Probabilités et observations économiques

C. Tuleau-Malot

Université de Nice - Sophia Antipolis

Plan

Contexte

Exemple :

numéro de l'observation	X population étudiante	Y ventes trimestrielles
1	2	58
2	6	105
3	8	88
4	8	118
5	12	117
6	16	137
7	20	157
8	20	169
9	22	149
10	26	202

Contexte

Exemple :

Ces données sont des données relevées sur le terrain par une chaîne de fast food qui souhaite estimer le chiffre d'affaire d'un nouveau restaurant implanté aux alentours d'un campus universitaire.

L'idée est d'essayer de regarder s'il existe un lien entre les ventes trimestrielles et la population étudiante.

Vocabulaire

Cadre :

- On cherche à expliquer un phénomène Y (variable à expliquer) à l'aide d'une variable X (variable explicative).

Vocabulaire

Cadre :

- On cherche à expliquer un phénomène Y (variable à expliquer) à l'aide d'une variable X (variable explicative).
- On dispose d'observations $(x_1, y_1), \ldots, (x_n, y_n)$ qui sont des réalisations de variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ qui sont des respectivement des variables indépendantes et identiquement distribuées de même loi que X et Y.

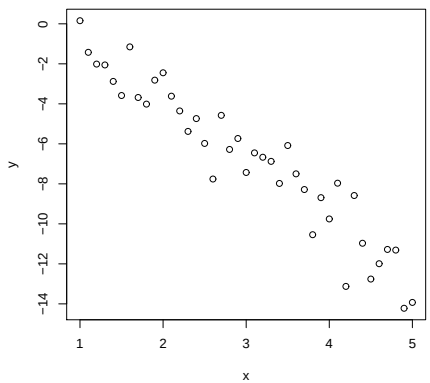
Vocabulaire

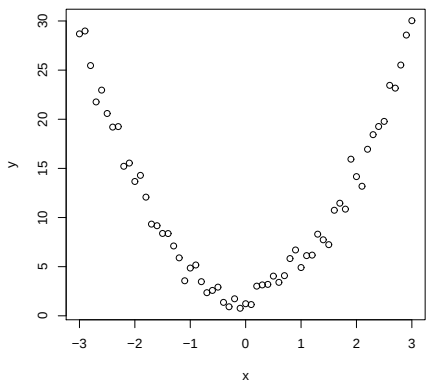
Cadre :

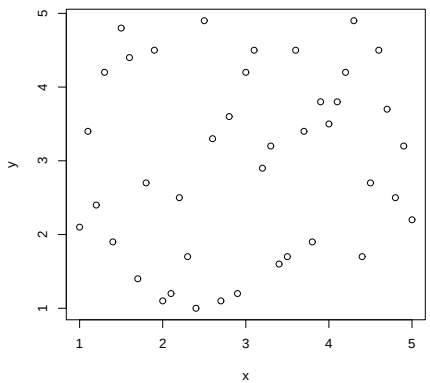
- On cherche à expliquer un phénomène Y (variable à expliquer) à l'aide d'une variable X (variable explicative).
- On dispose d'observations $(x_1, y_1), \ldots, (x_n, y_n)$ qui sont des réalisations de variables $(X_1, Y_1), \ldots, (X_n, Y_n)$ qui sont des respectivement des variables indépendantes et identiquement distribuées de même loi que X et Y.
- on veut regarder s'il existe une fonction f telle que :

$$Y \approx f(X)$$

Vocabulaire (2)



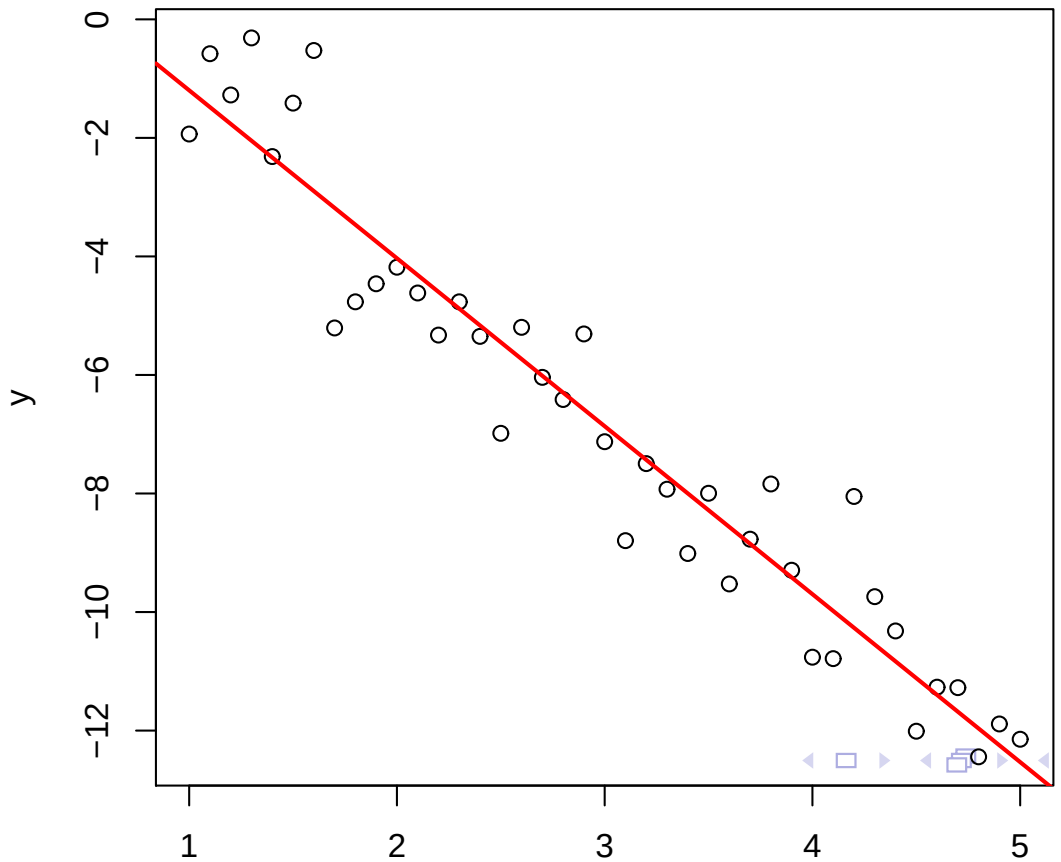




Régression linéaire simple

Cadre :

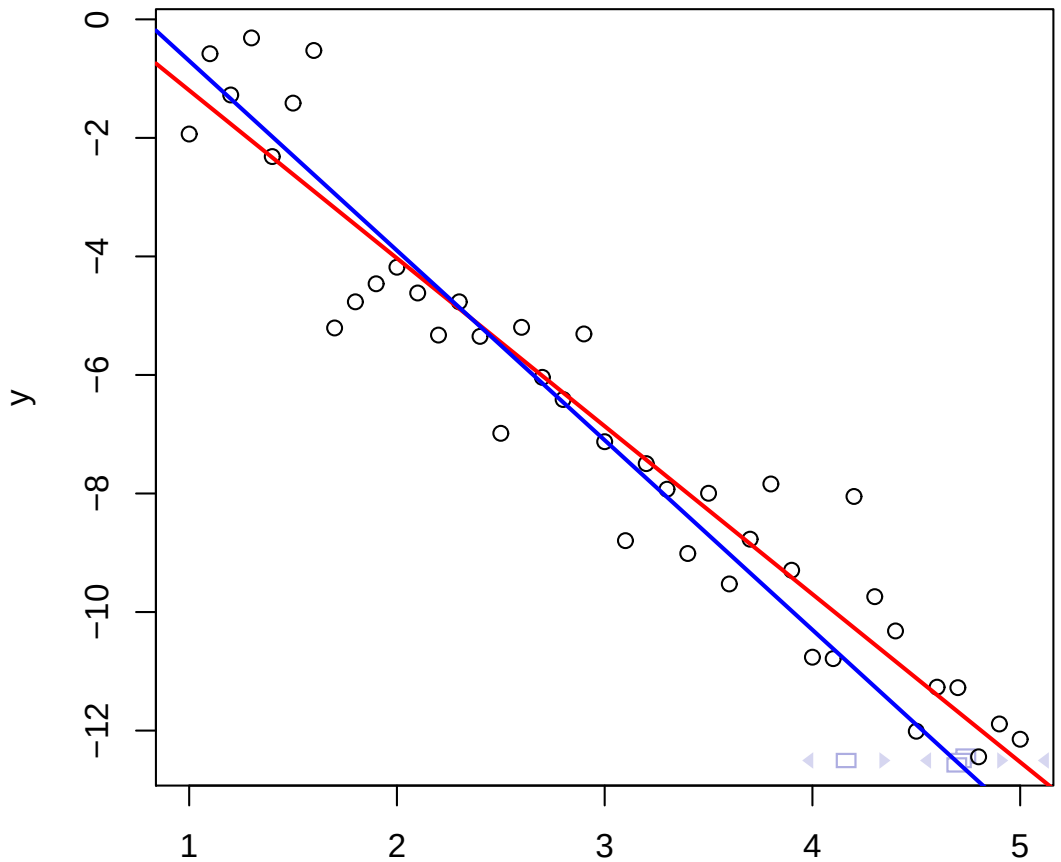
- on se limite au cadre des fonctions f qui sont linéaires $\Rightarrow$ $f(x) = a.x + b$



Régression linéaire simple

Cadre :

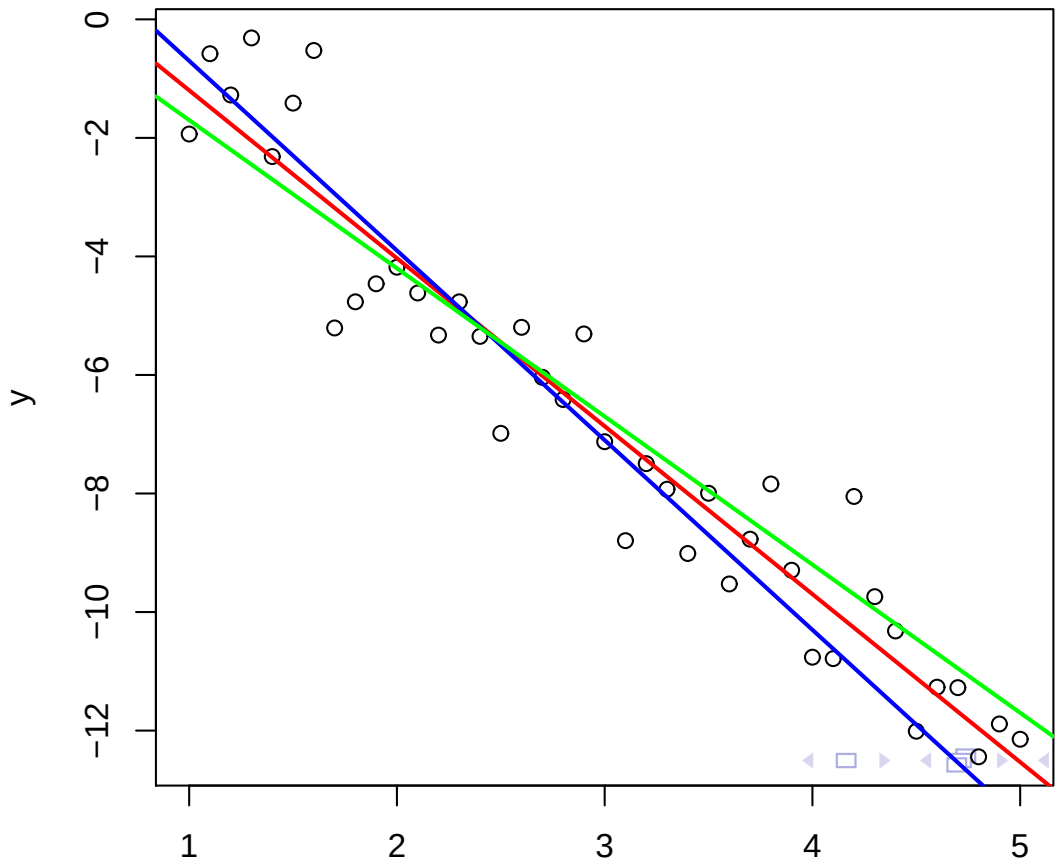
- on se limite au cadre des fonctions f qui sont linéaires $\Rightarrow$ $f(x) = a.x + b$



Régression linéaire simple

Cadre :

- on se limite au cadre des fonctions f qui sont linéaires $\Rightarrow$ $f(x) = a.x + b$



Régression linéaire simple

Cadre :

- on se limite au cadre des fonctions f qui sont linéaires $\Rightarrow$ $f(x) = a.x + b$
- comment trouver a et $b \Rightarrow$ estimation

Régression linéaire simple (2)

Modèle :

$$Y_i = a + b.x_i + \varepsilon_i$$

- a et b : inconnus
- $\varepsilon_1, \ldots, \varepsilon_n$: variables aléatoires d'espérance nulle (bruit)
  $\Rightarrow \mathbb{E}[Y_i] = a + b.x_i$

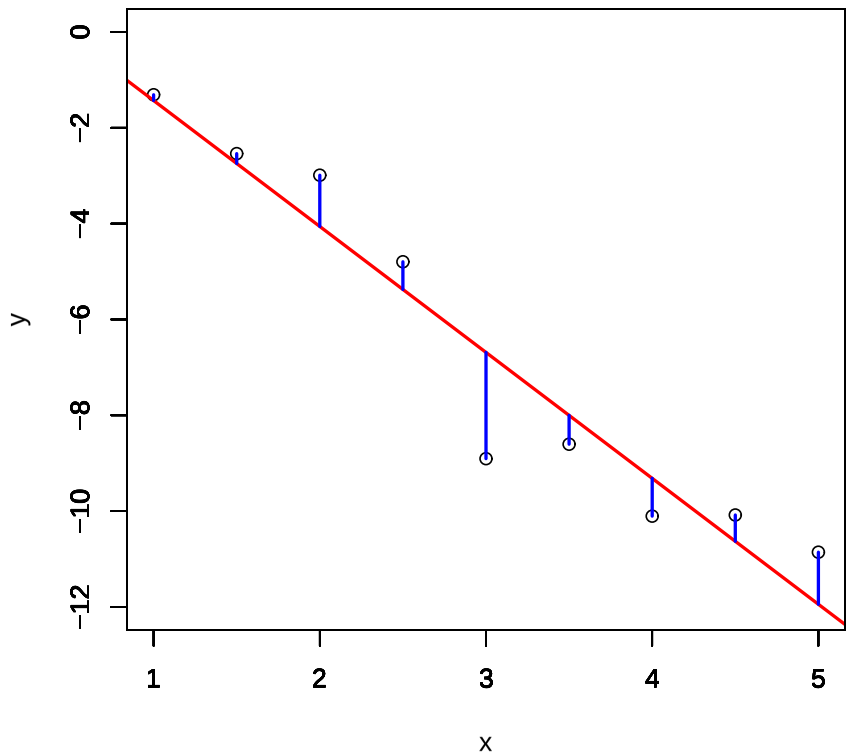
Régression linéaire simple (2)

Objectifs :

- estimer a et b
- donner un intervalle de confiance pour b
- tester $b = 0$
- faire des prévisions

Moindres carrés

Principe :



Moindres carrés

Formule :

$$F(a,b) = \sum_{i=1}^{n} (Y_i - a - b.x_i)^2$$

- on veut minimiser F fonction de deux variables

Moindres carrés

Formule :

$$F(a,b) = \sum_{i=1}^{n} (Y_i - a - b.x_i)^2$$

- on veut minimiser F fonction de deux variables
- solution :

$$\begin{cases} \hat{B}_n = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n).(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} \\ \hat{A}_n = \bar{Y}_n - \hat{b}_n.\bar{x}_n \end{cases}$$

Moindres carrés

Formule :

$$F(a,b) = \sum_{i=1}^{n} (Y_i - a - b.x_i)^2$$

- on veut minimiser F fonction de deux variables
- solution :

$$\begin{cases} \hat{B}_n = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n).(Y_i - \bar{Y}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} \\ \hat{A}_n = \bar{Y}_n - \hat{b}_n.\bar{x}_n \end{cases}$$

- $\hat{B}_n$ et $\hat{A}_n$: estimateurs de b et a (variables aléatoires)
-

$$\begin{cases} \hat{b}_n = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n).(y_i - \bar{y}_n)}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} \\ \hat{a}_n = \bar{y}_n - \hat{b}_n.\bar{x}_n \end{cases}$$

estimations de b et a

Coefficient de détermination

- somme des carrés des résidus :

$$\mathrm{SCR}_{\mathrm{res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Coefficient de détermination

- somme des carrés des résidus :

$$\mathrm{SCR}_{\mathrm{res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- somme des carrés totaux :

$$\mathrm{SCT} = \sum_{i=1}^{n} (y_i - \bar{y}_n)^2$$

Coefficient de détermination

- somme des carrés des résidus :

$$\mathrm{SCR}_{\mathrm{res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- somme des carrés totaux :

$$\mathrm{SCT} = \sum_{i=1}^{n} (y_i - \bar{y}_n)^2$$

- somme des carrés de la régression :

$$\mathrm{SCR}_{\mathrm{reg}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_n)^2$$

Coefficient de détermination

- somme des carrés des résidus :

$$\mathrm{SCR}_{\mathrm{res}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- somme des carrés totaux :

$$\mathrm{SCT} = \sum_{i=1}^{n} (y_i - \bar{y}_n)^2$$

- somme des carrés de la régression :

$$\mathrm{SCR}_{\mathrm{reg}} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_n)^2$$

$$\mathrm{SCT} = \mathrm{SCR}_{\mathrm{res}} + \mathrm{SCR}_{\mathrm{reg}}$$

Coefficient de détermination (2)

Coefficient de détermination :

$$R^2 = \frac{\text{SCR}_{\text{reg}}}{\text{SCT}} = 1 - \frac{\text{SCR}_{\text{res}}}{\text{SCT}}$$

Coefficient de détermination (2)

Coefficient de détermination :

$$R^2 = \frac{\text{SCR}_{\text{reg}}}{\text{SCT}} = 1 - \frac{\text{SCR}_{\text{res}}}{\text{SCT}}$$

$$R^2 = (r_{xy})^2$$

avec : $r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n).(y_i - \bar{y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 . \sum_{i=1}^{n}(y_i - \bar{y}_n)^2}}$.

Coefficient de détermination (2)

Coefficient de détermination :

$$R^2 = \frac{\text{SCR}_{\text{reg}}}{\text{SCT}} = 1 - \frac{\text{SCR}_{\text{res}}}{\text{SCT}}$$

$$R^2 = (r_{xy})^2$$

avec : $r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_n).(y_i - \bar{y}_n)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 . \sum_{i=1}^{n}(y_i - \bar{y}_n)^2}}$.

- le coefficient de détermination permet de dire à quel point les points sont proches de la droite
- le coefficient de détermination ne permet pas de conclure sur la significativité des résultats obtenus $\Rightarrow$ intervalle de confiance; tests

intervalle de confiance et test sur b

Pour faire un intervalle de confiance pour b ou un test sur b, il faut des hypothèses supplémentaires sur la variable de loi!

- ε_i : variable centrée
- la variance des ε_i est la même pour toutes. On la note σ^2
- les ε_i sont des variables indépendantes
- les ε_i ont une distribution normale

intervalle de confiance et test sur b

Pour faire un intervalle de confiance pour b ou un test sur b, il faut des hypothèses supplémentaires sur la variable de loi!

- ε_i : variable centrée
- la variance des ε_i est la même pour toutes. On la note σ^2
- les ε_i sont des variables indépendantes
- les ε_i ont une distribution normale

$\Rightarrow$ les Y_i sont des variables $\mathcal{N}(a + b.x_i, \sigma^2)$!

intervalle de confiance sur b

- On admet le résultat suivant :

$$\hat{B}_n \sim \mathcal{N}(b, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}$$

intervalle de confiance sur b

- On admet le résultat suivant :

$$\hat{B}_n \sim \mathcal{N}(b, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}$$

- il faut estimer σ^2

$$\hat{\sigma}_n^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2} = \frac{\mathrm{SCR}_{\mathrm{res}}}{n-2}$$

intervalle de confiance sur b

- On admet le résultat suivant :

$$\hat{B}_n \sim \mathcal{N}(b, \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2})$$

- il faut estimer σ^2

$$\hat{\sigma}_n^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{\text{SCR}_{\text{res}}}{n-2}$$

on peut montrer que

$$\frac{(n-2)\hat{\sigma}_n^2}{\sigma^2} \sim \chi^2(n-2)$$

intervalle de confiance sur b (2)

- intervalle de confiance pour b : $[\hat{B}_n - c; \hat{B}_n + c]$
  $\Rightarrow$ identifier c

intervalle de confiance sur b (2)

- intervalle de confiance pour b : $[\hat{B}_n - c; \hat{B}_n + c]$
  $\Rightarrow$ identifier c
- on utilise le fait que :

$$\frac{\hat{B}_n - b}{\frac{\hat{\sigma}_n}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}} \sim \mathcal{T}(n-2)$$

intervalle de confiance sur b (2)

- intervalle de confiance pour b : $[\hat{B}_n - c; \hat{B}_n + c]$
  $\Rightarrow$ identifier c
- on utilise le fait que :

$$\frac{\hat{B}_n - b}{\frac{\hat{\sigma}_n}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2}}} \sim \mathcal{T}(n-2)$$

- on trouve $c = t_{1-\alpha/2} \cdot \frac{\hat{\sigma}_n}{\sqrt{\sum_{i=1}^n (x_i - \bar{x}_n)^2}}$

Test sur b (2)

Test :

- $\mathcal{H}_0 : b = 0$
- $\mathcal{H}_a : b \neq 0$

Test sur b (2)

Test :

- $\mathcal{H}_0 : b = 0$
- $\mathcal{H}_a : b \neq 0$

Méthode :

- on utilise le fait que sous $\mathcal{H}_0$, $\dfrac{\hat{B}_n}{\dfrac{\hat{\sigma}_n}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}} \sim \mathcal{T}(n-2)$

Test sur b (2)

Test :

- $\mathcal{H}_0 : b = 0$
- $\mathcal{H}_a : b \neq 0$

Méthode :

- on utilise le fait que sous $\mathcal{H}_0$, $T_n = \dfrac{\hat{B}_n}{\dfrac{\hat{\sigma}_n}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}} \sim \mathcal{T}(n-2)$
- région de rejet : quand $|T_n|$ trop grand $\Rightarrow$ $\{T_n \in]-\infty, t_{1-\alpha/2}] \cup [t_{1-\alpha/2}, +\infty[\}$

Test sur b (2)

Test :

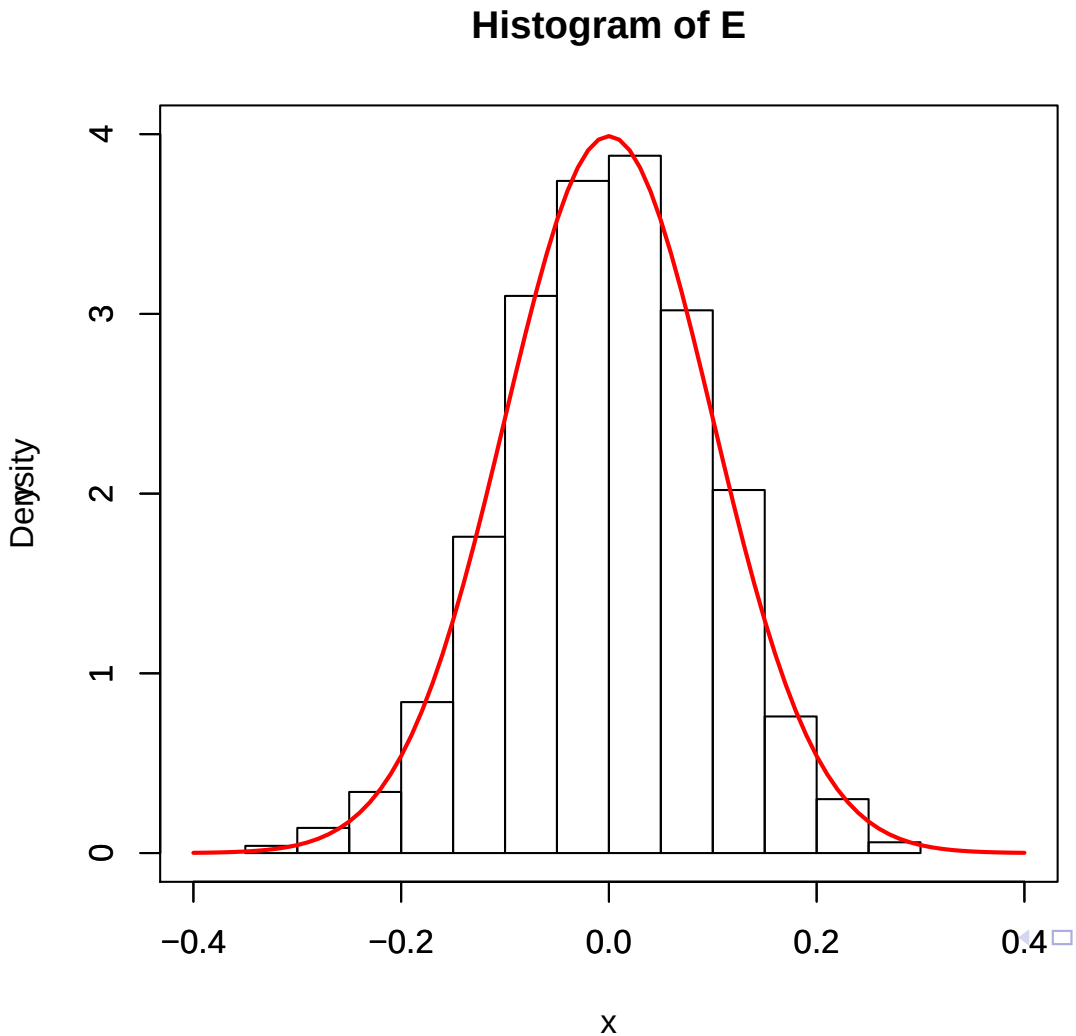
- $\mathcal{H}_0 : b = 0$
- $\mathcal{H}_a : b \neq 0$

Méthode :

- on utilise le fait que sous $\mathcal{H}_0$, $T_n = \frac{\hat{B}_n}{\frac{\hat{\sigma}_n}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_n)^2}}} \sim \mathcal{T}(n-2)$
- région de rejet : quand $|T_n|$ trop grand $\Rightarrow$ $\{T_n \in]-\infty, -t_{1-\alpha/2}] \cup [t_{1-\alpha/2}, +\infty[\}$
- p-valeur : $P(|\mathcal{T}(n-2)| >= t_n)$ avec t_n la valeur observée de T_n

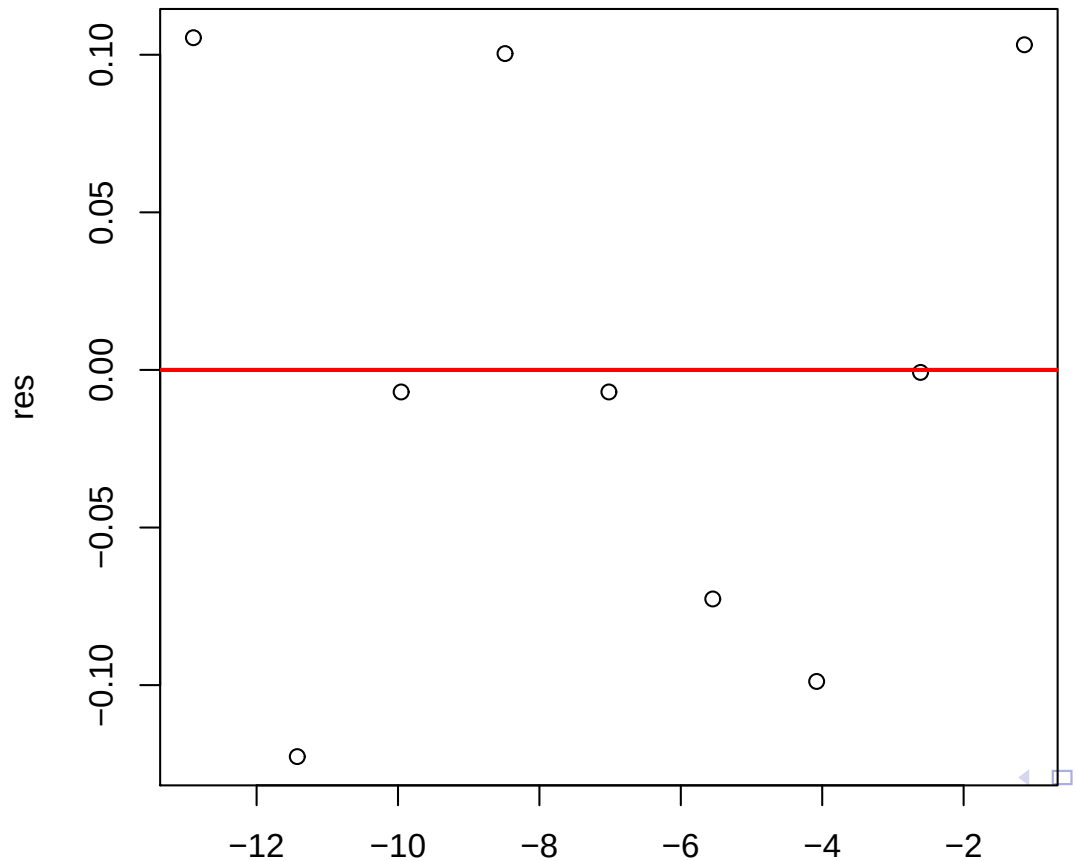
Vérification des hypothèses

- Caractère de normalité : on trace l'histogramme des résidus et on le compare à la densité d'une loi normale



Vérification des hypothèses

- Caractère de normalité : on trace l'histogramme des résidus et on le compare à la densité d'une loi normale
- caractère centré



Vérification des hypothèses

- Caractère de normalité : on trace l'histogramme des résidus et on le compare à la densité d'une loi normale
- caractère centré
- caractère homoscédastique

Test d'adéquation

Objectif :

Tester une hypothèse sur la distribution complète d'une population.

Test d'adéquation

Objectif :
Tester une hypothèse sur la distribution complète d'une population.

Cadre :
Variable discrète

Test d'adéquation (2)

Exemple :

La Roma a encaissé 38 buts cette saison. L'entraineur aimerait savoir si la répartition des buts au cours des matchs est compatible avec une probabilité uniforme au cours du temps d'encaisser un but.

$\mathcal{H}_0$: 1/3 des buts la 1ère 1/2h, 1/3 des buts la 2nd 1/2h et 1/3 des buts la dernière 1/2h

$\mathcal{H}_a$: autre répartition

Test d'adéquation (2)

Exemple :

La Roma a encaissé 38 buts cette saison. L'entraineur aimerait savoir si la répartition des buts au cours des matchs est compatible avec une probabilité uniforme au cours du temps d'encaisser un but.

$\mathcal{H}_0$: 1/3 des buts la 1ère 1/2h, 1/3 des buts la 2nd 1/2h et 1/3 des buts la dernière 1/2h

$\mathcal{H}_a$: autre répartition

Cadre général :

On considère une variable discrète X qui prend les valeurs $a_1, \ldots, a_K$.

$\mathcal{H}_0$: les probabilités que X prennent les valeurs $a_1, \ldots, a_K$ sont $p_1, \ldots, p_K$

$\mathcal{H}_a$: ces probabilités ne sont pas $p_1, \ldots, p_K$

Test d'adéquation (3)

Nos données :

valeurs	a_1	a_2	$\dots$	a_K
effectifs	n_1	n_2	$\dots$	n_K

Test d'adéquation (3)

Nos données :

valeurs	a_1	a_2	$\ldots$	a_K
effectifs	n_1	n_2	$\ldots$	n_K

Si on pose n le nombre d'observations, on devrait observer sous $\mathcal{H}_0$

valeurs	a_1	a_2	$\ldots$	a_K
effectifs	n_1	n_2	$\ldots$	n_K
théorie	e_1	e_2	$\ldots$	e_K

avec : $e_i = n * p_i$

Test d'adéquation (3)

Nos données :

valeurs	a_1	a_2	$\dots$	a_K
effectifs	n_1	n_2	$\dots$	n_K

Si on pose n le nombre d'observations, on devrait observer sous $\mathcal{H}_0$

valeurs	a_1	a_2	$\dots$	a_K
effectifs	n_1	n_2	$\dots$	n_K
théorie	e_1	e_2	$\dots$	e_K

avec : $e_i = n * p_i$

$\Rightarrow$ on va privilégier $\mathcal{H}_0$ si tous les n_i sont proches des e_i