Unité Fondamentale : Statistique 2 1ère année
Notes de Cours
ENSAT

J. Gergaud

Septembre 2003

Table des matières

[1] Exemple provenant de B. Scherrer, "Biostatistique", p. 598, ed. G. Morin, 1984

Chapitre 1

Régression linéaire

1 Introduction

Exemple 1.0.1. [1] On désire savoir comment le taux de cholestérol sérique dépend de l'âge chez l'homme. Pour cela on a pris 5 échantillons d'hommes adultes d'âges bien déterminés 25,35,45,55,et 65 ans. On a obtenu les données suivantes :

Ages	25	25	25	25	25	25	25	35	35	35
Taux	1.8	2.3	2	2.4	2	2.5	2.6	2.6	2.9	2.3
Ages	35	35	35	35	45	45	45	45	45	45
Taux	2.4	2.1	2.5	2.7	2.7	3	3.1	2.3	2.5	3
Ages	45	45	55	55	55	55	55	65	65	65
Taux	3.3	2.7	3.1	2.9	3.4	2.4	3.4	3.7	2.8	3.3
Ages	65	65	65							
Taux	3.5	3.3	2.6							

La figure 1.1 donne la représentation graphique de ces données.

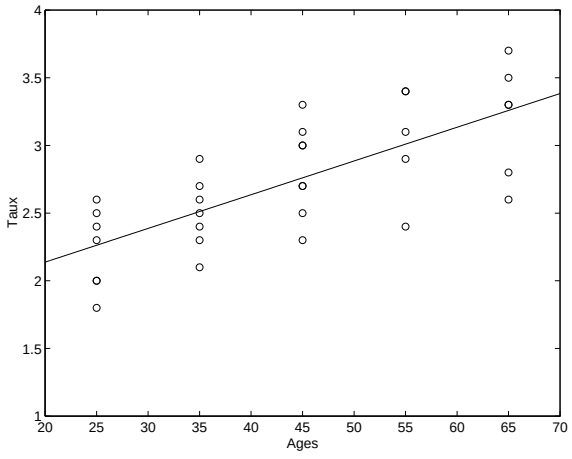

FIG. 1.1 – Taux de cholestérol en fonction de l'âge

Que peut-on conclure de ces données ?

En pratique nous sommes souvent amenés à rechercher une relation entre deux variables x et y. Pour cela, dans un premier temps, nous collectons des données $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Ensuite nous représentons graphiquement ces données. Nous pouvons par exemple avoir les cas suivants :

[1] Exemple provenant de l'ouvrage de Grémy et Salmon, "Bases statistiques", page 122.

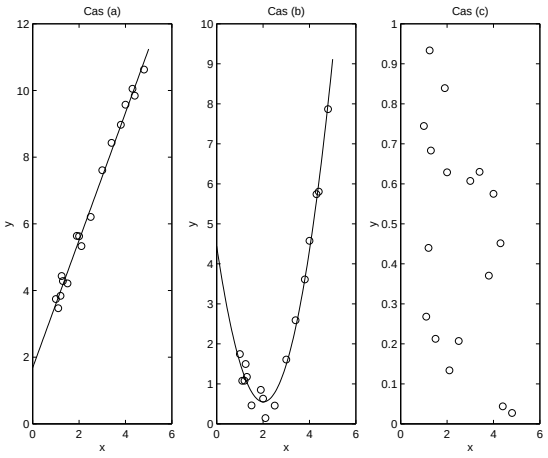

FIG. 1.2 – Différentes formes de graphes

Suivant les cas de la figure 1.2, nous pouvons penser aux modèles :

Cas (a) $y(x) = \beta_0 + \beta_1 x$;

Cas (b) $y(x) = \beta_0 + \beta_1 x + \beta_2 x^2$;

Cas (c) pas de modèle.

L'objet de la régression linéaire simple est l'étude du cas (a). L'un des buts sera de prédire la "meilleure" valeur de y connaissant x. Nous aurons donc dans un premier temps à estimer les valeurs des paramètres β_0 et β_1, puis à estimer la valeur prédite. Ceci nécessite bien évidemment de mettre dans le modèle de l'aléatoire.

2 Formalisation mathématique du problème

2.1 Modèle

Pour des raisons de notation nous noterons dans la suite Z la variable aléatoire définissant le modèle Nous considérerons dans ce chapitre le cas où :

(i) x sera une variable parfaitement connue ;

(ii) Z sera une variable aléatoire réelle.

Nous pouvons alors écrire le modèle suivant :

$$E(Z) = \beta_0 + \beta_1 x \tag{1.1}$$

$$\begin{aligned} \text{ou encore } Z &= \beta_0 + \beta_1 x + \eta \\ \text{avec } E(\eta) &= 0 \end{aligned}$$

Remarque 2.1.1. le modèle choisi ici est appelé le modèle linéaire. La régression linéaire concerne en fait le cas où les variables X et Z sont toutes les deux des variables aléatoires. Ce qui nous intéresse ici c'est de pouvoir prédire, tout du moins en parti, la valeur de Y connaissant la valeur de X. Dans ce cas que X soit une variable aléatoire ou une variable parfaitement connue ne change rien dans les formules. Nous nous garderons cependant de parler de corrélation (cf. le chapitre suivant) lorsque X est une variable parfaitement connue. Nous avons choisi ici le modèle linéaire pour simplifier un peu la présentation.

2.2 Postulats

En pratique nous avons un n-échantillon $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Si nous répétons l'expérience une infinité de fois nous définissons ainsi le n-échantillon aléatoire $Y = (Y_1, Y_2, \ldots, Y_n)$ où :

$$\begin{aligned} Y_i : \mathcal{P}^n &\longrightarrow \mathbf{R} \\ 1 \text{ expérience} &\longmapsto y_i \end{aligned}$$

avec

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad \forall i \in \{1, \ldots, n\} \tag{1.2}$$

Si nous posons :

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix}; \qquad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}; \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \text{et} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

alors le modèle s'écrit :

$$Y = X\beta + \varepsilon \tag{1.3}$$

Remarque 2.2.1. Nous supposerons toujours dans la suite que la matrice X est de rang 2, c'est-à-dire que tous les x_i ne sont pas égaux.

Pour pouvoir avoir des propriétés intéressantes, en particulier sur les estimateurs des paramètres β_0 et β_1, nous aurons besoin de supposer que les postulats suivants sont vrais :

Postulats :

(i) les variables aléatoires ε_i sont d'espérance nulle et de variance constante : $E(\varepsilon_i) = 0$ et $Var(\varepsilon_i) = \sigma^2$;

(ii) les variables aléatoires ε_i sont indépendantes ;

(iii) les variables aléatoires ε_i sont de loi normale.

Remarque 2.2.2.
- (i) $\Leftrightarrow E(Y_i) = \beta_0 + \beta_1 x_i$ et $Var(Y_i) = \sigma^2$;
- (ii) $\Leftrightarrow (Y_i)_i$ indépendantes ;
- (i) et (iii) $\Leftrightarrow Y_i$ est de loi $\mathcal{N}(\beta_0 + \beta_1 x_i, \sigma^2)$;
- Ecrire $Var(Y_i) = \sigma^2$ implique en particulier que la variabilité de doit pas dépendre de la variable x.

Remarque 2.2.3. Les postulats sont équivalents à

(i) $\varepsilon : \mathcal{N}(0, \Gamma)$

(ii) $\Gamma = \sigma^2 I$

2.3 Visualisation

On peut visualiser le modèle et les postulats (i) et (iii) par le graphique (1.3)

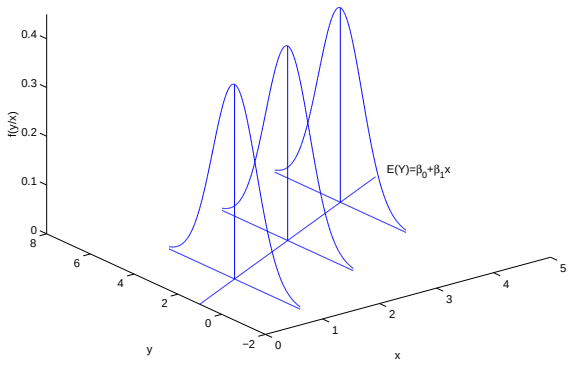

FIG. 1.3 – Visualisation du modèle linéaire

2.4 Questions

Les questions qui se posent maintenant sont en autre :

(i) Comment, à partir des données estimer les paramètres β_0 et β_1 ;

(ii) Comment avoir une estimation de la prévision de Y connaissant x ;

(iii) Comment détecter des écarts aux postulats de départ.

Nous allons dans les sections suivantes répondre à ces questions.

3 Estimation des paramètres

3.1 Calcul des estimateurs

Une droite sera d'autant plus proche des points $M_i(x_i, y_i)$ que les écarts entre ces points et la droite seront faibles. L'un des critères les plus utilisés est le critère des moindres carrés qui est la somme des carrés des écarts $r_i = y_i - \hat{y}_i$ (cf figure (1.4)).

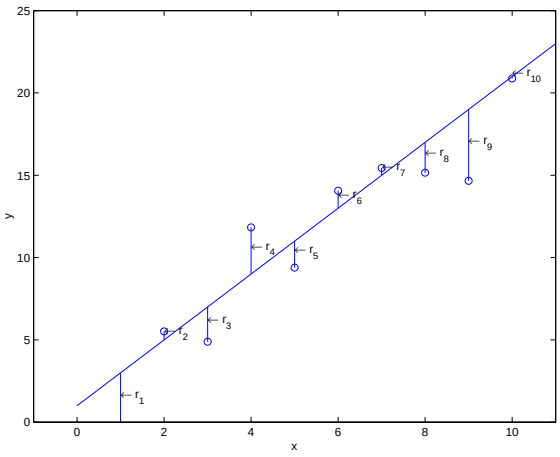

FIG. 1.4 – Moindres carrés.

La question est maintenant de trouver les valeurs des paramètres qui rendent la valeur du critère la plus faible possible. Nous sommes ainsi ramené au problème d'optimisation suivant :

$$(P)\begin{cases} Min & f(\beta) = \frac{1}{2}\sum_{i=1}^{n} r_i^2 = \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 = \frac{1}{2}\|X\beta - y\|^2 \\ \beta \in \mathbf{R}^2 \end{cases}$$

Théorème 3.1.1. *La solution du problème* (P) *est :*

$$\begin{cases} b_0 = \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \dfrac{SPE}{SCE_x} \end{cases}$$

Démonstration

Nous pouvons appliquer ici la condition nécessaire et suffisante de solution du problème d'optimisation. Ceci nous conduit aux équations normales

$$\nabla f(\beta) = {}^tXX\beta - {}^tXy = 0$$

D'où le système linéaire suivant :

$$\Leftrightarrow \begin{cases} n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

Ce système admet une seule solution et une seule car la matrice X est de rang 2 et donc la matrice ${}^tXX$ est inversible.

$$\Leftrightarrow \begin{cases} \bar{y} = \beta_0 + \beta_1 \bar{x} \\ n\beta_0 \bar{x} + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

$$\Leftrightarrow \begin{cases} \bar{y} = \beta_0 + \beta_1 \bar{x} \\ n(\bar{y} - \beta_1 \bar{x})\bar{x} + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

$$\Leftrightarrow \begin{cases} \bar{y} = \beta_0 + \beta_1 \bar{x} \\ \beta_1 (\sum_{i=1}^{n} x_i^2 - n\bar{x}^2) = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \end{cases}$$

$$\Leftrightarrow \begin{cases} \bar{y} = \beta_0 + \beta_1 \bar{x} \\ \beta_1 = \dfrac{SPE}{SCE_x} \end{cases}$$

□

Nous noterons dans la suite $\hat{\beta}_0$ et $\hat{\beta}_1$ ces solutions.

Exemple 3.1.2. Reprenons l'exemple (1.0.1). Lorsque l'on effectue les calculs à la main il est utile de calculer le tableau préliminaire 1.2.

	y	x	xy	y^2	x^2
1	y_1	x_1	$x_1 y_1$	y_1^2	x_1^2
⋮	⋮	⋮	⋮	⋮	⋮
i	y_i	x_i	$x_i y_i$	y_i^2	x_i^2
⋮	⋮	⋮	⋮	⋮	⋮
n	y_n	x_n	$x_n y_n$	y_n^2	x_n^2
Totaux	$Y_.$	$X_.$	$\sum_i x_i y_i$	$\sum_i y_i^2$	$\sum_i x_i^2$
Moyennes	$\bar{y}_.$	$\bar{x}_.$			

TAB. 1.1 – Calculs préliminaires

	x	y	xy	x^2	y^2
1	25	1.8	45.0	625	3.24
2	25	2.3	57.5	625	5.29
3	25	2.0	50.0	625	4.00
4	25	2.4	60.0	625	5.76
5	25	2.0	50.0	625	4.00
6	25	2.5	62.5	625	6.25
7	25	2.6	65.0	625	6.76
8	35	2.6	91.0	1225	6.76
9	35	2.9	101.5	1225	8.41
⋮	⋮	⋮	⋮	⋮	⋮
33	65	2.6	169.0	4225	6.76
Totaux	1445	90.1	4103.5	69625	253.31
Moyennes	43.79	2.73			

TAB. 1.2 – Calculs préliminaires sur l'exemple

Les estimations ponctuelles sont alors

$$\hat{\beta}_1 = \frac{4103.5 - \dfrac{1445 \times 90.1}{33}}{69625 - \dfrac{1445^2}{33}} = \frac{158.2}{6351.5} = 0.025$$

$$\hat{\beta}_0 = 2.73 - 0.025 \times 43.79 = 1.64$$

Remarque 3.1.3. Nous noterons r_i le résidu d'indice i :

$$r_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) = y_i - \hat{y}_i$$

On vérifie alors que

$$\begin{array}{rcl} \sum_{i=1}^n r_i & = & \sum_{i=1}^n (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)) \\ & = & \sum_{i=}^n y_i - \sum_{i=1}^n \hat{\beta}_0 - \sum_{i=1}^n \hat{\beta}_1 x_i \\ & = & n\bar{y} - n\hat{\beta}_0 - n\bar{x}\hat{\beta}_1 = 0 \end{array}$$

3.2 Propriétés des estimateurs

Si nous refaisons complètement la collecte de données, avec les mêmes valeurs pour les x_i, nous obtenons de nouveaux points M_i et donc de nouvelles estimations ponctuelles $\hat{\beta}_0$ et $\hat{\beta}_1$ de β_0 et β_1. Nous pouvons ainsi définir :
le couple de variables aléatoires (B_0, B_1) :

$$\begin{array}{rcl} (B_0, B_1) : \mathcal{P}^n & \longrightarrow & \mathbf{R}^2 \\ 1 \text{ expérience} & \longmapsto & (\hat{\beta}_0, \hat{\beta}_1) \end{array}$$

La variable aléatoire

$$\begin{array}{rcl} S^2 : \mathcal{P}^n & \longrightarrow & \mathbf{R} \\ 1 \text{ expérience} & \longmapsto & \hat{\sigma}^2 = \dfrac{\sum_{i=1}^n r_i^2}{n-2} \end{array}$$

Théorème 3.2.1. *(i) Si les postulats (i) et (ii) sont vérifiés :*

(a) les variables aléatoires B_0 et B_1 sont de "bons" estimateurs[2] de β_0 et β_1 de variances :

$$\begin{array}{rcl} \sigma^2_{B_0} = Var(B_0) & = & \sigma^2 \left(\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\ \sigma^2_{B_1} = Var(B_1) & = & \dfrac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \end{array}$$

(b) La variable aléatoire S^2 est un "bon" estimateur de σ^2 ;

(c) Les covariances entre S^2 et B_0 et entre S^2 et B_1 sont nulles.

(d) La covariance entre B_0 et B_1 n'est pas nulle :

$$Cov(B_0, B_1) = -\sigma^2 \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

(ii) Si nous ajoutons maintenant le postulat (iii) nous pouvons alors démontrer que :

(a) (B_0, B_1) suit une loi normale à deux dimensions ;

(b) $\dfrac{(n-2)S^2}{\sigma^2}$ suit une loi du Khi-2 à $\nu = (n-2)$ ddl ;

(c) S^2 et B_0 sont indépendantes comme S^2 et B_1 ;

Démonstration
Nous n'allons pas démontrer toutes les assertions du théorème. Notre objectif n'est ici que de montrer sur quelques exemples comment les résultats sont obtenus et où interviennent les postulats. Démontrons tout d'abord que l'estimateur B est un estimateur sans biais de β, de loi normale et calculons la matrice de variance covariance de cet estimateur.

La théorie des probabilités nous dit que si Y est un vecteur aléatoire de loi normale $\mathcal{N}(\mu, \Gamma)$ et si A est une matrice de dimension (p, n) alors AY est un vecteur aléatoire de dimension p de loi normale $\mathcal{N}(A\mu, A\Gamma^t A)$. Or, ici, nous avons

$$\hat{\beta} = ({}^tXX)^{-1\,t}Xy$$

Par suite, nous pouvons écrire

$$B = ({}^tXX)^{-1\,t}XY$$

[2] C'est-à-dire sans biais, convergeants et de variances minimales.

Posons alors $A = ({}^tXX)^{-1}{}^tX$, nous avons bien $B = AY$. Donc B suit une loi normale de dimension 2 et on a :

$$E(B) = AE(Y) = ({}^tXX)^{-1}{}^tXX\beta = \beta$$

Donc B est un estimateur sans biais.

Quant-à la matrice de variance, covariance de B il suffit d'écrire

$$Var(B) = AVar(Y){}^tA = (({}^tXX)^{-1}{}^tX)\sigma^2 I^t(({}^tXX)^{-1}{}^tX) = \sigma^2({}^tXX)^{-1}$$

Or

$${}^tXX = \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}$$

Donc

$$({}^tXX)^{-1} = \frac{1}{n\sum_i x_i^2 - (\sum_i)^2}\begin{pmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{pmatrix} = \frac{1}{SCE_x}\begin{pmatrix} \frac{\sum_i x_i^2}{n} & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

On en déduit immédiatement les valeurs des variances et covariance des estimateurs B_0 et B_1.

Pour $(n-2)S^2/\sigma^2$ c'est plus délicat. On pourraît penser en effet écrire

$$\frac{(n-2)S^2}{\sigma^2} = \frac{\sum_{i=1}^n \varepsilon_i^2}{\sigma^2}$$

et donc écrire cette variable comme une somme de n carrés de loi normales centrées réduites ce qui donne une loi du Khi-2 à n ddl. Mais ceci est faux. On sait en effet que $\sum_{i=1}^n r_i = 0$, alors que les variables aléatoires $(\varepsilon_i)_i$ sont elles indépendantes. Il faut en fait dans la démonstration décomposer proprement la variable aléatoire $(n-2)S^2/\sigma^2$ et utiliser ensuite le théorème de Cochran (qui est hors programme ici). Nous retiendrons cependant que cette variable aléatoire est en lien avec des sommes de carrés de loi normales centrées réduites, ce qui conduit à une loi du Khi-2. Quant-au degré de liberté il est de $(n-2)$ car nous avons estimé, à partie des données 2 paramètres β_0 et β_1. Nous reviendrons sur ce point dans la section analyse de la variance. □

3.3 Estimations par intervalle

Théorème 3.3.1. *Si les postulats sont vérifiés alors les intervalles de confiances de β_0, β_1 et σ^2 au seuil $(1-\alpha)$ sont :*

(i)

$$\sigma^2 \in \left[\frac{\sum_{i=1}^n r_i^2}{\chi_{1-\alpha/2}}; \frac{\sum_{i=1}^n r_i^2}{\chi_{\alpha/2}}\right]$$

où les valeurs de $\chi_{1-\alpha/2}$ et $\chi_{\alpha/2}$ sont lues dans le table du Khi-2 à $\nu = n-2$ ddl.

(ii)

$$\begin{aligned} \beta_1 &\in [\hat{\beta}_1 - t_{1-\alpha/2}\hat{\sigma}_{B_1}; \hat{\beta}_1 + t_{1-\alpha/2}\hat{\sigma}_{B_1}] \\ \beta_0 &\in [\hat{\beta}_0 - t_{1-\alpha/2}\hat{\sigma}_{B_0}; \hat{\beta}_0 + t_{1-\alpha/2}\hat{\sigma}_{B_0}] \end{aligned}$$

où

$$\begin{aligned} \hat{\sigma}^2_{B_1} &= \frac{\hat{\sigma}^2}{SCE_x} \\ \hat{\sigma}^2_{B_0} &= \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SCE_x}\right) \\ t_{1-\alpha/2} &\quad \text{est lu dans la table de Student à } n-2 \text{ ddl} \end{aligned}$$

Démonstration

Le théorème (3.2.1) nous dit que si les postulats sont vérifiés alors $(n-2)S^2/\sigma^2$ suit une loi du Khi-2 à $(\nu = n-2)$ ddl. Nous avons donc :

$$P\left(\chi_{\alpha/2} < \frac{(n-2)S^2}{\sigma^2} < \chi_{1-\alpha/2}\right) = 1-\alpha$$

soit

$$P\left(\frac{(n-2)S^2}{\chi_{1-\alpha/2}} < \sigma^2 < \frac{(n-2)S^2}{\chi_{\alpha/2}}\right) = 1-\alpha$$

d'où l'assertion pour la variance.

Pour le paramètre β_1, nous savons, toujours par application du théorème 3.2.1 que la variable aléatoire $\frac{B_1 - \beta_1}{\sqrt{\frac{\sigma^2}{SCE_x}}}$ suit une loi normale centrée réduite. Le problème est que l'on ne connaît pas ici la variance σ^2. Nous allons donc remplacer celle-ci par son estimateur. Ceci nous conduit à la variable aléatoire suivante :

$$T = \frac{(B_1 - \beta_1)}{\sqrt{\frac{S^2}{SCE_x}}} = \frac{(B_1 - \beta_1)/\sqrt{\frac{\sigma^2}{SCE_x}}}{\sqrt{\frac{(n-2)S^2/\sigma^2}{n-2}}} = \frac{A}{B}$$

où :
- A est une variable aléatoire de loi normale centrée et réduite $\mathcal{N}(0,1)$;
- B est la racine carrée d'une variable aléatoire suivant une loi du Khi-2 à $(n-2)$ ddl divisée par son degré de liberté ;
- A et B sont indépendantes.

Par suite (cf chapitre sur la théorie de l'échantillonnage) T suit une loi de Student à $\nu = (n-2)$ ddl. Nous en déduisons alors immédiatement que :

$$P(-t_{1-\alpha/2} < T < t_{1-\alpha/2}) = 1 - \alpha$$

où encore

$$P\left(B_1 - t_{1-\alpha/2}\sqrt{\frac{S^2}{SCE_x}} < \beta_1 < B_1 + t_{1-\alpha/2}\sqrt{\frac{S^2}{SCE_x}}\right) = 1 - \alpha$$

d'où le résultat.

On démontre de la même façon le résultat pour β_0. □

Remarque 3.3.2. Nous donnerons les résultats numérique sur notre exemple (1.0.1) après le paragraphe suivant. Nous allons en effet dans celui-ci calculer entre autre la Somme des Carrés Ecarts des Résidus : $SCE_R = \sum_{i=1}^{n} r_i^2$. Ceci sans calculer tous les résidus.

4 Analyse de la variance

4.1 Introduction

Lorsque la pente de la droite β_1 est nulle cela signifie que la valeur de la variable aléatoire Y est en moyenne la même pour tout x, c'est donc dire que le modèle linéaire n'apporte aucune information sur Y. Nous sommes donc ramener à tester les hypothèses nulle et alternative suivantes :
- $H_0 : \beta_1 = 0$;
- $H_1 : \beta_1 \neq 0$.

Il s'agit ici d'un test bilatéral. L'objet de l'analyse de la variance pour une régression linéaire simple est de réaliser ce test statistique.

Remarque 4.1.1. Attention accepter H_0 signifie que l'on ne peut prédire Y en fonction de x par un modèle linéaire. cela ne signifie pas qu'il n'y a pas de modèle permettant de prédire Y en fonction de x.

L'approche, via l'analyse de la variance, est dans sa conception différente. Il s'agira de savoir en fait si la part de variabilité expliquée par le modèle linéaire (et donc par la pente de la droite) est significative par rapport à la variance résiduelle

Partons des relations suivantes vraies pour tout i :

$$\begin{array}{rcccl} y_i & = & \hat{\beta}_0 + \hat{\beta}_1 x_i & + & r_i \\ & = & \hat{y}_i & + & r_i \\ & = & \hat{y}_i & + & (y_i - \hat{y}_i) \end{array}$$

ou encore

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \quad \forall i$$

Elevons au carré et sommons :

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

en effet la somme des doubles produits est nulle :

$$\begin{array}{rcl}\sum_{i=1}^n(\hat{y}_i-\bar{y})(y_i-\hat{y}_i) & = & \sum_{i=1}^n \hat{\beta}_1(x_i-\bar{x})(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)\\ & = & \sum_{i=1}^n \hat{\beta}_1(x_iy_i-\bar{x}y_i)-\hat{\beta}_0\hat{\beta}_1\sum_{i=1}^n(x_i-\bar{x})-\hat{\beta}_1^2\sum_{i=1}^n(x_i^2-x_i\bar{x})\\ & = & \hat{\beta}_1(\sum_{i=1}^n x_iy_i-n\bar{x}\bar{y})-\hat{\beta}_1^2(\sum_{i=1}^n x_i^2-n\bar{x}^2)\\ & = & \frac{SPE}{SCE_x}SPE-\left(\frac{SPE}{SCE_x}\right)^2SCE_x=0\end{array}$$

On en déduit ce que nous appellerons **l'équation l'analyse de la variance** :

$$\begin{array}{rcccc}\sum_{i=1}^n(y_i-\bar{y})^2 & = & \sum_{i=1}^n(\hat{y}_i-\bar{y})^2 & + & \sum_{i=1}^n(y_i-\hat{y}_i)^2\\ & = & \hat{\beta}_1^2\sum_{i=1}^n(x_i-\bar{x})^2 & + & \sum_{i=1}^n r_i^2\\ SCE_T & = & SCE_{y(x)} & + & SCE_R\\ n-1 & = & 1 & + & n-2\end{array}$$

Nous avons ainsi décomposé la somme des carrés des écarts des données y_i appelé Somme est Carrés des Ecarts Totale en la somme de deux termes. Le premier représente la somme des carrés des écarts expliquée par la droite de régression et le deuxième la somme des carrés des écarts résiduelle (cf figure 1.5).

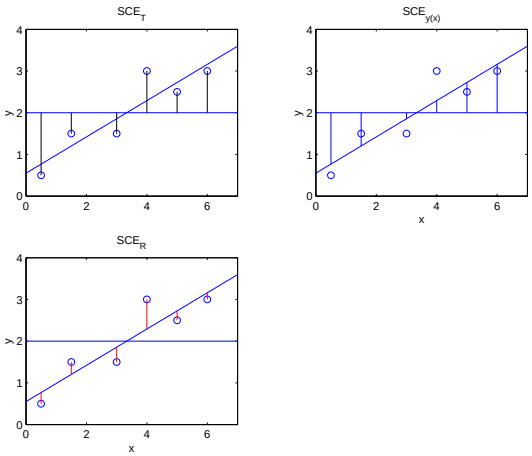

FIG. 1.5 – Equation d'analyse de la variance.

La question est maintenant de savoir si l'on peut juger que la quantité $SCE_{y(x)}$ est grande devant la quantité SCE_R. Nous avons pour cela le théorèrme suivant :

Théorème 4.1.2. *Si les postulats sont vérifiés alors :*

(i) une "bonne" estimation de σ^2 est

$$CM_R=\frac{SCE_R}{n-2}=\hat{\sigma}^2$$

(ii) si l'hypothèse nulle H_0 est vraie alors une autre estimation de σ^2, indépendante de la précédante est :

$$CM_{y(x)}=\frac{SCE_{y(x)}}{1}$$

(iii) si l'hypothèse nulle H_0 est vraie alors la statistique

$$\begin{array}{rcl}F:\mathcal{P}^n & \longrightarrow & \mathbf{R}\\ 1\text{ expérience} & \longmapsto & \frac{CM_{y(x)}}{CM_R}\end{array}$$

suit une loi de Fisher à $(1,n-2)$ ddl.

On présente en pratique les résultats sous la forme d'un **tableau d'analyse de la variance** (cf table 1.3).

Source de variation	SCE	ddl	carrés moyens	Statistique F	
(1) Totale	$\sum_i (y_i - \bar{y})^2$ $= \sum_i y_i^2 - c_y$	$n-1$			
(2) Régression	$\sum_i (\hat{y}_i - \bar{y})^2$ $= \hat{\beta}_1^2 (\sum_i x_i^2 - c_x)$	1	$CM_{y(x)}$	$F_{obs} = \frac{CM_{y(x)}}{CM_R}$	F_{crit}
Résiduelle	(1)-(2)	$n-2$	$CM_R = \frac{SCE_R}{n-2}$		

TAB. 1.3 – Tableau d'analyse de la variance

4.2 Tableau d'analyse de la variance

Remarque 4.2.1. (i) dire que F_{obs} est très suppérieure à 1 est équivalent à dire que la variabilité expliquée par la régression est très supérieure à la variabilité résiduelle.

(ii) Attention ici le test est un test unilatéral, il n'y a en effet qu'une seule zone de rejet, celle ou $F > F_{crit}$;

(iii) On note R^2 la quantité, appelé coefficient de détermination, $\frac{SCE_{y(x)}}{SCE_T}$. Une autre façon d'exprimer que le modèle linéaire apporte de l'information est de dire que R^2 est différent de 0. Il existe un test statistique pour cela. Ce test est équivalent à celui effectué dans l'analyse de la variance.

Exemple 4.2.2. Reprenons notre exemple (1.0.1). En utilisant les calculs préliminaires du tableau (1.2) nous obtenons le tableau l'analyse de la variance suivant :

Source de variation	SCE	ddl	carrés moyens	Statistique F	
(1) Totale	$253.31 - 90.1^2/33 = 7.310$	32			
(2) Régression	$0.025^2 \times (69625 - 1445^2/33) \simeq 3.941$	1	3.941	36.266	$\simeq 4.171$
Résiduelle	3.369	31	0.109		

Conclusion : $F_{obs} > F_{crit}$ donc on rejette l'hypothèse nulle $H_0 : \beta_1 = 0$ au risque $\alpha = 0.05$. Le modèle linéaire apporte de l'information sur Y.

Nous pouvons maintenant facilement calculer les intervalles de confiance à 95% de la variance et des paramètres β_0 et β_1.

- $\sigma^2 \in \left[\frac{\sum_{i=1}^n r_i^2}{\chi_{1-\alpha/2}}; \frac{\sum_{i=1}^n r_i^2}{\chi_{\alpha/2}}\right] = \left[\frac{3.369}{48,232}; \frac{3.369}{17,539}\right] = [0.070; 0.192]$ au niveau 0.95 ;

$$\begin{aligned} \hat{\sigma}^2_{B_1} &= \frac{\hat{\sigma}^2}{SCE_x} = \frac{0.109}{69625 - 1445^2/33} = 0.0000171 \\ \hat{\sigma}^2_{B_0} &= \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{SCE_x}\right) = 0.109 \left(\frac{1}{33} + \frac{43.79^2}{6351.51}\right) = 0.0361 \end{aligned}$$

$$\begin{aligned} \beta_1 &\in [b_1 - t_{0.975}\hat{\sigma}_{b_1}; b_1 + t_{0.975}\hat{\sigma}_{b_1}] = [0.0165; 0.0333] \\ \beta_0 &\in [b_0 - t_{0.975}\hat{\sigma}_{b_0}; b_0 + t_{0.975}\hat{\sigma}_{b_0}] = [1.252; 2.027] \end{aligned}$$

au niveau 0.95.

4.3 Interprétation géométrique

Si nous notons $x_{.1}$ et $x_{.2}$ les deux vecteurs colonnes de la matrice X et si P est le plan engendré par ces deux vecteurs, nous avons :

$$P = \{v = X\beta; \beta \in \mathbf{R}^2\}$$

Par suite minimiser $\frac{1}{2}\|X\beta - y\|^2$, c'est trouver le vecteur du plan $(O, x_{.1}, x_{.2}$ le plus proche de y. La solution est la projection orthogonale de y sur P. Cette projection est le vecteur $X\hat{\beta}$. Nous avons alors :

$$X\hat{\beta} = X({}^tXX)^{-1t}Xy = Hy$$

$H = X({}^tXX)^{-1t}X$ est donc l'opérateur, dans $\mathbf{R}^n$ de la projection orthogonale sur le plan P (cf. la figure (1.6))

Le vecteur des résidus est alors $r = y - Hy = (I - H)y$ et ce vecteur est orthogonal au vecteur $Hy = X\hat{\beta}$. Par suite le théorème de Pythagore donne

$$\begin{aligned} \|y\|^2 &= \|X\hat{\beta}\|^2 + \|r\|^2 \\ &= \|\hat{y}\|^2 + \|r\|^2 \end{aligned}$$

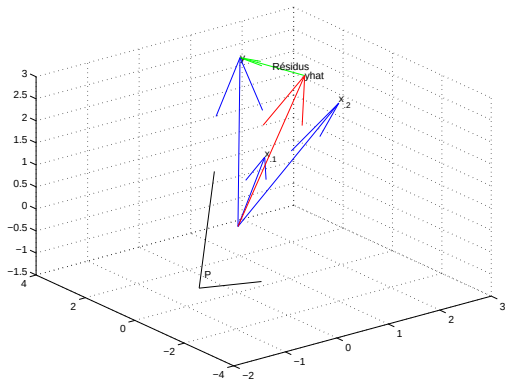

FIG. 1.6 – Interprétation géométrique de l'équation d'analyse de la variance

Nous en déduisons alors

$$\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{n} \hat{y}_i^2 - n\bar{y}^2 + \sum_{i=1}^{n} r_i^2$$

C'est-à-dire l'équation d'analyse de la variance.

La dimension le l'espace des résidus, qui est l'orthogonal de P, est $n-2$. C'est le degré de liberté associé aux résidus.

5 Test sur les pamarètres β_0 et β_1

5.1 Test sur β_1

Dans la démonstration du théorème (3.3.1) nous avons vu par exemple que la variable aléatoite

$$T_1 = \frac{(B_1 - \beta_1)}{\sqrt{\frac{S^2}{SCE_x}}} = \frac{(B_1 - \beta_1)/\sqrt{\frac{\sigma^2}{SCE_x}}}{\sqrt{\frac{(n-2)S^2/\sigma^2}{n-2}}} = \frac{A}{B}$$

suivait une loi de Student à $\nu = (n-2)$ ddl. Nous pouvons donc aisément réaliser un test d'égalité à une constante de ce paramètre. Notons T_{1obs} la valeur observée de cette variable aléatoire, c'est-à-dire la valeur :

$$T_{1obs} = \frac{\hat{\beta_1} - \beta_1}{\hat{\sigma}_{B1}}$$

. Nous avons comme règle de décision :

- Pour le test bilatéral $H_0 : \beta_1 = \beta_1^0$ contre $H_1 : \beta_1 \neq \beta_1^0$:
  - Si $\mid T_{1obs} \mid < t_{1-\alpha/2}$ alors on accepte l'hypothèse nulle H_0 d'égalité de β_1 à β_1^0 au risque α ;
  - Si $\mid T_{1obs} \mid > t_{1-\alpha/2}$ alors on rejette l'hypothèse nulle H_0 d'égalité de β_1 à β_1^0 au risque α ;
- Pour le test unilatéral à droite $H_0 : \beta_1 = \beta_1^0$ contre $H_1 : \beta_1 > \beta_1^0$:
  - Si $T_{1obs} < t_{1-\alpha}$ alors on accepte l'hypothèse nulle H_0 d'égalité de β_1 à β_1^0 au risque α ;
  - Si $T_{1obs} > t_{1-\alpha}$ alors on rejette l'hypothèse nulle H_0 d'égalité de β_1 à β_1^0 au risque α ;
- Pour le test unilatéral à gauche $H_0 : \beta_1 = \beta_1^0$ contre $H_1 : \beta_1 < \beta_1^0$:
  - Si $T_{1obs} > t_{\alpha}$ alors on accepte l'hypothèse nulle H_0 d'égalité de β_1 à β_1^0 au risque α ;
  - Si $T_{1obs} < t_{\alpha}$ alors on rejette l'hypothèse nulle H_0 d'égalité de β_1 à β_1^0 au risque α ;

  où les valeurs critiques $t_\alpha, t_{1-\alpha}$ et $t_{1-\alpha/2}$ sont lues dans la table de Student à $\nu = (n-2)$ ddl.

Remarque 5.1.1. Lorsque $\beta_1^0 = 0$ nous avons $T_{1obs} = \dfrac{\hat{\beta}_1}{\hat{\sigma}_{B_1}} = \dfrac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/SCE_x}}$. Par suite $T_{1obs}^2 = \dfrac{\hat{\beta}_1^2 SCE_x}{\hat{\sigma}^2}$ est égale à la valeur de F_{obs} trouvée dans le tableau d'analyse de la variance. Or le carré d'une loi de Student à ν ddl est une loi de Fisher à $(1, \nu)$ ddl et l'inégalité $\mid T_{1obs} \mid < t_{1-\alpha/2}$ est équivalente à l'inégalité $F_{obs} < F_{crit}$. En conséquence nous réalisons exactement le même test d'égalité de la pente à 0 ; mais attention le test via Student est bilatéral, il est unilatéral via l'analyse de la variance, cela provient du carré.

5.2 Test sur β_0

De façon identique nous pouvons faire un test d'égalité à une constante pour le paramètre β_0. Posons

$$T_{0obs} = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{B0}}$$

. Nous avons alors comme règle de décision :

- Pour le test bilatéral $H_0 : \beta_0 = \beta_0^0$ contre $H_1 : \beta_0 \neq \beta_0^0$:
  - Si $\mid T_{0obs} \mid < t_{1-\alpha/2}$ alors on accepte l'hypothèse nulle H_0 d'égalité de β_0 à β_0^0 au risque α ;
  - Si $\mid T_{0obs} \mid > t_{1-\alpha/2}$ alors on rejette l'hypothèse nulle H_0 d'égalité de β_0 à β_0^0 au risque α ;
- Pour le test unilatéral à droite $H_0 : \beta_0 = \beta_0^0$ contre $H_1 : \beta_0 > \beta_0^0$:
  - Si $T_{0obs} < t_{1-\alpha}$ alors on accepte l'hypothèse nulle H_0 d'égalité de β_0 à β_0^0 au risque α ;
  - Si $T_{0obs} > t_{1-\alpha}$ alors on rejette l'hypothèse nulle H_0 d'égalité de β_0 à β_0^0 au risque α ;
- Pour le test unilatéral à gauche $H_0 : \beta_0 = \beta_0^0$ contre $H_1 : \beta_0 < \beta_0^0$:
  - Si $T_{0obs} > t_{\alpha}$ alors on accepte l'hypothèse nulle H_0 d'égalité de β_0 à β_0^0 au risque α ;
  - Si $T_{0obs} < t_{\alpha}$ alors on rejette l'hypothèse nulle H_0 d'égalité de β_0 à β_0^0 au risque α ;

où les valeurs critiques $t_\alpha, t_{1-\alpha}$ et $t_{1-\alpha/2}$ sont lues dans la table de Student à $\nu = (n-2)$ ddl.

6 Prévision

6.1 Introduction

L'un des objectifs lorsque l'on effectue une régression linéaire est ensuite de pouvoir estimer la valeur de la variable Y pour une valeur de x donnée, disons x_0. Il s'agit donc bien de prédire. Pour ceci il est nécessaire, bien évidemment, que le modèle linéaire soit exact, ce que nous supposerons dans cette section.

Notre modèle linéaire et les postulats nous disent que la valeur prédite par x_0 est en lien avec la variable aléatoire $Y_0 = \beta_0 + \beta_1 x_0 + \varepsilon$ avec ε de loi normale $\mathcal{N}(0, \sigma^2)$. Mais en pratique, nous n'avons accès qu'à des estimations de ces paramètres. Par conséquent la valeur prédite par le modèle $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ a deux raisons d'être erronée :

(i) la variable aléatoire Y_0 a une variance non nulle. Pour x_0 fixé les valeurs que peut prendre cette variable aléatoire fluctuent avec une variance de σ^2 autour le la moyenne $E(Y_0) = \beta_0 + \beta_1 x_0$.

(ii) l'estimation de $E(Y_0) = \beta_0 + \beta_1 x_0$, qui sera donné par $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$ est elle même sujette aux erreurs d'échantillonnage qui affecte $\hat{\beta}_0$ et $\hat{\beta}_1$.

Nous allons en fait construire deux intervalles de confiance, l'un pour la prévision, l'autre pour la moyenne de la prévision, c'est-à-dire pour $E(Y_0)$.

6.2 Intervalle de confiance de la moyenne d'une prévision

Théorème 6.2.1. *Si les postulats (i,ii,iii) sont vérifiés alors :*

(i) $\bar{Y}_0 = B_0 + B_1 x_0$ est une variable aléatoire de loi normale

$$\mathcal{N}\left(\beta_0 + \beta_1 x_0, \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)\right)$$

(ii)

$$T = \frac{\bar{Y}_0 - (\beta_0 + \beta_1 x_0)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}}$$

suit une loi de Student à $\nu = (n-2)$ ddl.

Remarquons une nouvelle fois ce qu'est la variable aléatoire $\bar{Y}_0$. A partir de n unités expérimentales, nous avons accès à n couples de données $(x_i, y_i)_i$. A partir de ces données nous pouvons calculer les estimations $\hat{\beta}_0$ et $\hat{\beta}_1$ et ensuite calculer la valeur de $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

$$\begin{array}{rcl} \bar{Y}_0 : \mathcal{P}^n & \longrightarrow & \mathbf{R} \\ \text{1 expérience} & \longmapsto & \hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \end{array}$$

Démontrons maintenant le résultat. Dans l'expression $\bar{Y}_0 = B_0 + B_1 x_0$, B_0 et B_1 sont des variables aléatoire et x_0 est un réel. $\bar{Y}_0$ est donc une combinaisons linéaire de loi normale, par suite $\bar{Y}_0$ suit une loi normale. Calculons sont espérance mathématique et sa variance.

$$E(\bar{Y}_0) = E(B_0) + x_0 E(B_1) = \beta_0 + x_0\beta_1$$

$$\begin{aligned} Var(\bar{Y}_0) &= Var(B_0) + x_0^2 Var(B_1) + 2x_0 Cov(B_0, B_1) \\ &= \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{SCE_x} + \frac{x_0^2}{SCE_x} - \frac{2x_0\bar{x}}{SCE_x}\right) \\ &= \sigma^2\left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{SCE_x}\right) \end{aligned}$$

D'où le premier point. Comme d'habitude en pratique on ne connaît pas la valeur de la variance σ^2 et en remplaçant celle-ci par son estimation $\hat{\sigma}^2$, nous obtenons une loi de Student.

A partir de ce théorème on en déduit immédiatement l'intervalle de confiance de la moyenne de la prévision, c'est-à-dire de $\beta_0 + \beta_1 x_0$.

Théorème 6.2.2. *Si les postulats sont vérifiés alors l'intervalle de confiance au niveau* $1-\alpha$ *de la moyenne de la prévision en* x_0 *est :*

$$\beta_0 + \beta_1 x_0 \in \left[\hat{\beta}_0 + \hat{\beta}_1 x_0 - t_{1-\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{SCE_x}}; \hat{\beta}_0 + \hat{\beta}_1 x_0 + t_{1-\alpha/2}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{SCE_x}}\right]$$

6.3 Intervalle de confiance d'une prévision

Nous nous intéressons ici, non pas à la valeur moyenne, mais à une valeur quelconque de la variable aléatoire Y_0 pour x fixé à x_0. La différence avec ce qui a été fait précédemment réside dans le fait qu'il faut rajouter la source de variation résiduelle, c'est-à-dire que nous travaillons avec la variable aléatoire $Y_0 = \bar{Y}_0 + \varepsilon$. Nous obtenons le même type de formule exepté pour la variance ou

$$Var(Y_0) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)$$

d'où le théorème suivant :

Théorème 6.3.1. *Si les postulats sont vérifiés alors l'intervalle de confiance au niveau* $1-\alpha$ *de la prévision en* x_0 *est :*

$$\left[\hat{\beta}_0 + \hat{\beta}_1 x_0 - t_{1-\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{SCE_x}}; \hat{\beta}_0 + \hat{\beta}_1 x_0 + t_{1-\alpha/2}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{SCE_x}}\right]$$

Exemple 6.3.2. On souhaite avoir pour notre exemple (1.0.1) les intervalles de confiances pour la prévision d'une observation et pour la moyenne de la prévision pour un homme agé de 60 ans.

L'intervalle de confiance à 95% de la prévision est

$$\left[\hat{\beta}_0 + \hat{\beta}_1 x_0 - t_{0.975}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{SCE_x}}; \hat{\beta}_0 + \hat{\beta}_1 x_0 + t_{0.975}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{SCE_x}}\right]$$

$\hat{\beta}_0 + \hat{\beta}_1 x_0 = 1.64 + 0.025 \times 60 = 3.14$

$t_{0.975}\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{SCE_x}} = 2.04 \times \sqrt{0.109}\sqrt{1 + \frac{1}{33} + \frac{(60-43.79)^2}{6351.51}} = 0.70$

$[2.44; 3.83]$

L'intervalle de confiance à 95% de la prévision de la moyenne est

$$\left[\hat{\beta}_0 + \hat{\beta}_1 x_0 - t_{0.975}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{SCE_x}}; \hat{\beta}_0 + \hat{\beta}_1 x_0 + t_{0.975}\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x_0-\bar{x})^2}{SCE_x}}\right]$$

$\hat{\beta}_0 + \hat{\beta}_1 x_0 = 1.64 + 0.025 \times 60 = 3.14$

$t_{0.975}\hat{\sigma}\sqrt{\frac{34}{n} + \frac{(x_0-\bar{x})^2}{SCE_x}} = 2.04 \times \sqrt{0.109}\sqrt{\frac{34}{33} + \frac{(60-43.79)^2}{6351.51}} = 0.18$

$[2.95; 3.31]$

Nous pouvons calculer en pratique les intervalles de confiance pour toute valeur de x et les visualiser. Sur le graphique (1.7) est visualisé : le nuage de points, la droite de régression, les limites des intervalles de confiance pour une prévision d'une observation (en trait plein), les limites des intervalles de confiances pour la valeur moyenne d'une prévision (trait en pointillé).

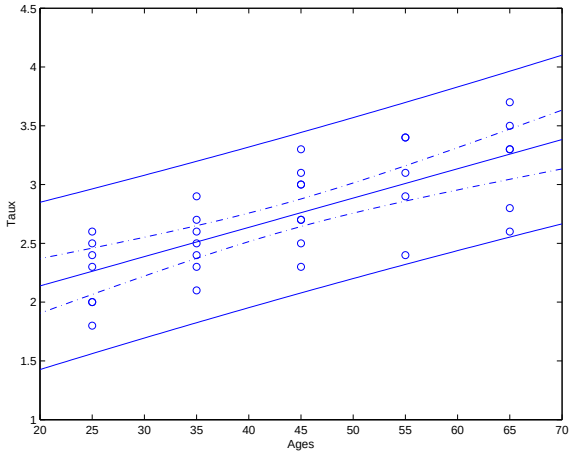

FIG. 1.7 – Intervalle de confiance

7 Test de la linéarité de la régression

7.1 Introduction

L'objectif est de tester si le modèle linéaire est adéquat. Il semblerait logique de réaliser ce test au tout début de l'analyse. Malheureusement, pour ceci nous avons besoin d'avoir des mesures répétées de Y pour chaque valeur de x (on parle alors de données groupées), ce qui est le cas de notre exemple, mais ceci est rarement vrai en pratique. Rajoutons tout d'abord un indice pour prendre en compte la structure des données groupées :

$$Y_{ij} = \beta_0 + \beta_1 x_i + \varepsilon_{ij}$$

où
- l'indice i varie de 1 à p ;
- l'indice j varie de 1 à n_i ;
- $\sum_{i=1}^{n} n_i = n$;

Nous allons dans ce cas pouvoir décomposer l'écart $y_{ij} - \bar{y}$ en 3 termes (cf. figure (1.8) :

$$y_{ij} - \bar{y} = (\hat{y}_{ij} - \bar{y}) + (\bar{y}_{i.} - \hat{y}_{ij}) + (y_{ij} - \bar{y}_{i.}) \tag{1.4}$$

où

(i) $\bar{y} = \dfrac{1}{n} \sum_{i=1}^{p} \sum_{i=1}^{n_i} y_{ij}$;

(ii) $\bar{y}_{i.} = \dfrac{1}{n_i} \sum_{i=1}^{n_i} y_{ij}$;

(iii) $\hat{y}_{ij} = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Intuitivement on peut voir que si les valeurs des $(\bar{y}_{i.} - \hat{y}_{ij})$ sont faibles, c'est que le modèle linéaire est adapté aux données. Nous allons donc comparer $\sum_{ij}(\bar{y}_{i.} - \hat{y}_{ij})^2$ avec la Somme des Carrés des Ecarts Résiduelles $\sum_{ij}(y_{ij} - \bar{y}_{i.})^2$. Ceci va être fait grâce à l'analyse de la variance.

Remarque 7.1.1. Si $n_i = 1$ alors $y_{i1} = \bar{y}_{i.}$ et donc le terme $\sum_{ij}(y_{ij} - \bar{y}_{i.})^2$ sera nulle. Nous ne pourrons donc pas effectuer de test dans ce cas.

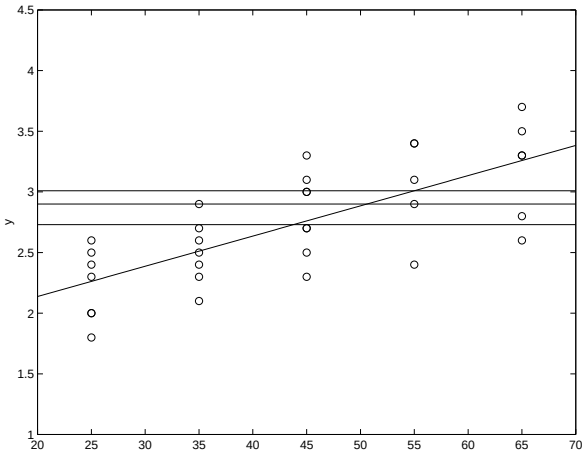

FIG. 1.8 – Test de la linéarité

7.2 Analyse de la variance

L'hypothèse nulle que va nous permettre de tester l'analyse de la variance est : H_0 : le modèle linéaire est correct.

Elevons au carré et sommons sur les indices i et j les équations (1.4). La sommes des doubles produits est, ici aussi, nulle et nous obtenons ainsi l'équation d'analyse de la variance suivante :

$$\begin{array}{lcllcllcl} \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 & = & \sum_i \sum_j (\hat{y}_{ij} - \bar{y})^2 & + & \sum_i \sum_j (\bar{y}_{i.} - \hat{y}_{ij})^2 & + & \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2 \\ SCE_T & = & SCE_{y(x)} & + & SCE_{ec} & + & SCE_R \\ n-1 & = & 1 & + & p-2 & + & n-p \end{array}$$

On démontre alors le théorème suivant :

Théorème 7.2.1. *Si les postulats sont vraies alors :*

(i) $CM_R = \dfrac{SCE_R}{n-p}$ *est une "bonne" estimation de la variance* σ^2 *et la variable aléatoire en lien avec* SCE_R/σ^2 *suit une loi du Khi-2 à* $(n-p)$ *ddl ;*

(ii) Si l'hypothèse nulle est vraie alors $CM_{ec} = \dfrac{SCE_{ec}}{p-2}$ *est une "bonne" estimation de la variance* σ^2*, la variable aléatoire en lien avec* SCE_{ec}/σ^2 *suit une loi du Khi-2 à* $(p-2)$ *ddl et cette variable aléatoire est indépendante de* SCE_R/σ^2*;*

(iii) Si l'hypothèse nulle est vraie alors $\dfrac{CM_{ec}}{CM_R}$ *est en lien avec une variable aléatoire qui suit une loi de Fisher à* $(p-2),(n-p)$ *ddl.*

En pratique on présentera les résultats sous la forme d'un tableau d'analyse de la variance. Le terme SCE_{ec} est en pratique long à calculer. Aussi on préfère obtenir cette valeur grâce à l'égalité $SCE_G = \sum_i \sum_j (\bar{y}_{i.} - \bar{y})^2 = SCE_{y(x)} + SCE_{ec}$. Une bonne façon de procéder est, lorsque l'on effectue les calculs à la main, d'effectuer les calculs préliminaire suivants :

	Variable expliquée 1 ⋯ j ⋯				Effectifs	Totaux	Moyennes	Produits	Produits		
x_1	y_{11}	⋯	y_{1j}	⋯	n_1	$Y_{1.}$	$\bar{y}_{1.}$	$Y_{1.}\bar{y}_{1.}$	n_1x_1	$n_1x_1^2$	$n_1x_1\bar{y}_{1.}$
⋮	⋮		⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_i	y_{i1}	⋯	y_{ij}	⋯	n_i	$Y_{i.}$	$\bar{y}_{i.}$	$Y_{i.}\bar{y}_{i.}$	n_ix_i	$n_ix_i^2$	$n_ix_i\bar{y}_{i.}$
⋮	⋮		⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮
x_p	y_{p1}	⋯	y_{pj}	⋯	n_p	$Y_{p.}$	$\bar{y}_{p.}$	$Y_{p.}\bar{y}_{p.}$	n_px_p	$n_px_p^2$	$n_px_p\bar{y}_{p.}$
Totaux					n	$Y_{..}$		$\sum_{i=1}^p Y_{i.}\bar{y}_{i.}$	$X_{..}$	$\sum_{i=1}^p n_ix_i^2$	$\sum_{i=1}^p n_ix_i\bar{y}_{i.}$
Moyenne							$\bar{y}_{..}$				

ainsi que

$$\sum_{i=1}^{p}\sum_{j=1}^{n_i} y_{ij}^2 \quad ; \quad c_y = Y_{..}\bar{y}_{..} \quad ; \quad c_x = X_{..}\bar{x}_{..}$$

On en déduit alors très rapidement le tableau d'analyse de la variance suivant :

Source de variation	SCE	ddl	Carrés Moyens	Statistiques F_{obs}	Statistiques F_{crit}
(1) Totale	$\sum_i\sum_j(y_{ij}-\bar{y}_{..})^2 = \sum_i\sum_j y_{ij}^2 - c_y$	$n-1$			
(2) Entre groupes	$\sum_i n_i(\bar{y}_{i.}-\bar{y}_{..})^2 = \sum_i Y_{i.}\bar{y}_{i.} - c_y$	$p-1$	$CM_A = \dfrac{SCE_A}{p-1}$	$F_{Gobs} = \dfrac{CM_A}{CM_e}$	$F_{p-1,n-p}$
(a) Régression $y(x)$	$b_1^2(\sum_i n_i x_i^2 - c_x)$	1	$CM_{y(x)} = \dfrac{SCE_{y(x)}}{1}$	$F_{y(x)obs} = \dfrac{CM_{y(x)}}{CM_e}$	$F_{1,n-p}$
(b) Ecarts/droites	(2) − (a)	$p-2$	$CM_{ec} = \dfrac{SCE_{ec}}{p-2}$	$F_{ecobs} = \dfrac{CM_{ec}}{CM_e}$	$F_{n-2,n-p}$
(3) Résiduelle	(1) − (2)	$n-p$	$CM_e = \dfrac{SCE_e}{n-p}$		

Exemple 7.2.2. Toujours sur le même exemple (1.0.1) nous obtenons

	Variable explicative 25	35	45	55	65	Totaux	Moyenne
Variable expliquée	1.8	2.6	2.7	3.1	3.7		
	2.3	2.9	3.0	2.9	2.8		
	2.0	2.3	3.1	3.4	3.3		
	2.4	2.4	2.3	2.4	3.5		
	2.0	2.1	2.5	3.4	3.3		
	2.5	2.5	3.0		2.6		
	2.6	2.7	3.3				
			2.7				
Effectifs	7	7	8	5	6	33	
Totaux	15.6	17.5	22.6	15.2	19.2	90.1	
Moyennes	2.229	2.5	2.825	3.04	3.2		2.730
Produits	34.766	43.75	63.845	46.208	61.44	250.009	
Produits	175	245	360	275	390	1445	
	4375	8575	16200	15125	25350	69625	
	390	612.5	1017	836	1248	4103.5	

Source de variation	SCE	ddl	Carrés Moyens	Statistiques F_{obs}	Statistiques F_{crit}
(1) Totale	7.31	32			
(2) Entre groupes	4.01	4	1.00	8.50	2.714
(a) Régression $y(x)$	3.94	1	3.94	33.43	4.196
(b) Ecarts/droites	0.067	3	0.022	0.19	2.947
(3) Résiduelle	3.30	28	0.12		

Conclusion : $F_{ecobs} = 0.19 < 2.947 = F_{eccrit}$ par suite on accepte l'hypothèse nulle H_0 le modèle linéaire est correct au risque α de 0.05.

8 Vérification des postulats

8.1 Les dangers de la régression

Observons les 5 graphiques (1.9) :

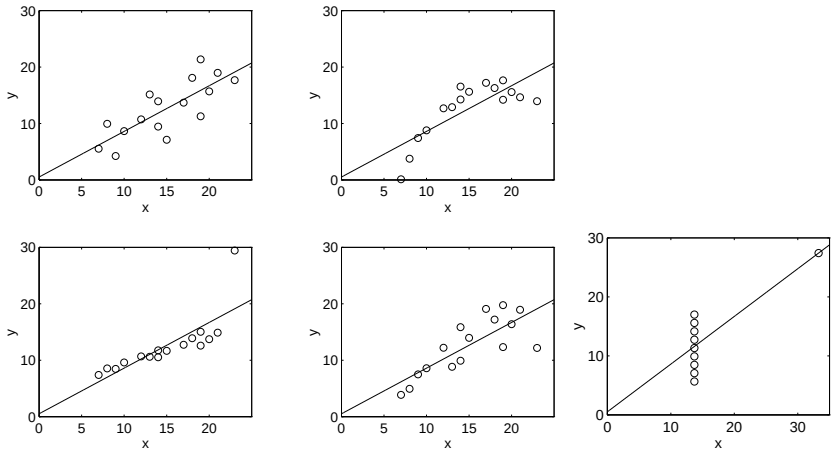

FIG. 1.9 – Les dangers de la régression : Données

Il est clair que c'est 5 graphiques 1.9 correspondent à des cas très différents :
Cas (a) : pas de problème a priori ;
Cas (b) : Il semble a priori qu'un modèle parabolique soit mieux adapté qu'un modèle linéaire ;
Cas (c) : il y a un point aberrant ;
Cas(d) : la dispersion en y semble augmenter en fonction de x ;
Cas(e) : il y a une forte influence d'un point à la définition de la droite de régression.
Pourtant pour ces 5 cas nous avons :

- $n = 16$;
- $\bar{x} = 14.938$;
- $\bar{y} = 12.600$;
- $SPE = 290.27$;
- $SCE_x = 358.9375$;
- $\hat{\beta}_0 = 0.52$;
- $\hat{\beta}_1 = 0.809$;
- $\hat{\sigma} = 3.226$;

Nous aurons donc pour ces 5 cas la même droite de régression, les mêmes résultats statistique, et les mêmes intervalles de confiance pour les prévisions (cf graphiques (1.10)) !!!

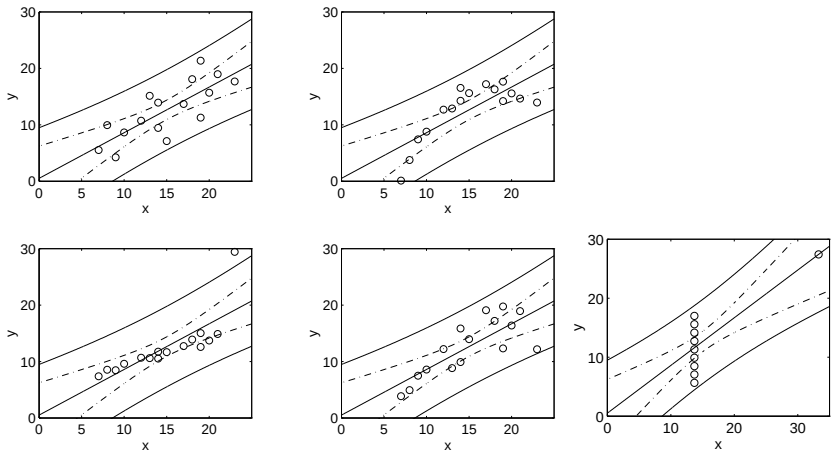

FIG. 1.10 – Les dangers de la régression : Prévisions

Ceci provient d'une part de l'existence de points aberrants et d'autres part du non respect des postulats. L'objet de cette section est de détecter ces problèmes. La grande majorité des méthodes sont basées sur l'étude des résidus.

8.2 Analyse des résidus

Rappelons le modèle et les postulats :

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \ldots, n$$

Postulats :

(i) les variables aléatoires ε_i sont d'espérance nulle et de variance constante : $E(\varepsilon_i) = 0$ et $Var(\varepsilon_i) = \sigma^2$;

(ii) les variables aléatoires ε_i sont indépendantes ;

(iii) les variables aléatoires ε_i sont de loi normale.

Calculs des résidus

Les résidus sont donnés par la formule : $r_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$. Ainsi toujours sur notre exemple nous avons si nous prenons le couple de données $(x_1, y_1) = (25, 1.8)$:

$$r_1 = 1.8 - 1.64 - 0.025 \times 25 = -0.465$$

Pour l'ensemble des résidus nous obtenons le tableau 1.4

Ages	Taux	Résidus
25	1.8	-0.46231
25	2.3	0.037691
25	2	-0.26231
25	2.4	0.13769
25	2	-0.26231
25	2.5	0.23769
25	2.6	0.33769
35	2.6	0.088597
35	2.9	0.3886
35	2.3	-0.2114
35	2.4	-0.1114
35	2.1	-0.4114
35	2.5	-0.011403
35	2.7	0.1886
45	2.7	-0.060496
45	3	0.2395
45	3.1	0.3395
45	2.3	-0.4605
45	2.5	-0.2605
45	3	0.2395
45	3.3	0.5395
45	2.7	-0.060496
55	3.1	0.09041
55	2.9	-0.10959
55	3.4	0.39041
55	2.4	-0.60959
55	3.4	0.39041
65	3.7	0.44132
65	2.8	-0.45868
65	3.3	0.041317
65	3.5	0.24132
65	3.3	0.041317
65	2.6	-0.65868

TAB. 1.4 – Résidus sur l'exemple (1.0.1)

Outils graphiques

Il s'agit ici d'outils descriptifs. nous traçons tout d'abord les résidus réduits en fonction de x (ou de $\hat{y}$) et en fonction de l'ordre des observations.

Nous cherchons à "voir" sur ces graphiques :

(i) l'indépendance des résidus ;

(ii) si la variance est constante ;

(iii) si on est dans le cas normale ;

(iv) si on peut détecter des valeurs aberrantes

Remarque 8.2.1. La méthode des moindres carrés utilisée pour estimer les paramètres implique que la somme des résidus e_i est nulle. Nous ne pouvons donc pas tester si $E(\varepsilon_i) = 0$.

Les quatres graphiques des figures 1.11 (respectivement **??**) qui correspondent aux résidus réduits suivant l'ordre des observations (respectivement suivant x) des quatres premiers cas des données de la figure 1.9 montrent :

(i) pour la cas (a) : il n'y a pas de problèmes ;

(ii) pour la cas (b) : il y a une structure dans les résidus. Ceci peut provenir de la non indépendance des résidus ou d'un mauvais choix de modèle. Ici un modèle parabolique serait sans doute plus adapté ;

(iii) pour le cas(c) : il y a un résidus très grand. La donnée correspondantes est sans doute suspecte ;

(iv) pour le cas (d) : la dispersion des résidus augmente en fonction de x. La variance n'est sans doute pas constante.

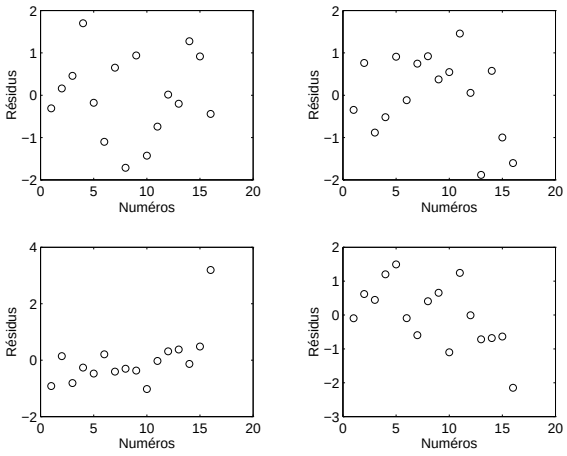

FIG. 1.11 – Résidus en fonction de l'ordre des observations

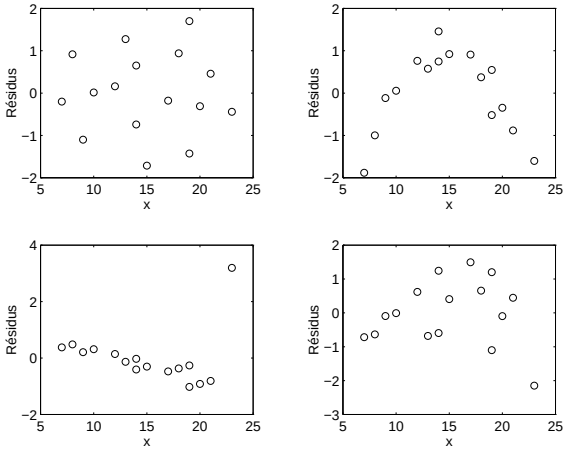

FIG. 1.12 – Résidus en fonction de x

Test sur les résidus

Les tests pour vérifier les postulats, en particulier l'indépendance, sont délicats à établir. Signalons cependant trois tests :

- le test de Bartlett pour tester l'égalité des variances. Celui-ci, pour notre cas, nécessite d'avoir des données groupées. Nous en reparlerons en deuxième année ;
- le test de Durbin-Waton couramment utilisé en économétrie. Il permet de s'assurer de la non corrélation des résidus. Les hypothèses nulle et alternative sont :
  H_0 : non corrélation des résidus ;
  H_1 : $\varepsilon_i = \rho\varepsilon_{i-1} + u_i$ avec $\rho > 0$.
  H_1 représente ce que l'on appelle un processus auto-régressif d'ordre 1. Ce test de pourra donc détecter que les cas où le résidu numéro i dépend du résidu $i-1$ et uniquement de lui d'une façon bien particulière. Il est surtout utilisé dans le cas où les observations sont ordonnées dans le temps.
  Il est basé sur le calcul de la quantité

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

  qui doit être voisin de 2 si l'hypothèse nulle H_o : non corrélation des résidus est vraie. Il existe des tables de valeurs critiques de d.

Variance des résidus

rappelons le modèle mathématique

$$Y = X\beta + \varepsilon$$

avec

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix}; \qquad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}; \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \text{et} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

et à partir des données nous avons obtenu

$$y = X\hat{\beta} + r$$

où

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{pmatrix}; \qquad \text{et} \quad r = \begin{pmatrix} r_1 \\ \vdots \\ r_i \\ \vdots \\ r_n \end{pmatrix}$$

Par suite nous pouvons définir le vecteur aléatoire des résidus :

$$\begin{array}{rcl} R : \mathcal{P}^n & \longrightarrow & \mathbf{R}^n \\ 1 \text{ expérience} & \longmapsto & r = \begin{pmatrix} r_1 \\ \vdots \\ r_i \\ \vdots \\ r_n \end{pmatrix} \end{array}$$

Nous avons alors le résultat suivant

Théorème 8.2.2. *Si les postulats sont vrais alors le vecteur aléatoire R suit une loi normale de dimension n de moyenne 0 et de matrice de variance, covariance $\sigma^2(I-H)$*

Démonstration
Nous avons $r = y - X\hat{\beta} = y - X({}^tXX)^{-1t}Xy = (I-H)y$, par suite, nous pouvons écrire $R = (I-H)Y$. Or Y suit une loi $\mathcal{N}(X\beta, \sigma^2 I)$, donc R suit une loi normale et

$$E(R) = (I-H)E(Y) = (I-H)X\beta = X_\beta - X({}^tXX)^{-1t}XX\beta = 0$$

$$Var(R) = (I-H)Var(Y)^t(I-H) = (I-H)\sigma^2 I^t(I-H)$$

Mais $(I-H)$ est un projecteur orthogonal, donc ${}^t(I-H) = (I-H)$ et $(I-H)^2 = (I-H)$. On en déduit immédiatement le résultat pour la matrice de variance, covariance. □

Corollaire 8.2.3. *Si les postulats sont vrais alors les variables aléatoires R_i liées aux résidus r_i suivent des lois normales $\mathcal{N}(0, \sigma^2(1-h_{ii}))$, où h_{ii} est le $i^{\text{ième}}$ terme diagonal de la matrice H.*

Définition 8.2.4 (Résidus standardisés). On appelle résidus standardisés[3] les résidus :

$$rstand_i = \frac{r_i}{\hat{\sigma}\sqrt{1-h_{ii}}}$$

Notos T_i la variable aléatoire suivante

$$\begin{array}{rcl} T_i : \mathcal{P}^n & \longrightarrow & \mathbf{R} \\ \texttt{1 exp} & \longmapsto & rstand_i \end{array}$$

Corollaire 8.2.5. *Si les postulats sont vérifiés alors les vriables aléatoires T_i suivents une loi de Student à $\nu = n - 2 ddl$.*

En Conclusion si les postulats sont vérifiés alors on doit avoir

$$0 \in [r_i - t_{1-\alpha/2}\hat{\sigma}\sqrt{1-h_{ii}}; r_i + t_{1-\alpha/2}\hat{\sigma}\sqrt{1-h_{ii}}]$$

avec une probabilité $1-\alpha$.

8.3 Mesures d'influences

Introduction

Les résultats d'une régression linéaire peuvent être modifiés par la suppression ou la perturbation d'une donnée. Différentes statistiques ont été définies pour quantifier ces influences. Nous allons ici voir les méthodes qui comparent les résultats obtenus lorsque l'on ajuste le modèle à l'ensemble des données et ceux obtenus lorsqu'on ajuste le modèle après avoir supprimé une ou plusieurs observations.

Mesures basés sur les résidus

L'idée est de regarder si la suppression d'une observation a une influence sur sa prédiction, ou encore de regarder si l'observation y_i est suffisemment proche de la valeur prédite $\hat{y}_{i((i)}$ obtenue en n'utilisant pas la $i^{\text{ième}}$ observation dans le calcul.

Notons $\hat{Y}_{i(i)}$ la variable aléatoire "prévision de y pour $x = x_i$" obtenue sans l'observation i. On a alors le théorème suivant.

Théorème 8.3.1. *Si les postulats sont vérifiés alors la variable aléatoire $Y_i - \hat{Y}_{i(i)}$ suit une loi normale*

$$\mathcal{N}\left(0, \sigma^2\left(1 + \begin{pmatrix}1 & x_i\end{pmatrix}({}^tX_{(i)}X_{(i)})^{-1}\begin{pmatrix}1\\ x_i\end{pmatrix}\right)\right)$$

où $X_{(i)}$ est la matrice extraite de X en supprimant la $i^{\text{ième}}$ ligne.

Démonstration
Admise □

Posons maintenant

$$\begin{array}{rcl} T_i^* : \mathcal{P}^n & \longrightarrow & \mathbf{R} \\ \texttt{1 exp} & \longmapsto & \dfrac{y_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{\left(1 + \begin{pmatrix}1 & x_i\end{pmatrix}({}^tX_{(i)}X_{(i)})^{-1}\begin{pmatrix}1\\ x_i\end{pmatrix}\right)}} \end{array}$$

où $\hat{\sigma}_{(i)}$ est l'estimation de l'écart-type obtenu sans la donnée i. Nous avons le théorème suivant

[3]Ces résidus sont appelés dans la littérature anglo-saxonne suivant les ouvrage et logiciels standardized residuals or studentized residuals

Théorème 8.3.2. *Si les postulats sont vrais et si la matrice $X_{(i)}$ est toujours de rang 2 alors T_i^* suit une loi de Student à $n-3ddl$*

Démonstration
Admise □

Définition 8.3.3. On appelle résidus par validation croisée[4] les résidus :

$$rvalcrois_i = \frac{y_i - \hat{y_{i(i)}}}{\hat{\sigma_{(i)}}\sqrt{\left(1 + \begin{pmatrix}1 & x_i\end{pmatrix}({}^tX_{(i)}X_{(i)})^{-1}\begin{pmatrix}1\\ x_i\end{pmatrix}\right)}}$$

Remarque 8.3.4. On démontre que l'on a

$$rvalcrois_i = \frac{r_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}}$$

Les intervalles de confiance des résidus sont alors

$$0 \in [r_i - t_{1-\alpha/2}\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}; r_i + t_{1-\alpha/2}\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}]$$

au niveau $1-\alpha$.

Exemple 8.3.5. La figure (1.13) donne ces intervalles de confiance à 95% pour les cas (a),(b),(c) et (d)

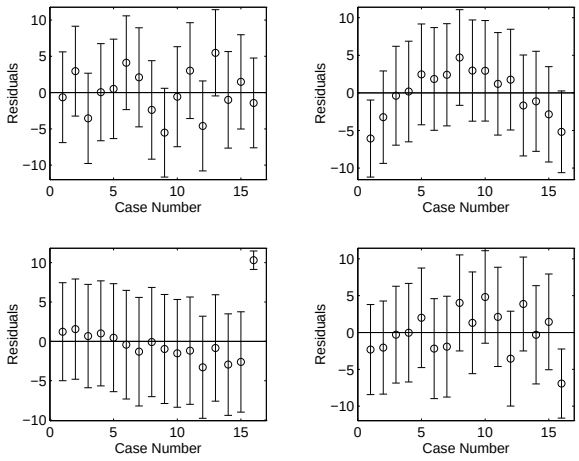

FIG. 1.13 – Détection de résidus aberrants : cas (a), (b), (c), (d).

On constate que l'on détecte bien la donnée aberrante dans la cas (c)

Exemple 8.3.6. Pour notre exemple du taux de cholestérol nous obtenons les résultats de la figure (1.14) où l'on ne détecte pas de résidu abberant.

Remarque 8.3.7. Il est préférable d'utiliser les résidus par validation croisée plutôt que les résidus standardisés. Cependant en pratique, on ne constate pas de grandes différence entre les deux.

Grâce à ces résidus nous avons aussi accès à des graphiques utiles pour vérifier la normalité. On représente les résidus standardisés (ou par validation croisée) et fonction des quantiles de la loi normale correspondante. Ces graphiques sont appelés dans la terminologie anglo saxone *normal probability plot*. La figure (1.15) donne ce graphique pour notre exemple du taux de cholestérol.

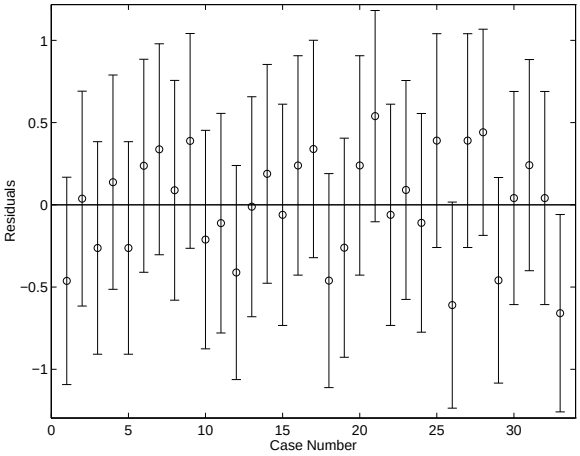

FIG. 1.14 – Détection de résidus aberrants : données de l'exemple

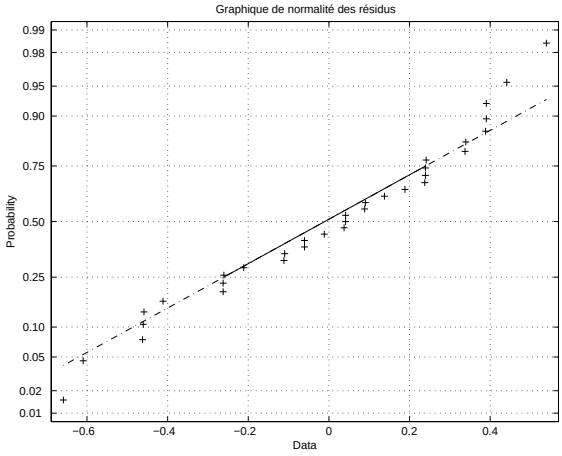

FIG. 1.15 – Graphique de normalité des résidus

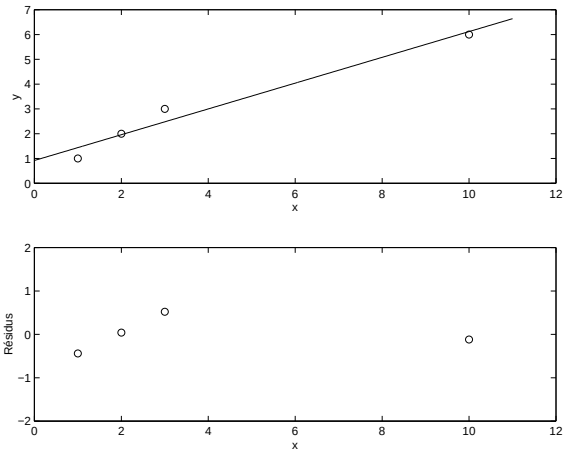

FIG. 1.16 – Contribution forte d'une observation

Mesure d'influence sur la variance résiduelle

Nous avons vu au paragraphe précédent qu'une forte valeur pour un résidu pouvait indiquer une valeur aberrante. Mais la réciproque est fausse (voir le graphique 1.16).

[4] appeés dans certains ouvrages studentized residuals

Nous allons ici étudier l'influence de chaque observation sur les la variance résiduelle ou, ce qui revient au même à une constante près sur la somme des carrés des écarts résiduelle. Supprimons l' observation i (x_i, y_i) et faisons la régression avec les autres observations. Nous obtenons une nouvelle somme des carrés des écarts des résidus notée $SCE_{R_{(i)}}$. La quantité $C_i = SCE_R - SCE_{R_{(i)}}$ est la contribution de la $i^{\text{ième}}$ observation à la somme des carrés. Nous pouvons effectuer ce calcul pour toutes les observations. Pour le cas simple correspondant au graphique 1.16 nous obtenons :

Observation no	x_i	y_i	$\hat{y}_i$	r_i	C_i	C_i/SCE_R en %
1	1	1	1.44	-0.44	0.3396	70.8
2	2	2	1.96	0.04	0.0024	0.5
3	3	3	2.48	0.52	0.3704	77.2
4	10	6	6.12	-0.12	0.4800	100.0

Nous observons alors que la quatrième observation qui ne donne pas le résidu le plus élevé est celle qui apporte la contribution la plus forte. Lorsque nous supprimons cette quatrième observation les trois points restant sont alignés!!!

Remarque 8.3.8. On peut aussi étudier l'influence d'une observation sur les valeurs des estimations des paramètres. On définit alors ce que l'on appelle la distance de Cook.

9 Les transformations de variables

9.1 Introduction

Que peut-on faire lorsque les postulats ne sont pas vérifiés et/ou que le modèle n'est las linéaire. Nous pouvons parfois par un simple changement de variables nous ramener au cas linéaire.

9.2 Linéarisation du modèle

Si nous avons par exemple comme modèle $y = \beta_0 e^{\beta_1 x}$, alors en prenant le logarithme népérien nous obtenons le modèle linéaire : $z = \ln y = \ln \beta_0 + \beta_1 x = \beta_0^{'} + \beta_1 x$.

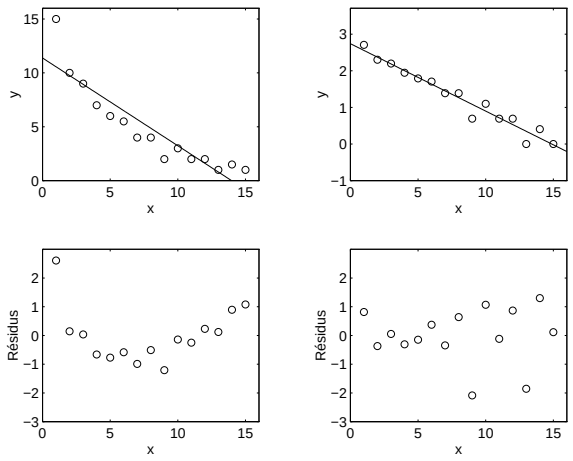

FIG. 1.17 – $y = \beta_0 e^{\beta_1 x}; z = \ln y = \ln \beta_0 + \beta_1 x$

9.3 Stabilisation de la variance

Lorsque la variance σ^2 n'est pas constante nous pouvons aussi parfois par un bon changement de variable stabiliser la variance. C'est le cas si la variance dépend de x. Il y a en particulier deux cas importants : le cas de la loi de poisson et celui de la loi binomiale.

Cas de la loi de Poisson

On considère ici le cas où l'on effectue des comptages, par exemple un nombre de cellules dans un carré de longueur fixé, ou un nombre de défauts de fabrication dans une pièce de tissu. Et on suppose qu'en moyenne cette quantité dépende de façon linéaire d'une variable x. Nous avons donc comme modèle : $E(Y) = \beta_0 + \beta_1 x$ et Y suit une loi de Poisson. On pourrai dans certains cas approximer cette loi par une loi normale, mais le principal problème est dans ce cas que la variance n'est pas une constante. En effet on sait que pour une loi de Poisson de paramètre λ, on a $E(Y) = Var(Y) = \lambda = \beta_0 + \beta_1 x$. Par suite la variance ne sera pas constante. On démontre dans ce cas que le changement de variable $z = \sqrt{y}$ stabilise la variance.

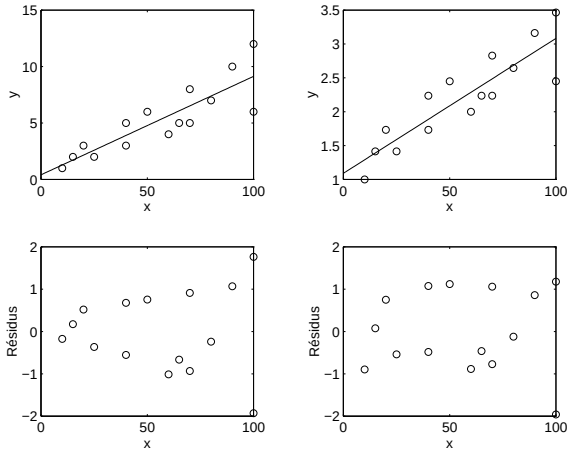

FIG. 1.18 – $z = \sqrt{y}$

Cas de la loi binomiale

Supposons que l'on s'interresse au taux de germination et que l'on pense que celui-ci dépend de façon linéaire d'une variable x. Concrètement pour différentes valeurs de x_i nous allons mettre à germer n graines et compter le nombre de graines qui on germé : n_i. Nous aurons ainsi comme données des couples $(x_i, y_i = \frac{n_i}{n})$. La loi de la variable aléatoire N_i en lien avec n_i est une loi binomiale de paramètres (n, p_i) . Nous aurons donc $E(Y_i) = p_i = \mu_i$ et $Var(Y_i) = \frac{p_i(1-p_i)}{n}$. Par suite, ici encore la variance des variables aléatoires $(Y_i)_i$ ne sera pas constante. On démontre que dans ce cas le bon changement de variable pour stabiliser la variance est de prendre $z = \arcsin\sqrt{y}$.

Remarque 9.3.1. (i) Pour que le changement de variable ci-dessus fonctionne il faut que n soit constant (ou en pratique varie peu). en effet on démontre que la variance de Z est $0.25/n$. Donc si n varie elle n'est pas constante ;

(ii) On rencontre parfois le changement de variable arcsin $\sqrt{y}$ sur des données exprimées en pourcentage. Ceci n'est justifié théoriquement que si ce pourcentage provient de données de notre type $\frac{n_i}{n}$.

Chapitre 2

Corrélation linéaire

1 Modèle

La figure (2.1) montre les fonctions de densité et courbes de niveaux associées pour des variables aléatoires de lois normales de paraètres $\mu = (1, 3)$,

$$\Gamma = \begin{pmatrix} 9 & 0 \\ 0 & 4 \end{pmatrix}$$

et $\mu = (1, 3)$,

$$\Gamma = \begin{pmatrix} 9 & 4.5 \\ 4.5 & 4 \end{pmatrix}$$

Pour la deuxième loi La figure (2.2) montre elle l'intersection de la fonction de densié avec un plan vertical d'équation $x = cte$. Quant-à la figure (2.3) elle montre les fonctions de densité de la loi de Y conditionnellement à X.

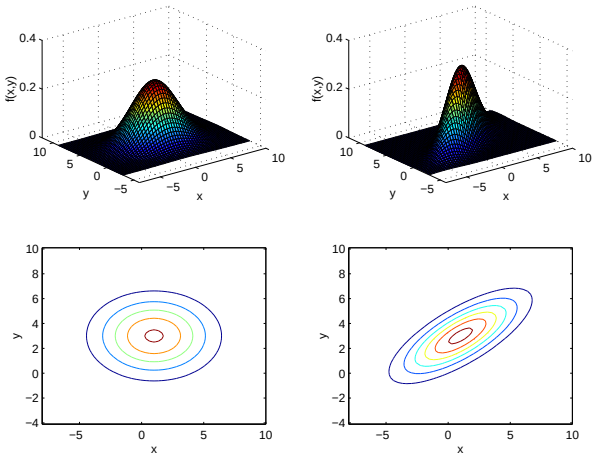

FIG. 2.1 – Fonctions de densité

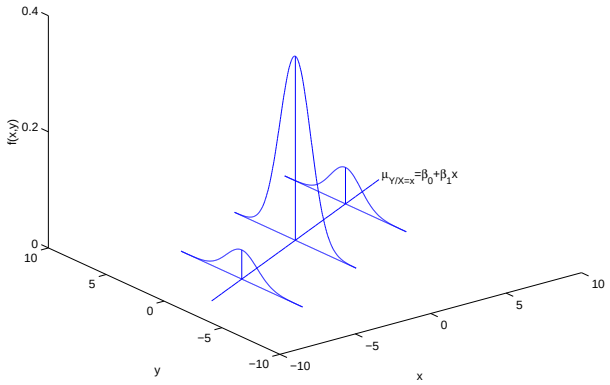

FIG. 2.2 – Intersection avec le plan $x = cte$

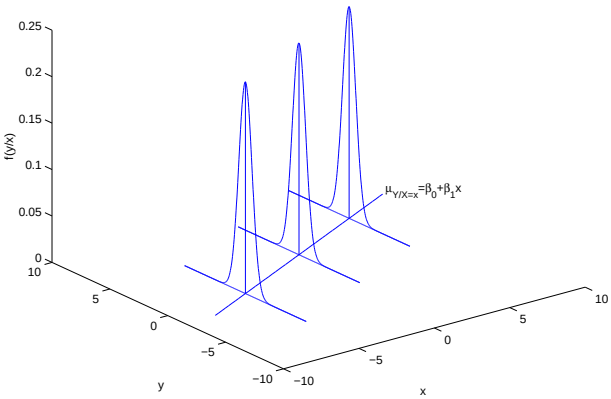

FIG. 2.3 – Fonction de densité conditionnelle

2 Corrélation de rang

2.1 Application

- La loi de probabilité jointe n'est pas normale ;
- Les données sont des rangs

2.2 Exemple[1]

Spherophoria philanthus est un *syrphidae* qui, à l'état larvaire, est un grand prédateur de pucerons. Cependant, son action est importante si ses oeufs sont pondus près des colonies de pucerons.

Dans une étude réalisée à l'époque de la ponte du *syrphidae*, Coderre (1983), en a recensé les oeufs ainsi que le nombre de pucerons sur 30 plants de maïs-grains.

La question est de savoir s'il y a une relation entre le nombre d'oeufs du prédateur et le nombre de proies (pucerons).

N° de plant	Nbre pucerons	Nbre d'oeufs	N° plant	Nbre pucerons	Nbre d'oeufs
1	25	1	16	45	2
2	10	1	17	55	2
3	75	3	18	0	0
4	250	6	19	125	4
5	10	1	20	35	1
6	50	2	21	10	0
7	100	3	22	75	3
8	85	2	23	50	2
9	0	0	24	0	0
10	0	0	25	30	1
11	65	2	26	20	1
12	30	1	27	200	5
13	5	0	28	40	1
14	400	8	29	70	3
15	35	1	30	15	1

2.3 Corrélation de rang de Spearman

Définition

On note d_i la différence des rangs des observations. S'il n'y a pas d'ex aequo le coefficient de corrélation de Spearman est donné par :

$$r_s = 1 - 6\sum_{i=1}^{n} \frac{d_i^2}{n(n^2-1)}$$

S'il y a des ex aequo il faut corriger par :

$$r_s = \frac{\sum x^2 + \sum y^2 - \sum_i d_i^2}{2\sqrt{\sum x^2 \sum y^2}}$$

avec :

- $\sum x^2 = \dfrac{n^3 - n}{12} \sum_{i=1}^{k} T_{x_i}$;
- $\sum y^2 = \dfrac{n^3 - n}{12} \sum_{i=1}^{l} T_{y_i}$;
- $T_{x_i} = \dfrac{t_{x_i}^3 - t_{x_i}}{12}$;
- $T_{y_i} = \dfrac{t_{y_i}^3 - t_{y_i}}{12}$;
- k est le nombre de série d'ex aequo pour la variable x ;
- l est le nombre de série d'ex aequo pour la variable y ;
- t_{x_i} est le nombre d'éléments ex aequo dans la i^e série d'ex aequo de x ;
- t_{y_i} est le nombre d'éléments ex æquo dans la i^e série d'ex aequo de y ;

[1] Exemple provenant de B. Scherrer, "Biostatistique", p. 598, ed. G. Morin, 1984

Application

N° d'éléménts	Nbre pucerons (x)	Nbre d'oeufs (y)	Différence (d_i)	d_i^2
1	11	11.5	0.5	0.25
2	6.5	11.5	5	25
3	23.5	24.5	1	1
4	29	29	0	0
5	6.5	11.5	5	25
6	18.5	19.5	1	1
7	26	24.5	1.5	2.25
8	25	19.5	5.5	30.25
9	2.5	3.5	1	1
10	2.5	3.5	1	1
11	21	19.5	1.5	2.25
12	12.5	11.5	1	1
13	5	3.5	1.5	2.25
14	30	30	0	0
15	14.5	11.5	3	9
16	17	19.5	2.5	6.25
17	20	19.5	0.5	0.25
18	2.5	3.5	1	1
19	27	27	0	0
20	14.5	11.5	3	9
21	6.5	3.5	3	9
22	23.5	24.5	1	1
23	18.5	19.5	1	1
24	2.5	3.5	1	1
25	12.5	11.5	1	1
26	10	11.5	1.5	2.25
27	28	28	0	0
28	16	11.5	4.5	20.25
29	22	24.5	2.5	6.25
30	9	11.5	2.5	6.25
Total				165.75

- pour ($x = 0$) $T_{x_1} = \frac{4^3-4}{12} = 5$;
- pour ($x = 10$) $T_{x_2} = \frac{3^3-3}{12} = 2$;
- pour ($x = 30$) $T_{x_3} = \frac{2^3-2}{12} = 0.5$;
- pour ($x = 35$) $T_{x_4} = \frac{2^3-2}{12} = 0.5$;
- pour ($x = 50$) $T_{x_5} = \frac{2^3-2}{12} = 0.5$;
- pour ($x = 75$) $T_{x_6} = \frac{2^3-2}{12} = 0.5$;

Donc $\sum T_x = 9$.
De même $\sum T_y = 122.5$
Par suite :

$$r_s = \frac{2238.5 + 2125 - 165.75}{2\sqrt{2238.5 \times 2125}} = 0.962$$

2.4 Test

On démontre que si les deux variables aléatoires X et Y sont indépendantes et si n est assez grand ($n > 30$) alors la variable aléatoire R_s associée à r_s suit une loi normale $\mathcal{N}(0, \frac{1}{n-1})$.

Application

$$Z_{obs} = r_s/\sqrt{n-1} = 0.962/29 = 5.18$$

$Z_{crit} = 1.96$ pour $\alpha = 0.05$ et un test bilatéral.

Conclusion : La relation entre le nombre d'oeufs du prédateur et le nombre de proies (pucerons) est hautement significative.

Chapitre 3

Compléments

1 Tests d'indépendance et d'homogénéité

1.1 Test d'homogénéité

Données

Evolution de l'âge de la population agricole familiale dans un canton du Loiret.

Année :Age	< à 25 ans	25 à 34 ans	35 à 44 ans	45 à 54 ans	55 à 64 ans	> à 65 ans	Total
1970	88	24	27	61	20	25	245
1979	63	17	20	39	27	25	191
1988	41	15	18	22	31	17	144
Total	192	56	65	122	78	67	580

Représentation graphique

Tableau des profils lignes

Année :Age	< à 25 ans	25 à 34 ans	35 à 44 ans	45 à 54 ans	55 à 64 ans	> à 65 ans
1970	0.3592	0.0980	0.1102	0.2490	0.0816	0.1020
1979	0.3298	0.0890	0.1047	0.2042	0.1414	0.1309
1988	0.2847	0.1042	0.1250	0.1528	0.2153	0.1181

Diagramme en bâtons

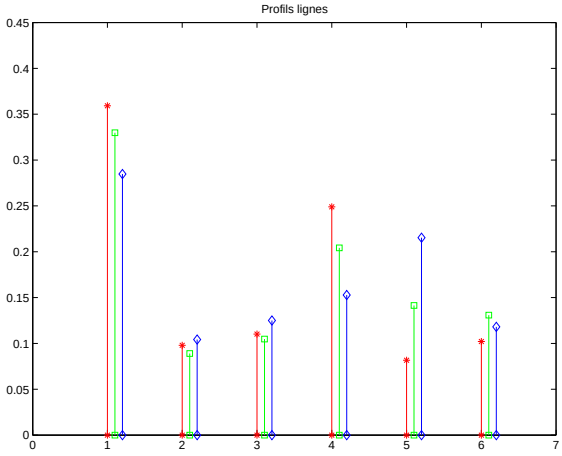

FIG. 3.1 – Profils lignes

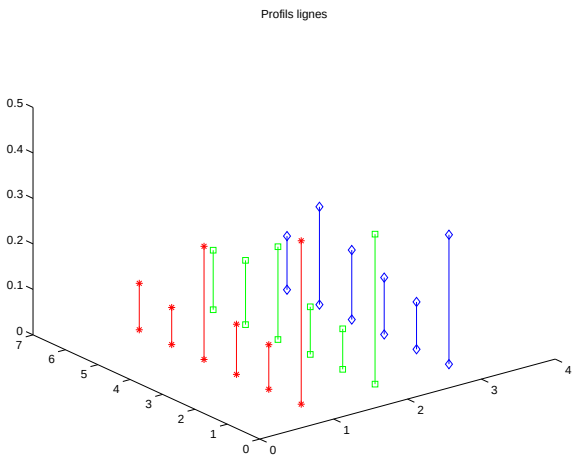

FIG. 3.2 – Profils lignes

Test du Khi2

$K_{obs} = 19.01$
$\alpha = 0.05; \nu = 10 ddl$
$K_{crit} = 18.31$
Conclusion : On rejette l'homogénéité des populations au risque α de 0.05 ; c'est-à-dire que la répartition des populations en fonction de l'âge a changé entre 1970, 1979 et 1988.

Contributions

Année :Age	< à 25 ans	25 à 34 ans	35 à 44 ans	45 à 54 ans	55 à 64 ans	> à 65 ans
1970	0.0308	0.0003	0.0004	0.0915	0.2677	0.0203
1979	0.0000	0.0059	0.0049	0.0018	0.0035	0.0206
1988	0.0491	0.0045	0.0113	0.1193	0.3677	0.0004

1.2 Test d'indépendance

Données

Répartition de 10000 étudiants en fonction de leur domaine d'étude et de la catégorie socio-professionnelle du père

CSP : Etude	Droit	Sc-éco	Lettres	Sciences	Méd Dent	Phar	Plur	IUT	Total
Explo agri	80	36	134	99	65	28	11	58	511
Sal agri	6	2	15	6	4	1	1	4	39
Patron	168	74	312	137	208	53	21	62	1035
Prof lib									
Cadre sup	470	191	806	400	876	164	45	79	3031
Cadre moy	236	99	493	264	281	56	36	87	1552
Employé	145	52	281	133	135	30	20	54	850
Ouvrier	166	64	401	193	127	23	28	129	1131
Pers serv	16	6	27	11	8	2	2	8	80
Autres	305	115	624	247	301	47	42	90	1771
Total	1592	639	3093	1490	2005	404	206	571	10000

Représentation graphique

Tableau des fréquences relatives

CSP : Etude	Droit	Sc-éco	Lettres	Sciences	Méd Dent	Phar	Plur	IUT
Explo agri	0.0080	0.0036	0.0134	0.0099	0.0065	0.0028	0.0011	0.0058
Sal agri	0.0006	0.0002	0.0015	0.0006	0.0004	0.0001	0.0001	0.0004
Patron	0.0168	0.0074	0.0312	0.0137	0.0208	0.0053	0.0021	0.0062
Prof lib								
Cadre sup	0.0470	0.0191	0.0806	0.0400	0.0876	0.0164	0.0045	0.0079
Cadre moy	0.0236	0.0099	0.0493	0.0264	0.0281	0.0056	0.0036	0.0087
Employé	0.0145	0.0052	0.0281	0.0133	0.0135	0.0030	0.0020	0.0054
Ouvrier	0.0166	0.0064	0.0401	0.0193	0.0127	0.0023	0.0028	0.0129
Pers serv	0.0016	0.0006	0.0027	0.0011	0.0008	0.0002	0.0002	0.0008
Autres	0.0305	0.0115	0.0624	0.0247	0.0301	0.0047	0.0042	0.0090

Diagrammes en bâtons

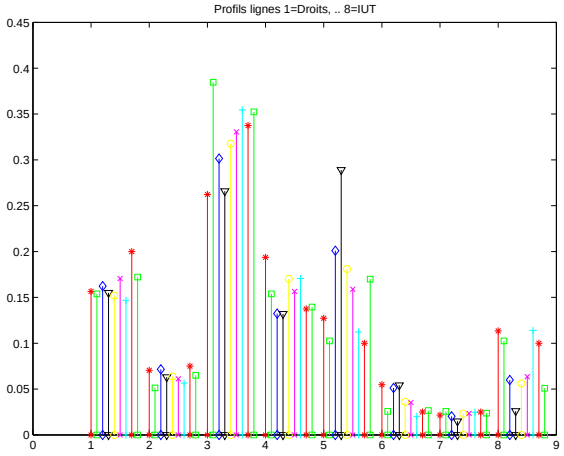

FIG. 3.3 – Profils lignes

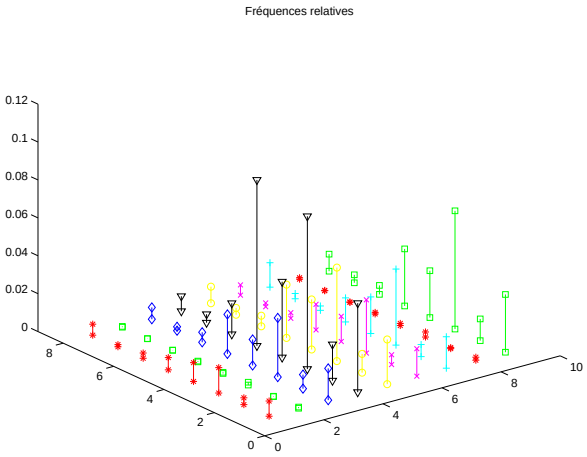

FIG. 3.4 – Fréquences relatives

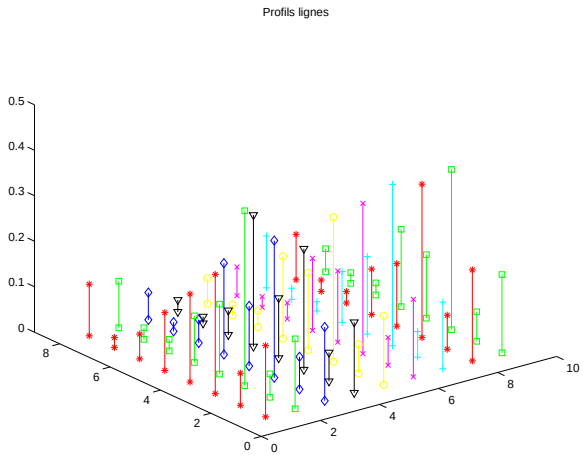

FIG. 3.5 – Profils lignes

Test du Khi2

$K_{obs} = 474.67$
$\alpha = 0.05; \nu = 56 ddl$
$K_{crit} = 74.47$
Conclusion : On rejette l'indépendance au risque α de 0.05 ; c'est-à-dire le choix du type d'étude dépend de la CSP du père

Contributions

CSP : Etude	Droit	Sc-éco	Lettres	Sciences	Méd Dent	Phar	Plur	IUT
Explo agri	0.0000	0.0007	0.0077	0.0145	0.0288	0.0055	0.0000	0.0600
Sal agri	0.0000	0.0002	0.0015	0.0000	0.0039	0.0004	0.0001	0.0030
Patron	0.0001	0.0020	0.0004	0.0040	0.0000	0.0063	0.0000	0.0003
Prof lib								
Cadre sup	0.0007	0.0001	0.0389	0.0124	0.2495	0.0297	0.0103	0.1077
Cadre moy	0.0010	0.0000	0.0007	0.0098	0.0062	0.0015	0.0011	0.0001
Employé	0.0015	0.0002	0.0026	0.0007	0.0155	0.0012	0.0007	0.0013
Ouvrier	0.0023	0.0020	0.0158	0.0075	0.0925	0.0237	0.0020	0.1354
Pers serv	0.0018	0.0003	0.0004	0.0001	0.0085	0.0010	0.0002	0.0054
Autres	0.0040	0.0001	0.0223	0.0023	0.0174	0.0177	0.0018	0.0026

2 Simulation loi de Fisher

5-échantillons de $\mathcal{N}(2,4)$					3-échantillons de $\mathcal{N}(1,4)$			F
1.1349	-1.3312	2.2507	2.5754	-0.2929	1.9572	-0.5277	-1.9969	0.6936
4.3818	4.3783	1.9247	2.6546	2.3493	3.4774	0.8699	0.5281	0.5201
1.6266	3.4516	0.8234	6.3664	1.7272	1.0733	1.9152	0.4796	9.4064
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0.9801	-0.6693	1.7307	2.8375	6.3461	2.6155	1.8116	1.3095	15.8573

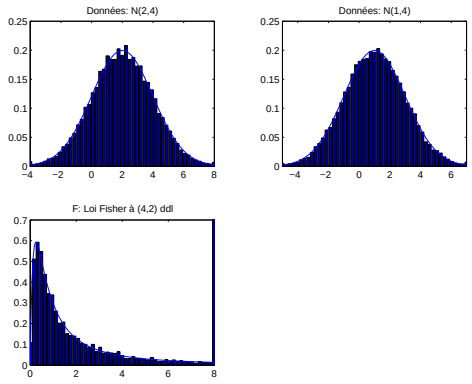

FIG. 3.6 – Simulation de la loi de Fisher à $\nu_1 = 4$ et $\nu_2 = 2$ ddl

3 Quelques remarques sur la collecte des données

3.1 Exemple 1 (Tomassone)

- On choisit 3 cercles au hasard (en jetant une bille par exemple) puis on fait la moyenne des surfaces. On répète l'expérience 50 fois, puis on fait l'histogramme des 50 moyennes.
- On choisit au hasard 3 nombres compris entre 1 et 60, on prend les 3 cercles qui correspondent, puis on fait la moyenne des surfaces. On répète l'expérience 50 fois, puis on fait l'histogramme des 50 moyennes.

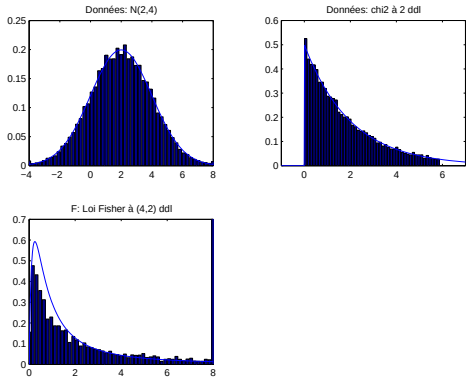

FIG. 3.7 – Non loi de Fisher

Question : Quel processus choisir ?

Réponse :
- Dans le processus 1 un gros cercle a plus de chance d'être choisi. Tous les individus (les cercles ici) n'ont pas la même probabilité d'être choisis. Par suite l'estimation de la surface moyenne est biaisée.
- Dans le deuxième processus, chaque individu a la même probabilité d'être choisi.

Remarque 3.1.1. Dans certains cas les individus n'ont pas le même poids, il faut alors en tenir compte dans l'échantillonnage.

3.2 Exemple 2

On désire savoir quelle est la meilleure marque de yaourt parmi 5 marques A, B, C, D, E. On demande pour cela à un goûteur de donner une note à chaque yourt. On réalise cette expérience pendant 10 jours de suite, juste après le petit déjeuner. On donne toujours les yourts dans le même ordre.

Question : Quels problèmes posent cette expérience ?

Réponse :
- On ne pourra avoir une réponse que pour l'appéciation de ce goûteur. A-t-on tous le même goût ? C'est un problème de définition de la population : $\mathcal{P} = \{$les yaourts mangés par le goûteur$\}$.
- On ne pourra avoir une réponse que pour les yaourts mangés après le petit déjeuner. C'est un problème de définition de la population : $\mathcal{P} = \{$les yaourts mangés après le petit déjeuner$\}$.
- Problème d'indépendance. Si les yaourts de la marque A sont plus sucrés que ceux de la marques B et que l'on mange toujours celui de la marque A avant celui de la marque B, alors les résultats pour les yaourts de la marque A influencent ceux de la marque B. Le modèle est ici :

$$\begin{array}{lcl} Y_{i,j} : \mathcal{P}^n & \longrightarrow & \{0, 1, \ldots, 10\}^5 \\ 5 \text{ yaourts} & \longmapsto & \text{note du yourt de la marque } i \text{ la } j \text{ ième fois} \end{array}$$

$\mathcal{P} = \{$échantillons de 5 yaourts (1 de marque A,..., 1 de marque E)$\}$ et n =le nombre d'expérience.
Les variables aléatoire $Y_{i,j}$ ne sont pas indépendantes.
Conclusion : il faut répartir aléatoirement à chaque fois l'ordre de dégustation des yaourts.

3.3 Les sondages

(i) Avoir une bonne base de sondage : la population.
Exemple 3.3.1. le vote de paille au USA organisé par le "Literary Digest" auprès de 10 millions de lecteurs le 3 novembre 1936 prédit la victoire de Lindon alors que 3 sondages (Crossley, Roger et Gallup) par la méthode des quotas prédirent la victoire de Roosevelt ;
Exemple 3.3.2. Pour une étude sur les boulangeries, il faut la liste de toutes les boulangeries.

(ii) Cas des personnes absentes et des refus de réponses ;

(iii) Le questionnaire :
- bien formuler les questions, pas de question avec une double négation, ...
- rythme, longueur, ...

- ...

(iv) La méthode de collecte des données : par courrier, téléphone, enquêteur ?

(v) Non influence de l'enquêteur ;

(vi) Qualité, fiabilité des réponses.

4 Rédaction

4.1 Introduction

- bien poser le problème ou l'étude ;
- formulation des questions posées sous forme statistique ;
- choix de la méthode statistique utilisée et de la façon de collecter les données.

4.2 Pour la collecte des données :

- bien définir la population étudiée ;
- bien définir les individus, unités expérimentales ;
- bien définir les variables étudiées et mésurées ;
- bien définir la méthode dans le choix des individus ;
- bien définir la méthode pour les mesures.

Pour l'analyse des données :

- faire les statistiques descriptives (en particulier les graphiques pour visualiser les données) ;
- donner la méthode employée ;
- vérifier la validité de l'emploi de la méthode ;
- présenter les résultats, dans la mesure du possible, sous forme de graphique ;
- ne pas oublier la **variabilité** des variables étudiées (en particulier dans les graphiques).

Remarque 4.2.1. La question de savoir s'il faut mettre les calculs statistiques dans le rapport ou en annexe dépend de l'étude.