Cours de biostatistique ⋈ Illustrations dans 

Principes du modèle linéaire

A.B. Dufour et D. Chessel

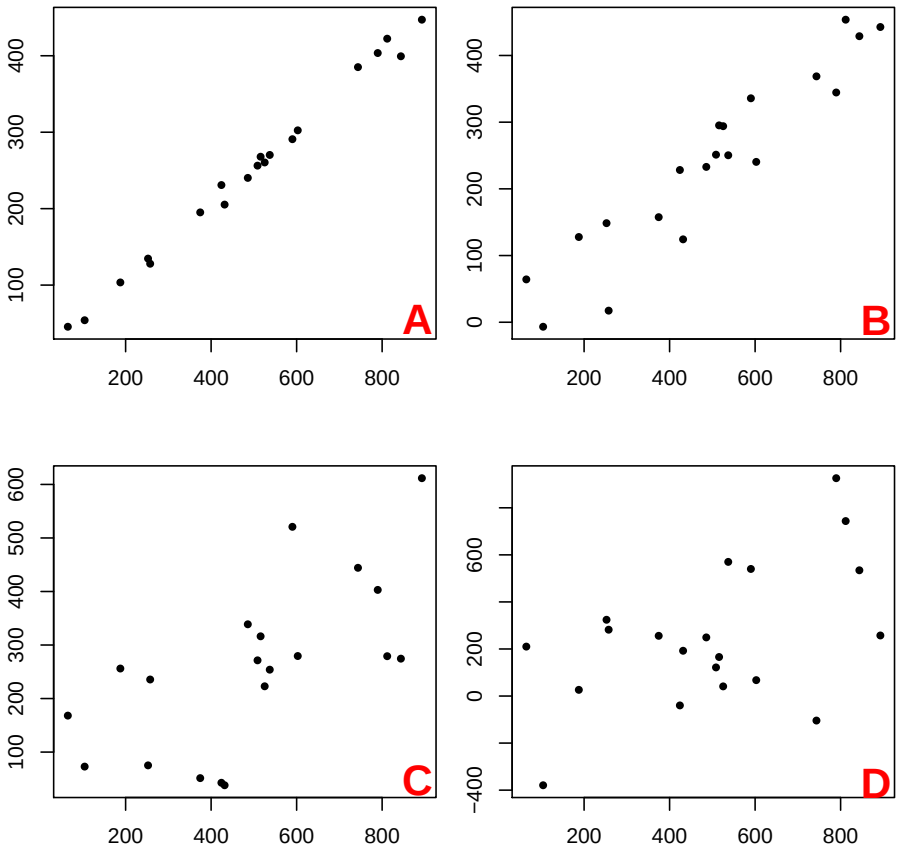
24 novembre 2015

La fiche donne des indications sur le contenu d'un tableau d'analyse de la variance. On utilise la régression simple comme support des illustrations. Orthogonalité géométrique et indépendance, théorème de Pythagore et décomposition de somme de carrés sont mis en parallèle.

Table des matières

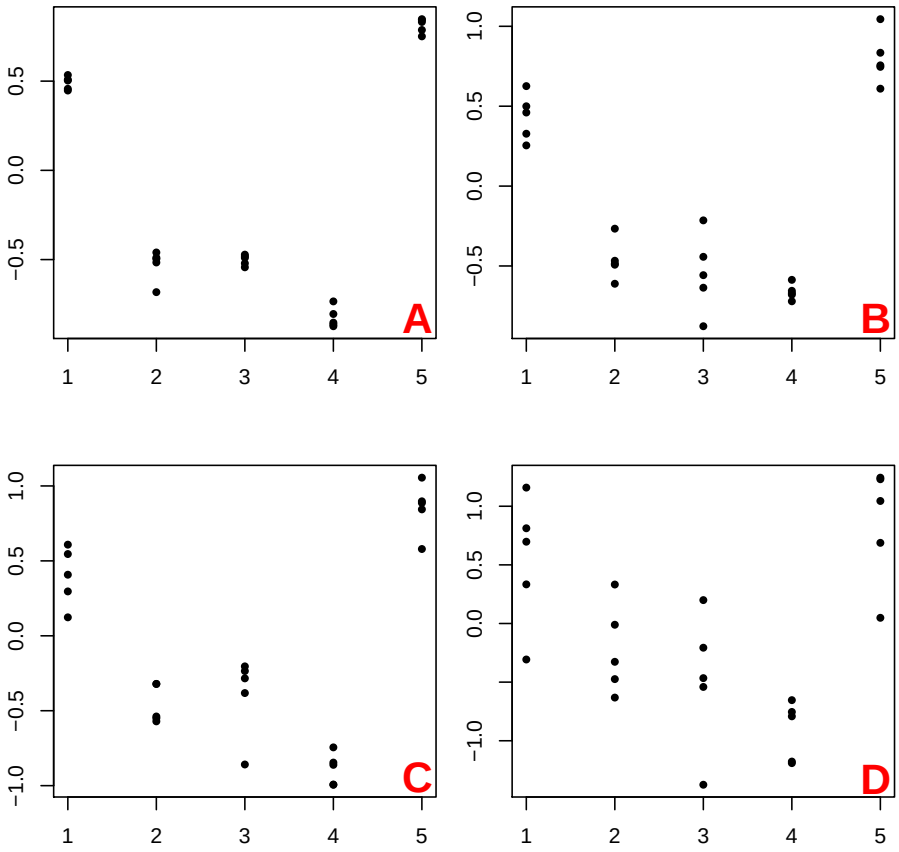
1 Introduction

On considère la relation entre deux variables quantitatives **x** et **y** et étudions les quatre situations suivantes.



Le cas **A** est clair : **y** croît linéairement avec **x**. Le cas **B** l'est aussi : **y** croît linéairement avec **x**, bien qu'une erreur entoure la relation. Le cas **C** l'est moins. L'erreur est largement de l'ordre de l'effet. Le cas **D** ne l'est plus du tout. Quand peut-on dire que **y** croît linéairement avec **x** ?

On considère maintenant la relation entre une variable qualitative **x** et une variable quantitative **y** et étudions les quatre situations suivantes.



Le cas **A** est clair : **y** dépend de la modalité de **x**. Le cas **B** l'est aussi : **y** dépend de la modalité de **x**, bien qu'une erreur entoure la relation. Le cas **C** l'est moins. L'erreur est largement de l'ordre de l'effet. Le cas **D** ne l'est plus du tout. Quand peut-on dire que **y**dépend de la modalité de **x** ?

2 Le modèle de l'erreur

Jusqu'à présent, les deux variables **x** et **y** sont des listes de valeurs vues comme deux vecteurs de $\mathbb{R}^n$. On peut alors les écrire $\mathbf{x}^\top = (x_1\ x_2\ \cdots\ x_n)$ et $\mathbf{y}^\top = (y_1\ y_2\ \cdots\ y_n)$.

Un modèle de **y** à partir de **x** s'écrit :

$$\hat{\mathbf{y}} = a\mathbf{x} + b \quad \text{ou encore} \quad \forall i = 1, n \quad \hat{y}_i = ax_i + b$$

Ajouter un test au modèle proposé, c'est introduire la notion de modèle probabiliste. La situation la plus simple, qui sert de support à ce cours, est le modèle à effet fixe.

1. Une valeur x_i de **x** est connue sans erreur. La variable est dite contrôlée.

2. Une valeur y_i de $\mathbf{y}$ est inconnue et répond à un processus aléatoire. y_i est une réponse dans un ensemble de réponses possibles. On parle alors de tirage aléatoire dans une loi de probabilité notée Y_i.

Le modèle linéaire est caractérisé par l'hypothèse suivante : Y_i est une loi normale de moyenne μ et d'écart-type σ :

$$Y_i \mapsto \mathcal{N}(\mu, \sigma).$$

y_i est alors un échantillon aléatoire simple de cette loi normale Y_i.

Un modèle de $\mathbf{y}$ à partir de $\mathbf{x}$ est un ensemble d'hypothèses sur la moyenne μ et la variance σ^2. Dire que $\mathbf{y}$ dépend linéairement de $\mathbf{x}$ s'écrit :

$$Y_i \mapsto \mathcal{N}(\mu, \sigma) \text{ avec } \left\{ \begin{array}{l} \mu = ax_i + b \\ \sigma^2 = \text{constante} \end{array} \right.$$

Il y a trois paramètres et l'hypothèse fondamentale de la normalité des résidus. Dans le cadre de la statistique descriptive classique, le modèle a deux paramètres a et b, aucune loi de probabilité. La solution est obtenue par le critère des moindres carrés :

$$\underset{a,b}{Min} \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - ax_i - b)^2 \right) = \underset{a,b}{Min} \left(\|\mathbf{y} - a\mathbf{x} - b\mathbf{1}_n\|^2 \right)$$

Dans le cadre actuel, le modèle possède trois paramètres a, b et σ^2. Comme il est probabilisé, le critère d'estimation est celui de la vraisemblance :

$$\underset{a,b,\sigma^2}{Max} \left(LL(a, b, \sigma^2) \right) = \underset{a,b,\sigma^2}{Max} \left(Log \left(P_{(a,b,\sigma^2)}(y_1, y_2, \ldots, y_n) \right) \right)$$

Dans le cas des moindres carrés, on obtient :

$$a = \frac{c(\mathbf{x}, \mathbf{y})}{v(\mathbf{x})} \quad \text{et} \quad b = m(\mathbf{y}) - am(\mathbf{x}).$$

Dans le cas de la vraisemblance, la densité de probabilité de Y_i est :

$$d(y_i) = \frac{1}{\sqrt{2\pi}\sigma} exp \left(-\frac{1}{2} \left(\frac{y_i - ax_i - b}{\sigma} \right)^2 \right)$$

La vraisemblance de l'échantillon est :

$$L(a, b, \sigma^2) = \prod_{i=1}^{n} d(y_i) \Leftrightarrow LL(a, b, \sigma^2) = \sum_{i=1}^{n} log \left[\frac{1}{\sqrt{2\pi}\sigma} exp \left(-\frac{1}{2} \left(\frac{y_i - ax_i - b}{\sigma} \right)^2 \right) \right]$$

$$LL(a, b, \sigma^2) = -nlog(\sigma) - \frac{1}{2} \sum_{i=1}^{n} \left(\frac{y_i - ax_i - b}{\sigma} \right)^2 + Cte$$

Pour maximiser cette quantité, on calcule les trois dérivées partielles.

- $\frac{\partial LL(a,b,\sigma^2)}{\partial b} = \sum_{i=1}^{n}\left(\frac{y_i - ax_i - b}{\sigma}\right) = 0 \Rightarrow b = m(\mathbf{y}) - am(\mathbf{x})$

- $\frac{\partial LL(a,b,\sigma^2)}{\partial a} = \sum_{i=1}^{n}\left(\frac{y_i - ax_i - b}{\sigma}\right) x_i = 0 \Rightarrow \sum_{i=1}^{n} y_i x_i - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i = 0$

$$a = \frac{\sum_{i=1}^{n}\left(y_i - m(\mathbf{y})\right)\left(x_i - m(\mathbf{x})\right)}{\sum_{i=1}^{n}\left(x_i - m(\mathbf{x})\right)^2}$$

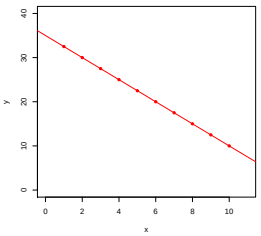
- $\frac{\partial LL(a,b,\sigma^2)}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n}\frac{(y_i - ax_i - b)^2}{\sigma^3} = 0 \Rightarrow \sigma^2 = E(a,b)$

Les estimations au maximum de vraisemblance de a et b sont les mêmes valeurs obtenues que par les moindres carrés. Cela souligne qu'on peut donner à un calcul des significations bien différentes.

3 Estimateur et estimation par l'expérience

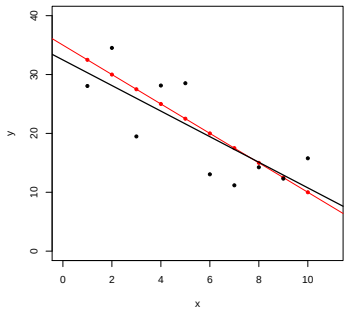
On considère la situation théorique suivante :

```
x <- 1:10
y <- -2.5*x+35
plot(x,y,xlim=c(0,11),ylim=c(0,40),type="n")
points(x,y,cex=1,pch=20, col="red")
abline(35,-2.5, col="red",lwd=1.5)
```

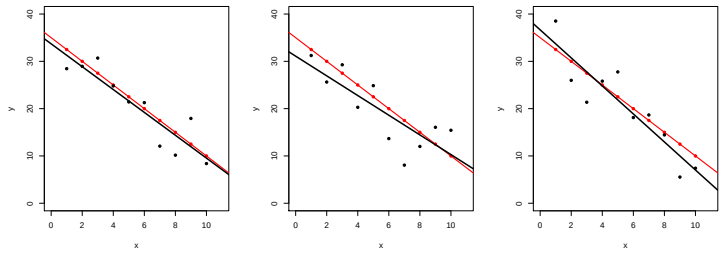


On introduit un bruit gaussien $\mathcal{N}(0,4)$ sur la variable $\mathbf{y}$ et on calcule le modèle associé.

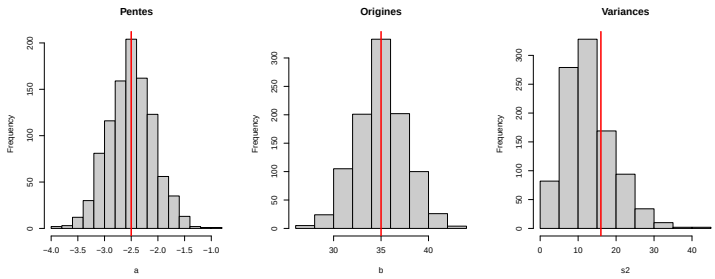
```
plot(x,y,xlim=c(0,11),ylim=c(0,40),type="n")
points(x,y,cex=1,pch=20, col="red")
abline(35,-2.5, col="red",lwd=1.5)
z <- y+rnorm(10,m=0,sd=4)
points(x,z,cex=1,pch=20)
abline(lm(z~x),lwd=2)
```



On recommence l'expérience $n = 3$ fois.



On recommence l'expérience $n = 1000$ fois et on regarde la distribution des trois paramètres : pente, ordonnée à l'origine et variance associée à la loi de Gauss.



A chaque expérience, le hasard donne des résultats autour du modèle et on calcule l'estimation des paramètres a, b, σ^2. On obtient ainsi une **une distribution d'échantillonnage** pour chaque paramètre visualisée par les histogrammes ci-dessus.

- a varie autour de la vraie valeur -2.5.
- b varie autour de la vraie valeur 35.
- La variance varie autour de la vraie valeur 16.

Une véritable expérience est par contre **unique**. Elle donne une seule valeur estimée qui est **fausse**. Les données $\mathbf{y}^\top = (y_1\ y_2\ \cdots\ y_n)$ sont des **réalisations** de n variables aléatoires $Y = (Y_1, Y_2, \ldots, Y_n)$. Le calcul de l'estimation se fait par une

formule du type $\hat{\theta} = f(y_1, y_2, \ldots, y_n)$. La vraie valeur est θ. $\hat{\theta}$ est l'**estimation**. θ est une réalisation d'une variable aléatoire $\Theta = f(Y_1, Y_2, \ldots, Y_n)$. Θ est une variable aléatoire : c'est l'estimateur. $\hat{\theta}$ et θ sont des nombres.

Ce qui est essentiel, en pratique, c'est la relation entre θ (**inconnue**) et $\hat{\theta}$ (**calculée**). Cette relation dépend des propriétés de l'estimateur c'est-à-dire des caractéristiques de sa distribution d'échantillonnage.

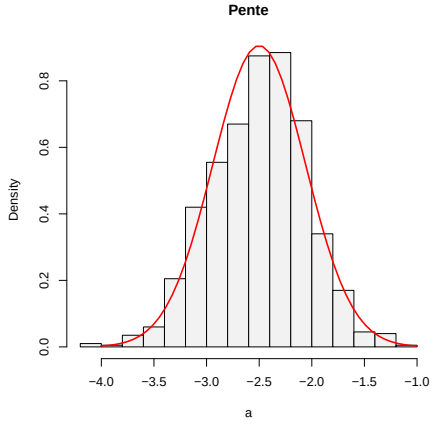
Si on prend par exemple la pente associée à une régression linéaire. Il faut distinguer :

1. la formule de calcul qui représente une **fonction** : $\theta = f(y_1, y_2, \ldots, y_n)$ c'est-à-dire ici
$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$
2. l'**estimation** du paramètre pour les vraies données qui représente la **valeur de la fonction** :
$$\hat{a} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$
3. l'**estimateur** qui représente une variable **aléatoire** :
$$A = \frac{n\sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

4 Distributions d'échantillonnage

On retient que l'estimation est fausse mais que sa qualité dépend de la loi de probabilité de l'estimateur. Dans l'exemple de la régression simple, on a les résultats ci-dessous.

• La pente :
$$A = \frac{n\sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$
A est une combinaison linéaire de lois normales qui suit une loi normale de paramètres :
$$\mathbb{E}(A) = a \quad \text{et} \quad V(A) = \frac{\sigma^2}{nv(\mathbf{x})}$$

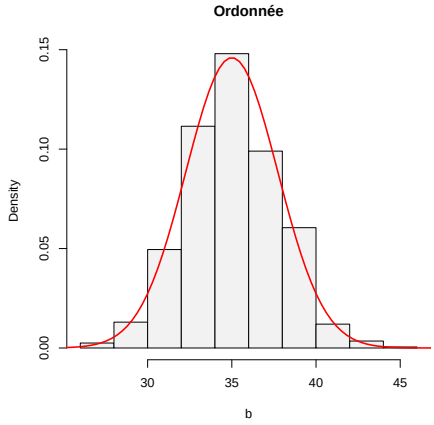


• L'ordonnée à l'origine

$$B = \frac{1}{n}\sum_{i=1}^{n} Y_i - A\frac{1}{n}\sum_{i=1}^{n} x_i$$

B est une combinaison linéaire de lois normales qui suit une loi normale de paramètres :

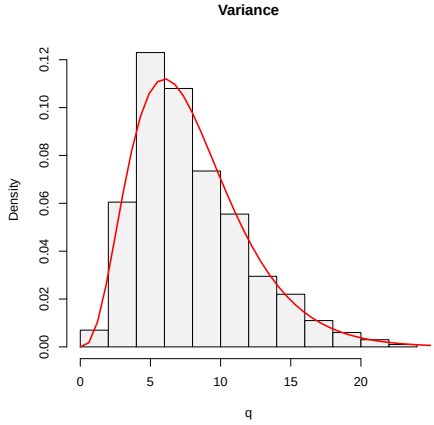
$$\mathbb{E}(B) = b \quad \text{et} \quad V(B) = \sigma^2\left(\frac{1}{n} + \frac{m(\mathbf{x})^2}{nv(\mathbf{x})}\right)$$



• La variance $S2$ est plus complexe. On va montrer que $\mathbb{E}(S2) = \frac{(n-2)\sigma^2}{n}$. La variable aléatoire

$$\frac{nS2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{Y_i - Ax_i - B}{\sigma}\right)^2$$

est une somme de carrés de lois normales c'est-à-dire une loi du Chi-Deux à $n-2$ degrés de liberté.



5 Les 3 propriétés des estimateurs

5.1 Le biais

La première propriété d'un estimateur est d'être **juste** c'est-à-dire de tourner autour de la vraie valeur. Ceci se formalise par : $\mathbb{E}(\Theta) = \theta$, la moyenne de l'estimateur est la vraie valeur.
Si l'estmateur n'est pas juste, sa moyenne s'écarte de la vraie valeur. On dit alors qu'il est biaisé. Ce biais est défini par : $Biais(\Theta) = \theta - \mathbb{E}(\Theta)$

5.1.1 A est un estimateur sans biais de a

$$A = \frac{n\sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

L'espérance de A est donc :

$$\begin{aligned}
\mathbb{E}(A) &= \frac{n\sum_{i=1}^{n} x_i \mathbb{E}(Y_i) - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} \mathbb{E}(Y_i)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\
&= \frac{n\sum_{i=1}^{n} x_i (ax_i + b) - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} (ax_i + b)}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\
&= \cdots \\
&= a
\end{aligned}$$

5.1.2 B est un estimateur sans biais de b

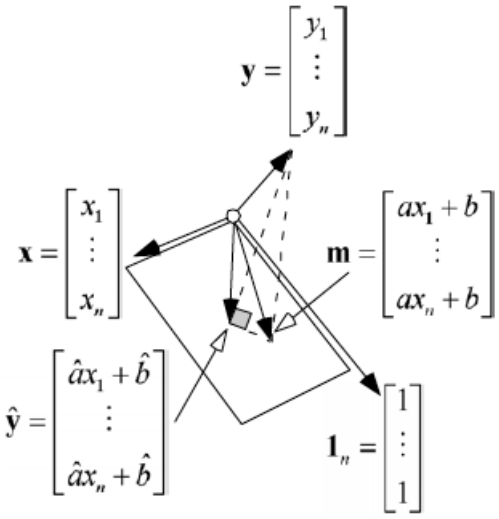
$$B = \frac{1}{n}\sum_{i=1}^{n} Y_i - A\frac{1}{n}\sum_{i=1}^{n} x_i$$

L'espérance de B est donc :

$$\begin{aligned}
\mathbb{E}(B) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(Y_i) - \mathbb{E}(A)\frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \cdots \\
&= b
\end{aligned}$$

5.1.3 S2 est un estimateur biaisé

Pour le démontrer, on commence par regarder la représentation algébrique de la régression simple.



$\mathbf{x}$, $\mathbf{m}$ et $\mathbf{1}_n$ sont fixes. $\mathbf{y}$ et $\hat{\mathbf{y}}$ sont aléatoires. L'estimation de la variance est :

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{a}x_i - \hat{b}\right)^2$$

L'estimateur est : $S2 = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - Ax_i - B\right)^2$
Le vecteur $\hat{\mathbf{y}}$ est le projeté de $\mathbf{y}$ sur le plan $\mathcal{H}$ engendré par $\mathbf{x}$ et $\mathbf{1}_n$. $\mathbf{y} - \hat{\mathbf{y}}$ est orthogonal à $\mathcal{H}$ et tous les vecteurs qui sont dedans, en particulier $\hat{a}\mathbf{x} + \hat{b}\mathbf{1}_n$ (le modèle estimé) et $a\mathbf{x} + b\mathbf{1}_n$ (le modèle vrai) ainsi qu'à leur différence. Donc, le projeté de $\mathbf{y} - \mathbf{m}$ sur le sous-espace orthogonal au plan est $\mathbf{y} - \hat{\mathbf{y}}$:

$$\mathbf{y} - \hat{\mathbf{y}} = P_{\mathcal{H}^\top}(\mathbf{y} - \mathbf{m})$$

Pour poursuivre cette démonstration, ne pas oublier les deux points suivants :

1. Si P est un projecteur orthogonal, alors :

$$\langle P(\mathbf{z}) | \mathbf{z} - P(\mathbf{z}) \rangle = 0 \Rightarrow \|P(\mathbf{z})\|^2 = \langle \mathbf{z} | P(\mathbf{z}) \rangle$$

2. Si P est un projecteur orthogonal sur $\mathcal{K}$ de dimension p, on peut construire une base d'éléments de $\mathcal{K}$ complétée par une base d'éléments de l'orthogonal de $\mathcal{K}$. Au total, on obtient une base orthogonale de $\mathbb{R}^n$ pour laquelle la matrice P est du type :

$$\mathbf{A} = \begin{pmatrix} \mathbf{I}_p & \mathbf{0}_{p(n-p)} \\ \mathbf{0}_{(n-p)p} & \mathbf{0}_{n-p(n-p)} \end{pmatrix}$$

parce que les projetés de $\mathcal{K}$ sont eux-mêmes et les projetés de l'orthogonal de $\mathcal{K}$ soit nuls. La somme des éléments diagonaux, notée $Trace(\mathbf{A})$ vaut

p. Dans une autre base, la matrice du projecteur serait obtenue par le changement de base $\mathbf{B} = \mathbf{HAH}^{-1}$ dont la somme des éléments diagonaux ($Trace(\mathbf{B})$) vaut encore p. En effet,

$$Trace(\mathbf{XY}) = Trace(\mathbf{YX}) \Rightarrow Trace(\mathbf{HAH}^{-1}) = Trace(\mathbf{H}^{-1}\mathbf{HA}) = Trace(\mathbf{A})$$

$$\|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \|P_{\mathcal{H}^\perp}(\mathbf{y} - \mathbf{m})\|^2 = \langle \mathbf{y} - \mathbf{m} | P_{\mathcal{H}^\perp}(\mathbf{y} - \mathbf{m}) \rangle$$

On utilise la matrice $\mathbf{A}$ du projecteur :

$$\begin{aligned} \|\mathbf{y} - \hat{\mathbf{y}}\|^2 &= \langle \mathbf{y} - \mathbf{m} | P_{\mathcal{H}^\perp}(\mathbf{y} - \mathbf{m}) \rangle \\ &= \frac{1}{n}(\mathbf{y} - \mathbf{m})^\top \mathbf{A}(\mathbf{y} - \mathbf{m}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(y_i - ax_i - b)(y_i - ax_i - b) \end{aligned}$$

d'où :

$$\begin{aligned} \mathbb{E}(S2) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}(Y_i - ax_i - b)(Y_i - ax_i - b)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\mathbb{E}\left((Y_i - ax_i - b)(Y_i - ax_i - b)\right) \end{aligned}$$

Si les Y_i sont indépendants (échantillons aléatoires simples), les termes en ij sont nuls et les termes en ii tous égaux à σ^2. Il reste donc :

$$\mathbb{E}(S2) = \frac{1}{n}\sum_{i=1}^{n} a_{ii}\sigma^2 = \frac{\sigma^2}{n}\sum_{i=1}^{n} a_{ii} = \frac{\sigma^2}{n}Trace((A) = \frac{(n-2)\sigma^2}{n}$$

Du point de vue fondamental, le $n-1$ de l'estimation de la variance est la dimension du sous-espace vectoriel orthogonal au seul vecteur $\mathbf{1}_n$, le $n-2$ de la régression linéaire simple est la dimension du sous-espace orthogonal à $\mathbf{x}$ et $\mathbf{1}_n$. Cette propriété se retrouve en régression multiple, en analyse de la variance à un facteur, ...

Ce résultat peut s'expérimenter. Le modèle a des vraies valeurs et nous avons fait 1000 expériences indépendantes.

- a a pour vraie valeur 2.5

```
mean(a)
[1] -2.501802
```

- b a pour vraie valeur 35

```
mean(b)
[1] 35.04594
```

- σ^2 a pour vraie valeur 16

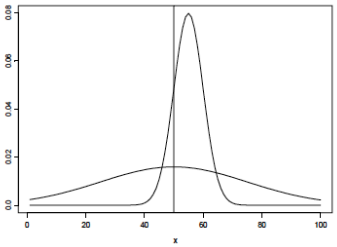
```
mean(s2)
[1] 12.55911
mean(10*s2/8)
[1] 15.69889
```

5.2 La précision

La seconde propriété importante d'un estimateur concerne sa précision. On la mesure par l'erreur quadratique moyenne :
$EQM(\Theta) = \mathbb{E}(\Theta - \theta)^2 = \mathbb{E}\left(\Theta - \mathbb{E}(\Theta)\right)^2 + \mathbb{E}\left(\mathbb{E}(\Theta) - \theta\right)^2$ soit

$$EQM(\Theta) = Variance + Biais^2$$

Pour des estimateurs justes, EQM et variance sont confondues. Comparer deux estimateurs n'est pas simple. Par exemple, un estimateur peut être biaisé de variance faible, un autre peut être juste de grande variance. Il est préférable d'utiliser un estimateur biaisé car ne moyenne l'erreur sera systématique mais faible d'un côté, aléatoire mais pouvant être forte de l'autre.



Pour deux estimateurs sans biais, le meilleur est celui qui a la variance minimale.

5.3 La convergence

La troisième propriété des estimateurs et peut-être la plus importante est la **convergence**.

$$\mathbb{E}(A) = a\,V(A) = \frac{\sigma^2}{nv(\mathbf{x})} \quad \text{et} \quad \mathbb{E}(B) = b\,V(B) = \sigma^2\left(\frac{1}{n} + \frac{m(\mathbf{x})^2}{nv(\mathbf{x})}\right)$$

Dans les deux cas, quand $n \to \infty$, les variances de ces estimateurs tendent vers 0. Donc, pour un niveau d'erreur quelconque, il existe un n à partir duquel Θ_n (l'estimateur avec n valeurs) sera aussi près que l'on veut de θ (la vraie valeur). Cette propriété s'écrit :

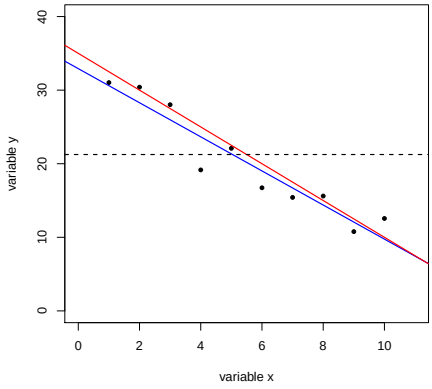
$$\forall \alpha \in \mathbb{R}^+ \quad \lim_{x \to +\infty} P\left(|\Theta_n - \theta| < \alpha\right) = 1$$

convergence en probabilités.

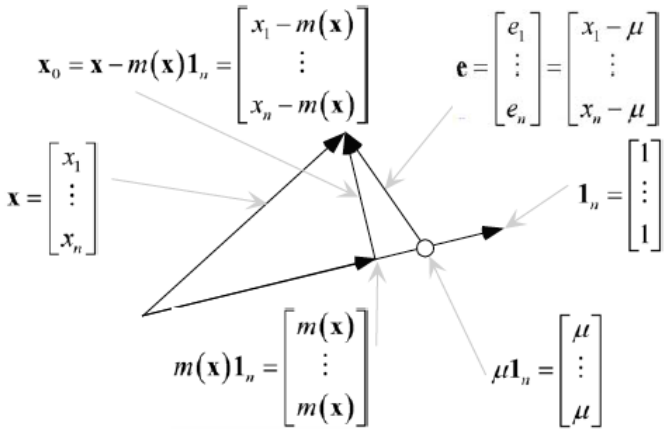
Un estimateur sans biais dont la variance tend vers 0 est convergent.

6 Théorème de Cochran

En résumé, on a créé un jeu de données (points noirs) à partir d'un modèle (droite rouge). La variable Y décroit linéairement avec X (droite bleue). C'est dans le modèle.



Mais quand on fait une expérience, comment le prouver ? La base est déjà en oeuvre dans l'étude d'un échantillon aléatoire simple d'une loi normale.



Soit $\mathbf{x} = (x_1, \ldots, x_n)$ un échantillon aléatoire simple d'une loi normale de moyenne μ et de variance σ^2. Chacun des x_i est la réalisation d'une variable aléatoire X_i de même loi. Ces variables sont indépendantes deux à deux. Le vecteur $\mathbf{x}$ est donc lui-même la réalisation d'une variable aléatoire $X = (X_1, \ldots, X_n)$ dont chaque composante est gaussienne de moyenne $(\mu_1, \ldots, \mu_n)$. Sa matrice de variances-covariances est $\sigma^2\mathbf{I}_n$ où $\mathbf{I}_n$ est la matrice identité de rang n. La figure représente, ensemble, le vecteur des erreurs $\mathbf{e}$ et la variable centrée $\mathbf{x}_0$, le premier utilisant la vraie moyenne inconnue μ, le second utilisant la moyenne empirique $m(\mathbf{x})$, estimation de la précédente. On a donc géométriquement, de par le théorème de Pythagore :

$$\|\mathbf{e}\|^2 = \|\mathbf{x}_0\|^2 + (m(\mathbf{x}) - \mu)^2\|\mathbf{1}_n\|^2$$

Jusque-là, on utilisait, dans l'espace des variables, le produit scalaire associé à la pondération uniforme (ou plus généralement une pondération quelconque). Maintenant, on utilise le produit scalaire **canonique** :

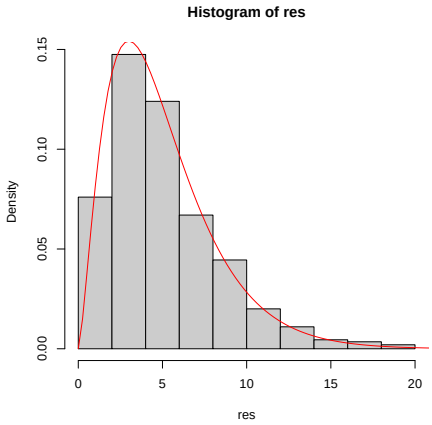
$$\langle \mathbf{x}|\mathbf{y}\rangle = \sum_{i=1}^{n} x_i y_i \;\Rightarrow\; \|\mathbf{x}\|^2 = \sum_{i=1}^{n} x_i^2$$

Au facteur $\frac{1}{n}$ près, cela ne change pas grand chose mais est indispensable pour intégrer l'aspect probabiliste.

• $\|\mathbf{e}\|^2$ est maintenant une variable aléatoire qui prend une valeur. Cette variable aléatoire est la somme des carrés de n variables normales indépendantes de moyenne 0 et de variance σ^2.

$$\frac{\|\mathbf{e}\|^2}{\sigma^2} = \sum_{i=1}^{n} \left(\frac{x_i - \mu}{\sigma}\right)^2$$

Cette quantité est la réalisation d'une variable aléatoire, somme des carrés de n lois normales centrées réduites, donc la réalisation d'une loi du Chi-Deux à n degrés de liberté (χ^2_n).



Cette loi a pour moyenne n et pour variance $2n$.

• La quantité $(m(\mathbf{x})-\mu)^2\|\mathbf{1}_n\|^2$ est en partie aléatoire. $\|\mathbf{1}_n\|^2$ vaut n et $m(\mathbf{x})-\mu$ est la réalisation d'une variable aléatoire de moyenne 0 et de variance $\frac{\sigma^2}{n}$. $\frac{(m(\mathbf{x})-\mu)^2\|\mathbf{1}_n\|^2}{\sigma^2}$ est donc la réalisation du carré d'une loi normale centrée réduite soit une loi du Chi-Deux à 1 degré de liberté (χ^2_1).

• La quantité $\sum_{i=1}^{n}(x_i - m(\mathbf{x}))^2$ pose un problème beaucoup plus délicat. C'est une réalisation d'une variable aléatoire qui s'écrit :

$$Z = \sum_{i=1}^{n} \left(X_i - \frac{1}{n}\sum_{k=1}^{n} X_k\right)^2$$

$*$ Lorsque $n = 2$, on a :

$$Z = \left(X_1 - \frac{(X_1+X_2)}{2}\right)^2 + \left(X_2 - \frac{(X_1+X_2)}{2}\right)^2 = \frac{(X_1-X_2)^2}{2}$$

C'est le carré d'une loi normale de variance σ^2. Donc, $\frac{Z}{\sigma^2}$ suit une loi du Chi-Deux à un degré de liberté.

$*$ Lorsque $n = 3$, on a :

$$\begin{aligned} Z &= \left(X_1 - \frac{(X_1+X_2+X_3)}{3}\right)^2 + \left(X_2 - \frac{(X_1+X_2+X_3)}{3}\right)^2 + \left(X_3 - \frac{(X_1+X_2+X_3)}{3}\right)^2 \\ &= \left(\frac{2}{3}X_1 - \frac{1}{3}X_2 - \frac{1}{3}X_3\right)^2 + \left(-\frac{1}{3}X_1 + \frac{2}{3}X_2 - \frac{1}{3}X_3\right)^2 + \left(-\frac{1}{3}X_1 - \frac{1}{3}X_2 + \frac{2}{3}X_3\right)^2 \\ &= \frac{2}{3}X_1^2 + \frac{2}{3}X_2^2 + \frac{2}{3}X_3^2 - \frac{2}{3}X_1X_2 - \frac{2}{3}X_1X_3 - \frac{2}{3}X_2X_3 \end{aligned}$$

On fait alors une opération très simple et très particulière :

$$\begin{aligned} \frac{3}{2}Z &= \left(X_1^2 - X_1X_2 - X_1X_3\right) + \left(X_2^2 + X_3^2 - X_2X_3\right) \\ &= \left(X_1 - \frac{1}{2}X_2 - \frac{1}{2}X_3\right)^2 - \frac{1}{4}X_2^2 - \frac{1}{4}X_3^2 - \frac{1}{2}X_2X_3 + (X_2^2 + X_3^2 - X_2X_3) \end{aligned}$$

Finalement,

$$Z = \frac{2}{3}\left(X_1 - \frac{1}{2}X_2 - \frac{1}{2}X_3\right)^2 + \frac{1}{2}(X_2 - X_3)^2$$

Le premier terme contient le carré d'une loi normale de variance σ^2 et le second aussi. La covariance entre les deux est nulle et on a donc que $\dfrac{Z}{\sigma^2}$ suit une loi du Chi-Deux à deux degrés de liberté. Ceci introduit au fait qu'il s'agit d'un problème d'algèbre. Z est la somme de trois carrés de lois normales qui devient la somme de carrés de deux lois normales indépendantes. Ces écritures sont des polynômes homogènes de degré 2 c'est-à-dire des formes quadratiques.

$Z = \sum_{i=1}^{n}(X_i - \mu)^2$ est une somme de n carrés de lois normales indépendantes mais $Z = \sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{i=1}^{n}X_i\right)^2$ est une somme de $n-1$ carrés de lois normales indépendantes.

Finalement,

$$\|\mathbf{e}\|^2 = \|\mathbf{x}_0\|^2 + (m(\mathbf{x}) - \mu)^2\|\mathbf{1}_n\|^2 \quad \text{vecteurs}$$

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 \quad \text{valeurs observées}$$

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{i=1}^{n}X_i\right)^2 + n\left(\frac{1}{n}\sum_{i=1}^{n}X_i - \mu\right)^2 \quad \text{variables aléatoires}$$

$$\underbrace{\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2}}_{\chi^2_n} = \underbrace{\frac{\sum_{i=1}^{n}\left(X_i - \frac{1}{n}\sum_{i=1}^{n}X_i\right)^2}{\sigma^2}}_{\chi^2_{n-1}} + \underbrace{\frac{n\left(\frac{1}{n}\sum_{i=1}^{n}X_i - \mu\right)^2}{\sigma^2}}_{\chi^2_1} \quad \text{lois}$$

En résumé, soit $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ un échantillon aléatoire simple d'une variable normale centrée réduite dans un tirage de la variable aléatoire $Y = (Y_1, Y_2, \ldots, Y_n)$ dont les composantes sont normales centrées réduites, indépendantes deux à deux. Si $\mathcal{H}$ est un sous-espace vectoriel de dimension p, $\mathbf{y}$ se décompose en deux vecteurs orthogonaux :

$$\mathbf{y} = \mathbf{y}_{\mathcal{H}} + \mathbf{y}_{\mathcal{H}^\top}$$

D'après le théorème de Pythagore,

$$\|\mathbf{y}\|^2 = \|\mathbf{y}_{\mathcal{H}}\|^2 + \|\mathbf{y}_{\mathcal{H}^\top}\|^2$$

D'après la définition de la loi du Chi-Deux, $\|\mathbf{y}\|^2$ est la réalisation d'une variable aléatoire $\sum_{i=1}^{n} Y_i^2$ qui suit une loi χ^2_n. Le théorème de Cochran indique que $\|\mathbf{y}_{\mathcal{H}}\|^2$ est la réalisation d'une variable aléatoire qui suit une loi χ^2_p ; $\|\mathbf{y}_{\mathcal{H}^\top}\|^2$ est la réalisation d'une variable aléatoire qui suit une loi χ^2_{n-p} et ces deux variables sont indépendantes.

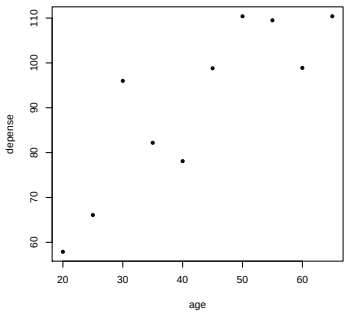
Généralisation.

> Soit $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ un échantillon aléatoire simple d'une variable normale centrée réduite.
> L'espace $\mathbb{R}^n$ se décompose en K sous-espaces vectoriels $\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_K$ de dimensions respectives $p_1, p_2, \ldots, p_K$ avec $p_1 + p_2 + \cdots + p_K = n$.
> $\mathbf{y}$ se décompose en K vecteurs orthogonaux deux à deux : $\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2 + \cdots + \mathbf{y}_K$.
> D'après le théorème de Pythagore, $\|\mathbf{y}\|^2 = \|\mathbf{y}_1\|^2 + \|\mathbf{y}_2\|^2 + \cdots + \|\mathbf{y}_K\|^2$.
> $\|\mathbf{y}\|^2$ suit une loi χ^2_n et le théorème de Cochran indique que $\|\mathbf{y}_j\|^2$ suit une loi $\chi^2_{p_j}$ et que les variables sont indépendantes.

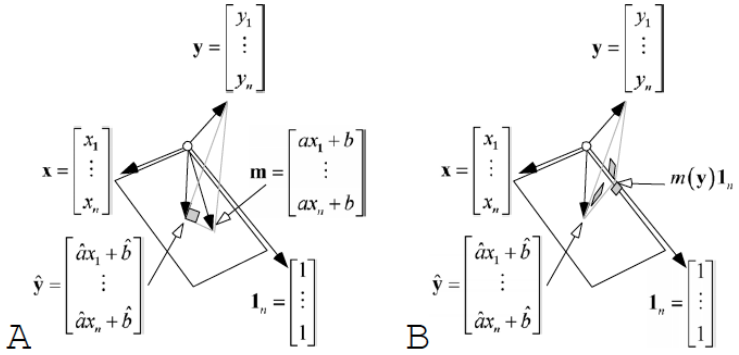
7 Application : la régression simple

On connait la dépense de santé moyenne pour le poste "maladie" de 20 à 65 ans.

```
age <- seq(20,65,by=5)
depense <- c(57.9,66.1,96,82.2,78.1,98.8,110.4,109.5,98.9,110.4)
plot(age,depense,pch=20)
```



On note $\mathbf{x}$ le vecteur âge et $\mathbf{y}$ le vecteur dépense pour le poste "maladie", tous deux vecteurs de $\mathbb{R}^n$ avec $n = 10$. $\mathbb{R}^n$ se décompose en trois sous-espaces vectoriels : $\mathbf{1}_n$ de dimension 1, $\mathbf{x}_0$ de dimension 1 et le complémentaire $\mathcal{R}$ de dimension $n - 1 - 1$ soit $n - 2$.
On applique le théorème de Cochran (partie A de la figure ci-dessous) :



Si le modèle de la régression linéaire simple est vérifié, on a :

$$Y_i \to \mathcal{N}(\mu, \sigma) \text{ avec } \mu_i = ax_i + b$$

C'est le **modèle vrai**.

La somme des carrés des écarts au modèle vrai suit une loi du Chi-Deux :

$$\frac{\sum_{i=1}^{n} (Y_i - (ax_i + b))^2}{\sigma^2} \to \chi^2_n$$

La somme des carrés des écarts au **modèle estimé** suit une loi du Chi-Deux :

$$\frac{\sum_{i=1}^{n} (Y_i - (Ax_i + B))^2}{\sigma^2} \to \chi^2_{n-2}$$

```
options(show.signif.stars=FALSE)
lmrs <- lm(depense~age)
lmrs
```

```
Call:
lm(formula = depense ~ age)
Coefficients:
(Intercept)          age
      45.76         1.06
```

On peut alors calculer les valeurs prédites par le modèle :

```
predict(lmrs)
```

```
       1        2        3        4        5        6        7        8
66.96909 72.27152 77.57394 82.87636 88.17879 93.48121 98.78364 104.08606
       9       10
109.38848 114.69091
```

Le vecteur des observations et le vecteur des prédictions se projettent au même endroit sur le vecteur des constantes (théorème des trois perpendiculaires, partie B de la figure ci-dessus).

On a toujours la relation :

$$\sum_{i=1}^{n} (y_i - m(\mathbf{y}))^2 = \sum_{i=1}^{n} (y_i - \widehat{y_i})^2 + \sum_{i=1}^{n} (\widehat{y_i} - m(\mathbf{y}))^2$$

c'est-à-dire

$$\begin{array}{ccccc} \text{Somme} & & \text{Somme} & & \text{Somme} \\ \text{des} & = & \text{des} & + & \text{des} \\ \text{carrés totale} & & \text{carrés résiduelle} & & \text{carrés expliquée} \end{array}$$

soit encore

$$\text{Variance des données} = \text{Erreur} + \text{Variance du modèle}$$

```
 sum((depense-mean(depense))^2)
[1] 3202.401
 sum((depense-predict(lmrs))^2)
[1] 882.8555
 sum((predict(lmrs)-mean(depense))^2)
[1] 2319.545
```

L'hypothèse nulle H_0 se caractérise par le fait que la variable $\mathbf{x}$ est sans effet sur la variable $\mathbf{y}$.
Cela signifie que la pente est nulle ($a = 0$). La situation se décrit alors par :

$$Y_i \to \mathcal{N}(b, \sigma)$$

La somme des carrés des écarts au modèle vrai suit une loi χ^2_n :

$$\frac{\sum_{i=1}^{n} (Y_i - b)^2}{\sigma^2} \to \chi^2_n$$

La somme des carrés des écarts au modèle estimé suit une loi χ^2_-1 :

$$\frac{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}{\sigma^2} \to \chi^2_n$$

Que la vraie valeur de a soit nulle ou non, cela n'enlève rien au résultat :

$$\frac{\sum_{i=1}^{n} (Y_i - (Ax_i + B))^2}{\sigma^2} \to \chi^2_{n-2}$$

La décomposition en parties orthogonales - théorème de Cochran - implique que :

$$\frac{\sum_{i=1}^{n} \left(Ax_i + B - \bar{Y}\right)^2}{\sigma^2} \to \chi^2_1$$

Les deux variables sont indépendantes.

On a donc :

$$\frac{(n-2)\sum_{i=1}^{n}\left(Ax_i+B-\bar{Y})\right)^2}{\sum_{i=1}^{n}\left(Y_i-(Ax_i+B)\right)^2} \rightarrow F_{(1,n-2)}$$

que l'on retrouve dans le tableau de décomposition de la variance :

```
 anova(lmrs)
```

```
Analysis of Variance Table
Response: depense
          Df  Sum Sq Mean Sq F value   Pr(>F)
age        1 2319.55 2319.55  21.019 0.001791
Residuals  8  882.86  110.36
```

```
 plot(age,depense,pch=20)
 abline(lmrs, col="red")
 shapiro.test(lmrs$residuals)
```

```
        Shapiro-Wilk normality test
data:  lmrs$residuals
W = 0.91268, p-value = 0.2999
```