<td>4.6</td>
<td>32.0</td>
<td>41</td>
<td>2</td>
<td>67</td>
<td>2.7</td>
</tr>
<tr>
<td>2</td>
<td>53</td>
<td>0.7</td>
<td>20</td>
<td>9</td>
<td>1.1</td>
<td>63</td>
<td>2.5</td>
<td>12.0</td>
<td>25</td>
<td>11</td>
<td>74</td>
<td>1.1</td>
</tr>
<tr>
<td>3</td>
<td>79</td>
<td>0.6</td>
<td>47</td>
<td>8</td>
<td>4.0</td>
<td>18</td>
<td>7.4</td>
<td>59.0</td>
<td>50</td>
<td>3</td>
<td>64</td>
<td>3.9</td>
</tr>
<tr>
<td>4</td>
<td>68</td>
<td>17.0</td>
<td>46</td>
<td>13</td>
<td>3.3</td>
<td>21</td>
<td>7.2</td>
<td>80.0</td>
<td>49</td>
<td>4</td>
<td>62</td>
<td>3.3</td>
</tr>
<tr>
<td>5</td>
<td>90</td>
<td>4.4</td>
<td>23</td>
<td>7</td>
<td>1.7</td>
<td>41</td>
<td>3.1</td>
<td>12.3</td>
<td>33</td>
<td>9</td>
<td>75</td>
<td>1.6</td>
</tr>
<tr>
<td>6</td>
<td>60</td>
<td>7.7</td>
<td>45</td>
<td>8</td>
<td>3.7</td>
<td>19</td>
<td>7.4</td>
<td>54.0</td>
<td>51</td>
<td>3</td>
<td>67</td>
<td>3.7</td>
</tr>
<tr>
<td>7</td>
<td>80</td>
<td>1.9</td>
<td>34</td>
<td>3</td>
<td>3.2</td>
<td>22</td>
<td>4.4</td>
<td>19.0</td>
<td>40</td>
<td>1</td>
<td>72</td>
<td>3.1</td>
</tr>
<tr>
<td>8</td>
<td>80</td>
<td>3.3</td>
<td>30</td>
<td>8</td>
<td>2.2</td>
<td>32</td>
<td>3.8</td>
<td>52.0</td>
<td>38</td>
<td>5</td>
<td>65</td>
<td>2.2</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>1.3</td>
<td>47</td>
<td>14</td>
<td>3.3</td>
<td>21</td>
<td>7.1</td>
<td>117.0</td>
<td>44</td>
<td>3</td>
<td>52</td>
<td>3.3</td>
</tr>
<tr>
<td>10</td>
<td>86</td>
<td>0.3</td>
<td>34</td>
<td>4</td>
<td>3.0</td>
<td>23</td>
<td>5.6</td>
<td>42.0</td>
<td>34</td>
<td>2</td>
<td>69</td>
<td>3.0</td>
</tr>
<tr>
<td>11</td>
<td>72</td>
<td>14.8</td>
<td>39</td>
<td>7</td>
<td>3.1</td>
<td>22</td>
<td>6.9</td>
<td>79.0</td>
<td>37</td>
<td>2</td>
<td>63</td>
<td>3.2</td>
</tr>
<tr>
<td>12</td>
<td>49</td>
<td>11.3</td>
<td>47</td>
<td>9</td>
<td>3.8</td>
<td>18</td>
<td>7.2</td>
<td>59.0</td>
<td>49</td>
<td>4</td>
<td>63</td>
<td>3.8</td>
</tr>
<tr>
<td>13</td>
<td>46</td>
<td>51.4</td>
<td>30</td>
<td>9</td>
<td>2.1</td>
<td>33</td>
<td>4.0</td>
<td>92.0</td>
<td>36</td>
<td>4</td>
<td>62</td>
<td>2.1</td>
</tr>
<tr>
<td>14</td>
<td>81</td>
<td>1.4</td>
<td>30</td>
<td>4</td>
<td>2.6</td>
<td>27</td>
<td>5.9</td>
<td>38.0</td>
<td>30</td>
<td>1</td>
<td>68</td>
<td>2.6</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>6.5</td>
<td>53</td>
<td>19</td>
<td>3.4</td>
<td>20</td>
<td>7.8</td>
<td>137.0</td>
<td>49</td>
<td>3</td>
<td>47</td>
<td>3.4</td>
</tr>
<tr>
<td>16</td>
<td>40</td>
<td>2.4</td>
<td>47</td>
<td>17</td>
<td>3.0</td>
<td>23</td>
<td>7.3</td>
<td>135.0</td>
<td>48</td>
<td>3</td>
<td>48</td>
<td>3.0</td>
</tr>
<tr>
<td>17</td>
<td>16</td>
<td>14.2</td>
<td>48</td>
<td>22</td>
<td>2.6</td>
<td>27</td>
<td>7.6</td>
<td>182.0</td>
<td>46</td>
<td>4</td>
<td>39</td>
<td>2.6</td>
</tr>
<tr>
<td>18</td>
<td>13</td>
<td>107.1</td>
<td>44</td>
<td>17</td>
<td>2.7</td>
<td>26</td>
<td>6.2</td>
<td>140.0</td>
<td>44</td>
<td>4</td>
<td>50</td>
<td>2.7</td>
</tr>
<tr>
<td>19</td>
<td>5</td>
<td>1.5</td>
<td>38</td>
<td>18</td>
<td>2.0</td>
<td>34</td>
<td>5.5</td>
<td>142.0</td>
<td>40</td>
<td>3</td>
<td>46</td>
<td>2.0</td>
</tr>
<tr>
<td>20</td>
<td>25</td>
<td>800.3</td>
<td>33</td>
<td>12</td>
<td>2.1</td>
<td>33</td>
<td>4.3</td>
<td>101.0</td>
<td>38</td>
<td>4</td>
<td>55</td>
<td>2.1</td>
</tr>
<tr>
<td>21</td>
<td>51</td>
<td>50.4</td>
<td>45</td>
<td>13</td>
<td>3.2</td>
<td>21</td>
<td>6.3</td>
<td>113.0</td>
<td>44</td>
<td>3</td>
<td>57</td>
<td>3.2</td>
</tr>
<tr>
<td>22</td>
<td>26</td>
<td>0.2</td>
<td>48</td>
<td>10</td>
<td>3.8</td>
<td>18</td>
<td>7.1</td>
<td>68.0</td>
<td>45</td>
<td>2</td>
<td>51</td>
<td>3.8</td>
</tr>
<tr>
<td>23</td>
<td>7</td>
<td>17.8</td>
<td>42</td>
<td>17</td>
<td>2.5</td>
<td>28</td>
<td>6.1</td>
<td>112.0</td>
<td>41</td>
<td>3</td>
<td>52</td>
<td>2.5</td>
</tr>
<tr>
<td>24</td>
<td>28</td>
<td>104.6</td>
<td>44</td>
<td>15</td>
<td>2.9</td>
<td>24</td>
<td>6.6</td>
<td>125.0</td>
<td>45</td>
<td>4</td>
<td>50</td>
<td>2.9</td>
</tr>
<tr>
<td>25</td>
<td>22</td>
<td>16.3</td>
<td>25</td>
<td>7</td>
<td>1.8</td>
<td>38</td>
<td>3.7</td>
<td>29.8</td>
<td>35</td>
<td>4</td>
<td>70</td>
<td>1.8</td>
</tr>
<tr>
<td>26</td>
<td>64</td>
<td>0.3</td>
<td>30</td>
<td>4</td>
<td>2.6</td>
<td>26</td>
<td>3.6</td>
<td>12.0</td>
<td>38</td>
<td>3</td>
<td>62</td>
<td>2.6</td>
</tr>
</tbody>
</table>

Dans le tableau 3.5, deux valeurs de la variable NAT (taux de natalité) pour OBS=11 et OBS=30 sont légèrement modifiées (39 est remplacé par 40).

Les régressions effectuées avant et après modifications donnent les résultats suivants pour les coefficients de régression :

**Tableau 3.6**

<table>
<thead>
<tr>
<th>Résultat 1 (valeur 39)</th>
<th>Résultat 2 (valeur 40)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Avant</strong></td>
<td><strong>Après</strong></td>
</tr>
<tr>
<td>URBA =</td>
<td>URBA =</td>
</tr>
<tr>
<td>25.541</td>
<td>20.689</td>
</tr>
<tr>
<td>-0.026 POP87</td>
<td>-0.026 POP87</td>
</tr>
<tr>
<td>-6.661 NAT</td>
<td>-4.047 NAT</td>
</tr>
<tr>
<td>+2.681 MORT</td>
<td>-0.005 MORT</td>
</tr>
<tr>
<td>+64.506 ACCR</td>
<td>+39.832 ACCR</td>
</tr>
<tr>
<td>+0.019 DOUB</td>
<td>+0.015 DOUB</td>
</tr>
<tr>
<td>+7.834 FERTI</td>
<td>+7.307 FERTI</td>
</tr>
<tr>
<td>+0.101 MORTI</td>
<td>+0.128 MORTI</td>
</tr>
<tr>
<td>-1.132 AGE15</td>
<td>-1.157 AGE15</td>
</tr>
<tr>
<td>+2.709 AGE65</td>
<td>+2.848 AGE65</td>
</tr>
<tr>
<td>+0.910 ESPER</td>
<td>+0.969 ESPER</td>
</tr>
</tbody>
</table>

Les résultats "Avant" et "Après" (tableau 3.6) sont particulièrement instables pour les estimations des coefficients des 3 variables, NAT (taux de natalité), MORT (taux de mortalité) et ACCR (taux d'accroissement de la population).
Explication
Ces 3 variables **ne sont pas indépendantes**, elles sont liées entre elles par une relation quasi-linéaire $ACCR= (NAT-MORT)/10$. Dans le tableau 3.5, on peut comparer $ACCR$ avec la variable $ACCR_{CAL}$, en dernière colonne, calculée avec la formule exacte.
Lors de l'inversion de la matrice $X'X$, il y a une valeur propre qui est presque nulle. Conséquence une légère perturbation des données entraîne de grands changements dans les estimations des paramètres.

**Exemple sur données avec modèle théorique connu et régresseurs corrélés**

Cet exemple est de T. Foucart (2007). Pour étudier l’effet des corrélations entre les régresseurs sur les estimations des paramètres, T. Foucart a généré 100 observations d’un **vrai modèle théorique** :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon$$

Les vraies valeurs des paramètres du modèle théorique sont :

$$\beta_1 = 0 \quad \beta_2 = 0.5 \quad \beta_3 = \beta_4 = -0.5$$

Les 4 régresseurs $X_1$ à $X_4$ suivent des lois normales centrées et réduites. L’erreur $\varepsilon$ suit une loi normale $N(0, \sigma^2)$, On impose de plus des contraintes sur la matrice des corrélations entre les régresseurs.

**Tableau 3.7 Corrélations imposées**

<table>
<thead>
<tr>
<th></th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$X_1$</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_2$</td>
<td>0.5</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$X_3$</td>
<td>0.5</td>
<td>0.5</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>$X_4$</td>
<td>-0.5</td>
<td>0.4</td>
<td>0.3</td>
<td>1</td>
</tr>
</tbody>
</table>

Et on impose la valeur du coefficient de détermination $R^2 = 0.5$.

Les données générées RIDGE1 sont disponibles sur le site de T. Foucart[^19].

**Résultats avec SAS/INSIGHT**

- la matrice de corrélation
- Menu ➔ Analyze # Mutivariate avec les 5 variables $X1, X2, X3, X4, Y$ dans le rôle $Y$

On la trouve dans le (tableau 3.8) ci-dessous.

Tableau 3.8 Statistiques univariées et matrice de corrélation

On vérifie sur le tableau 3.8 que la matrice de corrélation a bien les valeurs imposées du Tableau 3.7.

- la matrice de scatter plot

Menu ➜ Analyze → Scatter Plot avec les 5 variables (X1, X2, X3, X4, Y) dans le rôle X et les mêmes dans le rôle Y

Figure 3.6 Matrice de diagrammes de dispersion

La matrice de diagrammes de dispersion permet de repérer les liaisons entre les régresseurs et la variable Y. Y est lié à X1, X2, avec un coefficient positif, et Y est lié à X3, X4 avec un coefficient négatif.

- La régression linéaire

Menu ➜ Analyze → Fit avec les 4 variables X1, X2, X3, X4, dans le rôle de X et Y dans le rôle de Y
Avec la méthode des moindres carrés, le coefficient de détermination $R^2=0.4907$ est bien calculé (le théorique vaut 0.50), par contre les coefficients sont très différents de ceux du modèle théorique.

Les estimations MCO sont respectivement : 1.6339, -0.1482, -1.0375, 0.4439 au lieu des vraies valeurs : 0.5, 0.5, -0.5, -0.5. Mêmes les signes ne sont pas respectés.

**Conclusion**

Les conséquences des colinéarités entre les variables régresseurs sont les suivantes :

- Les coefficients de régression sont instables ;
- Leur signe peuvent changer (positif $\rightarrow$ négatif) rendant les interprétations fausses, ce qui a de graves conséquences lors de la recherche des effets d'une variable régresseur ;
- Les variances des estimateurs sont élevées.


### 3.3. Conditions d'utilisation de la régression, les diagnostics

Les différents exemples présentés dans ce chapitre montrent l'importance des analyses et diagnostics effectués avant et après les premiers traitements. Pour réaliser les tests d'hypothèses de la régression on a supposé, en pure théorie, que le modèle linéaire postulé est correct, que les suppositions, *a priori* sur les variables et sur les erreurs, sont vraies. Ces suppositions sont nécessaires pour

\(^{20}\) Site [http://foucart.thierry.free.fr/colreglin/Regression bornee.pdf](http://foucart.thierry.free.fr/colreglin/Regression bornee.pdf)
définir les tests comme on l’a vu au Chapitre 1, car on ne peut calculer les variances que dans des cas gaussiens et sous certaines conditions.

**Suppositions sur les variables**
- Les observations $Y_i$ sont supposées indépendantes
- Les variables $X_j$ sont non aléatoires.

**Suppositions sur les erreurs**
- les erreurs sont d’espérance nulle ce qui est vérifié par construction si la constante $\beta_0$ existe dans le modèle.
- les erreurs sont de variance constante
- les erreurs suivent une distribution normale
- les erreurs sont indépendantes

Les erreurs sont inconnues, elles sont approchées par les résidus, **si le modèle est correct.**

Après examen des résidus, on peut conclure: les suppositions semblent ou ne semblent pas être violées. Ce qui ne signifie pas que les suppositions soient correctes. Cela veut dire que sur la base des données que l'on a étudiée, on n'a aucune raison de dire que les suppositions sont fausses.

### 3.3.1. Modèle Inadapté

C'est par l'examen des résidus que l'on peut voir si le modèle postulé est vraisemblablement correct, ou s'il est inadapté.
Les résidus contiennent à la fois des erreurs de mesure et des erreurs de spécification du modèle, comme des variables omises, ou des liaisons non linéaires. On peut avoir des tests satisfaisants, de bonnes précisions sur les estimateurs des paramètres, alors que le modèle est inadapté à l'étude.

En dehors de certaines visualisations il n'y a que le bon sens et la **pré-connaissance du problème** qui permettent de repérer l'inadéquation du modèle aux données.

### 3.3.2. L'influence de certaines données, les données atypiques -Outliers-

Certaines données atypiques peuvent fausser les résultats. Les visualisations graphiques permettent parfois de les identifier.

Le livre de **Belsley, Kuh et Welsh** (BKW) a popularisé une méthode rigoureuse de recherche des observations influentes. L’option INFLUENCE de Proc REG permet cette analyse.

On peut alors être amené à retirer ces points atypiques des analyses, ou à procéder à des techniques « **robustes** » (voir **Rousseeuw** et al. (2003)). La procédure ROBUSTREG de SAS disponible en V9 reprend ces techniques.
3.3.3. Corrélation et colinéarité entre les régresseurs

La colinéarité est un **gros problème**, lot quotidien du statisticien praticien lorsqu’il analyse des données réelles, principalement en sciences économiques et sociales. C’est également le trio "BKW" qui a proposé des indicateurs de détection de colinéarités.

Les options TOL, VIF et COLLIN, COLLINOINT de Proc REG sont des aides aux diagnostics de colinéarité.

Tous ces compléments à la régression, représentations graphiques, indicateurs techniques de BKW, etc., nécessitent de faire appel aux nombreuses options de Proc REG, qui seront présentées au chapitre 4 « Validation d’une régression ». 
4. Validation d’une régression

Dans ce chapitre, on présente les différents éléments nécessaires à la validation d’une régression, c’est-à-dire la vérification des suppositions de base du modèle, l’étude de la robustesse au niveau des observations (détectio des observations influentes et atypiques) et au niveau des variables (colinéarités, choix d’un sous-ensemble de régresseurs).

Les sorties de SAS-version 9 illustrant ce chapitre sont réalisées avec le même exemple que pour le chapitre 2 (§2.3.1), issu du livre de Tomassone et al. (1992).

4.1. Introduction

Un principe de base doit être appliqué : explorer les données par des graphiques et/ou des calculs numériques.

Les calculs préalables des caractéristiques des variables Y et des Xj, ainsi que des histogrammes et des tracés Box-Plot de ces différentes variables permettent en effet de mettre en évidence des problèmes. Bien sûr le calcul des corrélations entre Y et les Xj est nécessaire.

Il faut étudier les variables, et donc en particulier faire des graphiques, par exemple de Y contre les régresseurs Xj pour contrôler la linéarité des liaisons. Pour cela, on utilisera la procédure G PLOT, ou le menu « Scatter-Plot » de SAS/INSIGHT.

Dans PROC REG, l’instruction PLOT permet de faire des graphiques des variables entrant dans la régression, mais aussi des variables créées comme les valeurs résiduelles et ajustées.

De plus, une Analyse en Composantes Principales des régresseurs, avec Y en variable supplémentaire, peut aussi être utile pour visualiser comment Y se reconstruit à partir de l’ensemble des X.

4.1.1. Modèle et notations

Si on postule un modèle avec n observations (i) et p variables régresseurs (j) et une constante, on note :
- Y réponse ou variable dépendante
- X matrice des variables régresseurs
- β coefficients de régression
- ε erreurs.

Le modèle linéaire s’écrit : $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta_p X_{ip} + \varepsilon_i$ .

D’où l’ajustement:
\[
\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_p X_{ip} \quad \text{et les résidus } e_i = Y_i - \hat{Y}_i.
\]

**Suppositions sur les erreurs** : \( \varepsilon_j \) sont des aléas, indépendants, d’espérance nulle, de variance constante, et de même loi (cf. chapitre 1).
\( \rightarrow \) on dit « IID avec loi normale \( N(0, \sigma^2) \) ».

### 4.1.2. Problèmes à étudier

**Vérification des suppositions sur les erreurs** :

- **Robustesse de la régression** : détection des observations influentes, et de la colinéarité des régresseurs ;
- **Choix d'un sous-ensemble** de régresseurs.

### 4.2. Vérification des suppositions de base sur les erreurs

Les suppositions sur les erreurs (inconnues) doivent être vérifiées à partir de leurs observations (les résidus).

#### 4.2.1. Espérance nulle

Il faut vérifier que les résidus sont de moyenne nulle. Or les résidus construits par les moindres carrés sont centrés par construction, si la constante est dans le modèle (ce que l'on suppose ici).

#### 4.2.2. Indépendance

Il faudrait vérifier que le vecteur des résidus forme un échantillon tiré de \( n \) variables aléatoires indépendantes

On obtient \( \hat{Y} \) par projection du vecteur \( Y \) sur le sous-espace engendré par les régresseurs : il en résulte que les \( n \) composants du vecteur \( e \) des résidus, \( e_i = Y_i - \hat{Y}_i \), sont reliés par des relations ainsi que le montre la figure 4.1 :
Dans la représentation géométrique dans l’espace $\mathbb{R}^n$, le vecteur des $n$ résidus est situé dans le sous-espace orthogonal à celui des régresseurs ; celui-ci étant de dimension $(p+1)$, le vecteur des résidus est alors situé dans un espace de dimension $(n-(p+1))$. L’indépendance n’a donc de sens que si $n$ est grand par rapport à $p$.

Remarque : 
De façon générale, pour tester l’indépendance, on met en œuvre des tests non paramétriques (qui ne sont proposés dans SAS) basés sur des séquences : séquence des signes des différences successives $(e_{i+1} - e_i)$, ou séquence des signes des différences à la médiane $(e_i - \text{Mediane})$.

Cas particulier où les observations sont apparentées (cas des chroniques) : 
Alors le test de Durbin-Watson permet de vérifier si le résidu en $i$ est non-corrélé au résidu en $(i+1)$ : on parle d’auto-corrélation d’ordre 1. Il est obtenu par l’option DW de l’instruction MODEL de Proc REG.

On calcule ainsi le coefficient de Durbin-Watson à partir des résidus $e_i = Y_i - \hat{Y}_i$, 

$$DW = \frac{\sum_{i}(e_{i+1} - e_i)^2}{\sum_{i}e_i^2}$$

En notant $\rho = \left(\frac{\sum e_{i+1}.e_i}{\sum e_i^2}\right)$, si les résidus forment un processus autorégressif d’ordre 1, c’est-à-dire suivent le modèle $e_{i+1} = \rho e_i + \eta_i$, alors $DW$ vaut à peu près $2(1-\rho)$.
où $$DW \approx 2 \left(1 - \frac{\sum e_{i+1} e_i}{\sum e_i^2}\right).$$

**Liens entre les valeurs ρ et DW:**

Si $$0 < \rho < 1 \Rightarrow DW$$ compris entre 0 et 2
Si $$0 > \rho > -1 \Rightarrow DW$$ compris entre 2 et 4

S’il n’y a pas d’auto-corrélation d’ordre 1 ⇔ ρ proche de 0, donc DW proche de 2.

Il existe des tables dites de Durbin-Watson permettant de tester l’absence d’auto-corrélation d’ordre 1 en fonction du niveau de confiance α, et de n (nombre d’observations) et p (nombre de variables). On y lit deux valeurs d1 et d2:

<table>
<thead>
<tr>
<th>ρ &gt; 0</th>
<th>0</th>
<th>ρ &lt; 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>d1</td>
<td>4-d2</td>
<td>4-d1</td>
</tr>
<tr>
<td>d2</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

**Les causes de l’auto-corrélation** sont une mauvaise spécification du modèle, ou l’absence d’une variable importante. **Les remèdes** sont soit de travailler sur les différences premières en Y c’est-à-dire (Y_i - Y_i-1), soit d’appliquer la méthode de Cochran-Orcutt (voir les livres spécialisés en Econométrie). PROC AUTOREG, du module SAS/ETS, réalise des régressions où le problème de l’auto-corrélation des résidus est résolu.

**Remarque** : Beaucoup de logiciels donnent systématiquement cette statistique DW, mais interpréter le test de Durbin-Watson sur les résidus n’a aucun sens si les données ne sont pas apparentées.

4.2.3. **Egalité des variances (homoscédasticité)**

Les graphiques des résidus contre les différents régisseurs Xj, permettent de visualiser si les résidus sont répartis dans une bande de valeurs horizontale autour de 0, c’est à dire s’il y a homoscédasticité. Sinon on peut alors détecter quelle est la variable responsable de l’hétéroscédasticité.

**Figure 4.2 : Graphique « typique » des résidus contre Xj révélant une hétéroscédasticité**
Ces graphiques peuvent être tracés par la procédure GPLOT si on a stocké les résidus dans une table, mais peuvent aussi être faits à l'intérieur de la procédure REG, à l'aide de l'instruction PLOT, avec « R. » comme nom de variable en ordonnée car le résidu est stocké en interne dans la variable ayant le nom R. Dans SAS/INSIGHT, un graphique des résidus (dénommé R_Y) contre X est tracé si le modèle est une régression à un régresseur ; sinon c'est le tracé de R_Y contre l'estimation Ŷ (dénommée P_Y) qui est tracé par défaut. Ces deux variables sont automatiquement créées dans la table SAS active. Il est possible de réaliser les autres graphiques à l'aide de la variable R_Y.

L'instruction MODEL de Proc REG possède une option SPEC pour tester s'il y a un problème d'hétéroscédasticité : l'hypothèse nulle « homoscédasticité » est testée à l'aide d'une statistique suivant une loi du Chi2 (cf. White H., (1980)). Le test est global, et donc en cas de rejet de H0, on ne sait pas quelle est la variable responsable de l'hétéroscédasticité.

D'autres tests, comme ceux de Goldfeldt et Quandt, ou de Breush et Pagan (voir les publications spécialisées en Econométrie, par exemple Green (2005)), permettent de mettre en évidence l'hétéroscédasticité dû aux différentes variables. Mais ils sont assez lourds à mettre en œuvre, et ne sont pas faits par des procédures SAS.

Une méthode plus simple peut permettre d'avoir une idée préalable sur l'existence d'un problème. Celle-ci s'apparente au test de Chow(1960) pour une série chronologique, où on teste l'égalité des variances des résidus de 2 sous périodes de la chronique.

Ici, on trie les résidus selon les valeurs croissantes de la variable Xj suspecte (par PROC SORT), ensuite on partage le vecteur des résidus triés en 2 paquets (premiers, derniers) dont on calcule les variances. Puis on teste l'hypothèse nulle d'égalité de ces 2 variances c'est à dire la possibilité d'homoscédasticité, à l'aide de la procédure TTEST de SAS.

**Quelques remèdes en cas d'hétéroscédasticité :**

- transformer Y ou Xj par une fonction racine carrée, ou Log, ou carré, etc. pour « aplatir » les variances : l'échelle de Tukey donnée au chapitre 3 (§3.1.3), peut aider au choix de la transformation ;
- mettre en œuvre une régression pondérée avec l'instruction WEIGHT, en prenant comme poids \( \frac{1}{\sqrt{f(X_j)}} \) si la variance \( \sigma^2 \) est une fonction connue f de Xj ;
- mettre en œuvre les moindres carrés généralisés, ce qui peut se faire par PROC GLM.
4.2.4. Normalité des erreurs

Supposition : Les $\varepsilon_i$ sont indépendant, et suivent une loi N(0, $\sigma^2$).

Cette supposition de normalité est nécessaire pour effectuer les tests sur les coefficients et les tests sur les sommes de carrés à l'aide des statistiques de Student ou de Fisher vues aux chapitres 1 et 2.

Comme tout test, le test d'adéquation à une loi nécessite l'indépendance. Or les résidus sont liés. Donc réaliser un test de normalité à une loi N(0, $\sigma^2$) sur ces résidus n’a pas de sens.

Dans SAS, un tracé « QQ-Plot » pour les résidus permet de vérifier graphiquement l'adéquation à la loi normale (0,s²) ou s² est estimé par MSE (Mean Square Error du modèle). Le QQ-Plot est obtenu dans la procédure REG avec l’instruction PLOT : on demande le tracé de R (variable interne des résidus) contre NQQ (variable interne contenant les quantiles de la loi normale) : voir l’exemple ci-dessous (§4.2.5). Dans SAS/INSIGHT, une fois que l'on a exécuté le modèle, on peut ajouter aux sorties standards un graphique QQ-Plot appelé « Residual Normal QQ » dans le menu Graphs (penser à cocher « Reference lines »dans le menu contextuel du graphique pour tracer la droite).

4.2.5. Exemple

On utilise les données « Processionnaire du pin » issu du livre de Tomassone et al. (1983), déjà traitées au chapitre 2. On se limite dans ce paragraphe au modèle Y = log = f(X1 X2 X4 X5), dont on verra que c’est un « bon » modèle.

On trouvera au §2.3.2 les caractéristiques des variables, la matrice de corrélation et les graphiques de dispersion des variables.

**Modèle**

```sas
proc reg data=libreg.chenilles; title 'Modèle Y = X1 X2 X4 X5';
LOG : model LOG=X1 X2 X4 X5 ;
run ;
quilt;
```

21 On trouvera en annexe 5 le principe de construction des QQ-Plot pour l’adéquation à certaines lois.
Dessin des résidus contre les 4 régresseurs (avec SAS/INSIGHT)

On y affiche le numéro de certaines observations, qui ont des résidus un peu grands, ou qui seront détectés plus loin comme « atypiques » (§4.3.11)
Remarque : instruction de tracé des résidus dans PROC REG

/* dessin des residus contre les X dans PROC REG */
plot R.*X1='1' R.*X2='2' R.*X4='4' R.*X5='5' / vref = 0 ;

QQ-Plot (avec SAS/INSIHT)

Le tracé QQ-Plot montre un assez bon ajustement à la loi normale.
/*option SPEC + dessin QQ_Plot */
proc reg data=libreg.chenilles; title 'Modèle Y = X1 X2 X4 X5';
 LOG : model LOG=X1 X2 X4 X5 / SPEC ;
run ;
/* QQ Plot */
plot R.* NQQ. ;
quit;

L'homoscédasticité des résidus n'est pas rejetée.

![QQ-Plot](image)

Modèle Y = X1 X2 X4 X5

The REG Procedure
Model: LOG
Dependent Variable: log

<table>
<thead>
<tr>
<th>Test d'indication du Premier et du Second</th>
</tr>
</thead>
<tbody>
<tr>
<td>DF</td>
</tr>
<tr>
<td>-----</td>
</tr>
<tr>
<td>14</td>
</tr>
</tbody>
</table>

©WP CES 2007.47 - Premiers pas en régression linéaire avec SAS- Confais Le Guen -
4.3. Influence d'observations

Dans le but d'avoir une régression plus robuste, il faut détecter les observations influentes, détection qui commence là-aussi par des graphiques soit de Y contre les X_j, soit des résidus contre les X_j.
Dans la figure 4.3, e_i est nul car l'observation i est influente à cause de son caractère atypique.

\[ e_i = 0 ! \]

*Figure 4.3 : Exemple d'observation à fort effet de levier*


Ces indicateurs sont basés sur des détections de l'influence selon des mesures différentes, donc déteccion des influences de nature différente : on distinguera donc des observations influentes sur la régression, ou suspectes, ou atypiques, ce dernier terme étant plutôt recommandé.

Les mesures peuvent être classées en 3 groupes :

- détection d'un effet de levier de l'observation, donnant un résidu petit : *leverage* ;
- détection de résidu grand donc observation atypique ;
- détection d'un grand effet sur l’ajustement, ou les coefficients, ou la précision.

4.3.1. Hat matrice et leverages

On utilise ici le modèle sous sa forme matricielle :

\[
Y = \begin{bmatrix}
Y_1 \\
Y_2 \\
Y_3 \\
\vdots \\
Y_n
\end{bmatrix} = \begin{bmatrix}
1 & X_{11} & X_{12} & \cdots & X_{1p} \\
1 & \quad & \quad & \quad & \quad \\
1 & \quad & \quad & \quad & \quad \\
\vdots & \quad & \quad & \quad & \quad \\
1 & X_{n1} & X_{n2} & \cdots & X_{np}
\end{bmatrix} \begin{bmatrix}
\beta_0 \\
\beta_1 \\
\beta_2 \\
\beta_p \\
\varepsilon_1 \\
\varepsilon_2 \\
\varepsilon_3 \\
\varepsilon_n
\end{bmatrix}
\]
L'estimation des moindres carrés est le vecteur \( B = (X'X)^{-1}X'Y \).
D'où l'ajustement \( \hat{Y} = X(X'X)^{-1}X'Y \)
En notant \( H = X(X'X)^{-1}X' \), on obtient \( \hat{Y} = HY \) et \( e = (I-H)Y \)
Cette matrice est nommée \( H \) pour *Hat matrice* car hat se traduit par chapeau, et \( \hat{Y} \) se dit *Y chapeau*.
Dans l'espace \( \mathbb{R}^n \), \( H \) est la matrice de la projection de \( Y \) sur l'espace engendré par les variables régresseurs \( X \) (espace de dimension \((p+1)\)) : c'est donc la matrice d'un projecteur, dont deux propriétés sont : \( H' = H \), et \( \text{trace}(H) = p+1 \).
\( H \) est une matrice carrée \((n,n)\), dont la diagonale comporte les \( n \) coefficients \( h_{ii} \). De l'expression matricielle de \( H \), on déduit \( h_{ii} = x_i'(X'X)^{-1}x_i \).
Les coefficients \( h_{ii} \) ne comportent donc que des données relatives aux variables explicatives \( X_j \).
Les « *leverages* » (leviers) des observations sont ces \( n \) valeurs \( h_{ii} \).
Un levier représente l'influence de l'observation \( i \) sur la valeur ajustée \( \hat{Y}_i \), à cause des valeurs \( x_i \) prises par les variables en \( i \).
On peut montrer que \( h_{ii} = \left(\frac{1}{n}\right) + (x_i - x_c)'(X_c'X_c)^{-1}(x_i - x_c) \), où \( (x_i - x_c) \) est la différence entre le vecteur des valeurs des variables pour l'observation \( i \), et le vecteur des valeurs moyennes, et \( X_c \) la matrice de taille \((n,p)\) des valeurs centrées.
Le levier en \( i \) est donc une « distance » entre les valeurs des \( X \) prises en \( i \) et les valeurs moyennes calculées sur les \( n \) observations.
Des différentes propriétés de \( H \), on déduit:
\[
\begin{align*}
h_{ii} &= \sum_{j \neq i} h_{jj}^2 + h_{ii}^2 \\
\text{trace}(H) &= p + 1 \\&\Rightarrow \sum_{i=1}^{p} h_{ii} = p + 1 \\
&\Rightarrow \frac{1}{n} \leq h_{ii} \leq 1
\end{align*}
\]
et aussi des formules concernant les variances : \( \text{var}(\hat{Y}_i) = \sigma^2 h_{ii} \) et \( \text{var}(e_i) = \sigma^2 (1 - h_{ii}) \).
On en conclut que \( h_{ii} \) est toujours plus petit ou égal à 1.
**Règle** (colonne *Hat Diag H*): Si les leviers étaient tous égaux, la valeur commune serait \((p+1)/n\). De façon empirique, un levier supérieur à \((p+1)/n\) est suspect.
4.3.2. Résidus studentisés internes
Ils sont appelés en anglais *Standardized Residuals* ou STUDENT.

On connaît la variance de chaque résidu : \( \text{var}(e_i) = \sigma^2 (1-h_{ii}) \). Dans cette formule, on estime \( \sigma^2 \) par 
\[
\hat{\sigma}^2 = \frac{\sum e_i^2}{n-(p+1)} = \text{MSE};
\]
donc le résidu standardisé est :
\[
\hat{r}_i = \frac{e_i}{s \sqrt{1-h_{ii}}}
\]

Bien que numérateur et dénominateur ne soient pas indépendants, on considère que cette quantité \( \hat{r}_i \) suit une loi de Student à \( (n-1-(p+1)) = (n-p-2) \) ddl, d'où le nom STUDENT.

*Règle* (colonne Student) : \( \hat{r}_i \) sera suspect si \( |\hat{r}_i| > 2 \) (quantile de la loi de Student \( (1-\alpha/2) \), pour le seuil \( \alpha=5\% \), avec l'approximation par une loi normale si \( n \) grand).

4.3.3. Résidus studentisés externes
Ils sont appelés en anglais *Studentized Residuals* ou RSTUDENT.

On remplace dans l'expression de \( \hat{r}_i \) l'estimation \( s \) par \( s_{(-i)} \) qui est l'estimation de \( s \) obtenue en refaisant l'ajustement du modèle sans l'observation \( i \), ce qui rend \( e_i \) indépendant de \( s_{(-i)} \) : ceci donne
\[
\hat{r}_{(-i)} = \frac{e_i}{s_{(-i)} \sqrt{1-h_{ii}}}
\]

*Règle* (colonne RStudent) : \( \hat{r}_{(-i)} \) suit aussi une loi de Student à \( (n-1-(p+1)) \) ddl et sera donc aussi suspecte si \( |\hat{r}_{(-i)}| > 2 \). Certains auteurs préconisent d'autres quantiles à un seuil fixé \( \alpha \) pour \( \hat{r}_{(-i)} \) : quantile \( 1-\alpha/2n \) plutôt que \( 1-\alpha/2 \).

4.3.4. Mesure globale de l'influence sur le vecteur des coefficients: Distance de COOK
Pour chaque observation \( i \), on calcule une « distance » entre le vecteur \( B \) des coefficients de la régression et le vecteur \( B_{(-i)} \) obtenu en refaisant la régression sans l'observation \( i \) : la distance se mesure à l'aide de \( (X'X) \) et est normée par \( s^2 \), estimation de \( \sigma^2 \).

\[
(\text{COOKD}) = \frac{|B-B_{(-i)}|}{(X'X)|B-B_{(-i)}|} \left(\frac{r_i^2 h_{ii}}{(p+1)s^2} \right) = \frac{r_i^2 h_{ii}}{(p+1)(1-h_{ii})}
\]

*Règle* : (colonne Cook's D) : La distance de COOK étant normée, une valeur supérieure à 1 est suspecte.\(^{22}\)

\(^{22}\) Certains auteurs suggèrent une limite de \( 4/(n-p-1) \), la calibration à 1 pouvant laisser passer des valeurs influentes.
4.3.5. Influence sur chacun des coefficients : DFBETAS

Pour chaque variable, on calcule la différence entre le coefficient estimé $b_j$ et celui obtenu avec l'estimation sans l'observation i, $b_j(-i)$.
Avec standardisation, on obtient pour chaque variable explicative $j$ :

$$
(DFBETAS)_j = \frac{(b_j - b_j(-i))}{s(-i)\sqrt{((X'X)_j)}^{-1}}
$$

Règle : Empiriquement, un DFBETAS dont la valeur absolue est plus grande que
$$\frac{2}{\sqrt{n}}$$ est suspect.

Utilisation conjointe COOKD et DFBETAS : S'il y a beaucoup de variables, on regarde d'abord les observations globalement influentes (COOKD élevé), puis pour cette observation quelle(s) variable(s) cause(nt) cette influence (DFBETAS).

4.3.6. Précision des estimateurs : COVRATIO

La quantité Mean Square Error (MSE) mesure la précision globale de l'estimation : MSE petit indique une bonne précision. MSE est aussi la variance des résidus.

Ici, on mesure la précision en utilisant une variance « généralisée », évaluée par :

$$s^2|| (X'X)^{-1} ||$$ calculée avec et sans l'observation i (la notation $|| .. ||$ désigne le déterminant de la matrice) :

$$(COVRATIO)_i = \frac{s^2(-i)\sqrt{((X'X)^{-1})_{(-i)}}}{s^2\sqrt{((X'X)^{-1})}}
$$

Donc $(COVRATIO)_i$ plus grand que 1 indique que le fait de mettre l'observation i augmente la précision, alors qu'une valeur plus petite que 1 indique une diminution de la précision.

Règle : Belsley, Kuh et Welsh suggèrent qu'un écart à l'unité dépassant $3(p+1)/n$ est grand.

4.3.7. Influence sur la valeur ajustée : DFFITS

Pour chaque observation i, $(DFFITS)_i$ donne la différence entre la valeur ajustée pour l'observation i et la valeur prédite de Y pour i dans le modèle estimé sans cette observation i. Un grand écart indiquera une forte modification dans la valeur ajustée par le modèle quand l'observation i est retirée.
Avec une standardisation à $s(-i)$ :

$$(DFFITS)_i = \frac{\hat{Y}_i - \hat{Y}_i(-i)}{s(-i)\sqrt{h_{ii}}} = r(-i)\frac{h_{ii}}{\sqrt{1-h_{ii}}}
$$

Règle : DFFITS est déclaré suspect s'il est en valeur absolue plus grand que
$$2\sqrt{\frac{(p+1)}{n}}$$.
4.3.8. Coefficient global PRESS

Predicted Residuals Sum of Squares = \[ \text{PRESS} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i(-i))^2 \]

Ce coefficient (unique) est calculé en faisant \( n \) estimations \( \hat{Y}_i(-i) \) obtenues en enlevant une observation. Il devrait donc être égal à la somme des carrés des résidus du modèle avec toutes les observations (SSResidus) si aucune observation ne pose problème.

Des coefficients PRESS individuels \( (Y_i - \hat{Y}_i(-i))^2 \) peuvent être obtenus uniquement dans la table SAS créé par l'instruction OUTPUT. Ces coefficients peuvent être comparés aux coefficients DFFITS, comme le montre la figure 4.4 ci-dessous :

![Figure 4.4 : Illustration de PRESS et DFFITS](image)

4.3.9. Comment obtenir les mesures d'influence dans SAS

**Dans PROC REG**

Elles sont affichées en sortie par des options de l'instruction MODEL.
- Option R ➔ pour toutes les observations, le résidu et son écart-type, les résidus standardisés STUDENT, la distance de COOK et un dessin indiquant la position du résidu par rapport à l'intervalle [-2 ; +2] ;
- Option INFLUENCE ➔ pour toutes les observations, levier, résidu studentisé externe RSTUDENT, COVRATIO, DFFITS et DFBETAS de chaque coefficient.

Certaines mesures peuvent être stockées dans une table de sortie SAS (table possédant \( n \) lignes au moins), à l'aide de l'instruction OUTPUT : on crée des variables en utilisant les mots-clefs R, STUDENT, RSTUDENT, H, COOKD, DFFITS et PRESS.

L'option PRESS de l'instruction MODEL permet d'obtenir le coefficient PRESS « global », qui sera affiché en sortie, et stocké dans la table de l'instruction OUTPUT si cette instruction existe.
Dans SAS/INSIGHT

Une fois que l'on a exécuté le modèle, on peut ajouter à la table SAS sur laquelle on travaille, des variables à l'aide du menu Vars. Les variables ajoutées ont alors le nom indiqué entre parenthèses ci-dessous (où Y est le nom de la variable réponse du modèle) :

- Hat Diag, (H_Y)
- Residual (R_Y)
- Standardized residual (RS_Y)
- Studentized residual (RT_Y)
- Cook’s D (D_Y)
- Dffits (F_Y)
- Covratio (C_Y)
- Dfbetas (BY_Intercept, BY_X1, BY_X2, etc.)

| Tableau récapitulatif |
|-----------------------|----------------|----------------|-----------------|
|                        | Std Err Residual | Student Residual | Rstudent        | Hat Diag H      |
| signifiant             | estimateur de l'erreur-type du résidu i | résidus studentisés internes, appelés standardized residual dans SAS-Insight | résidus studentisés externes, appelés studentized residual dans SAS-Insight | levier de l'obs. i |
| objet                  | permet de calculer l'intervalle de confiance autour du résidu i | test de significativité du résidu i | à comparer avec Student Residual écart-type calculé en retirant l'obs. i | mesure l'influence de l'obs.i à cause des valeurs xi |
| valeurs critiques      |                | 2              | 2              | 2(p+1) / n      |
| Règle de décision      | [\text{Student residual}] > 2 \text{ alors le résidu i est significativement } \neq 0 | [\text{RStudent}] > 2 \text{ alors l'observation i nécessite une investigation !} | h_{ii} > \frac{2(p+1)}{n} | nécessite une investigation |
| Option de PROC REG     | R              | R              | Influence      | Influence       |

©WP CES  2007.47 - Premiers pas en régression linéaire avec SAS- Confais Le Guen - 93
<table>
<thead>
<tr>
<th>signifiant</th>
<th>Cook's D</th>
<th>Df betas</th>
<th>Cov Ratio</th>
<th>Dffits</th>
</tr>
</thead>
<tbody>
<tr>
<td>distance de Cook</td>
<td>DFBETAS relatif à chaque coefficient $\beta_j$</td>
<td>Ratio de MSE sans et avec l'observation $i$</td>
<td>statistique DFFITS</td>
<td></td>
</tr>
<tr>
<td>objet</td>
<td>mesure le changement en retirant l'obs. $i$, sur les estimations de l'ensemble des coefficients</td>
<td>mesure normalisée de l'effet de l'obs. $i$ sur l'estimation, pour chaque coefficient $\beta_j$</td>
<td>mesure l'effet de l'obs. $i$ sur la précision</td>
<td>mesure normalisée du changement dans la valeur prédite, avec et sans l'obs. $i$</td>
</tr>
<tr>
<td>valeurs critiques</td>
<td>$1$ ou $\frac{4}{(n-p-1)}$</td>
<td>$\frac{2}{\sqrt{n}}$</td>
<td>$\frac{3(p+1)}{n}$</td>
<td>$2\sqrt{\frac{(p+1)}{n}}$</td>
</tr>
<tr>
<td>Règle de décision</td>
<td>CookD &gt; 1 alors l'observation $i$ est influente globalement</td>
<td>$</td>
<td>Dfbetas</td>
<td>&gt; \frac{2}{\sqrt{n}}$ indique une influence de l'obs. $i$ sur l'estimation de $\beta_j$</td>
</tr>
<tr>
<td>Option de PROC REG</td>
<td>$R$, Influence</td>
<td>Influence</td>
<td>$R$</td>
<td>$R$</td>
</tr>
</tbody>
</table>

On trouvera dans le programme SAS ci-dessous une macro CRITIQUE (avec comme paramètres : $n$ nombre d'observations, $p$ nombre de régresseurs, et $b0$ indicateur de la présence d'une constante) permettant d'afficher les valeurs critiques des différentes mesures d'influence.
4.3.11. Exemple

On utilise les données « Processionnaire du pin » issu du livre de Tomassone et al.(1983) et on se limite de nouveau au modèle $Y = \log = f(X_1 \ X_2 \ X_4 \ X_5)$.

Voici les valeurs limites pour les coefficients d'influence :

<table>
<thead>
<tr>
<th>Obs</th>
<th>n</th>
<th>p</th>
<th>b0</th>
<th>dcook</th>
<th>hat_diag</th>
<th>covratio</th>
<th>dffits</th>
<th>dfbetas</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>44</td>
<td>4</td>
<td>1</td>
<td>0.10256</td>
<td>0.22727</td>
<td>0.34091</td>
<td>0.67420</td>
<td>0.30151</td>
</tr>
</tbody>
</table>

La table des valeurs des coefficients d'influence de toutes les observations est donnée ci-après. Mais plutôt que de rechercher les valeurs limites dans cette table, de simples Box-Plot permettent de les repérer rapidement.
Dessins BOX-PLOT des coefficients d’influence :

Résidus standardisés

<table>
<thead>
<tr>
<th>Internes</th>
<th>Externes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leviers</td>
<td>CovRatio</td>
</tr>
<tr>
<td>Dffits</td>
<td>DCook</td>
</tr>
</tbody>
</table>
La distribution des leviers ne révèle pas d'outliers, mais au vu des valeurs, on constate que quelques observations ont cependant un levier un peu trop grand (> 0.23) : 8, 10, 12, 21 et 32.

Les observations 20 et 33 ont des résidus « grands » (|STUDENT| et |RDSTUDENT| > 2) et donc un grand effet sur la précision (COVRATIO -1) > 0.34 car quand on les enlève de la régression, SSE diminue fortement.

Et elles influent également sur |DFFITS| >> 0.67. Les observations 10 et 21 ont une forte incidence sur l'ensemble des coefficients (DCOOK > 0.10) et donc sur certains coefficients (DFBETAS > 0.30), ceux de X1 et X2 pour les 2 observations, celui de X4 pour l'observation 10.

Pour résumer, les observations 10 20 21 et 33 seraient atypiques. Cependant, au vu des données, où ces observations sont des placettes issues d'un plan d'expérience, on peut penser qu'il n'est pas judicieux de les retirer de la régression.

Le dessin humoristique de la figure 4.5 ci-dessous (publié dans le manuel du logiciel OSIRIS, mais l'auteur nous est inconnu) illustre le mauvais réflexe que pourrait avoir le statisticien au vu de données atypiques !
4.4. Collinéarité des régresseurs

Différents symptômes sont révélateurs de problèmes de collinéarité :
- De grandes corrélations entre les régresseurs ;
- Un grand changement dans les coefficients quand on ajoute ou enlève un régresseur ;
- Des coefficients non significatifs alors que le test global d’analyse de variance sur tous les coefficients est significatif ;
- La non significativité et/ou une très grande variance pour le coefficient d’un régresseur théoriquement important dans le modèle ;
- Un coefficient de signe opposé à celui auquel on s’attendait.

Ce sont des problèmes d’inversion de $X'X$, qui entraînent une augmentation des variances des coefficients, et donc leur instabilité car $\text{Var}(\hat{\beta}) = \sigma^2 (X'X)^{-1}$. Et la non-inversion de $X'X$ se rencontre quand il existe des combinaisons linéaires entre les colonnes de $X$. 
L’exemple simple de la figure 4.6 illustre le phénomène d’instabilité dû à la liaison entre deux régresseurs.

![Figure 4.6 : Projection de Y dans l’espace de 2 régresseurs corrélés](image)

Dans l'espace $\mathbb{R}^n$, les cosinus d’angle entre variables sont les corrélations (voir chapitre 1, §1.2.5): Ici $X_1$ et $X_2$ étant très corrélées, l’angle entre les 2 vecteurs est petit, ce qui rend le sous-espace $(X_1,X_2)$, et donc la projection, instables.

### 4.4.1. Méthodes basées sur l’étude de la matrice $X'X$

La matrice $X$ possède $n$ lignes et $(p+1)$ colonnes :

$$ X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} $$

La matrice $(X'X)$ est une matrice carrée $(p+1)$, symétrique, dont les éléments sont calculés ainsi :

$$ (X'X)_{ij} = \begin{cases} n & \text{si } i = j \\ \sum_{k=1}^{n} x_{ik} x_{jk} & \text{si } i \neq j \end{cases} $$

En étudiant les leviers au §4.3.1, on a utilisé les variables régresseurs centrées. Si on considère la matrice $X_c$ de taille $(n,p)$ ayant en colonnes les $p$ régresseurs centrés, alors $(X_c'X_c) = n \cdot \text{COV}$ où COV est la matrice de variance-covariance des variables $X_j$. Quand les variables sont non-correlées, alors $(X_c'X_c)$ est une matrice diagonale puisque toutes les covariances sont nulles.

---

24 Le site de Chong Ho Yu illustre de façon très amusante ce problème d’instabilité [http://creative-wisdom.com/computer/sas/collinear_subject_space.html](http://creative-wisdom.com/computer/sas/collinear_subject_space.html)
Donc $X'X$ est égal à $nR$ si on travaille sur des régresseurs centrés et réduits, où $R$ est la matrice carrée symétrique des coefficients de corrélation entre les $p$ régresseurs.

**Etude de la matrice de corrélation des régresseurs**

Cette étude se révèle intéressante.

- Si les régresseurs ne sont pas corrélés entre eux, cette matrice, notée $R$, n’a que des 1 sur sa diagonale, et 0 ailleurs : c’est la matrice identité.

- La matrice $R$ est symétrique définie et positive, et de rang $p$. Les valeurs propres de $R$ sont donc en nombre de $p$, elles sont positives et leur somme vaut $p$ (trace de R). Quand il n’y a aucune liaison entre les régresseurs, elles sont toutes égales à 1 car $R$ est la matrice identité. Sinon, on aura des valeurs propres plus petites et même proches de 0 : l’examen des valeurs propres révèlera donc les problèmes de liaison entre les régresseurs.

- $R^{-1}$, matrice inverse de $R$, est également riche en informations. En effet on peut montrer que ses éléments diagonaux $(R_j^{-1})$ sont égaux à $1/(1 - R^2_j)$ où $R^2_j$ est le coefficient de corrélation multiple de la régression avec constante de $X_j$ sur les $(p-1)$ autres variables.

- Si on définit un modèle avec les régresseurs centrés et réduits, en remplaçant l’expression de $(X'X)$ dans l’estimateur des moindres carrés, on montre (cf. Woolridge (2000)) que le vecteur des coefficients $B_c$ de ce modèle est $B_c = \frac{1}{n} R^{-1} X' Y$, et la matrice de variance-covariance vaut $\text{Var}(B_c) = \frac{\sigma^2}{n} R^{-1}$.

Une étude préalable de la matrice de corrélation des régresseurs, complétée éventuellement par une Analyse en Composantes Principales, s'impose donc et permet de visualiser les liaisons entre variables.

REG possède des options TOL, VIF et COLLIN, COLLINOINT pour détecter des problèmes de colinéarité selon deux optiques différentes. SAS/ INSIGHT donne les indices TOL et VIF par défaut, et affiche la table COLLIN uniquement.

### 4.4.2. Variance Inflation Factor

On a vu au §4.4.1 que la matrice de variance–covariance des coefficients du modèle où les régresseurs sont centrés et réduits est $R^{-1}$, à un facteur près. Donc l’élément diagonal $j$ de cette matrice mesure comment la variance du coefficient de $X_j$ sera augmentée par la colinéarité.

Pour chaque variable $X_j$, on nomme cet élément $VIF_j$ (*Variance Inflation Factor* ou *inflation de variance*) $VIF_j = 1/(1 - R^2_j)$ où $R^2_j$ est le coefficient de corrélation multiple de la régression avec constante de $X_j$ sur les $(p-1)$ autres variables.
S'il y a colinéarité, alors $R^2_j$ est proche de 1, donc $VIF_j$ est grand. Comme la loi de ce coefficient n'est pas connue, Belsley et al. ont défini un seuil limite de façon empirique.

**Règle** : une valeur de $VIF$ plus grande que 10 révèle un problème.

Tomassone (1983) propose de calculer un indice global de colinéarité défini comme la somme des $VIF$ de tous les régresseurs :

$$I = \frac{1}{p} \sum_{j=1}^{p} VIF_j$$

Remarque : la tolérance (option TOL) est définie comme l'inverse de la variance inflation

$$TOL_j = \frac{1}{VIF_j}$$

### 4.4.3. Condition index et variance proportion

On a signalé au §4.4.1 que l'étude les valeurs propres de la matrice de corrélations révèle les problèmes de liaisons entre les régresseurs.

Cette étude se fait aussi par l'analyse en composantes principales (ACP), qui consiste à transformer des variables pour obtenir d'autres variables orthogonales, qui sont des combinaisons linéaires des premières, appelées composantes principales (cf. Saporta (2006) ou Tenenhaus (1994)).

Plus précisément si $Z$ est la matrice $(n,p)$ des variables initiales centrées et réduites , on construit la matrice $W (n,p)$ des variables orthogonales avec la relation $W = ZU$. En ACP, on démontre que la matrice $U$ est la matrice des vecteurs propres normés de $R$, associés aux $p$ valeurs propres $(\lambda_k, k = 1,2,...,p)$, qui sont positives car $R$ est symétrique définie positive. On les ordonne de la plus grande à la plus petite : une liaison parfaite entre les variables $Z$ entraîne une nullité des dernières valeurs propres.

On montre également que la variance d'une composante $W_k$ est égale à $\lambda_k$.

Si on construit le modèle avec les variables $W$ \(^25\), on trouve donc la solution des moindres carrés $Y = W\hat{c} + e$, avec $\hat{c} = (W'W)^{-1} W'Y$.

Et la matrice de variance-covariance des coefficients est $\text{Var}(\hat{c}) = \sigma^2 (W'W)^{-1}$ ; pour un coefficient : $\text{var}(c_k) = \frac{\sigma^2}{n\lambda_k}$.

Comme les variables sont orthogonales et de variance égale à la valeur propre, on en déduit $Y = W\hat{c} + e = \Omega^{-1}W'Y$, avec $\Omega^{-1} = \text{Diag}\left(\frac{1}{n\lambda_k}\right), k = 1,..,p$ matrice diagonale.

\(^{25}\) On trouvera dans le chapitre 6 du livre Tomassone et al (1992), le principe de cette méthode qu'il appelle « régression orthogonalisée »
On passe facilement de $\hat{c}$ à $\tilde{\alpha}$ car $Y = W\hat{c} + e = ZU\hat{c} + e = Z\tilde{\alpha} + e$, et donc $\tilde{\alpha} = U\tilde{c}$.
La matrice de variance-covariance est $\text{Var}(\tilde{\alpha}) = U\text{Var}(\tilde{c})U'$ ; pour le coefficient du régresseur $j$, $\text{var}(\tilde{\alpha}_j) = \frac{\sigma^2}{n} \sum_{k=1}^p \frac{U_k^2}{\lambda_k}$ : des valeurs propres faibles entraînent donc de grandes variances des coefficients.

**Les indices de colinéarité**

A - Tout d'abord, l'édition des valeurs propres donnera des informations sur l'existence de colinéarité.

De façon générale, on calcule les valeurs propres de la matrice $(X'X)$ du modèle, préalablement transformée pour avoir uniquement la valeur 1 sur les éléments diagonaux. Il y a donc $(p+1)$ valeurs propres : c'est l'option COLLIN.
Si on travaille sur un modèle avec les régresseurs centrés comme ci-dessus, il y aura $p$ valeurs propres : c'est l'option COLLINOINT.

On édite ces valeurs propres de la plus grande $\lambda_1$ à la plus petite $\lambda_L$, ($L=p$ ou $p+1$). Une valeur propre nulle révèle l'existence d'une dépendance linéaire entre les colonnes de $X$, donc une colinéarité.

On nomme « Condition Index » le rapport $\text{CI} = \sqrt{\frac{\lambda_1}{\lambda_k}}$, appelé aussi « indice de conditionnement ». Le dernier de ces rapports $\text{CI} = \sqrt{\frac{\lambda_1}{\lambda_L}}$ ($L=p$ ou $p+1$) est nommé « Condition Number ».

Comme la loi de ce coefficient n'est pas connue, Belsley et al. ont défini un seuil limite de façon empirique :

**Règle (colonne Condition Index)** : une valeur grande met en évidence un problème ; empiriquement $\text{CI} > 30$ avec l'option COLLINOINT, ou $\text{CI} > 100$ avec COLLIN.

Remarque : On peut définir une « indice de multicolinearité » en calculant la moyenne des inverses des valeurs propres (cf. Foucart 2006, 2007) : Cet indice serait calculé comme $I = \frac{1}{p} \left( \sum_{k=1}^p \frac{1}{\lambda_k} \right)$ si on considère que les régresseurs sont centrés et réduits ($p$ valeurs propres).

B – Ensuite, pour chaque valeur propre et donc chaque CI, sont données des « VARiance PROPortions », qui indiquent quelles variables sont responsables de la colinéarité révélée par cette valeur propre.
En effet, on a vu que la matrice de variance-covariance des coefficients $\tilde{\alpha}$ de la régression sur les variables centrées et réduites est $\text{Var}(\tilde{\alpha}) = U\text{Var}(\tilde{c})U'$ et que pour un coefficient $j$, $\text{var}(\tilde{\alpha}_j) = \frac{\sigma^2}{n} \sum_{k=1}^p \frac{U_k^2}{\lambda_k}$.
La colonne « proportion de variance » pour le coefficient d'une variable j est le vecteur \( \text{Var. Prop.}_j = \left( \frac{U_i^2}{\lambda_k}, k = 1, \ldots, p \text{ ou } (p+1) \right) \), normé pour que la somme de ses composantes soit égale à 1.

Règle : d'après Belsley, Kuh et Welsh, si les proportions de variance de plusieurs variables sont plus grandes que 0.50 pour un « condition index » grand, les variables correspondantes ont un problème de colinéarité entre elles.

Remarque : l'option COLLINOINT exclut la constante des estimateurs de coefficients; les p variables sont centrées et réduites et \((X'X)\) est donc, à un coefficient près, la matrice de corrélation entre les p variables explicatives. L'option COLLIN inclut la variable constante dans les estimations de coefficients. \((X'X)\) contient donc la variable constante égale à 1. La matrice \((X'X)\), de taille \((p+1)\), est normée pour avoir 1 sur la diagonale, mais les variables ne sont pas centrées. Belsley, Kuh et Welsh recommandent de n'utiliser l'option COLLIN que si la constante a une interprétation physique. Centrer les variables (option COLLINOINT) consiste à supposer que la constante n'a pas d'effet sur la colinéarité des autres variables régresseurs. De plus, ceci est cohérent avec les calculs du « Variance Inflation Factor ». On trouvera dans l'article d'Hélène Rousse-Erkel (1990) des précisions et des prolongements aux travaux de Belsley et al. sur la colinéarité.

4.4.4. Remèdes en cas de multi-colinéarité

- Retirer certains régresseurs, principaux « responsables » de la colinéarité ;
- Les transformer par des ratios si on identifie le facteur commun de liaison ;
- Augmenter la taille n de l'échantillon avec le recueil d'autres observations ;
- Sélectionner les régresseurs, si p est trop grand par rapport à n ;
- Faire une Ridge-Regression \(^{26}\) (transformer \((X'X)\) en \((X'X + kI)\) (Hoerl et Kennard (1970)) ;
- Travailler sur les composantes principales issues des régresseurs \(^{27}\) ;
- Faire une régression PLS ;
- Utiliser les méthodes de type « LASSO » de Tibshirani (1996) ;
- Etc. .

La régression RIDGE et la régression sur composantes principales peuvent être réalisées à l'aide des options RIDGE et PCOMIT de l'instruction Proc REG (voir l'annexe 1 pour la syntaxe).

Remarques :
- La régression PLS (Partial Least Square) semble une méthode plus efficace que la régression Ridge ou la régression sur composantes principales en cas de

\(^{26}\) Voir l’exemple au §4.4.5.
\(^{27}\) C’est la « régression orthogonalisée » de Tomassone (§4.4.3.)
colinéarité, et s'applique aussi au cas où p est très grand par rapport à n (voir les publications de Tenenhaus (1995,1998)).

PROC ORTHOREG de SAS propose d'autres solutions pour réaliser une régression sur données « mal conditionnées », c'est à dire en cas de colinéarités des variables.

4.4.5. Exemple

On utilise les données « Processionnaire du pin » issu du livre de Tomassone et al.(1983), avec le modèle \( Y = \log = f(X_1 \ X_2 \ X_4 \ X_5) \).

```plaintext
proc reg data=libreg.chenilles ; title 'option TOL VIF';
LOG : model log=X1 X2 X4 X5 /tol vif collinoint;
run;
quit;
```

Aucune des valeurs VIF ne sont trop grandes, et tous les coefficients sont significatifs.

Dans la colonne « condition Index » (traduit dans la version française de SAS par « Index de condition ») il n'y a pas de grandes valeurs. Sur la 4\textsuperscript{ème} et dernière ligne (c'est celle de « condition number »), en regardant les proportions de variance, on constate que les 2 variables X4 et X5 sont les « responsables » de la faiblesse de la 4\textsuperscript{ème} valeur propre : on avait vu au chapitre 2 (§2.3.2) que c'est le couple de régresses le plus corrélé.
**Regression RIDGE**

Cette méthode, due à Hoerl et Kennard (1970), consiste à modifier \((X'X)\) pour la rendre inversible. Pour cela on ajoute un terme constant \(k\) à la diagonale \((0 \leq k \leq 1)\). La solution des moindres carrés sera donc obtenue en inversant \((X'X + kI)\) : les coefficients obtenus sont appelés « coefficients ridge ». On trace ensuite la variation des coefficients ridge en fonction de \(k\) : c’est la « Ridge Trace ». On détermine la valeur de \(k\) à partir de laquelle les coefficients se stabilisent : ce sera la valeur choisie.

```sas
title 'Ridge-regression sur le modèle à 4 variables' ;
proc reg data=libreg.chenilles ridge = 0 to 1 by 0.05 outest = coeff_ridge ;
  LOG: model log=X1 X2 X4 X5 ;
  plot / ridgeplot ;
run;
quit;
proc print data = coeff_ridge ;
run ;
```

Ici les coefficients ridge se stabilisent pour \(k \approx 0.3\).
Les valeurs des coefficients sont alors lues dans la table `coeff_ridge`. Dans cette table, la première ligne est le modèle habituel, et la deuxième correspond à \(k=0\), ce qui est le même modèle.

Pour `_RIDGE_ = 0.3` ,
\(b_1 = -0.003281930\), \(b_2 = -0.047532\); \(b_4 = -0.40556\); \(b_5 = 0.05207\).
4.5. Choix des régresseurs

Ce choix s’avère nécessaire en particulier si le nombre d’observations est petit par rapport au nombre de régresseurs, à cause du rang de X’X qui peut devenir plus petit que p. Ceci peut entraîner une instabilité des coefficients comme on l’a vu au paragraphe précédent.

Soit un modèle avec n observations et p régresseurs ; on sélectionne dans les cas suivants (cette liste est non exhaustive) :

1. n petit par rapport à p ;
2. colinéarité des régresseurs ;
3. choix d’un modèle plus simple pour la prédiction (principe de PARCIMONIE). (1 et 2 entraînent des problèmes d’inversion de X’X).

Proc REG permet ce choix par l’option « SELECTION = method », de l’instruction MODEL. Il n’y a pas de sélection dans SAS/INSIGHT.

4.5.1. Utilisation des sommes de carrés

La formule de base est : $SS_{Totale} = SS_{Modèle} + SS_{Erreurs}$

**Rappel sur les somme de carrés apportés par un régresseur**
Les sommes de carrés apportés par les régresseurs peuvent être obtenues par les options SS1 et SS2 de l'instruction MODEL de REG, ce qui a déjà été vu dans le chapitre 2 (§2.4.1), ou bien par les « Type III Tests » dans SAS/INSIGHT.

- SS1(Xj) = somme des carrés apportée par la variable Xj introduite en séquence dans la régression, la régression contenant uniquement les variables qui la précèdent dans la liste de variables explicatives de l'instruction MODEL.

- SS2(Xj) = somme des carrés apportée par la variable Xj, lorsque l'ensemble des (p-1) autres régresseurs est déjà dans la régression. Ce sont les sommes de carrés données par la table « Type III Tests » de SAS/INSIGHT.

SS2(Xj) correspond au calcul du carré de la différence entre la valeur de Y estimée par la régression avec les p variables et celle estimée dans la régression à (p-1) variables, sans Xj. Pour le choix de régresseurs, le deuxième calcul d'apport de somme de carrés est le plus intéressant, car il ne dépend pas de l'ordre d'introduction des variables dans le modèle. On notera « SS apporté par j » cette quantité SS2(Xj).

### Tests des apports à SS_{Modèle d'une variable}

Les tests décrits dans le chapitre 2 (§2.4.3), ne sont pas faits par l'option SS2 de l'instruction MODEL de REG, mais sont donnés dans la table « Type III Tests » de SAS/INSIGHT. Plus généralement, comme on l’a vu au §2.4.4, un modèle sans r variables est appelé modèle restreint par opposition au modèle complet à p variables.

- RRSS (Restricted Residual Sum of Squares) = Somme des carrés des résidus du modèle restreint
- URSS (Unrestricted Residual Sum of Squares) = somme des carrés des résidus du modèle complet.

La valeur de la statistique du test est

\[
F = \frac{(RRSS - URSS)/r}{URSS/(n-p-1)}
\]

Dans le cas d'une seule variable, r vaut 1, et donc en passant aux sommes de carrés du modèle :

\[
F = \frac{(RRSS - URSS)/1}{URSS/(n-p-1)} = \frac{SS_{Modèle complet} - SS_{Modèle sans j}}{MSE}.
\]

Les tests de significativité de ces sommes de carrés sont donc réalisés à l'aide d'une statistique F, obtenue en divisant SS par la quantité MSE (Mean Square Error) du modèle avec constante contenant tous les régresseurs. De plus, la valeur de F associée à SS est aussi le carré du t de Student du coefficient de la variable j dans la régression à p régresseurs.
**Exemple d'élimination progressive**


**Modèle à 10 variables**

\[
\begin{align*}
\log &= 10.9984 - 0.0044 \times X_1 - 0.0538 \times X_2 + 0.0679 \times X_3 \\
&\quad - 1.2936 \times X_4 + 0.2316 \times X_5 - 0.3568 \times X_6 - 0.2375 \times X_7 \\
&\quad + 0.1811 \times X_8 - 1.2853 \times X_9 - 0.4331 \times X_{10}
\end{align*}
\]

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Stat</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>10</td>
<td>34.4662</td>
<td>3.4466</td>
<td>5.01</td>
<td>0.0008</td>
</tr>
<tr>
<td>Error</td>
<td>22</td>
<td>15.1299</td>
<td>0.6877</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C Total</td>
<td>32</td>
<td>49.5960</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Estimate</th>
<th>Std Error</th>
<th>t Stat</th>
<th>Pr &gt;</th>
<th>Tolerance</th>
<th>Var Inflation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>10.9984</td>
<td>3.0603</td>
<td>3.59</td>
<td>0.0016</td>
<td>0.5327</td>
<td>1.8774</td>
</tr>
<tr>
<td>X1</td>
<td>1</td>
<td>-0.0044</td>
<td>0.0016</td>
<td>-2.85</td>
<td>0.0094</td>
<td>0.8400</td>
<td>1.1904</td>
</tr>
<tr>
<td>X2</td>
<td>1</td>
<td>-0.0538</td>
<td>0.0219</td>
<td>-2.46</td>
<td>0.0223</td>
<td>0.5017</td>
<td>4.8708</td>
</tr>
<tr>
<td>X3</td>
<td>1</td>
<td>0.0679</td>
<td>0.0366</td>
<td>1.88</td>
<td>0.0624</td>
<td>0.0624</td>
<td>16.0219</td>
</tr>
<tr>
<td>X4</td>
<td>1</td>
<td>1.2936</td>
<td>0.5636</td>
<td>2.29</td>
<td>0.0317</td>
<td>0.0624</td>
<td>16.0219</td>
</tr>
<tr>
<td>X5</td>
<td>1</td>
<td>0.2316</td>
<td>0.1044</td>
<td>2.22</td>
<td>0.0371</td>
<td>0.0624</td>
<td>16.0219</td>
</tr>
<tr>
<td>X6</td>
<td>1</td>
<td>-0.3568</td>
<td>1.9665</td>
<td>-0.23</td>
<td>0.8219</td>
<td>0.0170</td>
<td>58.7558</td>
</tr>
<tr>
<td>X7</td>
<td>1</td>
<td>-0.2375</td>
<td>1.0010</td>
<td>-0.24</td>
<td>0.8156</td>
<td>0.0096</td>
<td>1.6491</td>
</tr>
<tr>
<td>X8</td>
<td>1</td>
<td>0.1811</td>
<td>0.2367</td>
<td>0.76</td>
<td>0.4525</td>
<td>0.0693</td>
<td>14.4226</td>
</tr>
<tr>
<td>X9</td>
<td>1</td>
<td>-1.2853</td>
<td>0.8648</td>
<td>-1.49</td>
<td>0.1514</td>
<td>0.0895</td>
<td>11.1696</td>
</tr>
<tr>
<td>X10</td>
<td>1</td>
<td>-0.4331</td>
<td>0.7348</td>
<td>-0.59</td>
<td>0.5616</td>
<td>0.0657</td>
<td>1.6509</td>
</tr>
</tbody>
</table>

C'est un modèle globalement bon (F d'analyse de variance significatif), mais les coefficients des régresseurs X3 X6 X7 X8 X9 et X10 sont non significatifs. De plus, l’inflation de variance est plus grande que 10 pour les régresseurs X3 X4 X6 et X9. Le modèle n’est donc pas un « bon modèle » : il faut éliminer des régresseurs. Pour cela on effectue les tests sur les apports de sommes de carrés.
Les variables X3, X6, X7, X8, X9, et X10 ont un apport non significatif : on élimine la variable X6 dont l’apport est le plus petit.

Remarque : on peut vérifier ici que $F(X6) = t^2(X6) \Rightarrow 0.0529 = (-0.23)^2$ (cf. chapitre 1, §1.3.4)

→ On élimine X7

→ On élimine X8

→ On élimine X10
Type III Tests

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Stat</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>1</td>
<td>6.6711</td>
<td>6.6711</td>
<td>10.81</td>
<td>0.0029</td>
</tr>
<tr>
<td>X2</td>
<td>1</td>
<td>4.2645</td>
<td>4.2645</td>
<td>6.91</td>
<td>0.0142</td>
</tr>
<tr>
<td>X3</td>
<td>1</td>
<td>0.9272</td>
<td>0.9272</td>
<td>1.50</td>
<td>0.2314</td>
</tr>
<tr>
<td>X4</td>
<td>1</td>
<td>4.7457</td>
<td>4.7457</td>
<td>7.69</td>
<td>0.0101</td>
</tr>
<tr>
<td>X5</td>
<td>1</td>
<td>4.3079</td>
<td>4.3079</td>
<td>6.98</td>
<td>0.0138</td>
</tr>
<tr>
<td>X9</td>
<td>1</td>
<td>1.4451</td>
<td>1.4451</td>
<td>2.34</td>
<td>0.1381</td>
</tr>
</tbody>
</table>

-On élimine X3

Type III Tests

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Stat</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>1</td>
<td>5.8582</td>
<td>5.8582</td>
<td>9.32</td>
<td>0.0051</td>
</tr>
<tr>
<td>X2</td>
<td>1</td>
<td>4.0243</td>
<td>4.0243</td>
<td>6.40</td>
<td>0.0176</td>
</tr>
<tr>
<td>X4</td>
<td>1</td>
<td>6.2247</td>
<td>6.2247</td>
<td>9.90</td>
<td>0.0040</td>
</tr>
<tr>
<td>X5</td>
<td>1</td>
<td>5.6406</td>
<td>5.6406</td>
<td>8.97</td>
<td>0.0058</td>
</tr>
<tr>
<td>X9</td>
<td>1</td>
<td>0.5259</td>
<td>0.5259</td>
<td>0.84</td>
<td>0.3685</td>
</tr>
</tbody>
</table>

-On élimine X9

Type III Tests

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Stat</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>X1</td>
<td>1</td>
<td>7.3067</td>
<td>7.3067</td>
<td>11.69</td>
<td>0.0019</td>
</tr>
<tr>
<td>X2</td>
<td>1</td>
<td>5.4683</td>
<td>5.4683</td>
<td>8.75</td>
<td>0.0062</td>
</tr>
<tr>
<td>X4</td>
<td>1</td>
<td>11.2399</td>
<td>11.2399</td>
<td>17.98</td>
<td>0.0002</td>
</tr>
<tr>
<td>X5</td>
<td>1</td>
<td>8.6123</td>
<td>8.6123</td>
<td>13.78</td>
<td>0.0009</td>
</tr>
</tbody>
</table>

On a obtenu un modèle dont les apports de toutes les variables sont tous significatifs : on arrête donc le processus.
La procédure réalisée donne donc comme modèle final $Y = f(X1 \times X2 \times X4 \times X5)$ ; qui est le modèle déjà étudié et qui s’est révélé être un modèle « correct », vérifiant les suppositions de base sur les erreurs (§4.2.5), où 4 observations sont atypiques (§4.3.11), et dans lequel les 4 régresseurs n’ont pas de problème de colinéarité (§4.4.5).

4.5.2. Différentes méthodes basées sur les sommes de carrés

*Méthode FORWARD (ascendante)*

On introduit les variables une par une : on commence par un modèle à une variable, et on ajoute à chaque étape une variable. Les $SS_{\text{Modèle}}$ augmentent forcément (gain) et le principe est de faire entrer à chaque pas la variable qui apportera l'augmentation la plus significative de la somme des carrés du modèle. Donc, la variable qui est introduite est celle qui a $SS_{\text{apporté par } j}$ maximum, donc qui possède le F le plus grand, et significatif avec une probabilité par défaut associée au F de 0.5 : ce seuil s'appelle le seuil «pour entrer» SLE de REG.
Il y a au plus $p$ modèles sélectionnés, qui sont affichés par ordre croissant de $k$ ($k=1$ à $L$, $L \leq p$).
Méthode BACKWARD (descendante)\textsuperscript{28}

On part de la régression à p régresseurs, et on élimine à chaque pas la variable la moins significative, c'est-à-dire qu'on élimine la variable ayant SS apporté par j minimum, c'est-à-dire le F ou le t de Student le plus petit (probabilité par défaut associée au F de 0.10 : seuil «pour sortir» SLS de REG). Il y a au plus p modèles sélectionnés, qui sont affichés par ordre décroissant de k (k=p à L, L ≥ 1.

Méthode STEPWISE (progressive)

C'est une combinaison FORWARD/BACKWARD : on effectue une sélection FORWARD, en laissant la possibilité de faire sortir du modèle à chaque pas une des variables devenue non significative (seuils de probabilité «pour entrer» 0.15, «pour sortir» 0.15 par défaut dans REG).

Remarque : les méthodes FORWARD, BACKWARD et STEPWISE ne donnent pas forcément le meilleur sous-ensemble à k variables. On peut le voir sur l'exemple ci-dessous, extrait de Brenot, Cazes et Lacourly (1975). On considère un modèle à 3 régresseurs. La figure 4.6 illustre graphiquement les liaisons dans R\textsuperscript{n} :

\textbf{Figure 4.6 : représentation géométrique dans R\textsuperscript{n} de y et x1 x2 x3}

Dans l'espace R\textsuperscript{n}, y, x1 et x3 sont coplanaires, et x1, qui n’appartient pas au plan, est telle que son coefficient de corrélation avec y est le plus fort des coefficients de corrélations de y avec les 3 régresseurs.

La meilleure régression avec 2 variables est donc y=f(x1). La meilleure régression à 2 variables est donc y=f(x2, x3).

Les méthodes de sélection donneront les choix successifs :

\begin{align*}
\text{BACKWARD :} & \quad (x_1, x_2, x_3) \rightarrow (x_2, x_3) \rightarrow x_2 \text{ ou } x_3 \\
\text{FORWARD :} & \quad x_1 \rightarrow (x_1, x_2) \text{ ou } (x_1, x_3) \rightarrow (x_1, x_2, x_3) \\
\text{STEPWISE :} & \quad x_1 \rightarrow (x_1, x_2) \text{ ou } (x_1, x_3) \rightarrow (x_1, x_2, x_3) \rightarrow (x_2, x_3)
\end{align*}

Les modèles à 1 ou à 2 variables trouvés par ces méthodes sont différents, et ne sont pas forcément les meilleurs : BACKWARD trouve le meilleur modèle à 2 variables, mais pas celui à 1 ; FORWARD et STEPWISE trouvent le meilleur modèle à 1 variable mais pas celui à 2 ; la méthode STEPWISE effectue un pas de plus que FORWARD.

\textsuperscript{28} C'est la méthode réalisée au §4.5.1.
Exemples de sélection STEPWISE

On sélectionne parmi les 10 variables candidates dans les données « Processionnaire du pin » issu du livre de Tomassone et al. (1983).

- Tout d’abord, on conserve les seuils par défaut SLE = SLS = 0.15 de la méthode.

```plaintext
title 'regression STEPWISE '; proc reg data =libreg.chenilles ;
log10 : model log = x1--x10 / selection = stepwise;
run ;
quit;
```

### Tableau de variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Somme des carrés</th>
<th>Carré moyen</th>
<th>Valeur F</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>17.49887</td>
<td>17.49887</td>
<td>15.99</td>
<td>0.0003</td>
</tr>
<tr>
<td>Error</td>
<td>31</td>
<td>32.09735</td>
<td>1.03540</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>32</td>
<td>49.59603</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Table de résultats des paramètres

<table>
<thead>
<tr>
<th>Variable</th>
<th>Résultat estimé des paramètres</th>
<th>Erreur std</th>
<th>Type II SS</th>
<th>Valeur F</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.77374</td>
<td>0.65375</td>
<td>7.62205</td>
<td>7.36</td>
<td>0.0108</td>
</tr>
<tr>
<td>X9</td>
<td>-1.30538</td>
<td>0.31753</td>
<td>17.49887</td>
<td>16.90</td>
<td>0.0003</td>
</tr>
</tbody>
</table>

Bounds on condition number: 1, 1

Stepwise Selection: Step 1

Variable X9 Entered: R-Square = 0.3528 and C(p) = 17.6721

Stepwise Selection: Step 2

Variable X1 Entered: R-Square = 0.4039 and C(p) = 11.2964

```
```
Résultats :
La sélection se fait en 4 étapes. À chaque étape, les résultats globaux sur le modèle sont donnés ($R^2$, $CP$, analyse de variance). Pour vérifier si les régresseurs n'ont pas de colinéarité « pathologique », le «Condition Number » (cf. 4.4.3) est borné : ici par exemple au 4ième pas, Condition Number a une borne supérieure très grande (52.20) donc on peut soupçonner l'existence d'une colinéarité entre les 4 régresseurs.

Le modèle trouvé en 4 étapes est le modèle $LOG=f(X_9, X_1, X_2, X_3)$. 
Puis on fixe les seuils SLE et SLS à 5% par les options SLE= 0.05 et SLS = 0.05.

```
title 'regression STEPWISE avec seuils à 0.05  ';  
proc reg data =libreg.chenilles  ;  
log10 : model log = x1--X10 / selection = stepwise SLE = 0.05 
SLS = 0.05;  
run ;  
quit;  
```

Dans ce cas, 2 pas seulement sont effectués et le modèle est LOG= f(X9, X1).

**Remarque**

On peut raisonner soit avec les sommes de carrés comme on vient de le faire aux §4.5.1 et 4.5.2, soit avec \( R^2 = \frac{SS_{Modèle}}{SS_{Total}} \), ce qui est équivalent compte-tenu de l’équation d’analyse de variance : \( SS_{Totale} = SS_{Modèle} + SS_{Erreurs} \)

D'où les autres méthodes présentées ci-dessous.

### 4.5.3. Amélioration de R²

**Maximum R² Improvement (MAXR)**

C’est une méthode qui procède par étape comme les précédentes. Elle tente de trouver le meilleur modèle au sens du \( R^2 \) pour chaque valeur \( k \) du nombre de régresseurs.

La méthode MAXR commence par choisir la variable donnant le plus grand \( R^2 \) (c’est à dire celle qui est la plus corrélée avec \( Y \)).

Puis est ajoutée celle qui provoque la plus grande augmentation du \( R^2 \).

Une fois ce modèle à 2 variables obtenu, tous les échanges possibles entre une des 2 variables présentes dans le modèle et une variable extérieure sont examinés, c’est-
à-dire que le $R^2$ de la régression est calculé, et l’échange qui est fait est celui qui fournit l’accroissement maximum de $R^2$.

La comparaison recommence alors avec 3 variables dans le modèle.
Ce processus continue jusqu’à ce qu’aucune permutation n’augmente $R^2$.

La différence entre les méthodes STEPWISE et MAXR est que toutes les permutations possibles sont évaluées dans MAXR avant le changement. Dans STEPWISE, seul le « moins bon » régresseur est retiré, sans vérifier si on pourrait ajouter la « meilleure » variable. En contre partie, MAXR demande évidemment beaucoup plus de calculs.

**Minimum $R^2$ Improvement (MINR)**

Il s’agit du même processus que le précédent sauf que la procédure d’«échange» fait appel au couple de variables associé au plus petit accroissement du $R^2$. L’objectif est ainsi d’explorer plus de modèles que dans le cas MAXR et donc, éventuellement, de tomber sur un meilleur optimum.

### 4.5.4. Autres méthodes basées sur $R^2$: RSQUARE et ADJRSQ

Ces méthodes ne fonctionnent pas par étapes. Elles affichent pour toute valeur de $k$, le *meilleur sous-ensemble* de $k$ régresseurs au sens de $R^2$ ou $R^2_{adj}$ (défini au chapitre 1 §1.2.5), ce que n’assurent pas les méthodes STEPWISE ou MAXR/MINR.

Elles demandent beaucoup plus de calculs car il faut examiner toutes les régressions possibles, mais les puissants moyens informatiques actuels rendent ces méthodes très rapides.

Pour $k=1$ à $p$, les modèles sont affichés dans l’ordre décroissant de $R^2$ (ou $R^2_{adj}$). Pour limiter le volume des sorties quand le nombre de variables est grand, l’option BEST = q limite l’affichage aux q premiers modèles pour chaque valeur de $k$.

Ceci permet d’explorer rapidement les sous-ensembles de variables, mais les modèles sélectionnés, optimaux au sens $R^2$ (ou $R^2_{adj}$), ne le sont pas forcément au niveau des données. Aussi il est recommandé d’afficher des critères supplémentaires de qualité comme les critères CP, AIC et BIC présentés ci-dessous. Il faudra aussi valider les modèles, comme on l’a vu dans ce chapitre aux §4.2, §4.3, §4.4.

### 4.5.5. Coefficient CP de Mallows

Ce coefficient proposé par Mallows en 1973, est basé sur la recherche des régresseurs ayant le meilleur pouvoir prédictif, c’est à dire l’erreur totale moyenne la plus petite. Il permet ainsi de choisir entre plusieurs régressions différant à la fois par le nombre de régresseurs et la précision atteinte.

En effet, si on ajoute des régresseurs on diminue, en général, le biais des estimations mais on risque d’augmenter les variances des estimations et «l’erreur totale» moyenne, car on a pu ajouter des variables très liées aux variables initialement introduites.
La précision étant mesurée par MSE ou bien SSE, on calculera le coefficient ainsi : 

- Si $q < p$, alors $\text{CP}(q) = \frac{\text{SSE}(q)}{\text{MSE}} - \left[ n - 2(q + 1) \right]$ 
- où SSE($q$) = somme des carrés des erreurs de la régression avec $q$ régresseurs et constante, 
- et MSE = précision $s^2$ calculée avec les $p$ régresseurs (soit $\text{SSE}(p)/(n-(p+1))$).
- Si $q = p$, alors $\text{CP}(p) = p + 1$.

Si le « bon » modèle est choisi, l’estimation est sans biais, et alors $\text{CP}(q)$ est proche de $(q+1)$ : voir plus de détails dans Daniel et Wood (1980).
Dans le cas où $\text{CP}$ vaut $(p+1)$, on a la même précision que la régression globale.

**Sélection suivant le coefficient CP**

En plus du $R^2$, REG affiche les sous-ensembles de régresseurs par ordre croissant de CP.
A partir du modèle complet, on choisit le premier sous-ensemble dont le CP approche la valeur $(p+1)$: c'est le « meilleur » sous-ensemble de régresseurs.
On peut aussi faire le graphique de CP($q$) en fonction de $q$ ($q=1$ à $p$) et regarder la position de l'optimum ainsi que Mallows recommande:

**Utilisation du coefficient CP dans une sélection de régresseurs**

On demande à afficher le coefficient de chaque modèle obtenu par la méthode de sélection choisie, en ajoutant l’option CP à l'instruction MODEL de REG. Pour un nombre $k$ donné de régresseurs, il est conseillé de choisir le sous-ensemble de $k$ variables ayant une valeur de CP la plus grande.

**4.5.6. Critères AIC et BIC**

Ces critères ne sont pas des critères de sélection des régresseurs, mais des indicateurs de la qualité du modèle.

De manière générale, ces critères mesurent la qualité d’un modèle statistique bâti sur $k$ paramètres sur un échantillon de taille $n$, à partir de la fonction de vraisemblance $L$.

Akaike Information Criterion (1969) : $\text{AIC} = -2 \log(L) + 2k$

Sawa's Bayesian information criterion (1978) : $\text{BIC} = -2 \log(L) + k \log(n)$

Le critère BIC est d’un autre critère utilisé nommé « critère de Schwartz ».

Dans le cas d’un modèle de régression à $p$ régresseurs avec constante :

$$\text{AIC} = n \log\left(\frac{\text{SSE}}{n}\right) + 2(p + 1)$$
$$\text{BIC} = n \log\left(\frac{\text{SSE}}{n}\right) + 2(p + 3)q \quad \text{avec} \quad q = \frac{ns^2}{\text{SSE}}$$
Ces indicateurs sont utilisés comme la règle habituelle consistant à choisir le modèle ayant la meilleure précision (SSE petit).

### 4.5.7. Exemple de sélection RSQUARE


```sas
title 'regression RSQUARE ';
proc reg data =libreg.chenilles ;
log10 : model log = X1--X10 / selection = RSQUARE AIC BIC CP BEST = 2 ;
run;
quit;
```

<table>
<thead>
<tr>
<th>Nombre dans le modèle</th>
<th>R-carre</th>
<th>C(p)</th>
<th>AIC</th>
<th>BIC</th>
<th>Variables du modèle</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.3526</td>
<td>17.6721</td>
<td>3.0840</td>
<td>3.7414</td>
<td>X9</td>
</tr>
<tr>
<td>1</td>
<td>0.2931</td>
<td>21.9604</td>
<td>5.9985</td>
<td>6.3390</td>
<td>X8</td>
</tr>
<tr>
<td>2</td>
<td>0.4690</td>
<td>11.2904</td>
<td>-1.4474</td>
<td>-0.3145</td>
<td>X1 X9</td>
</tr>
<tr>
<td>2</td>
<td>0.4249</td>
<td>14.4777</td>
<td>1.1911</td>
<td>1.8612</td>
<td>X1 X3</td>
</tr>
<tr>
<td>3</td>
<td>0.5368</td>
<td>8.4026</td>
<td>-3.9541</td>
<td>-2.0508</td>
<td>X1 X4 X5</td>
</tr>
<tr>
<td>3</td>
<td>0.6310</td>
<td>8.8262</td>
<td>-3.5382</td>
<td>-1.7348</td>
<td>X1 X2 X3 X9</td>
</tr>
<tr>
<td>4</td>
<td>0.6471</td>
<td>2.4513</td>
<td>-10.9259</td>
<td>-6.1357</td>
<td>X1 X2 X4 X5</td>
</tr>
<tr>
<td>4</td>
<td>0.5799</td>
<td>7.2958</td>
<td>-5.1757</td>
<td>-2.2951</td>
<td>X1 X2 X3 X9</td>
</tr>
<tr>
<td>5</td>
<td>0.6577</td>
<td>3.6866</td>
<td>-9.9325</td>
<td>-4.1182</td>
<td>X1 X2 X4 X5 X9</td>
</tr>
<tr>
<td>5</td>
<td>0.6550</td>
<td>3.8819</td>
<td>-9.6724</td>
<td>-3.9701</td>
<td>X1 X2 X4 X5 X10</td>
</tr>
<tr>
<td>6</td>
<td>0.6764</td>
<td>4.3383</td>
<td>-9.7858</td>
<td>-2.3328</td>
<td>X1 X2 X3 X4 X5 X9</td>
</tr>
<tr>
<td>6</td>
<td>0.6764</td>
<td>4.3392</td>
<td>-9.7846</td>
<td>-2.3322</td>
<td>X1 X2 X4 X5 X6 X9</td>
</tr>
<tr>
<td>7</td>
<td>0.6864</td>
<td>5.6125</td>
<td>-8.8205</td>
<td>0.0594</td>
<td>X1 X2 X3 X4 X5 X8 X9</td>
</tr>
<tr>
<td>7</td>
<td>0.6848</td>
<td>5.7284</td>
<td>-8.6597</td>
<td>0.1627</td>
<td>X1 X2 X3 X4 X5 X9 X10</td>
</tr>
<tr>
<td>8</td>
<td>0.6926</td>
<td>7.1765</td>
<td>-7.4679</td>
<td>2.8385</td>
<td>X1 X2 X3 X4 X5 X8 X9 X10</td>
</tr>
<tr>
<td>8</td>
<td>0.6901</td>
<td>7.3519</td>
<td>-7.2110</td>
<td>2.9101</td>
<td>X1 X2 X3 X4 X5 X7 X8 X9</td>
</tr>
<tr>
<td>9</td>
<td>0.6942</td>
<td>9.0519</td>
<td>-5.6569</td>
<td>5.7796</td>
<td>X1 X2 X3 X4 X5 X7 X8 X9 X10</td>
</tr>
<tr>
<td>9</td>
<td>0.6942</td>
<td>9.0567</td>
<td>-5.6612</td>
<td>5.7805</td>
<td>X1 X2 X3 X4 X5 X6 X8 X9 X10</td>
</tr>
<tr>
<td>10</td>
<td>0.6945</td>
<td>11.0000</td>
<td>-3.7346</td>
<td>8.7654</td>
<td>X1 X2 X3 X4 X5 X6 X7 X8 X9 X10</td>
</tr>
</tbody>
</table>
Dans le tableau des modèles, on constate que, parmi les modèles à 4 variables, celui qui a le meilleur $R^2$, et les plus petites valeurs de AIC et BIC est $Y = f(X_1 X_2 X_4 X_5)$, qui est le modèle étudié aux paragraphes §4.2.5, §4.3.11 et §4.4.5. Par contre le critère CP est plus petit que celui 2ième modèle sélectionné $Y = f(X_1 X_2 X_3 X_9)$, qui pourrait donc être également un modèle à étudier : c'est d'ailleurs le modèle final trouvé par la méthode STEPWISE (§4.5.2) avec les seuils par défaut de 15 %.
Conclusion

Arrivé au terme de ce tutoriel nous voulons insister sur deux points non abordés par la technique de la régression : la qualité des données et les difficultés d'interprétation. Ces deux points sont du ressort du spécialiste du domaine d'études sur lequel le statisticien applique la régression.

La qualité de l'information apportée par les données (observations) intervient dans la validité et la robustesse d'un modèle de régression. Mais cette qualité n'est pas appréhendable par le statisticien-praticien. Ce sont des connaissances externes à la statistique mais internes à l'étude qui doivent intervenir. Ces connaissances sont aussi indispensables pour déterminer le plan d'échantillonnage. Cette étape qui se situe au niveau de la collecte des données et donc en amont de l'analyse statistique des données mériterait à elle seule un long développement.

Une fois réglées toutes les difficultés reste un dernier point et non des moindres, l'interprétation. Prenons un exemple concret : parmi des enfants, on effectue une enquête permettant de mesurer l'étendue du vocabulaire et la taille de leurs pieds. La corrélation entre ces deux variables est nettement significative ! Le bon sens permet d'éviter d'en tirer des conclusions aberrantes. Sous cette corrélation se cache l'influence de la variable âge.

Autre exemple, dans un état des U.S.A on a corrélé sur les 20 dernières années, le taux de criminalité et le taux de fréquentation dans les églises. Là aussi la corrélation obtenue est très élevée, mais le bon sens ne vient que peu en aide. La variable cachée est l'immigration italienne et irlandaise.

A la lumière de ces deux exemples, gardons-nous de toutes interprétations hâtives.

Avoir toujours à l'esprit que sous une corrélation peut se cacher l'effet d'une autre variable, ou d'un autre facteur.

On peut cependant utiliser le modèle identifié, s'il est correct, dans un but de prévision mais surtout pas dans un but de "contrôle" (action sur les variables explicatives dans l'espoir d'agir sur Y) ou d'explication. Sinon, on pourrait augmenter l'intelligence de nos enfants en augmentant la taille de leurs pieds!

L'erreur qui perdure dans la littérature, est de donner le nom de variable dépendante ou variable expliquée à Y et de variables indépendantes ou variables explicatives à X, ce qui amène à déduire logiquement qu'il existe une idée de cause à effet entre X et Y. « Mécaniquement », ce n'est pas l'objet de la régression.

La régression sur données d'observations ne permet pas de déduire une quelconque relation de cause à effet de X sur Y et/ou de Y sur X. Il faut d'autres pratiques méthodologiques pour expliquer la causalité qui peut avoir des formes multiples.
La liaison entre 2 variables X et Y peut se rencontrer dans 5 situations:

- X cause Y
- Y cause X
- X et Y inter-agissent l'une sur l'autre, problème de circularité, ou de rétro-action
- X et Y évoluent ensemble sous l'effet d'une même variable
- X et Y sont liées par hasard

La causalité peut être validée par “l'outil” régression que si on peut faire des comparaisons sur des groupes comparables. C'est un débat historique qui est de plus en plus d'actualité. Les économètres tentent de pallier la faiblesse des techniques de régression appliquées à des données d'observations, en essayant de se rapprocher des techniques expérimentales. Ils ont introduit la méthode des variables instrumentales, avec l'idée de pouvoir comparer des individus qui ne diffèrent que sur une seule dimension : le traitement. Ils ont également proposé de traiter des données issues « d’expériences naturelles » et « d’expériences contrôlées », c'est un premier pas vers des interprétations causales (voir BEHAGHEL (2006).

Terminons sur ce constat, les méthodes de régression sont des méthodes très puissantes, mais qui doivent être utilisées avec beaucoup de discernement et de prudence. En toute honnêteté il ne faut pas se contenter d'un seul modèle et d'une seule procédure REG, il faut en tester plusieurs.

C'est un travail d'explorateur et de détective. C'est ce que nous avons tenté de mettre en lumière.
ANNEXES

ANNEXE 1.................................................................................................................. 123

SYNTAXE SIMPLIFIÉE DE LA PROCÉDURE REG DE SAS......................................................... 123

PROC REG options ;........................................................................................................... 123
MODEL dépendante = régresseurs / options ;...................................................................... 123
Instructions BY FREQ ID WEIGHT ;.................................................................................. 125
REWEIGHT expression / WEIGHT = valeur ;................................................................. 125
TEST equation(s) ; ............................................................................................................. 125
RESTRICT equation(s);...................................................................................................... 125
OUTPUT OUT = nomtab mot_clef = nom_var ; ............................................................... 126
PLOT Yvar1*Xvar1='s' Yvar2*Xvar2='s' / options ;..................................................... 126
PRINT mots-clefs;............................................................................................................ 126
Options RIDGE et PCOMIT des instructions PROC REG ou MODEL ............................... 127

ANNEXE 2......................................................................................................................... 128

MODE D'EMPLOI TRÈS SUCCINCT DE SAS/INSIGHT ........................................................... 128

Le lancement de SAS/INSIGHT .......................................................................................... 128
Rôle statistique des variables dans SAS/INSIGHT............................................................ 129
Menu principal de SAS/INSIGHT...................................................................................... 130
Graphiques standard en SAS/INSIGHT............................................................................. 130
Les Analyses Statistiques avec SAS/INSIGHT ................................................................... 132
Exemple de Régression linéaire sur la Table SAS : Chenille (processionnaire du pin du §2.3.1.) ... 132
Impression et Sauvegarde................................................................................................. 133
Pour imprimer.................................................................................................................... 133
Pour sauvegarder les résultats graphiques ou tableaux dans un fichier............................. 133
Pour insérer un fichier externe.bmp dans Word .............................................................. 134
Pour plus d’information sur les graphiques..................................................................... 135

ANNEXE 3.......................................................................................................................... 136

STATISTIQUES RELATIVES À L’ANALYSE DE LA VARIANCE .................................................. 136

STATISTIQUES SUR LES PARAMÈTRES............................................................................. 137

ANNEXE 4.......................................................................................................................... 138

RELATIONS ENTRE LA LOI NORMALE ET LES STATISTIQUES DE LOIS .............................. 138

ANNEXE 5.......................................................................................................................... 139

CONSTRUCTION D’UN QQ- PLOT.............................................................................................. 139

PRINCIPE DE LA DROITE DE HENRY ................................................................................ 139
1. Soit une variable X dont on veut vérifier l’adéquation à une loi normale (m,σ) ............... 139
2. Si X suit une loi normale (m,σ) alors : ............................................................... 139
3. En pratique, on ordonne les valeurs x_i de x_0 les valeurs ordonnées................................ 139

GÉNÉRALISATION .............................................................................................................. 140
QQ- PLOT AVEC SAS........................................................................................................... 140

BIBLIOGRAPHIE.................................................................................................................. 141
Annexe 1

Syntaxe simplifiée de la Procédure REG de SAS

La procédure REG est une procédure « interactive » permettant d’étudier plusieurs modèles en un seul appel de PROC REG. On donne ici son utilisation pour l’étude d’un seul modèle.

PROC REG options ;
   MODEL dépendante = régresseurs / options ;
   BY nom_var ;
   FREQ nom_var ;
   ID nom_var ;
   WEIGHT nom_var ;
   REWEIGHT expression / option ;
RUN ;
   TEST équation(s) ;
   RESTRICT équation(s) ;
   OUTPUT OUT = data_sas mot_clef = nom_var ;
   PLOT yvar*xvar='symbol' / options ;
QUIT ;

PROC REG options ;

DATA=NOMTAB data_set_option option commune à toutes les procédures
OUTEST = TAB permet de créer des tables SAS de résultats utiles, comme les coefficients estimés, et des résultats créés dans des options.

Autres options : ALL CORR NOPRINT SIMPLE USSCP

ALL Demande beaucoup d'impressions (induit l’option SIMPLE, USSCP, et CORR).
CORR Imprime la matrice de corrélation de toutes les variables du modèle.
NOPRINT Supprime les impressions.
SIMPLE Imprime somme, moyenne, variance, écart-type et somme des carrés non corrigée pour les variables utilisées dans REG.
USSCP Imprime les sommes de carrés non corrigées et la matrice des produits croisés pour toutes les variables utilisées dans REG.
PRESS Permet d’obtenir dans la table OUTEST le coefficient PRESS
RIDGE et PCOMIT pour les régressions Ridge et sur composantes principales (voir à la fin de cette annexe).
MODEL dépendante = régresseurs / options ;

dépendante : nom de la variable dépendante

régresseurs : liste des noms des p variables régresseurs

Remarque : on peut donner un label à l'instruction MODEL, label qui sera alors affiché dans les sorties.

**Quelques options de l'instruction MODEL:**

**Sélection de régresseurs**
Option sous la forme SELECTION = nom (où nom est un des mots-clefs de la liste ci-après) :

- NONE : pas de sélection (choix par défaut)
- FORWARD : sélection ascendante
- BACKWARD : sélection descendante
- STEPWISE : sélection progressive ascendante
- MAXR, MINR : sélection basée sur gain maximum/minimum en R²
- RSQUARE, ADJRSQ : sélection du meilleur sous-ensemble au sens de R², R² ajusté
- CP : sélection basée sur CP de Mallows

**Autres options associées à SELECTION :**
- INCLUDE = n : inclure les n premières variables explicatives dans les modèles explorés (n<p)
- SLE = valeur : seuil de significativité pour entrer
- SLS = valeur : seuil de significativité pour rester
- STOP = s : arrête l'exploration au meilleur sous-ensemble de s variables (avec s < p) ; STOP = p par défaut.
- BEST = k : arrête l'exploration après k modèles.

**Attention aux valeurs par défaut des seuils :**
- SLE = 0.50 en FORWARD et SLE = 0.15 en STEPWISE
- SLS = 0.10 en BACKWARD et SLS = 0.15 en STEPWISE

**Remarque** : les différentes valeurs du critère de sélection choisi sont stockées dans la table OUTEST, où on trouve aussi les 2 variables :

- _IN_ nombre de régresseurs hors constante
- _P_ nombre de régresseurs y compris la constante si elle existe dans le modèle

**Autres options de l'instruction MODEL :**

**Définir un modèle sans constante** : NOINT

**Afficher** :
- des résultats complémentaires pour les observations :
- P (prévisions) CLI CLM (intervalles de prévision à 95 % individuels et sur la moyenne) R(résidus) INFLUENCE (indices de détection des observations influentes)
- des coefficients: DW (Durbin-Watson), CP (Cp de Mallows), BIC, AIC, etc.;
- les sommes de carrés SS1 SS2 (carrés de type I ou II)

**Diagnostiquer des problèmes particuliers:**
- heteroscédasticité: SPEC (et ACOV)
- colinéarité: TOL VIF et COLLIN COLLINOINT

**Instructions BY FREQ ID WEIGHT :**

Ce sont des instructions communes à toutes les procédures. En particulier, WEIGHT permet de définir une régression pondérée.

**REWEIGHT expression / WEIGHT = valeur ;**

Cette instruction permet de redéfinir les poids, et en particulier d'omettre une observation de la régression.

*expression* est une comparaison sur une variable (on peut utiliser la variable OBS. qui contient le numéro de l'observation),

*WEIGHT = valeur* donne cette valeur de poids aux observations vérifiant l'expression.

Exemple : pour supprimer l'observation numéro 20, on écrit

REWEIGHT obs. = 20 / WEIGHT = 0 ;

**TEST equation(s) ;**

Cette instruction permet de tester une ou des hypothèses sur les estimations des paramètres (les équations doivent être séparées par des virgules). Chaque équation est une fonction linéaire formée de coefficients et de noms de variables (ici, INTERCEPT est le nom de la constante).

Exemples: 

TEST X1 = 0 , INTERCEPT = 0 ; tester ($\beta_1=0$) et ($\beta_0=0$)

TEST X3-X4 = 0 ; tester ($\beta_3=\beta_4$)

**RESTRICT equation(s);**

Elle permet de fixer des contraintes sur les coefficients, avec des équations identiques à TEST.
Exemple: RESTRICT X1-X3 = 0 ;  * modèle avec contrainte (β1= β3)

**OUTPUT OUT = nomtab mot_clef = nom_var ;**

Cette instruction permet de créer une table SAS contenant certaines des variables créées par la régression. Ce tableau contiendra aussi les variables du modèle (réponse et régresseurs).

**Liste des mots_clefs pour les variables créées par la régression**
(ces mots_clefs sont utilisables aussi pour les instructions PLOT et PRINT, sauf PRESS):

- PREDICTED (ou P) valeur prédite
- L95M U95M limites des intervalles à 95% sur la moyenne des valeurs prédites
- L95 U95  limites des intervalles à 95% sur une valeur prédite
- STD P  écart-type de la valeur moyenne prédite
- STDI écart-type de la valeur prédite
- RESIDUAL (ou R) résidu
- STDR écart-type du résidu
- NQQ quantile normal (pour le dessin QQPLOT)
- STUDENT résidu studentisé interne
- RSTUDENT résidu studentisé externe
- H levier
- PRESS coefficient Press (individuel)
- COOKD DFFITS COVRATIO mesures d'influence des observations

**PLOT Yvar1*Xvar1='s' Yvar2*Xvar2='s' / options ;**

Cette instruction permet de tracer des graphiques en désignant les variables ordonnée Yvar et abscisse Xvar, et le symbole associé.

Différentes options sont possibles pour définir les caractéristiques des graphiques (cf. la documentation SAS).

Attention : on peut utiliser une des variables créées par la régression, définies plus haut dans la liste des mots-clefs de l'instruction OUTPUT, à condition de faire suivre son nom par un point : par exemple P. ou R. On peut aussi utiliser la variable OBS. pour désigner le numéro de l’observation.

Remarque : l’option RIDGEPLOT permet de tracer le dessin des coefficients RIDGE (voir plus loin la description de l’option RIDGE).

**PRINT mots-clefs;**

Cette instruction permet d'imprimer certaines des variables créées avec la liste des mots-clefs vue plus haut.
Options RIDGE et PCOMIT des instructions PROC REG ou MODEL

On peut effectuer une régression Ridge ou une régression sur composantes principales, par une option de PROC REG ou de MODEL. La procédure travaille alors sur les données centrées (l'option NOINT est ignorée).

**RIDGE = liste**

liste est une liste de valeurs qui peut être définie par la syntaxe
kd to kf by p, où l'intervalle de variation du coefficient ridge est [kd,kf], la variation se faisant par pas de p.

Chaque valeur donne une estimation des coefficients Ridge, qui est placée dans une table SAS à définir avec l'option OUTEST = table de PROC REG. La colonne _TYPE_ indique quelle méthode on a employé : pour la méthode Ridge, _TYPE_=RIDGE, et les valeurs de la liste sont stockées sous le nom de variable _RIDGE_. On trouve ensuite les valeurs des coefficients Ridge de chaque régresseur.

**PCOMIT = k**

k est un entier positif ou nul.

La procédure calcule alors les paramètres estimés en utilisant les composantes principales à l'exclusion des k dernières ; L'estimation des coefficients est placée dans une table SAS à définir avec l'option OUTEST = table de PROC REG, avec ici _TYPE_ = IPC.

Remarque : k peut aussi être une liste d'entiers non négatifs, pour permettre de faire plusieurs essais d'élimination de composantes.

**Autres options de PROC REG tilisables en association avec les options RIDGE et PCOMIT :**

OUTSTB pour avoir les estimations standardisés des coefficients estimés par RIDGE ou IPC ;

OUTSEB pour avoir les erreurs standardisées des coefficients ;

OUTVIF pour avoir les Variance Inflation Factor des coefficients.

**Dessin « Ridge Trace »**

Le dessin des coefficients Ridge en fonction des valeurs du paramètre (définies par l'option RIDGE = liste) est obtenu par l'instruction PLOT avec l'option RIDGEPLOT, à condition que les coefficients soient stockés dans une table par l'option OUTEST. On écrit alors simplement l'instruction :

```
PLOT / RIDGEPLOT ;
```
Annexe 2

Mode d’emploi très succinct de SAS/INSIGHT

Le module de SAS/Insight est à la fois un tableur un grapheur et un analyseur. Il permet de faire de l’Analyse Exploratoire des Données et de l’analyse confirmatoire dans l’esprit de Tukey. Il est particulièrement bien adapté à la régression linéaire couplée à l’AED, grâce à ses possibilités de visualisation et d’interactivité.

Tableau des Grandes Fonctions de SAS/INSIGHT

<table>
<thead>
<tr>
<th>Analyse Exploratoire</th>
<th>Méthodes</th>
<th>Analyse confirmatoire</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Visualisation</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>graphiques</td>
<td>1D, 2D, 3D</td>
<td>Analyse des distributions</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- paramétrique</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- non-paramétrique</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- QQ plots</td>
</tr>
<tr>
<td><strong>Exploration des données</strong></td>
<td></td>
<td>Ajustement par modèles</td>
</tr>
<tr>
<td>Rechercher, trier et éditer</td>
<td>Analyse Composantes Principales</td>
<td></td>
</tr>
<tr>
<td>identifier/données</td>
<td>Rotation axes</td>
<td>Régression linéaire,</td>
</tr>
<tr>
<td>Brosse (Brushing et Slicing)</td>
<td></td>
<td>Analyse de</td>
</tr>
<tr>
<td>Colorier, Marquer / critère</td>
<td>Analyse Canonique des Corrélations</td>
<td>- variance,</td>
</tr>
<tr>
<td>Supprimer, cacher, exclure des calculs</td>
<td>Analysis Discriminante</td>
<td>- covaraiance</td>
</tr>
<tr>
<td>Animation etc.</td>
<td></td>
<td>OLM (MLO)</td>
</tr>
<tr>
<td><strong>Ré-expression des données</strong></td>
<td>Plus de 20 fonctions mathématiques ou statistiques</td>
<td></td>
</tr>
</tbody>
</table>

Nous ne présentons que quelques manipulations essentielles de SAS/INSIGHT. Pour une présentation plus complète voir l’ouvrage de Destandau S. & Le Guen M.

Le lancement de SAS/INSIGHT

> dans la barre de commande de SAS, taper : INSIGHT puis entrée

1. Si la table de données n’existe pas encore
Dans la boîte de dialogue, cliquer sur le bouton-poussoir New
Un tableau de données vide s'ouvre. Saisissez vos données.

2. Si la table existe

Affichage de la table SAS dans un tableur (cf. ci-dessous écran à gauche) et menu déroulant (cf. écran à droite).

La Table TAILPOID a 3 variables et 20 observations indiqué par
En cliquant sur la petit flèche en haut à gauche le menu déroulant –pop menu ou encore menu contextuel- s'affiche avec les actions possibles sur le tableur.

Rôle statistique des variables dans SAS/INSIGHT

Dans SAS/INSIGHT toute variable SAS définit en caractère est forcément une variable nominale. Par défaut une variable numérique n’est pas nominale, elle est d'intervalle. C’est à l’utilisateur de choisir le type d’échelles de mesures (Interval/Nominal) souhaitée, en cliquant et cochant la zone au dessus du nom de la variable.
Ce rôle statistique déterminera les types de graphiques à 1 dimension, 2 dimensions ou 3 dimensions et les types d’analyses.

Menu principal de SAS/INSIGHT

Graphiques standard en SAS/INSIGHT

- Graphiques pour les variable nominales : Bar Chart (1D), Mosaic Plot (1D)
- Graphiques pour les variables d'intervalle : Histogram (1D), Box Plot (1D), Line Plot (2D), Scatter Plot (2D), Contour Plot (3D), Rotating Plot (3D).

Pour réaliser un graphique il y a 2 possibilités : en utilisant les options par défaut, ou en passant par une boîte de dialogue pour modifier les options par défaut. C’est un principe général dans SAS/INSIGHT.

Choix 1 – avec options par défaut

Dans le tableur :

⇒ **Cliquer** sur le nom de la variable d'intervalle Y
⇒ **menu** : Analyze# Box Plot/Mosaic Plot(Y)

L’affichage est immédiat avec les options par défaut.
Sur le graphique, en cliquant sur la flèche en bas à gauche un menu déroulant s'affiche pour modifier les options. Par exemple, ajouter la moyenne avec **Means**, ajouter les valeurs des quantiles avec **Values** etc.

**Choix 2 – avec options modifiables**

⇒ menu : **Analyze** → **Box Plot/Mosaic Plot(Y)**

Dans la boîte de dialogue qui s'affiche sélectionner la variable Y (dans la liste à gauche) puis cliquer sur le bouton-poussoir Y, pour que la variable choisie Y soit sélectionnée. Les boutons poussoirs **Method** et **Output** permettent de modifier les options par défaut.
Pour plus d'information sur les graphiques voir les articles en ligne sur les sites internet donnés en fin de cette annexe (page 354)

Les Analyses Statistiques avec SAS/INSIGHT

- menu : Analyze # Distribution
  Etude d'une distribution (équivalent à Proc Univariate)

- menu : Analyze # Fit ➤ Analyse de régression linéaire, GLM, Régression logistique, Probit, Logit, ANOVA


Même principe que pour les graphiques :
- Soit on sélectionne la ou les variables puis le menu Analyze #….. Les analyses sont réalisées avec les options par défaut,
- Soit on choisit le menu Analyze #….., une boîte de dialogue s' affiche pour sélectionner la ou les variables, et choisir les nouvelles options.

Exemple de Régression linéaire sur la Table SAS : Chenille (processionnaire du pin du §2.3.1.)
Sélectionner dans la liste de gauche la variable Réponse (Log) puis cliquer sur le bouton de rôle Y, idem pour les variables régresseurs (X4,X2,X4,X5), en cliquant sur le bouton de rôle X. Si on veut la constante 1 (β₀) dans le modèle, cocher Intercept.

Cliquer sur le bouton poussoir Method :

Pour la régression linéaire les options à cocher sont :

- Response Dist : Normal
- Link Function : Canonical
- Scale : MLE

Impression et Sauvegarde

Nous présentons seulement quelques possibilités, pour imprimer un ou des éléments affichés, puis les sauvegarder dans un fichier externe, et enfin les insérer dans un document Word.

Pour imprimer

Sélectionner avec la souris, le graphique ou le tableau à imprimer, ou choisir Menu : Edit# Windows # Select all pour sélectionner tous les éléments affichés

File # Print

Pour sauvegarder les résultats graphiques ou tableaux dans un fichier

Sélectionner avec la souris la bordure du graphique ou le tableau à sauvegarder, ou choisir

Menu : Edit# Windows # Select all pour sélectionner tous les éléments affichés

File # save # Graphics File…

puis renseigner la boîte de dialogue en choisissant par exemple le format .bmp et en suffixant le nom du fichier par .bmp (SAS ne le fait pas).
Le fichier sera sauvegardé dans le répertoire courant qui est affiché en bas de l'écran de SAS

 Pour modifier l'emplacement, il suffit de double cliquer dessus et de changer le répertoire (fenêtre Change Folder).

**Pour insérer un fichier externe .bmp dans Word**

Dans un document Word : ➔ insertion # image # à partir d'un fichier.....
Pour plus d'information sur les graphiques

Consulter les articles en ligne sur les sites internet :


LE GUEN M. (2001), La boîte à moustaches de Tukey, un outil pour initier à la Statistique, Statistiquement Vôtre, n° 4, 14 pages.


Annexe 3

Statistiques relatives à l’analyse de la variance

<table>
<thead>
<tr>
<th>Statistique</th>
<th>Formule</th>
<th>Signification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Square</td>
<td>$MS = \frac{SS}{DF}$</td>
<td>C'est le rapport d'une somme des carrés des écarts (SS) divisée par le nombre de degrés de liberté (DF).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• pour SS model $DF=p$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• pour SS error $DF=n-p-1$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>La statistique Mean Square Error donne l'estimation $s^2$ de la vraie valeur inconnue de la variance des erreurs $\sigma^2$.</td>
</tr>
<tr>
<td>F Value</td>
<td>$F = \frac{MS Model}{MS Error}$</td>
<td>Statistique de Fisher-Snedecor pour tester si tous les paramètres $\beta$ sont nuls.</td>
</tr>
<tr>
<td>Prob&gt;F</td>
<td>Probf(F Value,ndf,ddf)</td>
<td>C'est la $p$-value ou niveau de significativité du test, associée à F Value. La $p$-value est calculée en utilisant la fonction SAS : Probf.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Probf : fonction SAS de la fonction de répartition d'une variable de Fisher-Snedecor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ndf : nombre de degrés de liberté du numérateur de F Value</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ddf : nombre de degrés de liberté du dénominateur de F Value.</td>
</tr>
<tr>
<td>Root MSE</td>
<td>$\sqrt{MS Error}$</td>
<td>Standard Deviation, soit l'écart moyen résiduel.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>C'est l'estimation de &quot;s&quot;, l'écart-type des erreurs</td>
</tr>
<tr>
<td>R-square</td>
<td>$R^2 = \frac{SS Model}{SS Total} = \frac{SS Total - SS Error}{SS total}$</td>
<td>$F$ et Rsquare ($R^2$) sont liés par la relation $F = \frac{n-p-1 \cdot Rsquare}{p \cdot (1-Rsquare)}$</td>
</tr>
<tr>
<td>Dep Mean</td>
<td>$\bar{Y} = \frac{\sum Y_i}{n}$</td>
<td>moyenne de la variable réponse $Y$</td>
</tr>
</tbody>
</table>
### Statistiques sur les paramètres

Pour chaque paramètre $\beta_j$, SAS donne: l’estimation du paramètre avec son erreur-type, le test de l’hypothèse nulle ($\beta_j = 0$) et la *p-value* associée.

**Remarque:** La variable notée Intercept correspond à la variable constante $X_0 = 1$.

<table>
<thead>
<tr>
<th>Statistique</th>
<th>Formule</th>
<th>Signification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>solution de : $\hat{\mathbf{b}}^* = (X'X)^{-1}X'Y$</td>
<td>estimation du paramètre $\beta_j$</td>
</tr>
<tr>
<td>Standard Error</td>
<td>$\sqrt{(X'X)_{ii}^{-1}.MSE}$</td>
<td>erreur-type de l’estimateur du paramètre $\beta_j$ calculé à partir du $i$ème élément de la diagonale de la matrice $(X'X)^{-1}$</td>
</tr>
<tr>
<td>T for $H_0 :$</td>
<td>$T = \frac{\text{Estimate}}{\text{Std Error of Estimate}}$</td>
<td>statistique T de Student pour tester l’hypothèse nulle: $H_0 : \beta_j = 0$ contre $H_a : \beta_j \neq 0$</td>
</tr>
<tr>
<td>Parameter=0</td>
<td>$T^2 = \text{Fvalue partiel}$</td>
<td></td>
</tr>
<tr>
<td>Prob &gt;</td>
<td>T!</td>
<td>$\text{Probt}(T, df)$</td>
</tr>
</tbody>
</table>

---

29 S’il n’y a pas de constante $b_0$ à l’origine, les statistiques relatives à l’analyse de la variance n’ont pas la même interprétation.
Annexe 4

Relations entre la loi normale et les statistiques de lois

Chi2, T de Student et F de Fisher-Snedecor

**Normale**

si \( X \sim N(0,1) \)

**Chi2**

alors \( Z = X^2 \) suit une loi de \( \chi^2 \) à 1 degré de liberté (ddl)

et \( Z = \sum_{i=1}^{n} X_i^2 \) suit un \( \chi^2 \) à \( n \) ddl, si les \( X_i \) sont **indépendants** et \( N(0,1) \)

**T de Student à n ddl**

Si \( Z \) suit une loi de \( \chi^2 \) à \( n \) ddl et si \( Z \) est indépendant de \( X \)

alors \( T = \frac{X}{\sqrt{\frac{Z}{n}}} \) suit une loi de Student à \( n \)ddl.

**F de Fisher-Snedecor**

Si \( Z_1 \) et \( Z_2 \) sont des variables aléatoires **indépendantes** suivant chacune une loi de

\( \chi^2 \) à \( \nu_1 \) et \( \nu_2 \) ddl

alors \( F = \frac{\frac{Z_1}{\nu_1}}{\frac{Z_2}{\nu_2}} \) suit une loi de Fisher-Snedecor à \((\nu_1, \nu_2)\) ddl.
Annexe 5

Construction d’un QQ-Plot

Ce graphique permet une visualisation de l’adéquation à une loi. Dans le cas de la loi normale, il est appelé « droite de Henry »

Principe de la droite de Henry

1. Soit une variable $X$ dont on veut vérifier l’adéquation à une loi normale $(m, \sigma)$

On dispose de $n$ observations de $X$ : $(x_i)$ pour $i = 1$ à $n$.

On note $F(x_i)$ la fonction de répartition empirique en $(x_i)$ :

$$F(x_i) = \text{prob}(X \leq x_i) = \text{prob}\left( \frac{X - m}{\sigma} \leq \frac{x_i - m}{\sigma} \right)$$

Soit $\phi$ la fonction de répartition de la loi normale $(0,1)$ : on peut trouver une valeur $u_i$ de même fonction de répartition : $\exists u_i$ tel que $\phi(u_i) = F(x_i) \iff u_i = \phi^{-1}(F(x_i))$.

2. Si $X$ suit une loi normale $(m, \sigma)$ alors :

$$F(x_i) = \text{prob}(X \leq x_i) = \text{prob}\left( \frac{X - m}{\sigma} \leq \frac{x_i - m}{\sigma} \right) = \phi\left( \frac{x_i - m}{\sigma} \right)$$

$$\implies u_i = \phi^{-1}(F(x_i)) = \phi^{-1}\left( \phi\left( \frac{x_i - m}{\sigma} \right) \right) = \frac{x_i - m}{\sigma}$$

Donc les points $(x_i, u_i)$ sont alignés sur la droite d’équation $u_i = (x_i - m) / \sigma$.

3. En pratique, on ordonne les valeurs $x_i$ : on note $x_{(i)}$ les valeurs ordonnées.

Bien sûr, on a alors $F(x_{(i)}) = i / n$.

Le « nuage de points » à représenter est donc défini ainsi : $[u_i = \phi^{-1}(i/n); x_{(i)}]$.

Attention : pour faire les mêmes représentations que SAS, on mettra les valeurs de $X$ en ordonnées.
La droite d’équation $Y = (X-m)/\sigma$ est celle dont les points du nuage doivent se rapprocher en cas d’adéquation à la loi normale : on la représente donc également sur le même plan.

Remarque : Proc UNIVARIATE du module SAS/Base, ainsi que le menu « Distribution » de SAS-INSIGHT, utilise un calcul légèrement différent de celui exposé ici :

$$u_i = \phi^{-1} \left( \frac{r_i - \frac{3}{8}}{n + \frac{1}{4}} \right)$$

où $r_i$ est le rang de l’observation $i$ ($r_i = i$ en général).

Ceci permet en particulier de ne pas « perdre » les points extrêmes.

Généralisation

Si on veut visualiser l’adéquation à une autre loi que la loi normale, il suffit de connaître la fonction de répartition $G$ de cette loi, et que celle-ci soit inversible.

On remplace alors $\phi$ par $G$ dans les formules du §2.

QQ-Plot avec SAS

SAS permet de représenter des QQ-Plot pour les lois suivantes :

*Normale, LogNormale, Exponentielle et Weibull.*

Dans le module SAS/Base, la procédure UNIVARIATE possède une instruction QQPLOT (voir la documentation SAS pour son utilisation un peu complexe).

Dans SAS/INSIGHT, on les trouve dans le menu « Distribution », rubrique « Graphs » → « QQ-Plot ». Pour tracer la droite de référence, dans le menu « Curves » demander « QQ ref line ».

Lorsque l’on a exécuté une régression linéaire avec le menu « Fit », on peut ajouter aux sorties standards un graphique QQ-Plot appelé « Residual Normal QQ » dans le menu « Graphs » (penser à cocher « Reference lines » dans le menu contextuel du graphique pour tracer la droite).
Bibliographie


DESROSIEURES A., (1993), *La politique des grands nombres, histoire de la raison statistique*, Editions la Découverte


http://foucart.thierry.free.fr/colreglin/Regression_bornee.pdf


GALTON F., (1886), « Regression towards mediocrity in hereditary stature», *Journal of the Anthropological Institute* 15 (1886), p246-263.
http://www.stat.ucla.edu/history/regression.gif

INDJEHAGOPIAN J.P., (1993), *Cours d'économétrie*, Polycopié ISUP.

LADIRAY D., (1990), « Autopsie d'un résultat : L'exemple des procédures Forecast, X11, Cluster », *Club SAS 1990*

LADIRAY D., (1997 et suivantes), *Analyse exploratoire des données*, Cours polycopié de l'ENSAE.


MALINVAUD E., (1966), *Méthodes statistiques de l'économétrie*, Dunod

MOLES A., (1990), *Les sciences de l'imprécis*, Le Seuil

NETER J., WASSERMAN W., KUTNER M. H., (1990), *Applied Linear Statistical Models*, Irwin 3e édition


S.A.S , (1990), *Stat User's Guide version 6*


SEN A., SRIVASTAVA M., (1990), *Regression Analysis, Theory, Methods, and Applications*, Springer-Verlag


