Corrélation - Régression linéaire

Michaël Genin

Université de Lille 2
EA 2694 - Santé Publique : Epidémiologie et Qualité des soins
michael.genin@univ-lille2.fr
Plan

1. Introduction à l'étude de deux variables quantitatives
2. Coefficient de corrélation
3. Régression linéaire simple
4. Régression linéaire multiple
5. Références
Introduction

Corrélation - Régression linéaire simple

- Croisement de deux variables quantitatives
  - âge et fréquence cardiaque
  - Consommation et poids d’un véhicule
  - Capacité d’épargne et revenus
  - ...

- Étude du lien entre ces deux variables
  - Représentation graphique
  - Indicateur statistique
  - Modèle de prédiction
Corrélation - Régression linéaire simple

Notations

On considère $n$ individus sur lesquels on mesure $X$ et $Y$ deux variables quantitatives.

Pour chaque individu $i$ ($1 \leq i \leq n$), on dispose d’un couple d’observations $(x_i, y_i)$ qui représente les valeurs prises par $X$ et $Y$ pour l’individu $i$. 
Corrélation - Régression linéaire simple

Représentation graphique

Graphique pour représenter deux variables quantitatives ⇒ nuage de points

1ère étape de toute analyse de liaison : apprécier la forme de la relation entre les deux variables

liaison linéaire  liaison polynomiale  pas de liaison
Rappel sur la covariance

Covariance

Mesure de la variation simultanée de deux variables aléatoires. La covariance permet d’évaluer l’importance et le sens de cette variation.

\[ \sigma_{XY} = \text{cov}[X, Y] = E[XY] - E[X]E[Y] \]

- si les variables sont liées, la covariance est importante.
- une covariance peut être positive, négative ou nulle.
- Si les variables sont indépendantes \( \rightarrow \sigma_{XY} = 0 \)
Rappel sur la covariance

Estimation de la covariance

\[ s_{xy} = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum x_i y_i - \bar{x}\bar{y} \]

Remarque : \( S_{XY} \) est un estimateur biaisé de \( \sigma_{XY} \).

\[ \rightarrow \] On lui applique la correction \( \frac{n}{n-1} S_{XY} \)

Remarques :

- \( \text{cov}(X, Y) = \text{cov}(Y, X) \)
- \( \text{cov}(aX, Y) = a\text{cov}(X, Y) = a\text{cov}(Y, X) \)
- \( \text{cov}(X, X) = \text{Var}(X) \)
- \( \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{cov}(X, Y) \)
Coefﬁcient de corrélation linéaire

Coefﬁcient de corrélation théorique

\[ \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \]

\[ \rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{E[XY] - E[X]E[Y]}{\sigma_X \sigma_Y} \in [-1; 1] \]

Si \( X \) et \( Y \) sont indépendantes alors \( E[XY] = E[X]E[Y] \) donc \( \rho = 0 \)

\( covariance \ nulle \)

Si \( \rho = 0 \) et \( X \) et \( Y \) sont distribuées normalement alors \( X \) et \( Y \) sont indépendantes.

Si \( |\rho| = 1 \) → il existe une relation linéaire parfaite entre \( X \) et \( Y \) du type

\[ Y = \beta_1 X + \beta_0 \]
Coefficient de corrélation linéaire

Coefficient de corrélation de Bravais-Pearson

$\rho$ est estimé par

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

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)}}$$
Interprétation du coefficient de corrélation linéaire

ρ mesure la relation linéaire entre deux variables quantitatives X et Y, ρ est toujours compris entre -1 et 1.

- si ρ = 0, les variations des variables X et Y sont indépendantes (si X et Y distribuées normalement).
- si ρ > 0, les valeurs prises par Y ont tendance à croître quand les valeurs de X augmentent.
- si ρ < 0, les valeurs prises par Y ont tendance à décroître quand les valeurs de X augmentent.

La liaison linéaire est d’autant plus forte que |ρ| est proche de 1.

Le coefficient de corrélation mesure de façon symétrique la relation entre les deux variables, sans notion de contrôle sur l’une des deux variables :

$$\rho_{XY} = \rho_{YX}$$
Remarques

- $r$ est très sensible aux valeurs extrêmes.
  - Intérêt représentation graphique
  - Statistiques descriptives univariées

- On peut avoir une liaison même si $r = 0$ ;
  $r$ mesure seulement le caractère linéaire d’une liaison.

\[ r = 0 \quad r > 0 \quad r < 0 \]
**Exemple** : Fréquence cardiaque maximale (FCM)

On souhaite étudier une relation éventuelle entre l’âge d’un individu, notée $X$ et sa FCM, variable notée $Y$

<table>
<thead>
<tr>
<th>Individu $i$</th>
<th>Age $x_i$</th>
<th>FCM $y_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>40</td>
<td>187</td>
</tr>
<tr>
<td>2</td>
<td>36</td>
<td>195</td>
</tr>
<tr>
<td>3</td>
<td>51</td>
<td>180</td>
</tr>
<tr>
<td>4</td>
<td>49</td>
<td>190</td>
</tr>
<tr>
<td>5</td>
<td>47</td>
<td>185</td>
</tr>
<tr>
<td>6</td>
<td>51</td>
<td>183</td>
</tr>
<tr>
<td>7</td>
<td>32</td>
<td>195</td>
</tr>
<tr>
<td>8</td>
<td>55</td>
<td>185</td>
</tr>
<tr>
<td>9</td>
<td>55</td>
<td>189</td>
</tr>
<tr>
<td>10</td>
<td>23</td>
<td>201</td>
</tr>
<tr>
<td>11</td>
<td>49</td>
<td>189</td>
</tr>
<tr>
<td>12</td>
<td>52</td>
<td>185</td>
</tr>
<tr>
<td>13</td>
<td>35</td>
<td>195</td>
</tr>
</tbody>
</table>

Questions :

1. Calculer $\bar{x}$ et $\bar{y}$
2. Calculer $s_{xy}$, $s_x^2$ et $s_y^2$
3. Calculer $r$
**Exemple :** Fréquence cardiaque maximale (FCM)

![FCM en fonction de l'âge](image-url)
**Exemple :** Fréquence cardiaque maximale (FCM)

\( \bar{x} = 44.23, \ \bar{y} = 189.15 \)

\[
s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x}\bar{y}
\]

\[
s_{xy} = \frac{1}{13} \times 108157 - 44.23 \times 189.15
\]

\[
s_{xy} = -46.65
\]

On applique la correction \( n/(n-1) \), \( s_{xy} = -50.54 \)

\[
s_x^2 = 100.69, \ s_y^2 = 35.14
\]

\[
r = \frac{-50.54}{\sqrt{100.69 \times 35.14}} = -0.85
\]
Test de la significativité de $\rho$

**Principe du test** : Test de la nullité du coefficient de corrélation.
- Si $\rho = 0$ alors il n’y a pas de liaison linéaire entre $X$ et $Y$
- Si $\rho \neq 0$ alors il existe une relation linéaire entre $X$ et $Y$

**Condition d’application** : $X \sim \mathcal{N}(\mu_1, \sigma_1)$ et $Y \sim \mathcal{N}(\mu_2, \sigma_2)$

**En pratique** : Vérification de $X$ et $Y$ ”a peu près normales” (symétrique) car test robuste ($n$ grand)

**Hypothèses du test**

\[ \begin{align*}
H_0 : \rho &= 0 \\
H_1 : \rho &\neq 0
\end{align*} \]
Test de la significativité de $\rho$

- Statistique de test

Sous $\mathcal{H}_0$,

$$T = \frac{R \sqrt{n - 2}}{\sqrt{1 - R^2}} \sim T_{n-2 \; ddl}$$

- Région critique

$$W = ] - \infty ; -t_{n-2 \; ddl, \alpha/2} [ \cup [ t_{n-2 \; ddl, \alpha/2} ; + \infty [$$

- Décision

Si $t \in W$ alors on rejette $\mathcal{H}_0$ au risque de première espèce $\alpha$. Il existe une relation linéaire entre $X$ et $Y$. 

 Michaël Genin (Université de Lille 2)  
 Corrélation - Régession linéaire  
 Version - 19 février 2015  
 20 / 123
**Exemple :** Fréquence cardiaque maximale (FCM)

\[
t = \frac{r\sqrt{n-2}}{\sqrt{1 - r^2}} = \frac{-0.85\sqrt{13 - 2}}{\sqrt{1 - (-0.85)^2}} = -5.35
\]

Région critique \( W : ] - \infty; -2.201] \cup [2.201; +\infty[ \)

\( t \in W \) donc on rejette \( H_0 \) au risque de première espèce \( \alpha = 5\% \). Il existe une liaison linéaire statistiquement significative entre \( X \) et \( Y \).
Remarques

- La loi de $R$ est aussi tabulée et permet de calculer des seuils de significativité pour une taille échantillon et un risque $\alpha$ donné
  - *Exemple* $\alpha = 0.05$ et $n = 30$, une liaison est significative si $|r| > 0.36$.

- Le test est robuste mais si les conditions d’application ne sont pas clairement vérifiées, on utilisera un test non paramétrique

  → Test sur le coefficient de corrélation de Spearman
Coefficient de corrélation de Spearman

- Étude de la relation entre les rangs des variables $X$ et $Y$
- Permet la détection de relations monotones (croissantes ou décroissantes)
- La relation n’est pas forcément linéaire (exponentiel, puissance,)
- Adapté aux formes curvilignes
- Utile lorsque la distribution des variables est asymétrique

Soient $X = (x_1, \ldots, x_n)$, $Y = (y_1, \ldots, y_n)$
et $R = (r_1, \ldots, r_n)$, $S = (s_1, \ldots, s_n)$ leurs rangs respectifs.

Le coefficient de corrélation de Spearman calculé entre $X$ et $Y$ est égal au coefficient de corrélation de Pearson calculé entre $R$ et $S$.

En l’absence d’*ex-equo* :

$$r_s = 1 - \frac{6 - \sum_{i=1}^{n} [r_i - s_i]^2}{n^3 - n}$$
Test du coefficient de corrélation de Spearman

Test non paramétrique : ne nécessite pas d’hypothèses *a priori* sur les distributions de $X$ et $Y$
Souvent utilisé lorsque $n < 30$

Hypothèses du test

$$\begin{align*}
\mathcal{H}_0 : \rho &= 0 \\
\mathcal{H}_1 : \rho &\neq 0
\end{align*}$$

Petits effectifs ($4 \leq n \leq 10$) : Les valeurs limites de $r_s$ sont tabulées de manière exacte en fonction du risque $\alpha$.

Grands effectifs ($n > 10$) sous $\mathcal{H}_0$ :

$$T = \frac{R_s \sqrt{n - 1}}{\sqrt{1 - R_s^2}} \sim \mathcal{T}_{n-2}$$
Corrélations partielles

En pratique, il arrive fréquemment que la liaison observée entre 2 variables soit en fait due aux variations d’une troisième variable appelée facteur de confusion.

On peut définir, connaissant les 3 corrélations, une corrélation partielle (ou conditionnelle) entre 2 variables, conditionnellement à la 3ème :

\[ r_{xy/z} = \frac{r_{xy} - r_{xz} r_{yz}}{\sqrt{1 - r_{xz}^2} + \sqrt{1 - r_{yz}^2}} \]
Le coefficient de corrélation permet de mesurer le lien linéaire entre deux variables quantitatives $X$ et $Y$. On peut également chercher à modéliser le lien entre $X$ et $Y$ afin de réaliser des prédictions :

- Exprimer $Y$ en fonction de $X$
- Ex : Prédire la FCM d’un patient en ne connaissant que son âge

Le coefficient de corrélation n’est pas suffisant

Recours à la régression linéaire :

$$Y = \beta_1 X + \beta_0 + \epsilon$$

Où $Y$ est la variable à expliquer et $X$ la variable explicative.
Cadre d’étude

- $Y$ est un caractère non contrôlé (caractère expliqué)
- $X$ est un caractère contrôlé (caractère explicatif)

Considérons un échantillon de $n$ observations i.i.d. : $I = \{1, \ldots, n\}$

- $y_i$ est la valeur observée pour l’individu $i$
- $x_i$ est la valeur fixée pour l’individu $i$

Objectif : Exprimer le lien entre $Y$ et $X$.

$$Y = f(X) + \epsilon$$

Il existe une infinité de liaisons fonctionnelles $\rightarrow$ la plus simple est linéaire

Régression linéaire simple

1. Modèle de régression
2. Droite de régression au sens des moindres carrés
Modèle de régression linéaire

\( \forall i \in I, y_i \) est la réalisation de la v.a.r. \( Y_i \) telle que

\[
Y_i = \beta_1 x_i + \beta_0 + \epsilon_i
\]

Avec

- \( \epsilon_i \) : erreur du modèle (v.a.r.) (part de variabilité de \( Y \) qui n’est pas expliquée par le lien fonctionnel linéaire)
- \( \beta_0, \beta_1 \) : coefficients du modèle, constantes (valeurs fixes dans la population).

Hypothèses du modèle

- \( \mathbb{E}[\epsilon_i] = 0, \forall [\epsilon_i] = \sigma^2 \) (hypothèse d’homoscédasticité)
- L’erreur est indépendante de \( X \) \( \rightarrow \) \( \text{Cov}(x_i, \epsilon_i) = 0 \)
- Les \( \epsilon_i, 1 \leq i \leq n \), sont mutuellement indépendantes (absence d’autocorrélations des résidus) \( \rightarrow \) \( \text{Cov}(\epsilon_i, \epsilon_j) = 0 \) si \( i \neq j \).
- \( \epsilon_i \sim \mathcal{N}(0, \sigma^2) \) (normalité des résidus) \( \rightarrow \) tests dans le modèle
**Droite de régression au sens des moindres carrés**

**Objectif** : estimer $\beta_0$ et $\beta_1$ grâce à leur estimateurs $B_0$ et $B_1$ et leur réalisations $b_0$ et $b_1$ sur un échantillon d’observations i.i.d. de taille $n$.

Trouver $b_0$ et $b_1$ qui minimisent l’erreur.

**Figure** : Erreur importante

**Figure** : Erreur minimisée
Droite de régression au sens des moindres carrés

Objectif : estimer $\beta_0$ et $\beta_1$ grâce à leur estimateurs $B_0$ et $B_1$ et leur réalisations $b_0$ et $b_1$ sur un échantillon d’observations i.i.d. de taille $n$.

Trouver $b_0$ et $b_1$ qui minimisent un critère d’ajustement.

$\Rightarrow$ Méthode des moindres carrés ordinaires

\[
S(\beta_0, \beta_1) = \sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_1 x_i + \beta_0))^2
\]

$\Rightarrow \min S(\beta_0, \beta_1)$

Dérivées partielles $\Rightarrow$ Systèmes aux équations normales

\[
\text{Solutions : } b_1 = \frac{s_{xy}}{s_x^2} \text{ et } b_0 = \bar{y} - b_1 \bar{x}
\]
Droite de régression au sens des moindres carrés

La droite de régression au sens des moindres carrés a pour expression :

\[ \hat{y}_i = b_1 x_i + b_0 \]

C’est une estimation du modèle de régression par la méthode des moindres carrés.

Les erreurs observées sur l’échantillon sont appelés résidus.

\[ e_i = (y_i - \hat{y}_i) = y_i - b_1 x_i - b_0 \]
Remarques

- $b_1$ peut être estimé *via* le coefficient de corrélation de Pearson :

$$b_1 = r_{yx} \frac{s_y}{s_x}$$

- $b_0$ et $b_1$ sont des estimations de $\beta_0$ et $\beta_1$.

- $b_0$ et $b_1$ sont des réalisations des v.a.r. $B_0$ et $B_1$

  → Estimateurs des MCO de $\beta_0$ et $\beta_1$

$$B_0 = \bar{Y} - B_1 \bar{x}$$

$$B_1 = \frac{S_{xy}}{s_x^2}$$

- $\beta_0$ et $\beta_1$ peuvent être également estimés par la méthode de maximum de vraisemblance. On montre que les estimateurs de maximum de vraisemblance de $\beta_0$ et $\beta_1$ sont égaux aux estimateurs des MCO.
Propriétés des estimateurs - Moments (1)

\[ \mathbb{E}[B_0] = \beta_0 \quad \mathbb{E}[B_1] = \beta_1 \]

→ \( B_0 \) et \( B_1 \) sont des estimateurs *sans biais*

\[ \mathbb{V}[B_0] = \sigma^2_{B_0} = \sigma^2 \left( \frac{1}{n} + \frac{x^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right) \]

\[ \mathbb{V}[B_1] = \sigma^2_{B_1} = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma^2}{n} \times \frac{1}{s_x^2} \]

• \( B_0 \) et \( B_1 \) sont des estimateurs *convergents*

\[ \mathbb{V}[B_0] \xrightarrow{n \to \infty} 0 \]
\[ \mathbb{V}[B_1] \xrightarrow{n \to \infty} 0 \]

• \( B_0 \) et \( B_1 \) sont de *variance minimale* (Théorème de Gauss-Markov)
Propriétés des estimateurs - Moments (2)

Les variances $\sigma_{B_0}^2$ et $\sigma_{B_1}^2$ font intervenir la variance de l’erreur $\sigma^2$.

- Cette variance est inconnue.
- Nous devons l’estimer. Naturellement, on peut utiliser les écarts entre $Y_i$ et $\hat{Y}_i$ :

$$S^2 = \frac{1}{n} \sum_{i=1}^{n} \left( Y_i - \hat{Y}_i \right)^2$$

Or, cet estimateur est biaisé ($\mathbb{E}[S^2] = (n - 2)\sigma^2$).

On montre que

$$S_{n-2}^2 = \frac{1}{n-2} \sum_{i=1}^{n} \left( Y_i - \hat{Y}_i \right)^2,$$

est un estimateur non biaisé de $\sigma^2$. 
Propriétés des estimateurs - Distributions d’échantillonnage

De par les hypothèses du modèle et l’estimation de $\sigma^2$, on montre que

\[
\frac{B_0 - \beta_0}{\hat{\sigma}_{B_0}} \sim T_{n-2} \text{ d.d.l.}
\]

\[
\frac{B_1 - \beta_1}{\hat{\sigma}_{B_1}} \sim T_{n-2} \text{ d.d.l.}
\]

avec

\[
\hat{\sigma}^2_{B_0} = S_{n-2}^2 \left( \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right)
\]

\[
\hat{\sigma}^2_{B_1} = \frac{S_{n-2}^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

Ces distributions vont nous permettre de réaliser une inférence statistique sur les coefficients (tests de nullité et intervalles de confiance).
Qualité de l’ajustement

Equation d’analyse de la variance

\[ y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \]
\[ (y_i - \bar{y})^2 = (\hat{y}_i - \bar{y})^2 + (y_i - \hat{y}_i)^2 \]
\[ \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \]

\[ \begin{align*}
\sum_{i=1}^{n}(y_i - \bar{y})^2 & = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{Somme des carrés expliquée SCE}} \quad + \quad \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{Somme des carrés résiduelle SCR}}
\end{align*} \]
Qualité de l’ajustement

Evaluation de la qualité d’ajustement du modèle

⇒ Coefficient de détermination

\[ R^2 = \frac{SCE}{SCT} \in [0, 1] \]

Interprétation : Part de variabilité de Y expliquée par le modèle de régression linéaire.

Remarque importante : le carré du coefficient de corrélation linéaire est égal au coefficient de détermination :

\[ r^2 = R^2 \]

Conséquence : deux tests statistiques équivalents pour évaluer la liaison linéaire entre X et Y.
Etude de la validité du modèle

**Etude des résidus** : Vérification des hypothèses du modèle faites sur les erreurs

- Vérifier la normalité des résidus observés
- Vérifier que les résidus ne contiennent pas d’information structurée ($\forall [\epsilon] = \sigma^2$)
- Les résidus ne dépendent pas de $X$
- Vérifier que les résidus ne sont pas auto-corréllés entre eux (les $\epsilon_i$ sont mutuellement indépendantes)

$\rightarrow$ Ces hypothèses vont permettre par la suite de réaliser des tests dans le modèle linéaire.

**Observations aberrantes / influentes**

- Valeur aberrante de $x$
- Observation $i$ mal reconstituée par la régression $\rightarrow (e_i$ élevé)
- Observation $i$ présentant un poids exagéré dans la régression (prédiction / coefficients)
Etude des résidus

Vérification de la normalité des résidus

- histogramme $\Rightarrow$ la distribution doit être unimodale et symétrique autour de 0.
- Tests (Kolmogorov-Smirnov, Shapiro Wilks, ...) mais souvent tests peu puissants (peu aptes à rejeter $H_0$)
- Droite de Henry $\Rightarrow$ confronte les quantiles théoriques de la loi normale et la distribution cumulée estimée sur les données
Etude des résidus

Vérification de l’homoscédasticité des résidus

Les résidus sont dits **homoscédastiques** si leur dispersion est homogène et ne dépend pas des valeurs de la variable explicative (et donc pas non plus des valeurs prédites).
On vérifie que les résidus n’ont pas de structure particulière en traçant un graphe des résidus :

![Graphes de résidus](image1.png)
Etude des résidus

On peut localiser des points du nuage mal expliqués par la relation linéaire en traçant les deux droites $d_1 = 2s_{n-2}$ et $d_2 = -2s_{n-2}$

- on peut considérer ces points, si ils ne sont pas trop nombreux, comme des points exceptionnels, les éliminer et recalculer $b_1$ et $b_0$.
- on peut aussi attribuer un poids moindre aux points aberrants $\Rightarrow$ moindres carrés pondérés (fonction de l’écart $|y - \hat{y}|/2s_{n-2}^2$). Méthode plus robuste
- si il y a beaucoup de points mal expliqués (en dehors de la bande), c’est que le modèle est mal choisi.
Etude des résidus
Entude des résidus

Vérification de l’indépendance entre les résidus

Test de Durbin Watson

\[ H_0 : \text{il n’y a pas de corrélation entre } \epsilon_i \text{ et } \epsilon_{i-1} \]
\[ H_1 : \text{il y a de corrélation entre } \epsilon_i \text{ et } \epsilon_{i-1} \]

\[ d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \]

La valeur de \(d\) est toujours comprise entre 0 et 4, \(d = 2\) quand il n’y a pas d’autocorrélation.
La loi de \(d\) est tabulée : \(D_L\) et \(D_U\) bornes au risque \(\alpha\).
Observations aberrantes / influentes

**Valeur aberrante de x**

- Effet important sur l’estimation de la droite de régression
- Mauvais ajustement aux données
- **Solution** : descriptif univarié → boxplot
Observations aberrantes / influentes

Résidus studentisés internes

Idée : Mettre en évidence les observations dont le résidu $e_i$ est important. Pour une observation $i$, le résidu studentisé interne est défini par :

\[
t_i = \frac{e_i}{s_{n-2} \sqrt{1 - h_i}}
\]

avec $h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{j=1}^{n}(x_i - \bar{x})^2}$ (levier de l’observation $i$). On montre que

\[
T_i \sim T_{(n-2)}.
\]

D’où :

\[
RC : |t_i| > t_{n-2}^{1-\alpha/2}
\]

Problème : L’observation évaluée a participé à la construction de la droite (Juge et partie).
Résidus studentisés externes

Idée : Estimer le modèle sans l’observation $i$ ($-i$) et comparer la valeur observée de $y_i$ à celle prédite par le modèle ($\hat{y}_i(-i)$ : prédiction pour une nouvelle observation).

Pour une observation $i$, le résidu studentisé externe est défini par :

$$t_i^* = \frac{y_i - \hat{y}_i(-i)}{s_{n-2}(-i) \sqrt{1 - h_i(-i)}}$$

On montre que

$$T_i^* \sim T_{(n-3)}.$$ 

D'où :

$$RC : \left| t_i^* \right| > t_{n-3}^{1-\alpha/2}$$
Observations aberrantes / influentes

**Distance de Cook**

Idée : Évaluer l’influence d’une observation $i$ sur l’estimation des coefficients. Comparaison des prédictions du modèle complet et du modèle sans l’observation $i$. La distance de Cook pour une observation $i$ est définie par

$$D_i = \frac{\sum_{j=1}^{n}(\hat{y}_i - \hat{y}_i(-i))^2}{2s_{n-2}^2}$$

Règle de décision (cas régression simple) :

$$RC : D_i > 1$$

Si la différence entre les prédictions est élevée, l’observations $i$ joue un rôle sur l’estimation des coefficients.
Conclusion sur l’étude de la validité du modèle

Etape très importante !!

1. Descriptif univarié
2. Estimation des coefficients du modèle
3. Vérification des hypothèses sur les erreurs
4. Détection d’observations influentes
   • Si OUI : Correction ou suppression
   • Nouvelle estimation des coefficients

Après ces étapes : Inférence statistique
Evaluation globale de la régression

Tableau d’analyse de variance - Test de significativité globale

- $R^2$ permet d’évaluer la qualité de l’ajustement. L’information emmenée par la régression de $Y$ par $X$ traduit-elle une relation qui existe vraiment dans la population ?

**Table : Tableau ANOVA**

<table>
<thead>
<tr>
<th>Source de variation</th>
<th>Somme des carrés</th>
<th>DDL</th>
<th>Carrés moyens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expliquée</td>
<td>$SCE = \sum_i (\hat{y}_i - \bar{y})^2$</td>
<td>1</td>
<td>$CME = \frac{SC_E}{1}$</td>
</tr>
<tr>
<td>Résiduelle</td>
<td>$SCR = \sum_i (y_i - \hat{y}_i)^2$</td>
<td>$n-2$</td>
<td>$CMR = \frac{SCR}{n-2}$</td>
</tr>
<tr>
<td>Totale</td>
<td>$SCT = \sum_i (y_i - \bar{y})^2$</td>
<td>$n-1$</td>
<td>-</td>
</tr>
</tbody>
</table>

**Degrés de liberté :**

- $SCT$ nécessite l’estimation $\bar{y} \rightarrow n - 1$ ddl
- $SCR$ nécessite l’estimation de $\beta_0$ et $\beta_1$ pour $\hat{y}_i \rightarrow n - 2$ ddl
- $SCE$ par déduction : $(n - 1) - (n - 2) = 1$ ddl
Evaluation globale de la régression

Tableau d’analyse de variance - Test de significativité globale

Le test $F$ permet d’évaluer la significativité globale de la régression.

$$F = \frac{CME}{CMR} \sim \mathcal{F}_{1,n-2} \text{ ddl}$$

Interprétation :

$$\begin{align*}
\mathcal{H}_0 & : \text{"Le modèle est non explicatif"} \\
\mathcal{H}_1 & : \text{"Le modèle est explicatif"}
\end{align*}$$
Evaluation des coefficients - $\beta_1$

**Test de significativité de $\beta_1$**

Idée : tester la nullité de $\beta_1$.

\[
\begin{align*}
\mathcal{H}_0 : \beta_1 &= 0 \quad \text{"X n’a aucun pouvoir explicatif sur Y"} \\
\mathcal{H}_1 : \beta_1 &\neq 0 \quad \text{"X a un pouvoir explicatif sur Y"}
\end{align*}
\]

Nous savons que \( \frac{B_1 - \beta_1}{\hat{\sigma}_{B_1}} \sim \mathcal{T}_{n-2}, \) par conséquent sous $\mathcal{H}_0$

\[
\frac{B_1}{\hat{\sigma}_{B_1}} \sim \mathcal{T}_{n-2}
\]

**Intervalle de confiance de $\beta_1$**

\[
IC_{\beta_1}^{1-\alpha} = \left[ b_1 \pm t_{(1-\alpha/2; n-2)} \frac{s_{n-2}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \right]
\]
Evaluation des coefficients - $\beta_1$

**Équivalence avec le test de significativité globale**

Uniquement dans le cas de la régression linéaire simple, observons que

\[
F = \frac{SCE/1}{SCR/(n-2)} = \frac{\sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2}{S_{n-2}^2} = \frac{\sum_{i=1}^{n} (B_1 x_i + B_0 - \bar{Y})^2}{S_{n-2}^2}
\]

\[
F = \frac{\sum_{i=1}^{n} (\hat{Y}_i - (\bar{Y} - B_1 \bar{x}) - \bar{Y})^2}{s_{n-2}^2} = \frac{B_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{s_{n-2}^2} = \frac{B_1^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}
\]

\[
F = \frac{B_1^2}{\sigma_{\hat{B}_1}^2} = \left( \frac{B_1}{\sigma_{\hat{B}_1}} \right)^2.
\]

**Note** : Équivalence entre la loi de Student et la loi de Fisher.

\[
(T_{(n-2)})^2 \equiv F_{(1,n-2)}
\]
Evaluation des coefficients - $\beta_1$

**Equivalence avec le test de significativité de $\rho$**

Observons que

$$F = \frac{SCE/1}{SCR/(n-2)} = \frac{(n-2)SCE}{SCR} = \frac{(n-2)SCE}{SCT - SCR} = \frac{(n-2)R^2}{1 - R^2}$$

$$T = \frac{r\sqrt{n-2}}{\sqrt{1 - r^2}}$$

or $R^2 = r^2$ donc

$$F = (T)^2$$

Et comme $F = \left( \frac{B_1}{\hat{\sigma}_{B_1}} \right)^2$, on conclue que

$$F = (T)^2 = \left( \frac{B_1}{\hat{\sigma}_{B_1}} \right)^2$$
Evaluation des coefficients - $\beta_1$

Dans le cas d’une régression linéaire simple de type :

$$Y = \beta_1 X + \beta_0 + \epsilon$$

Tester la significativité globale du modèle

$$\equiv$$

Tester la significativité de $\beta_1$

$$\equiv$$

Tester la significativité de $\rho$
Evaluation des coefficients - $\beta_1$

Lien entre test et intervalle de confiance

\[
\text{IC}^{1-\alpha}_{\theta} = \{\theta_0 / \mathcal{H}_0 = \{\theta = \theta_0\} \text{ est accepté au niveau de confiance } 1 - \alpha\}
\]

Conséquences : pour tester $\mathcal{H}_0 : \beta_1 = 0$ au risque $\alpha$, on peut simplement vérifier si 0 appartient ou non à l’intervalle de confiance :

- Si $0 \in \text{IC}$ alors on conserve $\mathcal{H}_0$
- Si $0 \notin \text{IC}$ alors on rejette $\mathcal{H}_0$
Evaluation des coefficients - $\beta_0$

**Test de significativité de $\beta_0$**

Idée : tester la nullité de $\beta_0$.

\[
\begin{align*}
\mathcal{H}_0 & : \beta_0 = 0 \quad \text{"L’ordonnée à l’origine n’est pas significative"} \\
\mathcal{H}_1 & : \beta_0 \neq 0 \quad \text{"L’ordonnée à l’origine est significative"}
\end{align*}
\]

Nous savons que $\frac{B_0 - \beta_0}{\delta B_0} \sim \mathcal{T}_{n-2}$, par conséquent sous $\mathcal{H}_0$

\[
\frac{B_0}{\hat{\delta} B_0} \sim \mathcal{T}_{n-2}
\]

**Intervalle de confiance de $\beta_0$**

\[
IC_{\beta_0}^{1-\alpha} = \left[ b_0 \pm t_{(1-\alpha/2;n-2)} s_{n-2} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \right]
\]
Intervalle de prédiction d’une observation

**Contexte** : Le modèle de régression ayant été validé, il est possible d’estimer la valeur de \( y_{n+1} \) pour une observation \( x_{n+1} \) n’appartenant pas à l’échantillon :

\[
\hat{y}_{n+1} = b_1 x_{n+1} + b_0
\]

Or \( \hat{y}_{n+1} \) n’est qu’une estimation de la ”vraie droite de régression” dans la population. Une estimation ponctuelle ne suffit pas, il faut lui associer un intervalle de confiance :

\[
IC_{1-\alpha}^{y_{n+1}} = \left[ \hat{y}_{n+1} \pm t_{(1-\alpha/2; n-2)} s_{n-2} \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \right]
\]
Intervalle de prédiction d’une observation

\[
IC_{y_{n+1}}^{1-\alpha} = \left[ \hat{y}_{n+1} \pm t_{(1-\alpha/2; n-2)} s_{n-2} \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} \right]
\]

Quelques remarques

- La taille du rayon de l’intervalle de confiance sera d’autant plus faible que
  - \( s_{n-2}^2 \) est faible → la régression est de bonne qualité
  - \( n \) est élevé
  - \( x_{n+1} \) est proche de \( \bar{x} \) ↔ \( (x_{n+1} - \bar{x})^2 \) est faible
  - \( \sum_{i=1}^{n} (x_i - \bar{x})^2 \) est élevé → les \( x_i \) sont bien dispersés

- Attention : utiliser des valeurs de \( x \) qui sont dans le cadre d’étude (relativement proches de \( \bar{x} \) pour obtenir de bonnes prédictions.)
Mise en évidence d’un problème

Vente de CD et cas de grippe H1N1 en 2009

<table>
<thead>
<tr>
<th>Région</th>
<th>Nb cas de grippe H1N1</th>
<th>Nb ventes CD M.J.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Région 1</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Région 2</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>Région n</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>

Il existe une liaison linéaire significative …

Test de $\rho : p < 1.10^{-4}$, Test de $\beta_1 : p < 1.10^{-4}$

… mais pas de relation de cause à effet !
Mise en évidence d’un problème

Existence de **FACTEURS DE CONFUSION** (ici : la période de temps)
Exemple : vente de CD de MJ les plus fortes en automne alors

Pour la même période :

\[
\begin{align*}
\text{ventes} & \quad \text{cas de grippes}
\end{align*}
\]

**Figure** : Diagramme de corrélation
Mise en évidence d’un problème

- Nécessité de pouvoir ajuster sur des facteurs de confusion
- Dans de nombreuses situations, plusieurs facteurs peuvent expliquer un caractère

\[ Y = f(X_1, X_2, \ldots, X_p) \]

Régression linéaire multiple
Cadre d’étude

- $Y$ est un caractère non contrôlé (caractère expliqué)
- $X_1, X_2, \ldots X_p$ sont des caractères contrôlés (caractères explicatifs)

Considérons un échantillon de $n$ observations i.i.d. : $l = \{1, \ldots, n\}$
- $y_i$ est la valeur observée pour l’individu $i$
- $x_{ij}$ est la valeur fixée pour l’individu $i$ et la variable $j$, $j \in \{1, 2, \ldots, p\}$

Objectif : Exprimer le lien entre $Y$ et les $X_j$.

$$Y = f (X_1, X_2, \ldots, X_p) + \epsilon$$

Il existe une infinité de liaisons fonctionnelles → la plus simple est linéaire

**Régression linéaire multiple**

1. Modèle de régression
2. Hyperplan de régression au sens des moindres carrés
Modèle de régression linéaire multiple

\forall i \in I, y_i est la réalisation de la v.a.r. \( Y_i \) telle que

\[ Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_p x_{ip} + \epsilon_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i \]

Avec

- \( \epsilon_i \): erreur du modèle (v.a.r.) (part de variabilité de \( Y \) qui n'est pas expliquée par le lien fonctionnel linéaire)
- \( \beta_0, \beta_1, \ldots, \beta_p \): coefficients du modèle, constantes (valeurs fixes dans la population).

Hypothèses du modèle

- \( \mathbb{E}[\epsilon_i] = 0 \), \( \mathbb{V}[\epsilon_i] = \sigma^2 \) (hypothèse d'homoscédasticité)
- L'erreur est indépendantes des \( X_j \) \( \rightarrow \) \( \text{COV}(x_{ij}, \epsilon_i) = 0 \)
- \( \epsilon_i \sim \mathcal{N}(0, \sigma^2) \) (normalité des résidus) \( \rightarrow \) tests dans le modèle
- Les \( \epsilon_i \), \( 1 \leq i \leq n \), sont mutuellement indépendantes (absence d'autocorrélation des résidus) \( \rightarrow \) \( \text{Cov}(\epsilon_i, \epsilon_j) = 0 \) si \( i \neq j \).
**Ecriture du modèle**

Sur un échantillon de \( n \) observations i.i.d. :

\[
Y_1 = \beta_0 + \beta_1 x_{11} + \beta_2 x_{12} + \cdots + \beta_p x_{1p} + \epsilon_1 \\
Y_2 = \beta_0 + \beta_1 x_{21} + \beta_2 x_{22} + \cdots + \beta_p x_{2p} + \epsilon_2 \\
\vdots \\
Y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \cdots + \beta_p x_{np} + \epsilon_n
\]

**Ecriture matricielle**

\[
\begin{bmatrix}
Y_1 \\
\vdots \\
Y_n
\end{bmatrix} = \begin{bmatrix}
\beta_0 \\
\vdots \\
\beta_p
\end{bmatrix} = \begin{bmatrix}
1 & x_{11} & x_{12} & \cdots & x_{1p} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & x_{n1} & x_{n2} & \cdots & x_{np}
\end{bmatrix} \begin{bmatrix}
\epsilon_1 \\
\vdots \\
\epsilon_n
\end{bmatrix}
\]

\[
Y = X. \beta + \epsilon \quad \text{avec} \\
Y \in \mathbb{R}^{n \times 1}, \quad X \in \mathbb{R}^{n \times (p+1)}, \quad \beta \in \mathbb{R}^{(p+1) \times 1}, \quad \epsilon \in \mathbb{R}^{n \times 1}
\]
Hyperplan de régression au sens des moindres carrés

OBJECTIF : estimer $\beta_0, \beta_1, \ldots, \beta_p$ grâce à leur estimateurs $B_0, B_1, B_2, \ldots, B_p$ et leur réalisations $b_0, b_1, b_2, \ldots, b_p$ sur un échantillon d’observations i.i.d. de taille $n$.

$$
\beta = \begin{pmatrix}
\beta_0 \\
\vdots \\
\beta_p
\end{pmatrix} \quad B = \begin{pmatrix}
B_0 \\
\vdots \\
B_p
\end{pmatrix} \quad b = \begin{pmatrix}
b_0 \\
\vdots \\
b_p
\end{pmatrix}
$$

Trouver $b$ qui minimisent l’erreur :

$$S(\beta_0, \ldots, \beta_p) = \sum_{i=1}^{n} (\epsilon_i)^2 = \sum_{i=1}^{n} \left( y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 = ||\epsilon||^2$$

Solution : $b = (X^T X)^{-1} X^T Y$

Remarque : $B$ est appelé estimateur des moindres carrés de $\beta$. 
Hyperplan de régression au sens des moindres carrés

L’hyperplan de régression au sens des moindres carrés a pour expression :

\[ \hat{y}_i = b_0 + \sum_{j=1}^{p} b_j x_{ij} \]

C’est une estimation du modèle de régression multiple par la méthode des moindres carrés.

Les erreurs observées sur l’échantillon sont appelés résidus.

\[ e_i = (y_i - \hat{y}_i) = y_i - b_0 - \sum_{j=1}^{p} b_j x_{ij} \]
Propriétés des estimateurs - Moments (1)

\[ \mathbb{E}[\mathbf{B}] = \beta \]  
estimateur sans biais

\[ \mathbb{V}[\mathbf{B}] = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \]

\[ \mathbb{V}[\mathbf{B}]_{(p+1) \times (p+1)} \]  
est appelée matrice de variances - covariances des coefficients :

\[
\begin{pmatrix}
\hat{\sigma}_{B_0}^2 & \text{COV}(B_0, B_1) & \ldots & \text{COV}(B_0, B_p) \\
\cdot & \hat{\sigma}_{B_1}^2 & \ldots & \text{COV}(B_1, B_p) \\
\cdot & \cdot & \ldots & \cdot \\
\cdot & \cdot & \cdots & \hat{\sigma}_{B_p}^2
\end{pmatrix}
\]
Propriétés des estimateurs - Moments (2)

La matrice de variances - covariances $\nabla[B]$ fait intervenir la variance de l’erreur $\sigma^2$.

- Cette variance est inconnue
- Elle est estimée par $S^2_{n-p-1}$ au moyen du tableau d’ANOVA.

On montre que

$$S^2_{n-p-1} = \frac{SCR}{n-p-1} = \frac{\sum_{i=1}^{n} (E_i)^2}{n-p-1}$$

est un estimateur non biaisé de $\sigma^2$. 
Propriétés des estimateurs - Distributions d’échantillonnage

De par les hypothèses du modèle et l’estimation de $\sigma^2$, on montre que 
$$\forall j \in \{0, 1, \ldots, p\}$$

$$\frac{B_j - \beta_j}{\hat{\sigma}_{B_j}} \sim T_{n-p-1} \text{ d.d.l.}$$

avec $\hat{\sigma}_{B_j}$ tiré de la matrice de variances-covariances.

Ces distributions vont nous permettre de réaliser une inférence statistique sur les coefficients (tests de nullité et intervalles de confiance).
Qualité de l’ajustement

Equation d’analyse de la variance

\[ y_i - \bar{y} = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \]

\[ (y_i - \bar{y})^2 = (\hat{y}_i - \bar{y})^2 + (y_i - \hat{y}_i)^2 \]

\[ \sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]

\[ \sum_{i=1}^{n} (y_i - \bar{y})^2 \]

Somme des carrés totale

SCT

\[ \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 \]

Somme des carrés expliquée

SCE

\[ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \]

Somme des carrés résiduelle

SCR
Qualité de l’ajustement

Evaluation de la qualité d’ajustement du modèle

⇒ Coefficient de détermination

\[ R^2 = \frac{SCE}{SCT} \in [0, 1] \]

Interprétation : Part de variabilité de \( Y \) expliquée par le modèle de régression linéaire multiple.

Remarque importante : \( R^2 \) est fonction du nombre variables explicatives dans le modèle (même non pertinentes)

\[ p \uparrow \rightarrow R^2 \uparrow \]

Aussi : tests de significativité des coefficients, sélection de variables (modèle parcimonieux)
Etude de la validité du modèle

**Etude des résidus** : Vérification des hypothèses du modèle faites sur les erreurs
- Vérifier la normalité des résidus observés
- Vérifier que les résidus ne contiennent pas d’information structurée ($\forall [\epsilon] = \sigma^2$)
- Les résidus ne dépendent pas des $X_j$
- Vérifier que les résidus ne sont pas auto-corrélés entre eux (les $\epsilon_i$ sont mutuellement indépendantes)

→ Ces hypothèses vont permettre par la suite de réaliser des tests dans le modèle linéaire.

**Observations aberrantes / influentes**
- Valeur aberrante d’une observation (diagnostic univarié, multivarié)
- Observation $i$ mal reconstituée par la régression $\rightarrow (e_i$ élevé)
- Observation $i$ présentant un poids exagéré dans la régression (prédiction / coefficients)
Vérification de la normalité des résidus

- histogramme $\Rightarrow$ la distribution doit être unimodale et symétrique autour de 0.
- Tests (Kolmogorov-Smirnov, Shapiro Wilks, ...) mais souvent tests peu puissants (peu aptes à rejeter $\mathcal{H}_0$)
- Droite de Henry $\Rightarrow$ confronte les quantiles théoriques de la loi normale et la distribution cumulée estimée sur les données
Etude des résidus

Vérification de l’homoscédasticité des résidus

Les résidus sont dits homoscédastiques si leur dispersion est homogène et ne dépend pas des valeurs de la variable explicative $x_j$ (et donc pas non plus des valeurs prédites). On vérifie que les résidus n’ont pas de structure particulière en traçant un graphe des résidus :

![Diagramme des résidus non corrélés](#)

![Diagramme des résidus corrélés](#)
Etude des résidus

On peut localiser des points du nuage mal expliqués par la relation linéaire en traçant les deux droites $d_1 = 2s_{n-p-1}$ et $d_2 = -2s_{n-p-1}$

- on peut considérer ces points, si ils ne sont pas trop nombreux, comme des points exceptionnels, les éliminer et recalculer $b_1$ et $b_0$.
- on peut aussi attribuer un poids moindre aux points aberrants $\Rightarrow$ moindres carrés pondérés (fonction de l’écart $|y - \hat{y}|/2s_{n-p-1}^2$). Méthode plus robuste
- si il y a beaucoup de points mal expliqués (en dehors de la bande), c’est que le modèle est mal choisi.
Etude des résidus

Graphe des résidus

Michaël Genin (Université de Lille 2)
Etude des résidus

Vérification de l’indépendance entre les résidus

Test de Durbin Watson

\[ \mathcal{H}_0 : \text{il n'y a pas de corrélation entre } \epsilon_i \text{ et } \epsilon_{i-1} \]
\[ \mathcal{H}_1 : \text{il y a une corrélation entre } \epsilon_i \text{ et } \epsilon_{i-1} \]

\[ d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \]

La valeur de \( d \) est toujours comprise entre 0 et 4, \( d = 2 \) quand il n’y a pas d’autocorrélations.
La loi de \( d \) est tabulée: \( D_L \) et \( D_U \) bornes au risque \( \alpha \).
Observations aberrantes / influentes

**Valeur aberrante de \( x_j \) (Univarié) → Boxplot**

- Effet important sur l’estimation de l’hyperplan de régression
- Mauvais ajustement aux données
- **Solution** : descriptif univarié → boxplot

**Problème** : Ne prend pas en compte les interactions possibles entre les \( X_j \).
Observations aberrantes / influentes

**Valeur aberrante (Multivarié) → Levier**

**Principe :** Pour une observation $i$, mesure la distance avec le centre de gravité du nuage défini par les $X_j$, $j \in \{1, \ldots, p\}$

Le levier d’une observation $i$ se lit sur la diagonale de la matrice $H$ (*hat matrix*) :

$$H = X (X^T X)^{-1} X^T$$

**En pratique**

$$h_{ii} = h_i = x_i (X^T X)^{-1} x_i^T$$

avec $x_i$ la $i$ème ligne de la matrice $X$. 
Observations aberrantes / influentes

Valeur aberrante (Multivarié) → Levier

Règle de décision

\[ R.C. : h_i > 2 \times \frac{p + 1}{n} \]

- Permet de détecter des observations aberrantes / influentes d’un point de vue multivarié
- Effet néfaste sur l’estimation des \( \beta_j \) par les MCO
Observations aberrantes / influentes

Résidus studentisés internes

Idée : Mettre en évidence les observations dont le résidu $e_i$ est important. Pour une observation $i$, le résidu studentisé interne est défini par :

$$t_i = \frac{e_i}{s_{n-p-1} \sqrt{1 - h_i}}$$

avec $h_i$ levier de l’observation $i$. On montre que

$$T_i \sim T_{(n-p-1)}.$$ 

D’où :

$$RC : \ |t_i| > t_{n-p-1}^{1-\alpha/2}$$

Problème : L’observation évaluée a participé à la construction de la droite (Juge et partie).
**Observations aberrantes / influentes**

**Résidus studentisés externes**

Idée : Estimer le modèle sans l’observation \(i\) \((-i)\) et comparer la valeur observée de \(y_i\) à celle prédite par le modèle \((\hat{y}_i(-i) : \text{prédiction pour une nouvelle observation})\).

Pour une observation \(i\), le résidu studentisé externe est défini par :

\[
t_i^* = \frac{y_i - \hat{y}_i(-i)}{s_{n-p-1}(-i) \sqrt{1 - h_i(-i)}}
\]

On montre que

\[
T_i^* \sim T_{(n-p-1)}.
\]

D’où :

\[
RC : |t_i^*| > t_{n-p-1}^{1-\alpha/2}
\]

Remarque :

\[
t_i^* = \sqrt{\frac{n-p-2}{n-p-1-t_i^2}}
\]
Observations aberrantes / influentes

**Distance de Cook**

Idée : Évaluer l’influence d’une observation \( i \) sur l’estimation des coefficients. Comparaison des prédictions du modèle complet et du modèle sans l’observation \( i \).

La distance de Cook pour une observation \( i \) est définie par

\[
D_i = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_j(-i))²}{(p + 1)s_{n-p-1}²}
\]

Règle de décision

\[
RC : D_i > 1
\]

\[
RC : D_i > \frac{4}{n - p - 1} \quad \text{(Ajustement sur le nombre de variables)}
\]

Si la différence entre les prédictions est élevée, l’observations \( i \) joue un rôle sur l’estimation des coefficients.
DFBETAS

Idée: Si la distance de COOK a identifié une observation ayant une influence sur l’estimation des coefficients, on peut aller plus loin pour déterminer quel coefficient est affecté.
Pour une observation \( i \) et pour chaque coefficient \( \beta_j, j \in \{0, 1, \ldots, p\} \), le DFBETAS est défini par :

\[
DFBETAS_{i,j} = \frac{b_j - b_j(-i)}{s_{n-p-1}(-i) \sqrt{(X^T X)_j^{-1}}}
\]

Règle de décision

\[
RC : |DFBETAS_{i,j}| > \frac{2}{\sqrt{n}}
\]
**Evaluation globale de la régression**

**Tableau d’analyse de variance - Test de significativité globale**

- $R^2$ permet d’évaluer la qualité de l’ajustement. L’information emmenée par la régression de $Y$ par les $X_j$ traduit-elle une relation qui existe vraiment dans la population ?

<table>
<thead>
<tr>
<th>Source de variation</th>
<th>Somme des carrés</th>
<th>DDL</th>
<th>Carrés moyens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expliquée</td>
<td>$SCE = \sum_i (\hat{y}_i - \bar{y})^2$</td>
<td>$p$</td>
<td>$CME = \frac{CME}{p}$</td>
</tr>
<tr>
<td>Résiduelle</td>
<td>$SCR = \sum_i (y_i - \hat{y}_i)^2$</td>
<td>$n - p - 1$</td>
<td>$CMR = \frac{CMR}{n-p-1}$</td>
</tr>
<tr>
<td>Totale</td>
<td>$SCT = \sum_i (y_i - \bar{y})^2$</td>
<td>$n - 1$</td>
<td>-</td>
</tr>
</tbody>
</table>

**Degrés de liberté** :
- $SCT$ nécessite l’estimation $\bar{y} \rightarrow n - 1$ ddl
- $SCR$ nécessite l’estimation des $\beta_j$ pour $\hat{y}_i \rightarrow n - (p + 1)$ ddl
- $SCE$ par déduction : $(n - 1) - (n - p - 1) = 1$ ddl
Evaluation globale de la régression

Tableau d’analyse de variance - Test de significativité globale

Le test $F$ permet d’évaluer la significativité globale de la régression.

$$\begin{align*}
H_0 & : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \\
H_1 & : \exists j / \beta_j \neq 0
\end{align*}$$

Sous $H_0$

$$F = \frac{CME}{CMR} \sim \mathcal{F}_{p, n-p-1} \text{ ddl}$$

Interprétation :

$$\begin{align*}
H_0 & : "\text{Le modèle est non explicatif}" \\
H_1 & : "\text{Le modèle est explicatif}"\end{align*}$$
Evaluation des coefficients - $\beta_j$

**Test de significativité de $\beta_j$**

Idée : tester la nullité de $\beta_j$.

\[
\begin{align*}
\mathcal{H}_0 & : \beta_j = 0 \ " X_j \ n'a \ aucun \ pouvoir \ explicatif \ sur \ Y" \\
\mathcal{H}_1 & : \beta_j \neq 0 \ " X_j \ a \ un \ pouvoir \ explicatif \ sur \ Y"
\end{align*}
\]

Nous savons que $\frac{B_j - \beta_j}{\hat{\sigma}_{B_j}} \sim T_{n-p-1}$, par conséquent :

Sous $\mathcal{H}_0$

\[
\frac{B_j}{\hat{\sigma}_{B_j}} \sim T_{n-p-2}
\]

**Intervalle de confiance de $\beta_j$**

\[
IC_{\beta_j}^{1-\alpha} = \left[ b_j \pm t_{(1-\alpha/2; n-p-1)} \hat{\sigma}_{B_j} \right]
\]
Intervalle de prédiction d’une observation

Contexte : Le modèle de régression ayant été validé, il est possible d’estimer la valeur de $y_{n+1}$ pour une observation $X_{n+1}$ n’appartenant pas à l’échantillon :

$$X_{n+1} = (x_{n+1,1}, x_{n+1,2}, \ldots, x_{n+1,p})$$

$$\hat{y}_{n+1} = b_0 + \sum_{j=1}^{p} b_j x_{n+1,j}$$

$$\hat{y}_{n+1} = X_{n+1} \cdot b$$

Or $\hat{y}_{n+1}$ n’est qu’une estimation du ”vrai hyperplan de régression” dans la population. Une estimation ponctuelle ne suffit pas, il faut lui associer un intervalle de confiance :

$$IC_{1-\alpha}^{y_{n+1}} = \left[ \hat{y}_{n+1} \pm t(1-\alpha; n-p-1)s_{n-p-1}\sqrt{1 + \frac{X_{n+1} (X^T X)^{-1} X^T_{n+1}}{}}, \right]$$
**Définition**

Une variable explicative $X_j$ est **colinéaire** à une autre variable $X_k$ lorsque $r_{X_j, X_k} > 0.8$.

Une variable explicative $X_j$ est **multicolinéaire** aux autres variables lorsque

$$X_j = \beta'_0 + \beta'_1 X_1 + \ldots + \beta'_{j-1} X_{j-1} + \beta'_{j+1} X_{j+1} + \ldots + \beta'_p X_p,$$

elle peut s’écrire comme une combinaison linéaire des autres variables explicatives.

**Conséquences**

- Valeurs/Signes des coefficients contraires à l’intuition
- Variances estimées des coefficients trop importantes
- Coefficients non significatifs (inférence statistique)
- Instabilité du modèle
- Risque de passer à côté d’une variable importante (redondance)
Détection de la colinéarité

Variance Inflation Factor - VIF

Principe : Pour chaque $X_j$, réalisation de la régression de $X_j$ avec les autres variables explicatives. On note $R^2_j$ le coefficient de détermination associé à cette régression.

Pour une variable $X_j$, le VIF est défini par

$$VIF_j = \frac{1}{1 - R^2_j}$$

La valeur du VIF sera d’autant plus forte que $X_j$ est une combinaison linéaire des autres variables.

Règle de décision

$$VIF_j \geq 4$$

Rq : Détermination des variables incriminées dans la combinaison linéaire → Cercle des corrélations (ACP)
Variance Inflation Factor - VIF

Remarque :

\[ \sigma^2_{B_j} = \frac{\sigma^2}{n} VIF_j \]

Donc :

- Plus la valeur de $VIF_j$ est importante plus $\sigma^2_{B_j}$ sera importante
- Estimation instable
- Problème de significativité du coefficient
Traitement de la colinéarité

**Approche ”métier”**

- Mise en évidence de plusieurs variables explicatives colinéaires (VIF + ACP)
- Discussion et choix de la (ou les) variable(s) la (les) plus pertinente(s) pour l’analyse

**Approche statistique**

- Méthode de sélection de variables (*Forward, Backward, Stepwise*)
  
  Mais si toutes les variables sont pertinentes ?

- Régression sur les composantes principales de l’ACP
- Régression *ridge*
- Régression PLS
Motivations - Critère de sélection de variables

**Motivations**

- Sélection d’un sous-ensemble de $X_j$ pertinentes et non redondantes qui expliquent au mieux $Y$
- Principe du rasoir d’Occam → Modèles parcimonieux
  - Modèle plus simple, lisible, robuste, stable
  - Nombre restreint de variables explicatives (collecte des données)
- Traitement de la multicolinéarité

**Critère de sélection de variables**

Retrait ou ajout d’une variable $X_j$ dans le modèle en fonction de sa valeur du $F_j$ partiel de Fisher :

$$F_j = \left( \frac{B_j}{\hat{\sigma}_{B_j}} \right)^2 \sim \mathcal{F}_{1,n-p-1}$$
Motivations - Critère de sélection de variables

Autres critères de sélection de variables

- $R^2$ ajusté
  \[
  \bar{R}^2 = 1 - \frac{SCR/(n - p - 1)}{SCT/(n - 1)}
  \]

- AIC (à minimiser)
  \[
  AIC = n \ln \left( \frac{SCR}{n} \right) + 2(p + 1)
  \]

- BIC de Schwartz (à maximiser)
  \[
  BIC = n \ln \left( \frac{SCR}{n} \right) + \ln(n)(p + 1)
  \]

- etc. . .
Méthodes de sélection de variables

Méthode ascendante (Forward)

Principe : On part du modèle sans $X_j$. On ajoute successivement les $X_j$ qui sont significatifs au sens du $F$ partiel de Fisher et on s’arrête lorsqu’on ne peut plus ajouter de $X_j$ (NS dans le modèle)

TANT QUE Condition d’arrêt = FAUX FAIRE

- Calcul de $F_j$ pour chaque $X_j$ candidate (Si $\emptyset X_j \rightarrow$ Arrêt)
- Choix de $F_j^*$ tel que $F_j^* = \max\{F_1, \ldots, F_p\}$

SI $P(F_j^* > f_j^*) < \alpha_{sle}$ ALORS

- Ajout de $X_j$ au modèle

SINON Arrêt

Fin TANT QUE

Remarques :

- Plus $\alpha_{sle}$ est élevé plus le nombre de variables dans le modèle sera important
- En pratique : $\alpha_{sle} = 0.2$
Méthodes de sélection de variables

Méthode ascendante (Forward) - Exemple

Soient $Y$ et $X_1, X_2, X_3$ sur un échantillon de $n = 50$ observations

<table>
<thead>
<tr>
<th>Etape</th>
<th>Modèle</th>
<th>ddl</th>
<th>$F(p$-value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$Y = \beta_0$</td>
<td>1 ;50-1-1</td>
<td>$X_1 \rightarrow 43 \ (3,51.10^{-8})$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$X_2 \rightarrow 150 \ (2,22.10^{-16})$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$X_3 \rightarrow 12 \ (1,11.10^{-3})$</td>
</tr>
<tr>
<td>2</td>
<td>$Y = \beta_0 + \beta_2 X_2$</td>
<td>1 ;50-2-1</td>
<td>$X_1 \rightarrow 9 \ (4,31.10^{-3})$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$X_3 \rightarrow 2 \ (0.1639)$</td>
</tr>
<tr>
<td>3</td>
<td>$Y = \beta_0 + \beta_2 X_2 + \beta_1 X_1$</td>
<td>1 ;50-3-1</td>
<td>$X_3 \rightarrow 0.7 \ (0.4071)$</td>
</tr>
</tbody>
</table>

$X_3$ n’est plus significative dans le modèle. Modèle final :

$$Y = \beta_0 + \beta_2 X_2 + \beta_1 X_1$$
Méthodes de sélection de variables

Méthode descendante (Backward)

Principe : On part du modèle saturé (toutes les $X_j$). On retire successivement les $X_j$ qui sont non significatives au sens du $F$ partiel de Fisher et on s’arrête lorsqu’on toutes les $X_j$ sont significatives.

**TANT QUE** Condition d’arrêt = FAUX **FAIRE**

- Calcul de $F_j$ pour chaque $X_j$ candidate (Si $\emptyset X_j \rightarrow$ Arrêt)
- Choix de $F_j^*$ tel que $F_j^* = \text{min}\{F_1, \ldots, F_p\}$
  
  **SI** $P(F_j^* > f_j^*) > \alpha_{sls}$ **ALORS**
  - Retrait de $X_j$ du modèle

  **SINON** Arrêt

**Fin TANT QUE**

Remarques :
- Plus $\alpha_{sls}$ est faible plus le nombre de variables dans le modèle sera faible
- En pratique : $\alpha_{sls} = 0.2$
Méthodes de sélection de variables

Méthode Stepwise

Principe : "Mix" entre les méthodes ascendante et descendante. On début par le modèle sans $X_j$ et on choisit le $X_j$ le plus significatif au sens du $F$ partiel de Fisher ($\alpha_{sle}$). Dans les étapes suivantes, on vérifie que l’ajout d’une variable de conduit pas à la non significativité ($\alpha_{sls}$) des autres déjà présentes dans le modèle (phase descendante). Le processus se termine quand aucune $X_j$ n’est significative lors de la phase ascendante.

Avantage : Contrairement la méthode ascendant, une $X_j$ introduite dans le modèle peut être remise en cause lors des étapes suivantes.

En pratique :

$$\alpha_{sle} = \alpha_{sls} = 0.2$$
Méthodes de sélection de variables

Méthode Stepwise

On part du modèle : \( Y = \beta_0 \)

TANT QUE Condition d’arrêt = FAUX FAIRE

Calcul de \( F_j \) pour chaque \( X_j \) candidate (Si \( \emptyset X_j \rightarrow \) Arrêt)
Choix de \( F_j^* \) tel que \( F_j^* = \max\{F_1, \ldots, F_p\} \)

SI \( \mathbb{P}(F_j^* > f_j^*) < \alpha_{sle} \) ALORS

Ajout de \( X_j \) du modèle

POUR Chaque \( X_j \) inclue dans le modèle FAIRE

Calcul de \( F_j \) pour chaque \( X_j \)
Choix de \( F_j^* \) tel que \( F_j^* = \min\{F_1, \ldots, F_p\} \)

SI \( \mathbb{P}(F_j^* > f_j^*) > \alpha_{sle} \) ALORS

Retrait de \( X_j \) du modèle

FIN POUR

SINON Arrêt

Fin TANT QUE
Processus de modélisation

1. Estimation des coefficients (MCO)
2. Mesure de la qualité d’ajustement ($R^2$)
3. Étude la validité du modèle
   - Si hypothèses sur les erreurs non vérifiées → STOP
   - Si observations aberrantes/influentes →
     - Correction/Suppression
     - Retour à l’Étape 1
4. Inférence statistique
   - Test de significativité globale
   - Tests de significativité des coefficients
5. Evaluation de la multicollinéarité
6. Sélection de variables (Modèle parcimonieux)
   - Modèle restreint
   - Réitération des Étapes 1, 2, 3, 4
Références

Livres

- Probabilités Analyses des données et Statistique, G. Saporta, TECHNIP

Supports en ligne

- Econométrie - Régression linéaire simple et multiple, R. Rakotomalala
- Pratique de la régression linéaire multiple - Diagnostic et Sélection de variables, R. Rakotomalala
  http://eric.univ-lyon2.fr/~ricco/cours/cours/La_regression_dans_la_pratique.pdf
- Régression linéaire, A. Guyader
Annexe 1 - Système aux équations normales

\[
\min S(\beta_0, \beta_1) = \min \sum_{i=1}^{n} (y_i - (\beta_1 x_i + \beta_0))^2
\]

\[
= \min \sum_{i=1}^{n} [y_i^2 - 2x_i y_i \beta_1 - 2y_i \beta_0 + \beta_1^2 x_i^2 + 2\beta_0 \beta_1 x_i + \beta_0^2]
\]

S(\beta_0, \beta_1) est strictement convexe donc elle admet un minimum au point unique (b_0, b_1) déterminé en annulant les dérivées partielles de S :

\[
\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0 \Rightarrow \sum_{i=1}^{n} -2y_i + 2b_1 x_i + 2b_0 = 0 \tag{1}
\]

\[
\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0 \Rightarrow \sum_{i=1}^{n} -2x_i y_i + 2b_1 x_i^2 + 2b_0 x_i = 0 \tag{2}
\]
Annexe 1 - Système aux équations normales

De (1) on déduit que

\[ b_0 = \frac{\sum_{i=1}^{n} y_i - b_1 \sum_{i=1}^{n} x_i}{n} = \bar{y} - b_1 \bar{x} \] (3)

De (2) on déduit que

\[ b_1 \sum_{i=1}^{n} x_i^2 + b_0 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \] (4)

En utilisant (3) nous obtenons

\[ b_1 \sum_{i=1}^{n} x_i^2 + (\bar{y} - b_1 \bar{x}) \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \] (5)
Annexe 1 - Système aux équations normales

En divisant par $n$ les deux termes :

\[
b_1 \frac{1}{n} \sum_{i=1}^{n} x_i^2 + (\bar{y} - b_1 \bar{x}) \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i y_i \quad (6)
\]

\[
b_1 \left[ \frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2 \right] = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \bar{x} \bar{y} \quad (7)
\]

\[
b_1 = \frac{s_{xy}}{s_x^2} \quad (8)
\]
Annexe 2 - RLM : MCO

Observons que $||\epsilon||^2 = ||Y - X\beta||^2$. Aussi

$$||\epsilon||^2 = (Y - X\beta)^T(Y - X\beta) = Y^TY - Y^TX\beta - \beta^TX^TY + \beta^TX^TX\beta.$$

Or comme la transposée d’un scalaire est égale à lui-même :

$$(Y^TX\beta)^T = \beta^TX^TY,$$

nous avons donc

$$S(\beta_0, \ldots, \beta_p) = ||\epsilon||^2 = Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta$$

Minimiser la fonction $S$ revient à annuler les dérivées différentielles par rapport à $\beta$. L’annulation de la dérivation matricielle nous donne

$$\frac{\partial S}{\partial \beta} = -2(X^TY) + 2(X^TX)b = 0$$
Annexe 2 - RLM : MCO

Aussi

\[(X^T X) b = (X^T Y).\]

Or \(\text{rg}(X) = p \text{ et } p \leq n\) donc \((X^T X)\) est inversible. Aussi

\[b = (X^T X)^{-1} (X^T Y).\]
Annexe 3 - RLM : moments des estimateurs

**Espérance**

\[
\mathbb{E}[\mathbf{B}] = \mathbb{E} \left[ (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \right] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbb{E}[\mathbf{Y}]
\]

\[
\mathbb{E}[\mathbf{B}] = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbb{E}[\mathbf{X} \cdot \beta + \epsilon] = \beta
\]

**Variance**

\[
\mathbb{V}[\mathbf{B}] = \mathbb{V} \left[ (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \right] = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbb{V}[\mathbf{Y}] (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T
\]

\[
\mathbb{V}[\mathbf{B}] = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbb{V}[\epsilon] (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}
\]
Annexe 4 - $F$ partiel de Fisher

L’égalité

$$F = \left( \frac{B_j}{\hat{\sigma}_{B_j}} \right)^2 \sim \mathcal{F}_{1,n-p-1}$$

se base sur un cas particulier du test de nullité d’un bloc de $q$ coefficients.

\[ \begin{align*}
\mathcal{H}_0 : & \quad Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_j X_j + \beta_{j+q+1} X_{j+q+1} + \ldots + \beta_p X_p + \epsilon \\
\mathcal{H}_1 : & \quad Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon
\end{align*} \]

\[ \begin{align*}
\mathcal{H}_0 : & \quad \text{Modèle sans les } q \text{ variables (modèle restreint)} \\
\mathcal{H}_1 : & \quad \text{Modèle complet}
\end{align*} \]

Posons

$$R^2_{(0)} : \quad \text{Coefficient de détermination du modèle restreint (} \mathcal{H}_0 \text{)}$$

$$R^2_{(1)} : \quad \text{Coefficient de détermination du modèle complet (} \mathcal{H}_1 \text{)}$$
Annexe 4 - $F$ partiel de Fisher

Notons $F$ la statistique de test associée. On montre que sous $\mathcal{H}_0$,

$$F = \frac{\left( R^2_{(1)} - R^2_{(0)} \right)}{q} \left( 1 - R^2_{(1)} \right)/(n - p - 1) \sim F_{q, n - p - 1}$$

Autre interprétation : si l’accroissement

$$\left( R^2_{(1)} - R^2_{(0)} \right) > \frac{q}{n - p - 1} f_{q, n - p - 1}$$

alors la place des $q$ variables dans le modèle est justifiée.
Annexe 4 - *F* partiel de Fisher

**Cas particulier quand** \( q = 1 \)

On veut tester la nullité d’un coefficient \( \beta_j \)

\[
\begin{align*}
\mathcal{H}_0 & : Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} + \ldots + \beta_p X_p + \epsilon \\
\mathcal{H}_1 & : Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon
\end{align*}
\]

\[
\begin{align*}
\mathcal{H}_0 & : \text{Modèle sans } X_j \text{ (modèle restreint)} \\
\mathcal{H}_1 & : \text{Modèle complet}
\end{align*}
\]

Posons

\[
R^2_{(0)} : \text{Coefficient de détermination du modèle restreint } (\mathcal{H}_0)
\]

\[
R^2_{(1)} : \text{Coefficient de détermination du modèle complet } (\mathcal{H}_1)
\]
Annexe 4 - $F$ partiel de Fisher

**Cas particulier quand $q = 1$**

Notons $F$ la statistique de test associée. On montre que sous $\mathcal{H}_0$,

\[
F = \frac{(R^2_{(1)} - R^2_{(0)}) / 1}{(1 - R^2_{(1)}) / (n - p - 1)} \sim F_{1, n-p-1}
\]

Or

\[
T = \frac{B_j}{\hat{\sigma}_{B_j}} \sim T_{n-p-1}
\]

Donc

\[
F = T^2 = \left( \frac{B_j}{\hat{\sigma}_{B_j}} \right)^2
\]

**Remarque** : test de la significativité du coefficient et son apport au $R^2$. 
Annexe 4 - $F$ partiel de Fisher

**Cas particulier quand** $q = 1$

Autre interprétation : si l'accroissement

\[
\left( R^2_{(1)} - R^2_{(0)} \right) > \frac{1}{n - p - 1} f_{1, n-p-1}
\]

alors la place de $X_j$ dans le modèle est justifiée.

**Note** : Tests de type III sous SAS.