Probabilités et Statistiques

Année 2010/2011

laurent.carraro@telecom-st-etienne.fr

olivier.roustant@emse.fr

telecom Saint-Etienne

Cours n°14

Régression - fin

Prédicteurs influents

- Rappel modèle linéaire :

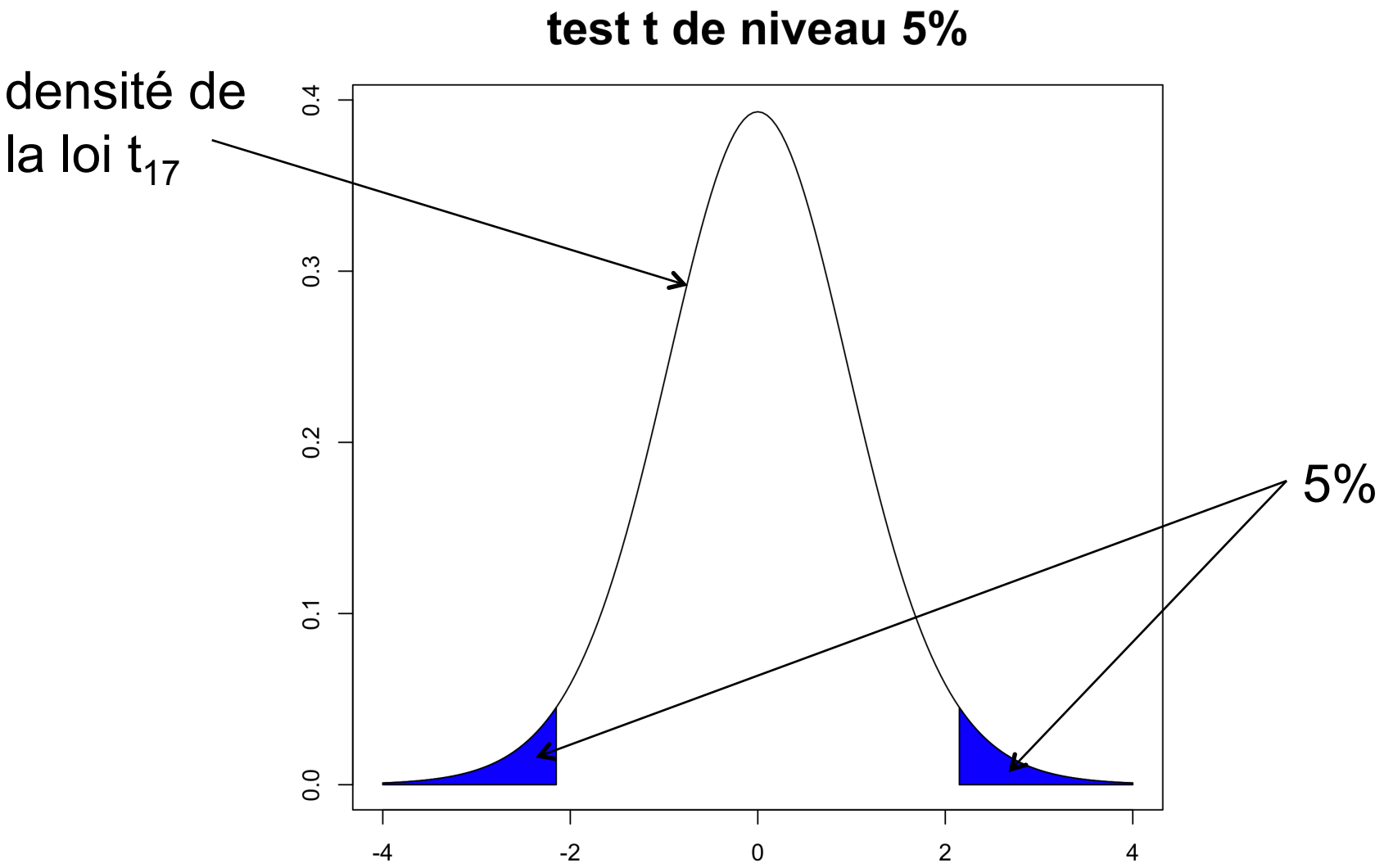
pour toute « expérience » n°i

$$y_i = \beta_0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i} + \varepsilon_i$$

- On teste la nullité du coefficient β_j
  - Hypothèse H_0 : $\beta_j = 0$
  - Au niveau α, si $q_{\alpha,n-p-1}$ est le quantile d'ordre α de la loi de Student t_{n-p-1}, on rejette H_0 si :

$$|T_{obs}| = \frac{|\hat{\beta}_{j,obs}|}{\hat{\sigma}_{obs}\sqrt{M_{j,j}}} \geq q_{\alpha,n-p-1}$$

Interprétation



La p-valeur

- Pour un niveau α fixé, on a :

$$\text{Pour } T = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{M_{j,j}}},\ P_{H_0}(|T| \geq q_{\alpha,n-p-1}) = \alpha$$

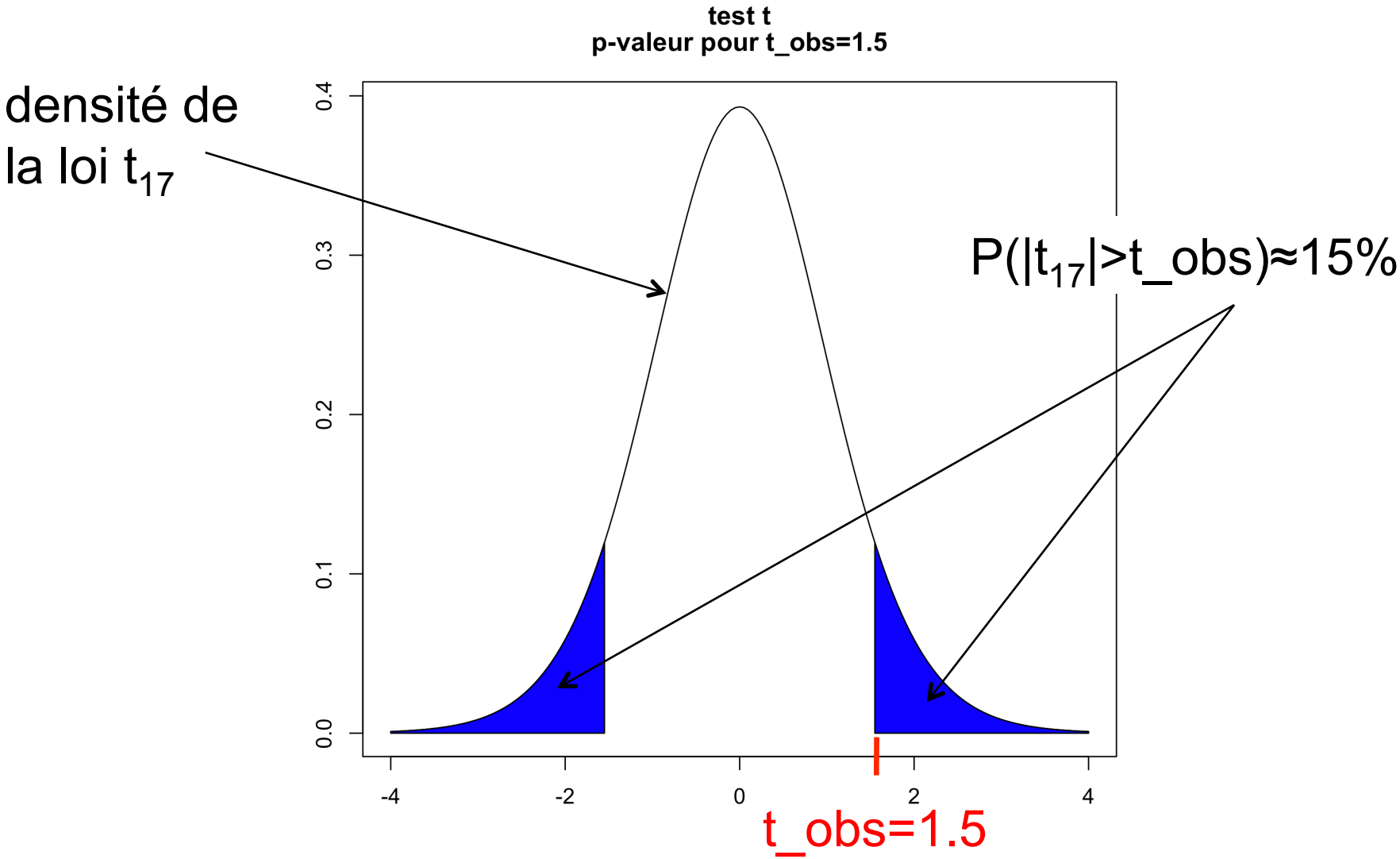
puis si $|T_{obs}| > q_{\alpha,n-p-1}$ on rejette H_0

- p-valeur :

$$\text{p} - \text{value} = P_{H_0}(|T| \geq |T_{obs}|)$$

  - si p-value < α on rejette H_0 au niveau α (i.e. x_j **influent**)
  - On fait donc le test à tous les niveaux à la fois

Idem avec p-valeur



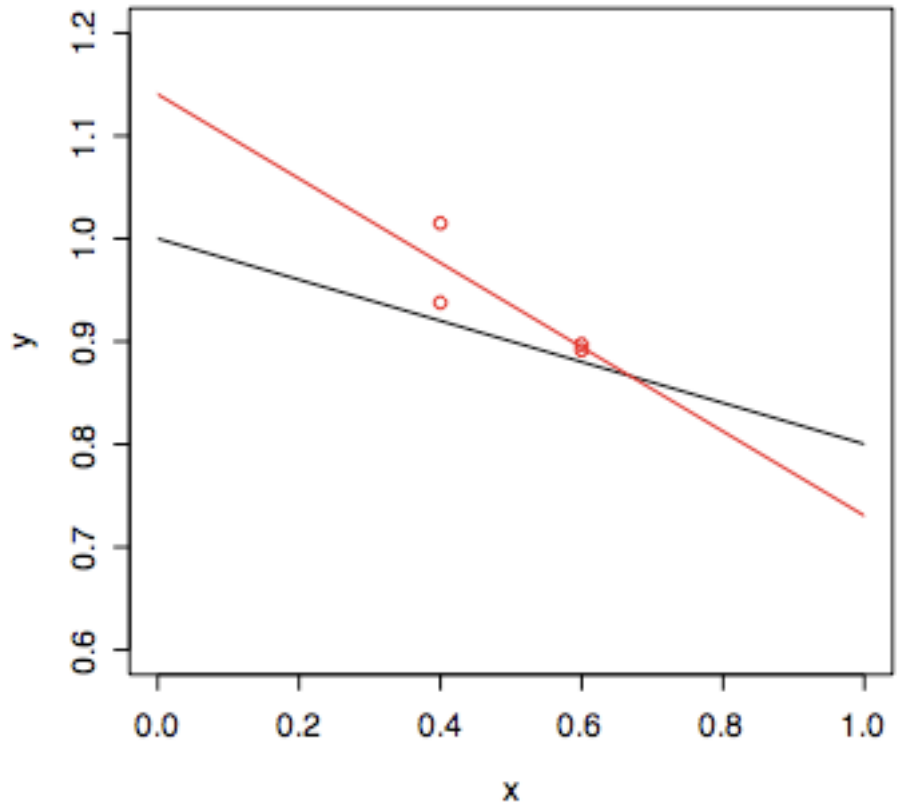
Test de signification : pratique

- En pratique, les logiciels donnent le tableau suivant :

prédicteur	estimation	erreur d'estimation	t – ratio	p – valeur
x_j	$\hat{\beta}_{j,obs}$	$\hat{\sigma}_{\text{j,obs}}$	$T_{obs} = \dfrac{\hat{\beta}_{j,obs}}{\hat{\sigma}_{j,obs}}$	$P_{H_0}(\lvert T\rvert > \lvert T_{obs}\rvert)$

Retour sur les simulations

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad \text{avec} \quad e_1, \ldots, e_4 \quad \text{i.i.d } N(0, 0.04^2)$$



Call:
lm(formula = ysim ~ exp

Residuals:
1 2 3
0.038612 -0.038612 0

Coefficients:

	Estimate	Std. Error	t	Pr(>|t|)
(Intercept)	1.1404	0.0987	11.554	0.00741 **
experiences	-0.4099	0.1936	-2.118	0.16838

Signif. codes: 0 '***' 0.001

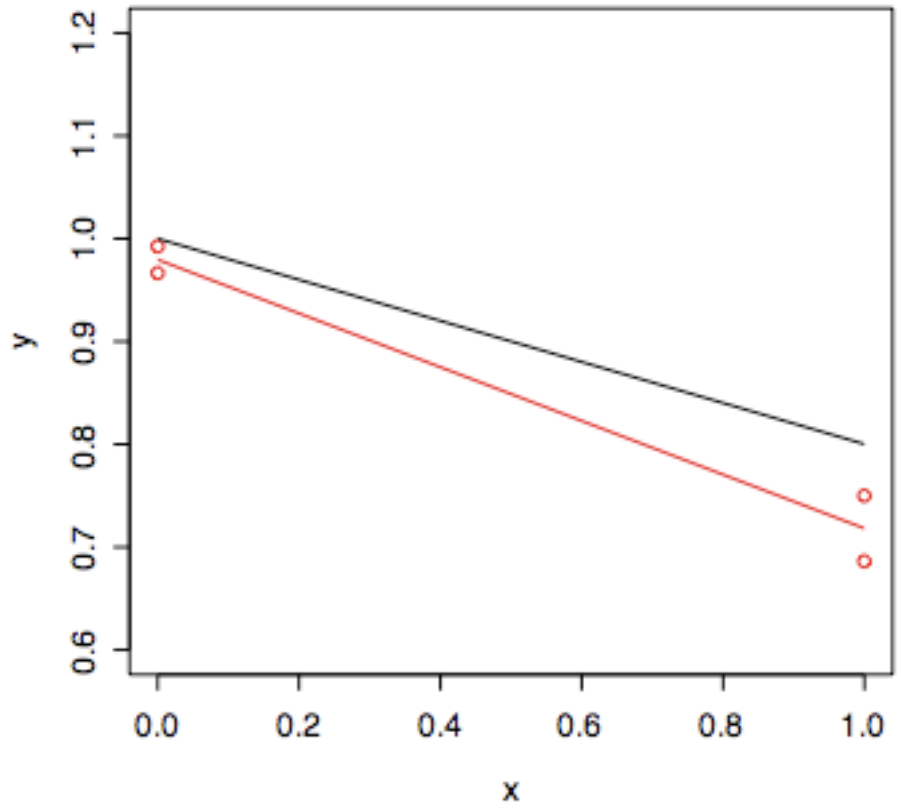
Residual standard error: 0.
Multiple R-Squared: 0.6916
F-statistic: 4.485 on 1 and

La t-valeur est > 1.96 en valeur absolue,
Pourtant on ne rejette pas H_0
Cela est dû au fait qu'on ne peut pas utiliser l'approximation normale (ici n-2=2 << 20)
La p-valeur est calculée à partir de la loi de Student

Moralité : la pente de la droite est négative,
Mais l'erreur d'estimation est trop importante
Et le paramètre est statistiquement non significatif au niveau 5% …rassurant !

Simulations (suite)

$$y_i = \beta_0 + \beta_1 x_i + e_i \quad \text{avec} \quad e_1, \ldots, e_4 \quad \text{i.i.d } N(0, 0.04^2)$$



Call:
lm(formula = ysim ~ experiences)

Residuals:
1 2 3 4
0.01300 -0.01300 0.03

Coefficients:
Estimate Std. Error t value |t|)
(Intercept) 0.97956 0.02436 40.22 0.0 0618 ***
experiences -0.26142 0.03444 -7.59 0.016921 *

Signif. co s: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

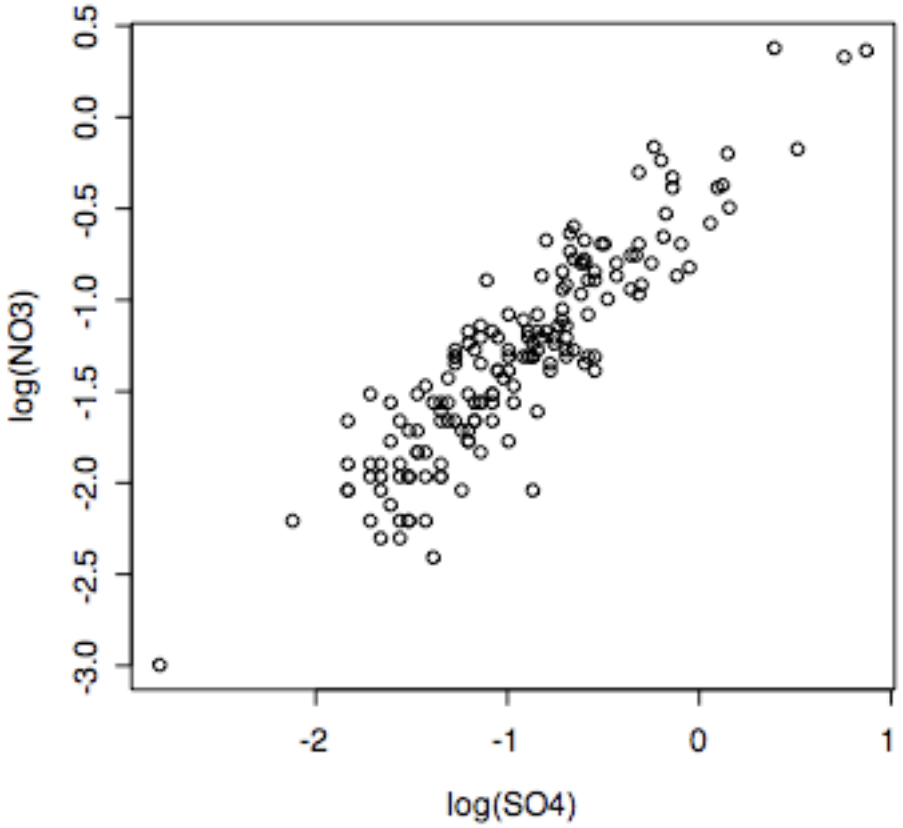
Re andard error: 0.03444 on 2 degrees of freedom
usted R-squared: 0.9497
alue: 0.01692

Moralité : la pente de la droite est négative,
Cette fois l’erreur d’estimation est assez faible
Et le paramètre est statistiquement
significatif au niveau 5% (mais pas 1%)

Remarque : la pente réelle (inconnue) est
- 0.2

Données de pollution

Rappel : on veut prévoir la teneur en NO_3 par celle en $S0_4$



Régression avec R

Le fichier de données : (format .txt)

NO3	SO4
0,45	0,78
0,09	0,25
1,44	2,39
…	…

lm : linear model

```
> pollution <- read.table("pollution.txt", header=TRUE, dec=",", sep="\t")
> modele_degre_1 <- lm(log(NO3)~log(SO4), data=pollution)
> summary(modele_degre_1)
> modele_degre_2 <- lm(log(NO3)~log(SO4)+I(log(SO4)^2), data=pollution)
> summary(modele_degre_2)
```

Sorties à commenter

Call:
lm(formula = log(NO3) ~ log

Residuals:
    Min   1Q Median   3
-0.80424 -0.14485 -0.01087

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.43642	0.03679	-11.86	<2e-16	***
log(SO4)	0.92168	0.03356	27.47	<2e-16	***

Signif. codes: 0 '***' 0.001 '*

Residual standard error: 0.2
freedom
Multiple R-Squared: 0.8205, Adjusted R-squared: 0.8195
F-statistic: 754.4 on 1 and 165 DF, p-value: < 2.2e-16

Comme n-p-1>20, on peut aussi se baser sur le fait que |t-ratio| > 2 ou que l'erreur d'estimation est < la moitié de l'estimation

p-valeur < 0.05
⇒ paramètres significatifs au niveau 5%
(on est même très large : p=2e-16 !)

Call:
lm(formula = log(NO3) ~ log(S

Residuals:
Min 1Q Median 3Q
-0.79819 -0.14085 -0.01470 0

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-0.42918	0.03955	-10.852	<2e-16	***
log(SO4)	0.95337	0.07098	13.432	<2e-16	***
I(log(SO4)^2)	0.01886	0.03720	0.507	0.613	

Signif. codes: 0 '***' 0.001 '**'

Residual standard error: 0.242
Multiple R-Squared: 0.8208,
F-statistic: 375.7 on 2 and 164

Somme n-p-1>20, on peut aussi se baser sur le fait que |t-ratio| < 2
ou
que l'erreur d'estimation est > la moitié de l'estimation

p-valeur > 0.05
⇒ paramètre non significatif au niveau 5%