10 Point de vue inférentiel

On peut supposer que x et y sont les observations d'un échantillon des variables X et Y. On écrit donc le modèle

$$Y = \beta_1 X + \beta_0 + \varepsilon$$ / erreur aléatoire.

Les valeurs β_1 et β_0 calculées ci-dessus sont en réalité les estimations $\hat{\beta}_1$ et $\hat{\beta}_0$ par la méthode des moindres carrés, *i.e.* minimisant la somme des carrés des écarts (par rapport à la droite)

$$SCE = \sum_i \varepsilon_i^2.$$

Echantillon: (X_i, Y_i) i.i.d.

avec $Y_i = \beta_1 X_i + \beta_0 + \varepsilon_i$ ← erreur aléatoire.

β_1, β_0 : paramètres

⚠ estimation $\Longrightarrow$ $Var(X) = \frac{1}{N-1} \sum_i (X_i - \bar{X})^2$

$Var(Y) = \frac{1}{N-1} \sum_i (Y_i - \bar{Y})^2$

$Cov(X,Y) = \frac{1}{N-1} \sum_i (X_i - \bar{X})(Y_i - \bar{Y})$

on s'en fout !

On a alors

$$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2},$$

et

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

On notera alors les valeurs prédites

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ valeurs sur la droite.

et $\hat{Y}_i$ est la valeur associée à X_i par la droite de régression (dite empirique). Il vient que

$$SCE = \sum_i (\hat{Y}_i - Y_i)^2.$$

prédite → observée.

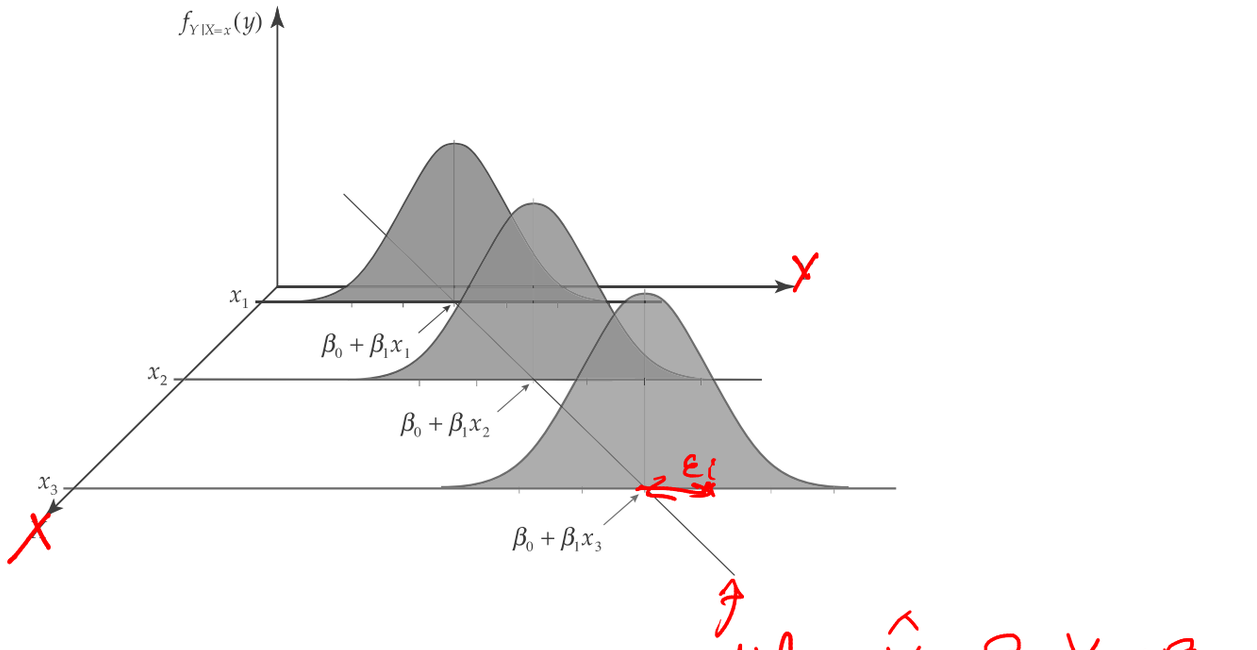
$Y_i = \beta_1 X_i + \beta_0 + \varepsilon_i$ → variable aléatoire.

Variable aléatoire ← Y_i considérée comme pas aléatoire car pas d'erreur.

$EY, VarY$ sont à estimer — $\bar{X}, VarX$ est à considérer au sens statistique.

10.1 Hypothèses sur les termes d'erreur ε

- Indépendance des erreurs : les $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$ sont indépendants.
- Éxogénéité : les variables explicatives $(X_1, \cdots X_n)$ ne sont pas corrélées au terme d'erreur. De plus, les erreurs sont centrées $E(\varepsilon_i) = 0$
- Homoscédasticité : les termes d'erreurs sont supposés de variance constante. $Var(\varepsilon_i) = \sigma^2$
- Normalité des termes d'erreur : les termes d'erreurs suivent une loi normale, centrées, de variance σ^2



modèle $\hat{Y} = \beta_1 X + \beta_0$

Lemma 10.1 *Les hypothèses du modèle montrent que*

$$Y_i = \beta_1 X_i + \beta_0 + \varepsilon_i$$

suit une loi normale $\mathcal{N}(\beta_1 X_i + \beta_0, \sigma^2)$. *De plus les* Y_i *sont indépendants.*

Note : $\overline{\hat{Y}} = \overline{\beta_1 X + \beta_0} = \beta_1 \bar{X} + \beta_0 + \bar{\varepsilon} = \bar{Y} \Rightarrow \bar{Y} = \bar{\hat{Y}}$

$\sum \varepsilon_i = 0$

10.2 Équation de la variance

D'après les hypothèse précédentes, il vient que

modèle erreur

$$Var(Y) = Var(\hat{Y}) + Var(\varepsilon),$$

indépendance car les termes d'erreur ε_i sont non corrélés aux X_i, donc non corrélés aux $\hat{Y}_i = \beta_1 X_i + \beta_0$

$Y = \hat{Y} + \varepsilon$.

grâce à l'éxongénéité. Il vient que

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

SCT

$Var(\hat{Y}) = \frac{1}{N-1} \sum_i (\hat{Y}_i - \bar{\hat{Y}})^2 = \frac{1}{N-1} \sum_i (\hat{Y}_i - \bar{Y})^2$

SCM = somme des carrés due au modèle

SCT = somme des carrés totale.

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

Cette équation a une signification intéssante.

- Le terme $\sum_i (Y_i - \bar{Y})^2$ représente la variation totale des valeurs des Y_i par rapport à leur moyenne $\bar{Y}$. On notera cette quantité SCT : *Somme des Carrés Totale*.
- Puisque le terme $\sum_i (\hat{Y} - \bar{Y})^2$ est l'écart de la valeur prédite par rapport à la moyenne, nous dirons que ce terme est la *Somme des Carrés due au Modèle*, notée SCM[4].

On a donc

$$SCT = SCM + SCE. \tag{10.1}$$

10.3 Coefficient de détermination

Le coefficient de détermination est le rapport de variance de Y expliquée par la régression :

$$R^2 = \frac{Var(\hat{Y})}{Var(Y)} = \frac{\frac{1}{N-1}\sum(\hat{Y}_i - \bar{Y})^2}{\frac{1}{N-1}\sum(Y_i - \bar{Y})^2} = \frac{SCM}{SCT}$$

Le coefficient R^2 est donc la proportion de variance de Y expliquée par le modèle. Dès lors que $\bar{\hat{Y}} = \bar{Y}$, on peut formuler

$$(r_{xy})^2 = \frac{(Cov(X,Y))^2}{Var X \; Var Y} = \frac{(Cov(X, B_1X + B_0 + \varepsilon))^2}{Var X \; Var Y} = \frac{(B_1 Cov(X,X) + Cov(X,\varepsilon))^2}{Var X \; Var Y}$$

$$= \frac{B_1^2 Var(X)^2}{Var X \; Var Y} = \frac{B_1^2 Var X}{Var Y} = \frac{Var(B_1X + B_0)}{Var Y} = \frac{Var\hat{Y}}{Var Y} = R^2.$$

$$Var(aX+b) = a^2 Var X$$

Rappels sur le modèle :

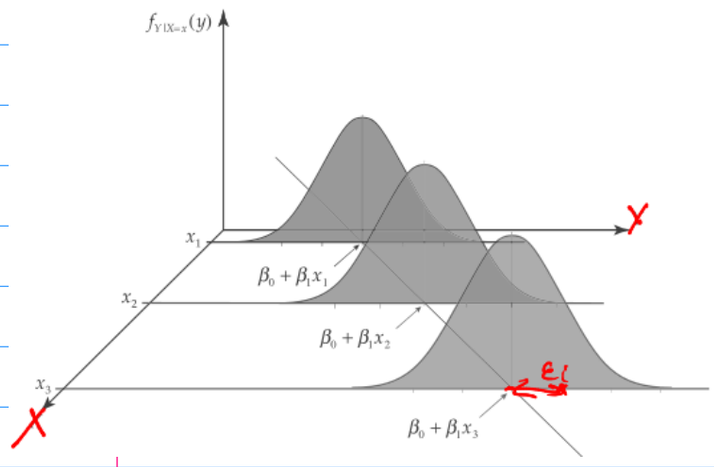
Sur échantillon : $(X_i, Y_i)_{1 \le i \le n}$.

Modèle : $Y_i = \beta_1 X_i + \beta_0 + \varepsilon_i$

- β_1, β_0 : paramètres du modèle. Non aléatoires.
- Pas d'aléa sur X_i
- ε_i → aléa sur l'erreur : $\varepsilon_i \sim \mathcal{N}(0, \sigma)$ i.i.d.

$$Y_i = \underbrace{\beta_1 X_i + \beta_0}_{\text{pas aléatoire}} + \underbrace{\varepsilon_i}_{\mathcal{N}(0,\sigma)} \Rightarrow Y_i \sim \mathcal{N}(\beta_1 X_i + \beta_0, \sigma)$$

indépendantes.



Estimons :

$$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}$$

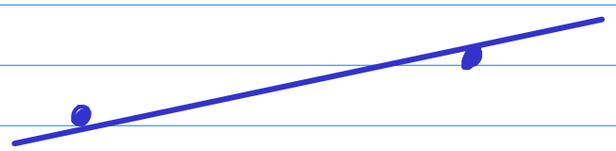
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

Estimons σ_ε^2 : variance de l'erreur.

$$S_\varepsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n}(\varepsilon_i - \bar{\varepsilon})^2 = \frac{1}{n-2}\sum_{i=1}^{n}\varepsilon_i^2 = \frac{1}{n-2}\sum_i (y_i - \hat{y}_i)^2 = \frac{SCE}{n-2}$$

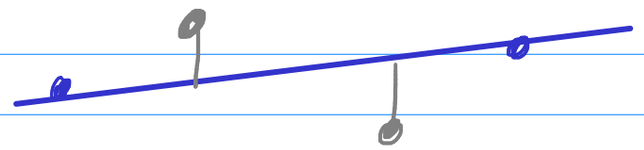
($\bar{\varepsilon_i} = 0$)

$n > 2$ (pour avoir une droite). Comptons le nombre d'erreurs.



$n = 2$ 0 erreurs

$n = 3$ 1 erreurs



$n = 4$ 2 erreurs

⋮

n $n-2$ erreurs

Note : l'estimateur de σ_ε^2 est $S_\varepsilon^2 = \frac{SCE}{n-2}$

10.4 Distribution de $\hat{\beta}_1$

La méthode des moindres carrés prend le parti de ne pas considérer d'erreur sur les valeurs x_i prises par X. Il vient qu'on peut considérer l'ensemble de valeurs $X_i = x_i$ qui seront non aléatoires et le modèle équivalent :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

Les termes ε_i sont des variables aléatoires normales identiques et indépendantes de moyenne nulle et variance σ_ε^2, quelque soit la valeur X_i. On en déduit le théorème suivant qui donne la distribution de $\hat{\beta}_1$.

Theorem 10.2 *Sous les hypothèses du modèle de régression linéaire simple, $\hat{\beta}_1$ suit une loi normale d'espérance β_1 et de variance*

$$\frac{\sigma_\varepsilon^2}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{X})^2$$

Lemma 10.3 *On pose*

$$a_i = \frac{X_i - \bar{X}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

$$\hat{\beta}_1 = \sum_{i=1}^{n} a_i Y_i,$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{S_{xx}}$$

$$= \sum_{i=1}^{n} \frac{(X_i - \bar{X})Y_i}{S_{xx}} - \bar{Y}\sum_i \frac{(X_i - \bar{X})}{S_{xx}}$$

$$= \sum_{i=1}^{n} a_i Y_i - \bar{Y} \sum_{i=1}^{n} a_i$$

Nous avons

et

$$\sum_{i=1}^{n} a_i = 0, \qquad \sum_{i=1}^{n} a_i^2 = \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \qquad \sum_{i=1}^{n} a_i X_i = 1.$$

$$\sum_i a_i = \sum_{i=1}^{n} \frac{(X_i - \bar{X})}{S_{xx}} = \frac{1}{S_{xx}} \sum_i (X_i - \bar{X})$$

$$= \frac{1}{S_{xx}} \sum_i X_i - \frac{\sum_i \bar{X}}{S_{xx}} = \frac{1}{S_{xx}} \sum_i X_i - \frac{n\bar{X}}{S_{xx}}$$

$$= \frac{1}{S_{xx}} \sum_i X_i - \frac{n}{S_{xx}} \frac{\sum_i X_i}{n} = 0$$

$$\sum_i a_i^2 = \sum_i \left(\frac{X_i - \bar{X}}{S_{xx}}\right)^2$$

$$= \frac{1}{S_{xx}^2} \sum_i (X_i - \bar{X})^2$$

$$= \frac{S_{xx}}{S_{xx}^2} = \frac{1}{S_{xx}}$$

$$\sum_i \frac{(X_i - \bar{X})}{S_{xx}} (X_i - \bar{X}).$$

$$= \sum_i \frac{(X_i - \bar{X})}{S_{xx}} X_i - \bar{X} \sum_i \frac{(X_i - \bar{X})}{S_{xx}}$$

$$= \sum_i a_i X_i$$

$$\frac{S_{xx}}{S_{xx}} = 1$$

$$\sum_i a_i = 0$$

Remarque : • les $a_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$ ne dépendent que de X, ne sont pas aléatoires.

• Pour rappel : $Y_i \sim \mathcal{N}(\beta_1 X_i + \beta_0, \sigma)$ indépendantes.

Proof. Notons que $\hat{\beta}_1 = \sum_{i=1}^{n} a_i Y_i$ suit une loi normale en tant que combinaison linéaire de $Y_1, \cdots Y_n$, variables normales indépendantes. Calculons son espérance.

espérance linéaire

$$E\hat{\beta}_1 = \sum_{i=1}^{n} a_i EY_i = \sum_{i=1}^{n} a_i \underbrace{(\beta_1 X_i + \beta_0)}_{EY_i}$$

$$= \beta_1 \sum_{i=1}^{n} a_i X_i + \beta_0 \underbrace{\sum_i a_i}_{0}$$

$$= \beta_1$$

$E\hat{\beta}_1 = \beta_1$

estimateur non biaisé.

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\left(\sum_i a_i Y_i\right)$$

$$= \sum_i a_i^2 \mathrm{Var}(Y_i)$$

$\hookrightarrow \sigma_\varepsilon^2$

$$= \sigma_\varepsilon^2 \sum_i a_i^2$$

$$= \sigma_\varepsilon^2 \times \frac{1}{\sum_i (X_i - \bar{X})^2}$$

les X_i sont indépendantes

a_i sont constantes.

$$\mathrm{Var}(aX+bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y)$$

si X sont indépendantes

Donc $\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1; \sqrt{\frac{\sigma_\varepsilon^2}{\sum_i (X_i - \bar{X})^2}}\right)$

On fait les remarques suivantes.

- Ce théorème montre que $\hat{\beta}_1$ est sans biais pour β_1. Cela signifie que d'un échantillon à l'autre, la valeur de $\hat{\beta}_1$ oscille autour de la valeur théorique β_1.

- Ces écarts par rapport à la moyenne β_1 sont distribués selon une loi normale dont la variance est

$$\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_\varepsilon}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

  On note, donc, que la variance de $\hat{\beta}_1$ croît avec σ^2_ε, mais qu'elle décroît lorsque $\sum_{i=1}^n (X_i - \bar{X})^2$ croît. Ainsi, plus les X_i sont nombreux et dispersés, plus notre estimation sera fiable.

- On acceptera que la distribution de $\hat{\beta}_0$ est normale et suit la loi

$$\mathcal{N}\left(\beta_0, \sqrt{\sigma^2_\varepsilon \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right]}\right).$$

En pratique nous ne connaissons pas σ_ε^2, la variance de la variable ε, qui est nécessaire dans le calcul de $\sigma_{\hat{\beta}_1}$. Cependant, nous disposons d'une estimation de celle-ci, à savoir :

$$s_\varepsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n}\varepsilon_i^2 = \frac{SCE}{n-2}.$$

Il vient que

$$(n-2)\frac{s_\varepsilon^2}{\sigma_\varepsilon^2} \backsim \chi^2_{n-2}.$$

On en conclut que le rapport $(\hat{\beta}_1 - \beta_1)/s_{\hat{\beta}_1}$ n'est pas distribué selon une loi $\mathcal{N}(0,1)$, mais plutôt selon une loi t de Student ayant $n-2$ degrés de liberté. On a donc les distributions suivantes.

Theorem 10.4 *En estimant σ_ε par s_ε, on obtient les distributions des estimateurs suivantes*

$$(\hat{\beta}_0 - \beta_0)/s_{\hat{\beta}_0} \backsim t_{n-2}, \qquad (\hat{\beta}_1 - \beta_1)/s_{\hat{\beta}_1} \backsim t_{n-2},$$

avec

$$s^2_{\hat{\beta}_1} = \frac{SCE}{(n-2)\sum_{i=1}^n (X_i - \bar{X})^2},$$

et

$$s^2_{\hat{\beta}_0} = \frac{SCE}{(n-2)}\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right].$$

Si le nombre de degrés de liberté est assez élevé (plus de trente), on peut faire une approximation de la loi t de Student par une loi $\mathcal{N}(0,1)$.

10.5 Intervalles de confiance

10.5.1 Intervalles de confiances des coefficients de la régression

Le dernier théorême de la section précédente, donne les distributions suivantes.

$$(\hat{\beta}_0 - \beta_0)/s_{\hat{\beta}_0} \backsim t_{n-2}, \qquad (\hat{\beta}_1 - \beta_1)/s_{\hat{\beta}_1} \backsim t_{n-2}, \qquad (n-2)\frac{s_\varepsilon^2}{\sigma_\varepsilon^2} \backsim \chi^2_{n-2}.$$

On en déduit les intervalles de confiance à $100(1-\alpha\%)$ suivants.

$$\frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \sim t_{n-2}$$

$q^{n-2}_{1-\alpha/2}$ quantile d'ordre $1-\alpha/2$ de t_{n-2}

$$\hat{\beta}_1 - q^{n-2}_{1-\alpha/2} s_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + q^{n-2}_{1-\alpha/2} s_{\hat{\beta}_1}$$

$$\hat{\beta}_1 - q^{n-2}_{1-\alpha/2} \sqrt{\frac{SCE}{(n-2)\sum_i (x_i - \bar{x})^2}} \leq \beta_1 \leq \hat{\beta}_1 + q^{n-2}_{1-\alpha/2} \sqrt{\frac{SCE}{(n-2)\sum_i (x_i - \bar{x})^2}}$$

$$\frac{(n-2)s_\varepsilon^2}{q_{\alpha/2}^{\chi^2(n-2)}} \geq \sigma_\varepsilon^2 \geq \frac{(n-2)s_\varepsilon^2}{q_{1-\alpha/2}^{\chi^2(n-2)}},$$

$$\hat{\beta}_0 - q_{1-\alpha/2}^{t_{n-2}} s_{\hat{\beta}_0} \leq \beta_0 \leq \hat{\beta}_0 + q_{1-\alpha/2}^{t_{n-2}} s_{\hat{\beta}_0},$$

$$\hat{\beta}_1 - q_{1-\alpha/2}^{t_{n-2}} s_{\hat{\beta}_1} \leq \beta_1 \leq \hat{\beta}_1 + q_{1-\alpha/2}^{t_{n-2}} s_{\hat{\beta}_1}$$

où la valeur $q_{1-\alpha/2}^{t_{n-2}}$ est le quantile d'ordre $1-\alpha/2$ d'une loi de Student à $n-2$ degrés de liberté (obtenu de la table de la loi de Student).

Exemple: X l'âge. Y x cholestérol.

Le modèle permet d'estimer:

- le taux de cholestérol moyen pour les individus de 50 ans: $\hat{Y}_{50} = \hat{\beta}_0 + \hat{\beta}_1 X_{50}$
- le taux de cholestérol pour l'individu de 50 ans face à moi $Y_{50} = \hat{\beta}_0 + \hat{\beta}_1 X_{50} + \hat{\varepsilon}_{50}$

on ne peut que l'estimer à 0 puisque l'on en est centré.

En terme d'estimation ponctuelle, de prédiction, les deux valeurs prédites sont égales. Mais en terme d'intervalle de confiance sur les prévisions, l'intervalle sur l'estimation moyenne est plus resserré que l'intervalle sur l'estimation d'une personne.

Valeurs prédites : $\hat{Y}_i = \hat{\beta}_1 X_i + \hat{\beta}_0$

10.5.2 Intervalles pour les prévisions

Lorsque nous substituons dans l'équation de la droite de régression une valeur donnée de X, soit X_0, nous obtenons une certaine valeur que nous notons :

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0.$$

Cette valeur de X_0 peut être utilisée à deux fins, car elle estime deux choses :

- $E[Y_0] = \beta_0 + \beta_1 X_0$, c'est-à-dire la moyenne de la variable Y en $X = X_0$;
- $E[Y_0] + \varepsilon = \hat{\beta}_0 + \hat{\beta}_1 X_0 + \varepsilon$, c'est-à-dire une observation de Y pour un individu ayant en $X = X_0$.

Lorsque l'on fait de telles prévisions, on préfère accompagner celles-ci de limites de confiance.

En $X = X_0$

$\hat{Y}_0 = \hat{\beta}_1 X_0 + \hat{\beta}_0$

(estimation de Y_0 moyen).

(i) Dans le premier cas, lorsque nous voulons estimer la moyenne de la variable Y lorsque la valeur de X demeure fixée à X_0, nous utilisons l'intervalle à $100\alpha\%$ de confiance suivant :

$$\hat{Y}_0 \pm q_{1-\alpha/2}^{t_{n-2}} \sqrt{s_\varepsilon^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right)}.$$

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

$t_{1-\alpha/2}$

Proof. On utilisera les notations et les résultat du Lemme 10.3. Pour rappel,

$$a_i = \frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Nous avons

$$\hat{\beta}_1 = \sum_{i=1}^n a_i Y_i,$$

et

$$\sum_{i=1}^n a_i = 0, \qquad \sum_{i=1}^n a_i^2 = \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2}, \qquad \sum_{i=1}^n a_i X_i = 1.$$

$\bar{Y} = \hat{\beta}_1 \bar{X} + \hat{\beta}_0 \Leftrightarrow \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$.

$\hat{Y}_i = \hat{\beta}_1 X_i + \hat{\beta}_0 = \hat{\beta}_1 X_i + \bar{Y} - \hat{\beta}_1 \bar{X} = \bar{Y} + \hat{\beta}_1 (X_i - \bar{X})$.

nous obtenons,

$$\hat{Y}_0 = \bar{Y} + \hat{\beta}_1(X_0 - \bar{X}).$$

Puisque

$$\hat{\beta}_1 = \sum_{i=1}^{n} a_i Y_i,$$

nous avons,

$$\hat{\beta}_1 = \sum_{i=1}^{n} a_i(\beta_1 X_i + \beta_0 + \varepsilon_i) = \beta_1 \sum_{i=1}^{n} a_i X_i + \beta_0 \sum_{i=1}^{n} a_i + \sum_{i=1}^{n} a_i \varepsilon_i.$$

Il suit des égalités sur les sommes des a_i que

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^{n} a_i \varepsilon_i.$$

Par conséquent,

$$\begin{aligned} Var(\hat{Y}_0) &= Var\left[\bar{Y} + \hat{\beta}_1(X_0 - \bar{X})\right] \\ &= Var\left[\beta_0 + \beta_1\bar{X} + \bar{\varepsilon} + (X_0 - \bar{X})\left(\beta_1 + \sum_{i=1}^{n} a_i\varepsilon_i\right)\right] \\ &= Var\left[\bar{\varepsilon} + (X_0 - \bar{X})\sum_{i=1}^{n} a_i\varepsilon_i\right]. \end{aligned}$$

On voit alors que le coefficient de ε_i est

$$\frac{1}{n} + (X_0 - \bar{X})a_i.$$

$$Var(\hat{Y}_0) = Var\left(\sum_{i=1}^{n}\left(\frac{1}{n} + (X_0 - \bar{X})a_i\right)\varepsilon_i\right)$$

$$= \sum_{i=1}^{n}\left(\frac{1}{n} + (X_0 - \bar{X})a_i\right)^2 Var\,\varepsilon_i = \sigma_\varepsilon^2 \sum_{i=1}^{n}\left(\frac{1}{n} + (X_0 - \bar{X})a_i\right)^2$$

$$Var(\hat{Y}_0) = \sigma_\varepsilon^2 \sum_{i=1}^{n} \left[\frac{1}{n^2} + 2\frac{(x_0 - \bar{x})}{n} a_i + (x_0 - \bar{x})^2 a_i^2 \right]$$

$$= \sigma_\varepsilon^2 \left(\sum_{i=1}^{n} \frac{1}{n^2} + 2\frac{(x_0 - \bar{x})}{n} \underbrace{\sum a_i}_{0} + (x_0 - \bar{x})^2 \underbrace{\sum a_i^2}_{1/\sum (x_i - \bar{x})^2} \right)$$

$$= \sigma_\varepsilon^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right]$$

Cette variance est estimée par

$$Var(\hat{Y}_0) = s_\varepsilon^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right).$$

□

D'où l'intervalle à α de confiance

$$\hat{Y}_0 \in \left[\hat{Y}_{0_{obs}} - q^{n-2}_{1-\alpha/2} \; s_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \; ; \; \hat{Y}_{0_{obs}} + q^{n-2}_{1-\alpha/2} \; s_\varepsilon \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \right]$$

quantile de t_{n-2}

s_ε : $\frac{SCE}{n-2}$

$\hat{Y}_0 = \hat{\beta}_1 X_0 + \hat{\beta}_0 + \hat{\varepsilon}_i$

$\hat{\varepsilon}$ s'ajoute dans la variance une variance d'erreur.

(ii) Dans le second cas, il s'agit de prévoir, pour un individu donné, la valeur de Y qui lui est propre, sachant que sa valeur en X est X_0. L'intervalle est

$$\hat{Y}_0 \pm q^{t_{n-2}}_{1-\alpha/2} \sqrt{s^2_\varepsilon \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right)}.$$

+1

Proof. Il s'agit maintenant de trouver la variance de $\hat{\beta}_0 + \hat{\beta}_1 X_0 + \varepsilon$. On note que ε est une réplication du terme d'erreur indépendante des autres. Il suit

$$Var(\hat{\beta}_0 + \hat{\beta}_1 X_0 + \varepsilon) = Var(\hat{\beta}_0 + \hat{\beta}_1 X_0) + Var(\varepsilon).$$

De manière analogue à la preuve précédente,

$$Var(\hat{\beta}_0 + \hat{\beta}_1 X_0 + \varepsilon) = \sigma_\varepsilon^2 \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right),$$

qui peut être estimée par

$$s_\varepsilon^2 \left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right).$$

□

Rappels

Modèle : $Y_i = \beta_1 X_i + \beta_0 + \varepsilon_i$ (non aléatoire)

$Y_i \sim \mathcal{N}(\beta_1 X_i + \beta_0, \sigma)$ indépendant

$\varepsilon_i \sim \mathcal{N}(0, \sigma)$ i.i.d

$\hat{\beta}_1$ $\hat{\beta}_0$ estimateurs des moindres carrés

$\bar{Y} = \hat{\beta}_1 \bar{X} + \hat{\beta}_0$

$$\hat{\beta}_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$SCT = \sum_i (Y_i - \bar{Y})^2$

$SCM = \sum_i (\hat{Y}_i - \bar{Y})^2$

$SCE = \sum_i \varepsilon_i^2 = \sum_i (\hat{Y}_i - Y_i)^2$

Valeurs prédites $\hat{Y}_i = \hat{\beta}_1 X_i + \hat{\beta}_0$ $\quad \bar{\hat{Y}} = \bar{Y}$

Équation de la variance

$$SCT = SCM + SCE.$$

$$\rightarrow R^2 = \frac{SCM}{SCT}$$

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma}{\sqrt{\sum_i (X_i - \bar{X})^2}}\right)$$

Pb : estimer σ :

Variance de l'erreur (estimateur)

$$S_\varepsilon^2 = \frac{SCE}{n-2} = \frac{\sum_i (\hat{Y}_i - Y_i)^2}{n-2}$$

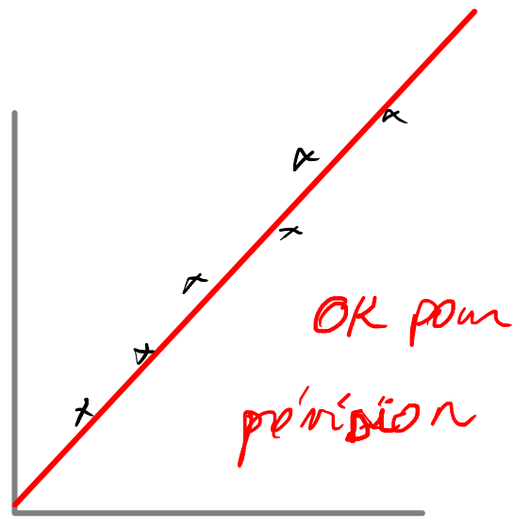
$\Rightarrow$ $\dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\dfrac{SCE}{(n-2)\sum_i (x_i - \bar{x})^2}}} \sim t_{n-2}$

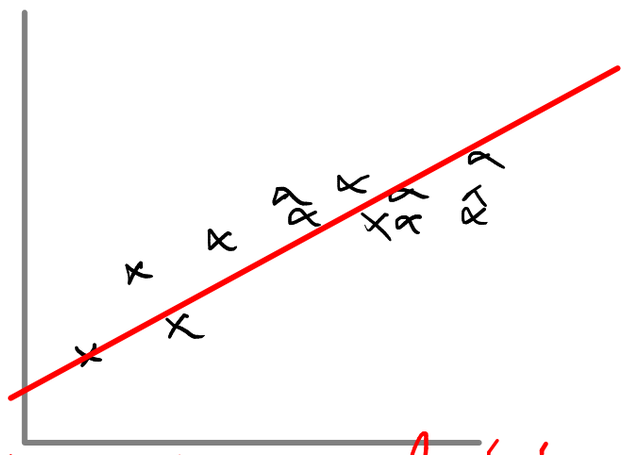
(d'où les intervalles de confiance).

↳ estimation de la variance de $\hat{\beta}_1$.

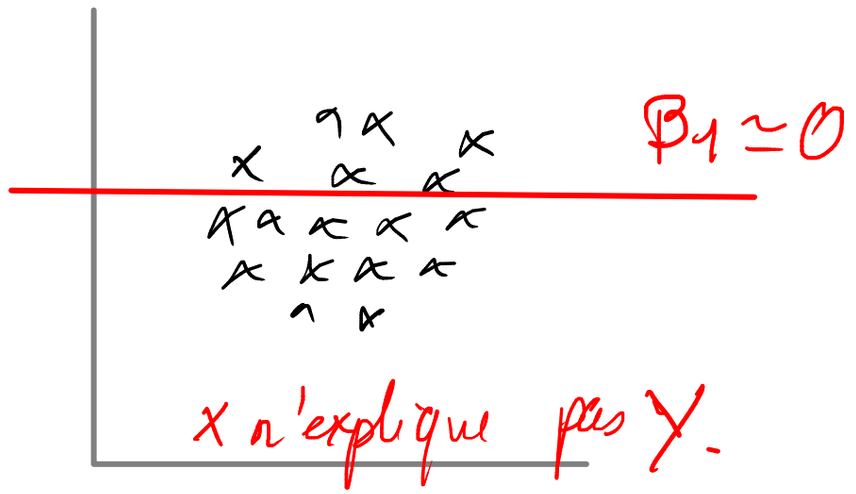
11 Tests sur la pente de la droite

Pour faire simple, les tests F de Fischer et t de Student testent l'hypothèse $\mathcal{H}_0$ sous laquelle le coefficient β_1 est nul, contre β_1 est non nul (ce qui permet d'affirmer que X explique Y, au moins en partie).





dynamique non linéaire
$\Rightarrow$ pas de prévision.
$\Rightarrow$ mais effet de X sur Y.



11.1 Test de Student

Notons l'hypthèse nulle

$$\mathcal{H}_0 = \text{“}\beta_1 = 0\text{”},$$

autrement formulée, $\mathcal{H}_0$ est équivalente à “X n'explique pas Y”. L'estimation de β_1 dans le théorème 10.4 montre que

$$\frac{\hat{\beta}_1 - \beta_1}{s_{\hat{\beta}_1}} \backsim t_{n-2},$$

et sous $\mathcal{H}_0$, nous garderons

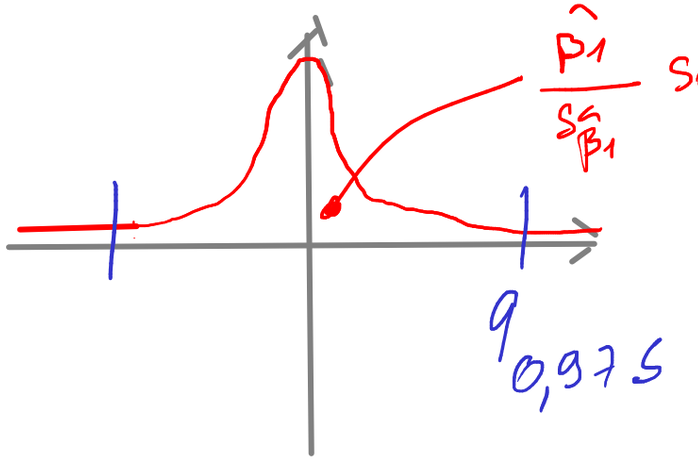
$$\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} \backsim t_{n-2}.$$

Rappelons l'estimation

$$s_{\hat{\beta}_1} = \sqrt{\frac{SCE}{(n-2)\sum(X_i - \bar{X})^2}}.$$

Nous ferons donc un test (bilatéral) de Student sur la statistique de test $\hat{\beta}_1 / s_{\hat{\beta}_1}$. On rejettera $\mathcal{H}_0$ au seuil α si

$$\frac{|\hat{\beta}_1|}{s_{\hat{\beta}_1}} \geq q^{t_{n-2}}_{1-\alpha/2}.$$



Si $\frac{|\hat{\beta}_1|}{s_{\hat{\beta}_1}} > q^{t_{n-2}}_{1-\alpha/2}$ on rejette $\mathcal{H}_0$ $\Rightarrow$ on conclut que X explique Y significativement.

Rappel : $X \sim t_{n-2}$ $\quad X = \frac{N}{\sqrt{\chi^2_{(n-2)}}}$ ← $\mathcal{N}(0,1)$ où $\chi \sim \chi^2(n-2)$.

Alors $X^2 = \frac{N^2/1}{\chi^2_{(n-2)}}$ $\quad N^2 \sim \chi^2(1)$. et $X^2 \sim F_{1,n-2}$

11.2 Table d'ANOVA

L'analyse de la variance, souvent présentée sous forme d'un tableau, permet d'éclairer sur l'influence de la variable X sur la variable Y grâce à l'étude de la décomposition de la variance (10.1). Notons, encore, l'hypthèse nulle

$$\mathcal{H}_0 = \text{“}\beta_1 = 0''.$$

On note que $SCM = \hat{\beta}_1^2 \sum (X_i - \bar{X})^2$. On a vu que

$$s_\varepsilon^2 = \frac{1}{n-2}\sum_{i=1}^{n} \varepsilon_i^2 = \frac{SCE}{n-2}.$$

et on rappelle que sous $\mathcal{H}_0$ $(\beta_1 = 0)$, $\hat{\beta}_1 / s_{\hat{\beta}_1} \backsim t_{n-2}$, avec

$$s_{\hat{\beta}_1}^2 = \frac{SCE}{(n-2)\sum (X_i - \bar{X})^2}.$$

Sous $\mathcal{H}_0$,

$$\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}} \sim t_{n-2}$$

donc

$$\frac{\hat{\beta}_1^2}{s^2_{\hat{\beta}_1}} \sim F_{1,n-2}$$

$SCM = \sum_i (\hat{Y}_i - \bar{Y})^2 = \sum_i \left(\hat{\beta}_1 X_i + \hat{\beta}_0 - (\hat{\beta}_1 \bar{X} + \hat{\beta}_0)\right)^2$

$= \sum_i \left(\hat{\beta}_1 (X_i - \bar{X})\right)^2 = \hat{\beta}_1^2 \sum_i (X_i - \bar{X})^2$

$H_0 = "\beta_1 = 0"$

$$\left(\frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}\right)^2 \sim F_{1,n-2}$$

$$= \frac{\hat{\beta}_1^2}{\frac{SCE}{(n-2)\sum_i (x_i - \bar{x})^2}} = \frac{\hat{\beta}_1^2 \sum_i (x_i - \bar{x})^2}{SCE/(n-2)}$$

$$= \frac{SCM}{SCE/(n-2)} \sim F_{1,n-2}$$

Il vient que

$$\left(\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}\right)^2 = \frac{\hat{\beta}_1^2 \sum (X_i - \bar{X})^2}{\frac{SCE}{n-2}} = \frac{SCM}{SCE/(n-2)}$$

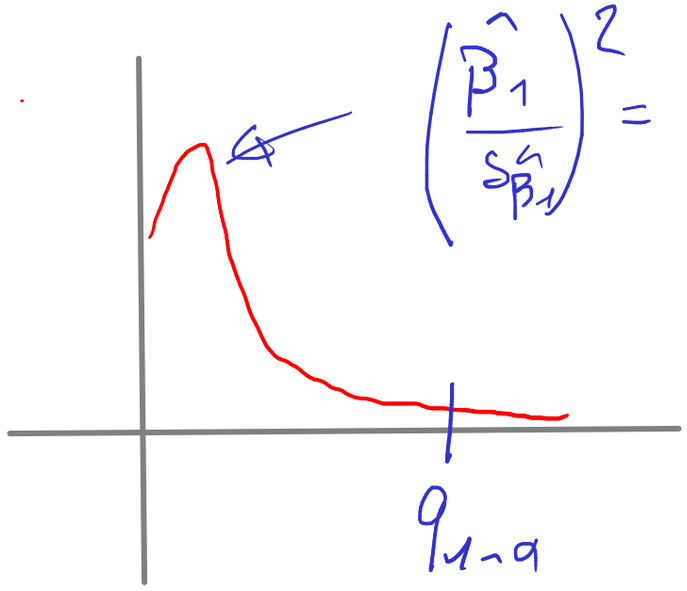
suit une loi de Fisher $\mathcal{F}_{1,n-2}$ en tant que carré de la loi de Student. La variable du test est donc

$$\frac{SCM}{SCE/(n-2)},$$

et on observe son éloignement (à droite) de zéro.

Ainsi, si la p-value dans la table d'ANOVA est proche de zéro (ou en dessous du seuil fixé), on rejettera la nullité de $\hat{\beta}_1$.

Sous H_0 .



Test de Student

$$t = \frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}$$

Test de Fisher (Anova)

$$F = \left(\frac{\hat{\beta}_1}{s_{\hat{\beta}_1}}\right)^2 = \frac{SCM}{SCE/(n-2)}$$

Sont le carré l'un de l'autre.

Dans le cadre de la régression linéaire simple, les deux tests sont pleinement équivalents.

⚠ plus le cas dans le cadre de la régression multiple (à voir en M1).

Régression multiple : $Y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \beta_0$.

Dans ce cadre, on voit qu'on peut utiliser le test de t de Student pour le rapport $\hat{\beta}_1 / s_{\hat{\beta}_1}$ ou F de Fisher pour le carré de ce rapport, sans distinction. Il est totalement équivalent en cas de régression simple (ce n'est pas le cas sur une régression multiple). On note d'ailleurs que la statistique F est le carré de la statistique t.

Dans le cadre d'une régression multiple (sur plusieurs variables explicatives), le test de Fischer teste l'effet global des variables sur la variable Y, les tests de Student testent l'effet de chaque variable explicative sur Y.

En général :

- tests t (student) testent "$\beta_i = 0$" ⇐ chaque variable n'explique pas Y
- ANOVA = F (Fisher) teste "$\beta_1 = \beta_2 = \ldots = \beta_n = 0$".

⇐ Dans leur globalité, les variables n'expliquent pas Y

12 Tests sur régression linéaire multiple

La question qu'on se pose est de savoir si la variable réponse est expliquée par les variables explicatives dans leur globalité, ou par telle ou telle variable explicative. Cela se traduit, mathématiquement, par la non nullité des coefficients de la régression. En effet, si le coefficient d'une des variable explicative est nul ou presque nul, cette variable explicative fait peu varier la régression linéaire, elle n'influence donc pas la variable réponse. Plaçons-nous dans le cadre de la régression linéaire multiple.

12.1 Les tests t de Student

Les tests de Student testent la nullité de chaque coefficient de la régression linéaire. Ainsi, on saura quelles variables explicatives ont un effet sur la variable expliquée.

12.1.1 Hypothèse nulle

L'hypothèse nulle de chaque test est $\mathcal{H}_0$ ="La variable X_i n'a pas d'effet sur la variable réponse"="$\beta_i = 0$".

12.2 Décision

- Si la $p-$value est inférieure au un niveau α choisi (en général 0.05), alors on rejette l'hypothèse nulle et on considère que la variable X_i a un effet sur la variable réponse.
- Si la $p-$value est supérieure au niveau α choisi (en général 0.05), alors on ne doit pas rejeter l'hypothèse nulle. La variable X_i n'a pas d'effet sur la variable réponse.

12.3 Les tests F de Fisher - ANOVA

Le test de Fisher teste l'effet de l'ensemble des variables explicatives sur la variable réponse. Ainsi, on saura si la variable réponse est expliquée par les variables explicatives. On appelle cela une ANalyse de la (Of) VAriance.

12.3.1 Hypothèse nulle

L'hypothèse nulle du test est $\mathcal{H}_0$ =“Les variables X_i n'ont pas d'effet, dans leur globalité, sur la variable réponse”=“la variance de l'erreur est très forte face à la variance expliquée par le modèle”.

12.3.2 Décision

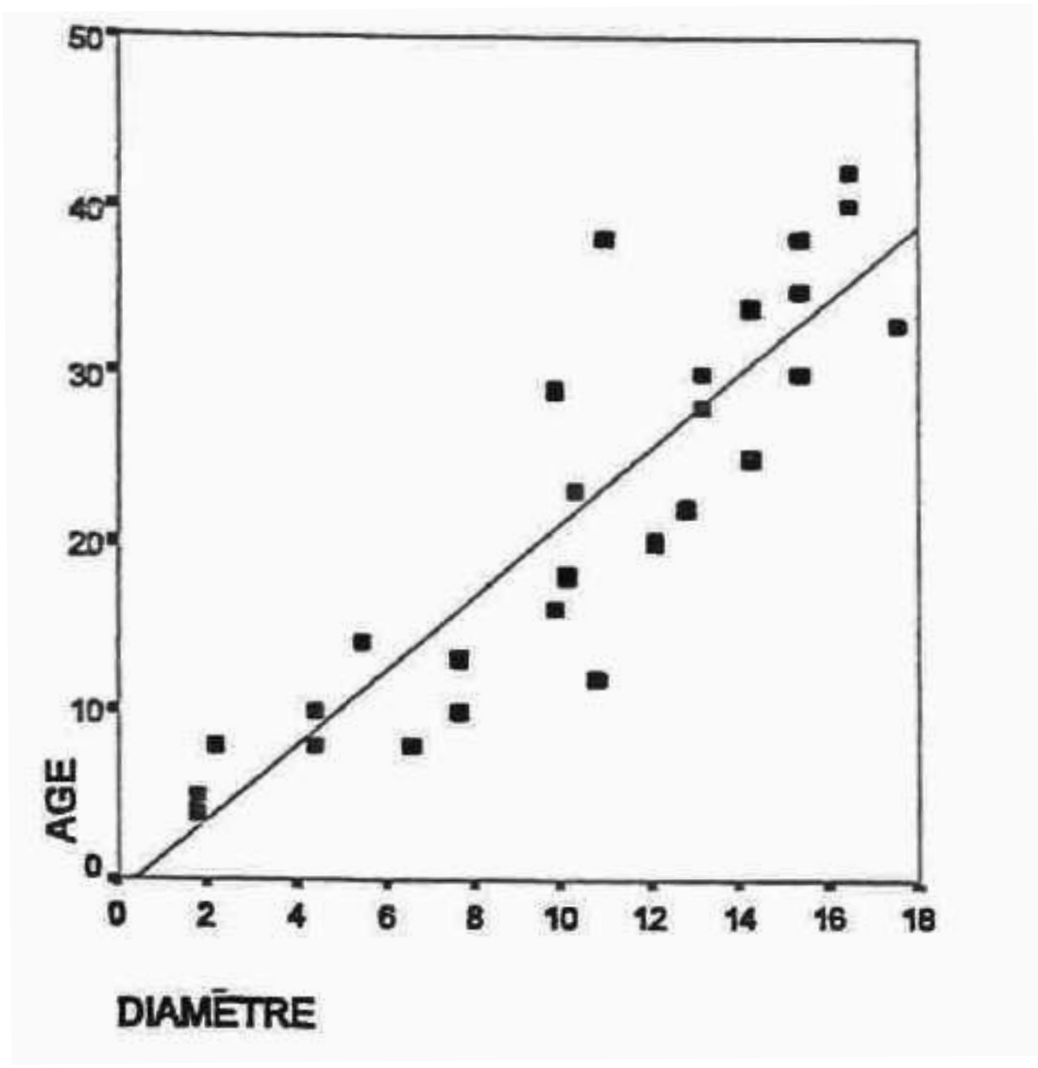
- Si la $p-$value est inférieure au un niveau α choisi (en général 0.05), alors on rejette l'hypothèse nulle et on considère que les variable X_i ont un effet global sur la variable réponse.
- Si la $p-$value est supérieure au niveau α choisi (en général 0.05), alors on ne doit pas rejeter l'hypothèse nulle. Les variables X_i n'ont pas d'effet sur la variable réponse.

12.4 Exemple : le cas de la régression linéaire simple

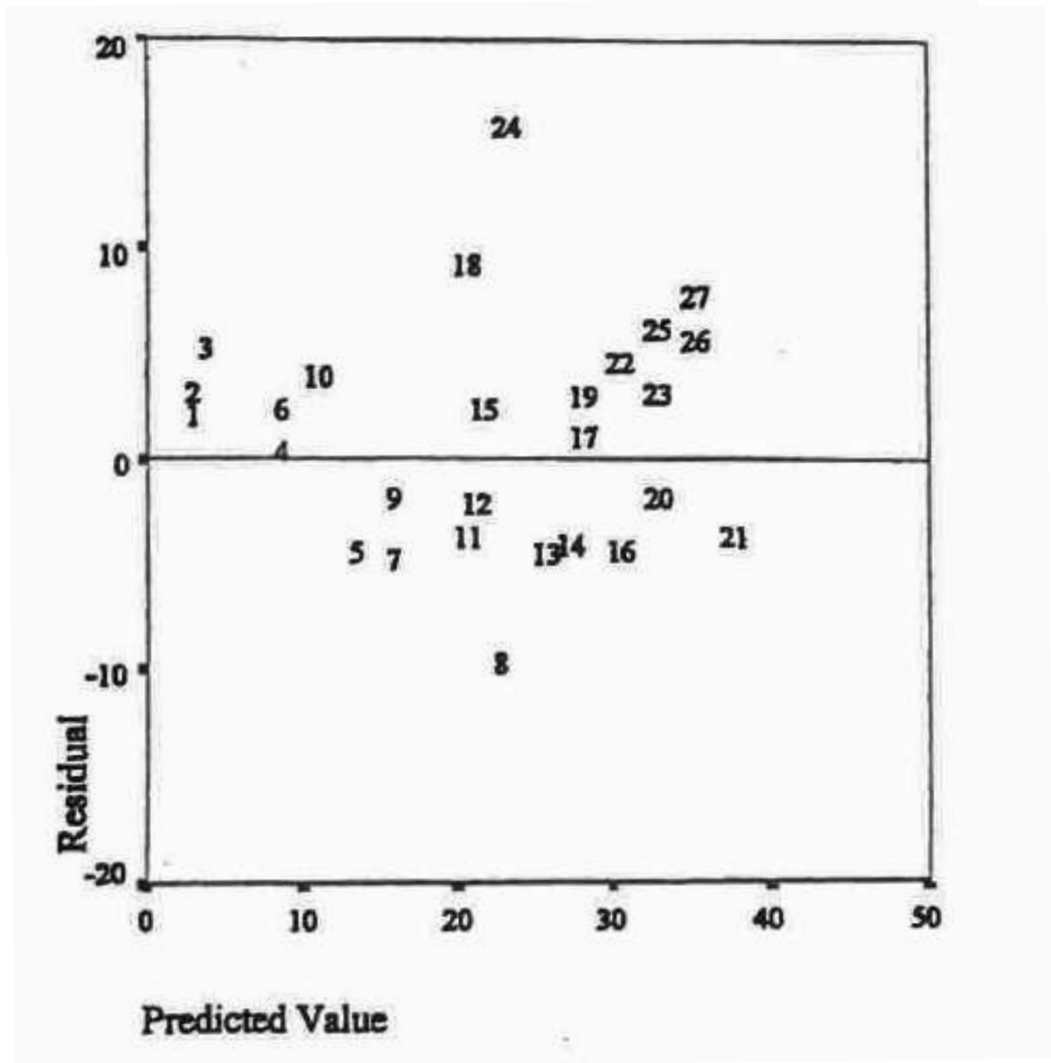
On sait que l'on mesure l'âge d'un arbre en comptant les anneaux sur une section transversale du tronc, mais cela nécessite de l'avoir abattu auparavant. Peut-on connaître l'âge à partir de la mesure de sa circonférence ?

Afin de répondre à cette question, on a effectué les mesures sur un échantillon de 27 arbres de la même espèce. À partir de ces données, on a effectué une régression de l'âge en fonction du diamètre . Les résultats ont été traités à l'aide du logiciel SPSS.

Nuage de points



Résidus



$SCT = SCE + SCM$

ANOVA

	Somme des carrés	ddl	Carré moyen	F	Signification
Régression	SCM = 2905,549	1	2905,55	93,44	,000
Résidu	SCE = 777,414	25	31,097		
Total	SCT = 3682,963	26			

$\frac{SCM}{1}$ $\frac{SCE}{n-2}$ $n-2$ $\frac{SCM}{SCE/n-2}$

$F = 93,44 = 9,67^2 = t^2$

on conclut que pr t.

Régression

	Coefficients non standardisés		Coefficients standardisés		
	B	Erreur standard	Bêta	t	Signification
(constante)	-,974	2,604		-,374	,711
DIAMETRE	2,206	,228	,888	9,67	,000

coeff sur régression de $\frac{Y-\bar{Y}}{\sigma_Y}$ sur $\frac{X-\bar{X}}{\sigma_X}$

Age = 2,206 diamètre − 0,974.

r_{XY}.

$t = \frac{\hat{\beta}_1}{S_{\beta_1}}$

H_0 = "$\beta_1 = 0$" = "X n'explique pas Y". p-value ≃ 0

⇒ On rejette H_0 - X explique significativement Y

12.5 Exemple : le cas de la régression linéaire multiple

On cherche à modéliser la relation entre poids des bébés à la naissance et l'âge, le poids et le statut tabagique de la mère durant la grossesse. (Exemple fictif) On pose

- y = poids de naissance en grammes (bwt),
- x_1 = âge de la mère (age),
- x_2 = poids de la mère en kilos (weight),
- x_3 = statut tabagique de la mère pendant la grossesse (smoke) codé par un score à une échelle de 1 à 20.

```
> modele=lm(bwt~age+weight+smoke)
> summary(modele)

Call:
lm(formula = bwt ~ age + weight + smoke)

Residuals:
    Min      1Q  Median      3Q     Max
-385.81  -65.83   -0.70   68.17  290.66

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3005.65519   22.99506  130.71   <2e-16 ***
age            0.02645    0.53148    0.05     0.96
weight         8.44845    0.30499   27.70   <2e-16 ***
smoke        -26.53764    1.82009  -14.58   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 97.58 on 996 degrees of freedom
Multiple R-squared:  0.4919,    Adjusted R-squared:  0.4904
F-statistic: 321.4 on 3 and 996 DF,  p-value: < 2.2e-16
```

les "β"

bwt = 3005 + 0,03 âge + 8,44 weight − 26 smoke.

pas intéressant

H_0 = "$\beta_i = 0$"

non rejet / rejet / rejet

Seul age n'a pas d'effet signif. sur bwt.

R^2 petit ne servira pas à la prévision.

H_0 = " les variable, dans leur globalité, n'expliquent pas bwt "

p-value proche de 0 → Rejet de H_0.

Dans leur globalité, les variables ont un effet significatif sur bwt.