Introduction à la régression
cours n°3
Influence d'un prédicteur

ENSM.SE – 1A

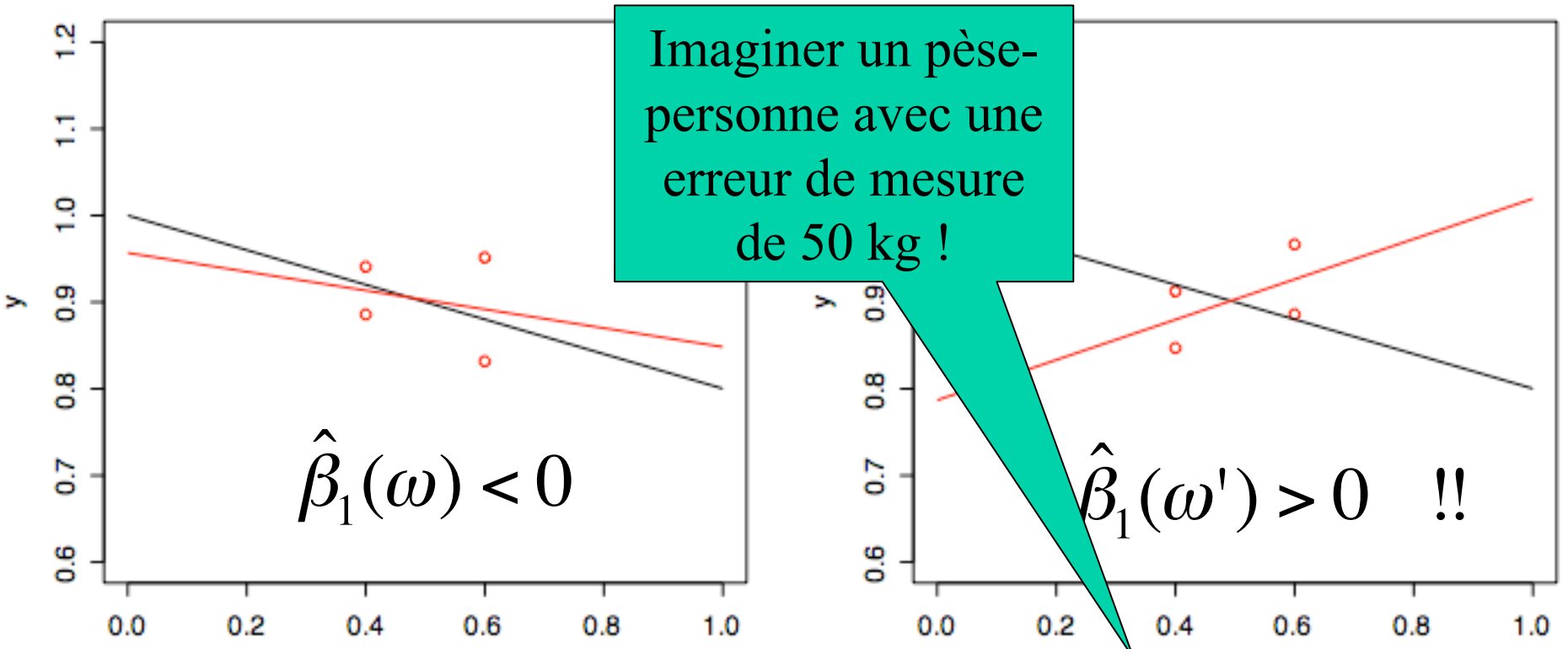
Olivier Roustant

Objectif du cours

- Utiliser les résultats théoriques sur l'estimation des paramètres pour savoir si un prédicteur d'un modèle linéaire est influent

Influent ou non influent ?

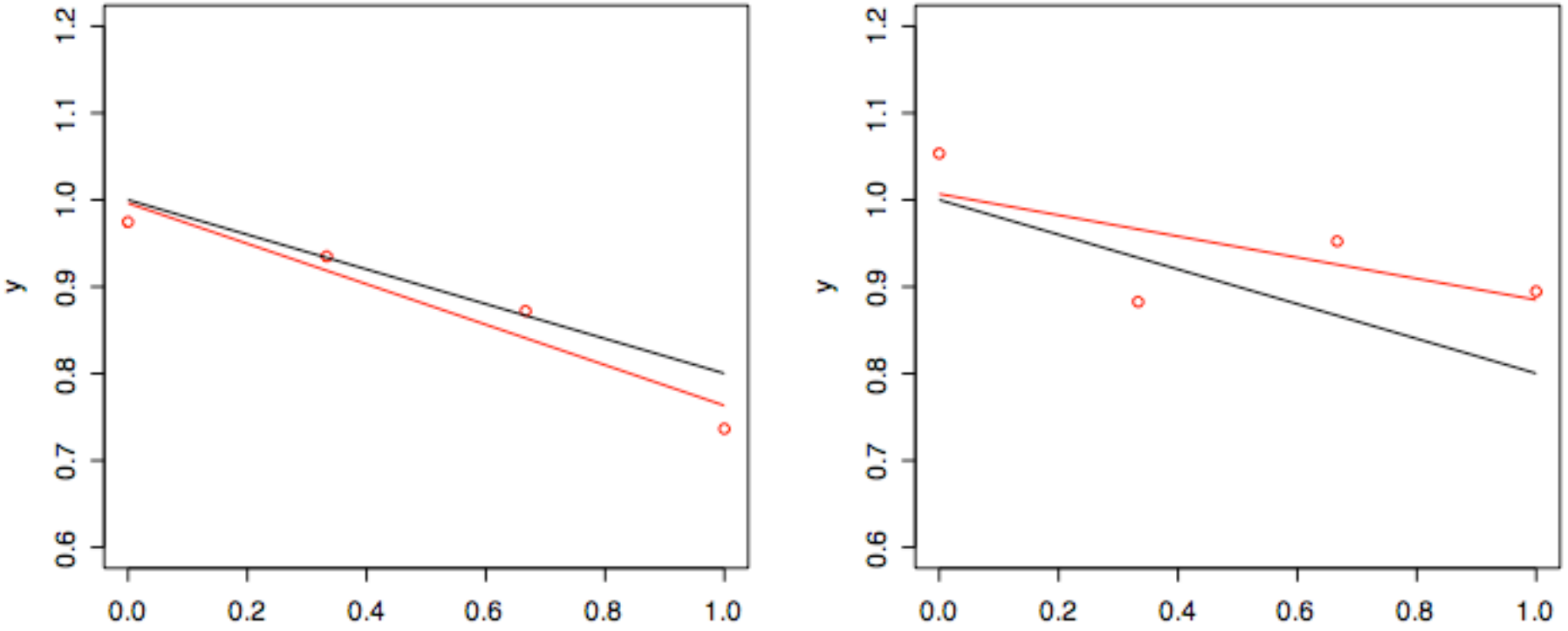
$y_i = \beta_0 + \beta_1 x_i + e_i$ avec $e_1, \ldots, e_4$ i.i.d $N(0, 0.04^2)$



En fait ici, l'erreur d'estimation sur $\beta_1 = |\beta_1|$!! (=0.2)

Influent ou non influent (suite)

Le même ex., mais en planifiant mieux les expériences



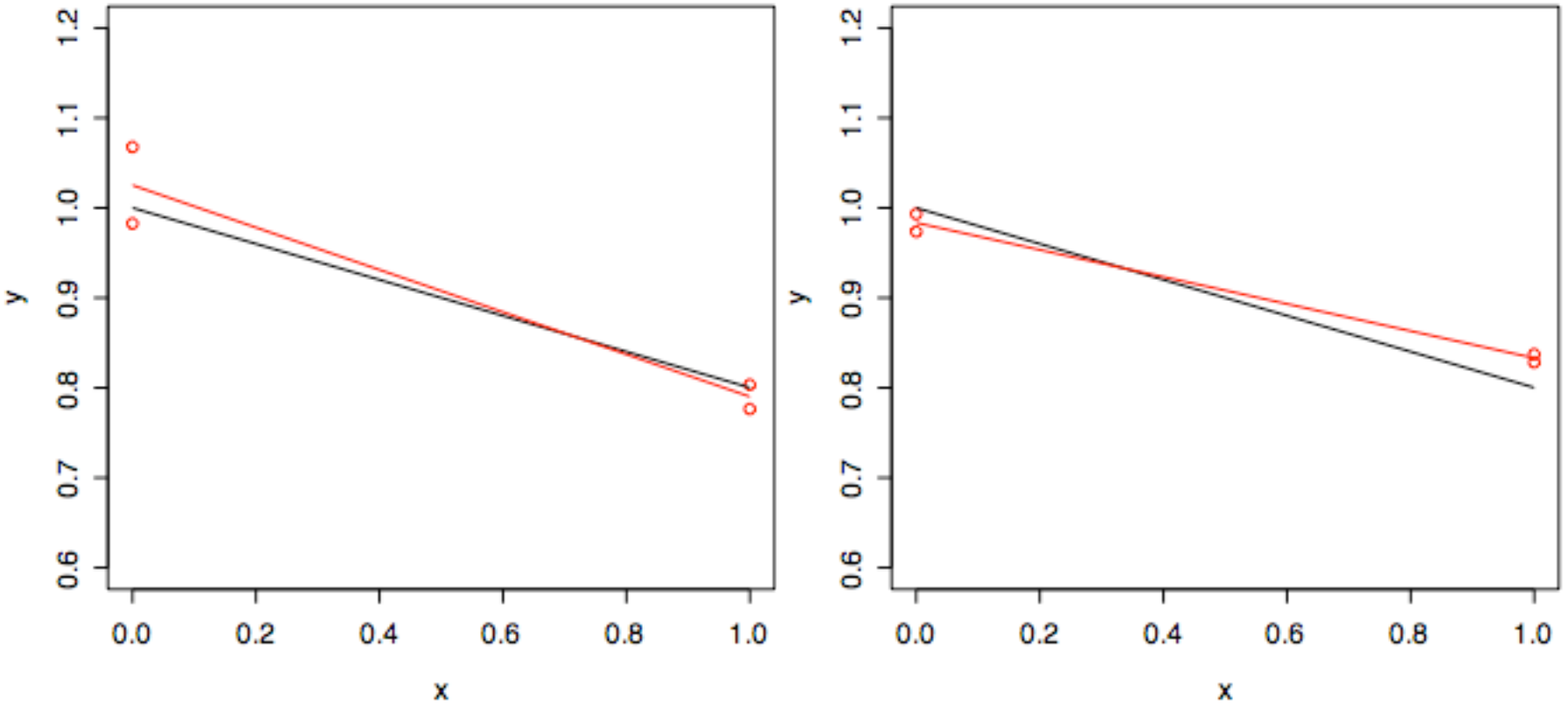
Cette fois, l'erreur d'estimation sur la pente = 0.0536

Exercice

- Vous pouvez réaliser *n* expériences pour estimer un phénomène linéaire sur [a,b] impliquant 1 prédicteur
  - Comment répartir les expériences dans le domaine expérimental [a,b] de façon à ce que l'estimation soit la plus précise possible ?

Influent ou non influent (suite)

Le même exemple, planification optimale (voir diapo. 26)



L'erreur d'estimation sur la pente = 0.04

Influent ou non influent ?
Morale de l'exemple

- Prendre en compte l'erreur d'estimation d'un paramètre pour savoir s'il est important ou pas

  → Décision en milieu incertain : test statistique

- L'impossibilité de décider peut venir d'une mauvaise planification des expériences

Formalisation

- Considérons le modèle linéaire

$$y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_p x_{p,i} + e_i$$

avec $e_1, \ldots, e_n$ i.i.d $N(0,\sigma^2)$

- Le prédicteur x_i est influent si $\beta_i \neq 0$
  - Test statistique opposant les hypothèses

$$\{\beta_i = 0\} \quad \text{et} \quad \{\beta_i \neq 0\}$$

Construction du test statistique

1ère étape : hypothèse H_0

- On veut contrôler le risque de décider qu'un prédicteur est influent alors qu'il ne l'est pas. Quelle est l'hypothèse H_0 ?

$$\boxed{H_0 = \{\beta_i = 0\}}$$

- Autre raison : pouvoir faire les calculs !

Construction du test (suite)

- 2ème étape : choix d'une statistique de décision
  - On part de l'estimateur des moindres carrés (EMC) de β_i
  - Matriciellement, on vérifie qu'on a :

$(Y-X\beta)'(Y-X\beta)$ minimum ssi $X'(Y-X\beta) = 0$

ssi $(X'X)\,\beta = X'Y$

D'où

$$\hat{\beta} = (X'X)^{-1}X'Y$$

Construction du test (suite)

- 2ème étape (suite) : loi de l'EMC sous H_0 ?
  - Une combinaison linéaire de v.a. de lois normales <u>indépendantes</u> est encore de loi normale (admis)
  - Ici, $\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta + e)$
    d'où $\boxed{\hat{\beta} = \beta + (X'X)^{-1}X'e}$
  - Chaque $\hat{\beta}_i$ est donc une combinaison linéaire des e_i, centrée sur $\beta_i \Rightarrow$ loi normale centré sur β_i

Construction du test (suite)

- 2ème étape (suite)
  - Exercice. Pour un vecteur u, n×1, on définit la matrice de covariance par $\text{cov}(u) = (\text{cov}(u_i,u_j))_{1\le i,j\le n}$
    a) Mq $\text{Cov}(u) = E((u-m)(u-m)')$, avec $m=E(u)$ (vect. n×1)
    b) Mq $\text{Cov}(e) = \sigma^2 I_n$
    c) Déduire de a) et b) que $\boxed{\text{cov}(\hat{\beta}) = \sigma^2 (X'X)^{-1}}$
    puis que $\boxed{\text{var}(\hat{\beta}_i) := \sigma_i^{\,2} = \sigma^2((X'X)^{-1})_{ii}}$

- Conclusion : sous H_0, l'EMC est de loi $N(0, \sigma_i^2)$

Construction du test (suite)

- 2ème étape (suite)
  - La loi de l'estimateur dépend des paramètres
  - Intuitivement, sous H_0 :

$$\hat{\beta}_i \sim N(0,\sigma^2((X'X)^{-1})_{ii}) \quad \Rightarrow \quad \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{((X'X)^{-1})_{ii}}} \approx N(0,1)$$

  - Résultat exact : remplacer la loi N(0,1) par la loi de Student t_{n-p-1}.

- Conclusion : choix de la statistique

$$T = \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{((X'X)^{-1})_{ii}}}$$

Construction du test (suite)

- 2ème étape (résumé)
  - Choix de la statistique de décision

$$T = \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{((X'X)^{-1})_{ii}}}$$

  - Interprétation : estimation du paramètre rapporté à son écart-type d'estimation
  - Vocabulaire : T est appelé t-ratio (à cause de la loi de Student, notée t)
    - Propriété : T est de loi de Student t_{n-p-1}

    → En pratique, dès que n-p-1≥20, on approche t_{n-p-1} par N(0,1)

Construction du test (suite)

- 3ème étape : détermination d'un seuil
  - Notation : $T_{obs} = \dfrac{\hat{\beta}_{i,obs}}{\hat{\sigma}_{obs}\sqrt{((X'X)^{-1})_{ii}}}$
  - Au niveau 5%, on rejette H_0
    - n-p-1≥20 : si T_{obs} dépasse 1.96 en valeur absolue
    - n-p-1<20 : utiliser les tables de la loi de Student
    - Mieux (dans tous les cas) : utiliser la p-valeur

Construction du test (fin)

- 3ème étape : p-valeur
  - On appelle p-valeur la probabilité d'obtenir pire que ce qu'on a :

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

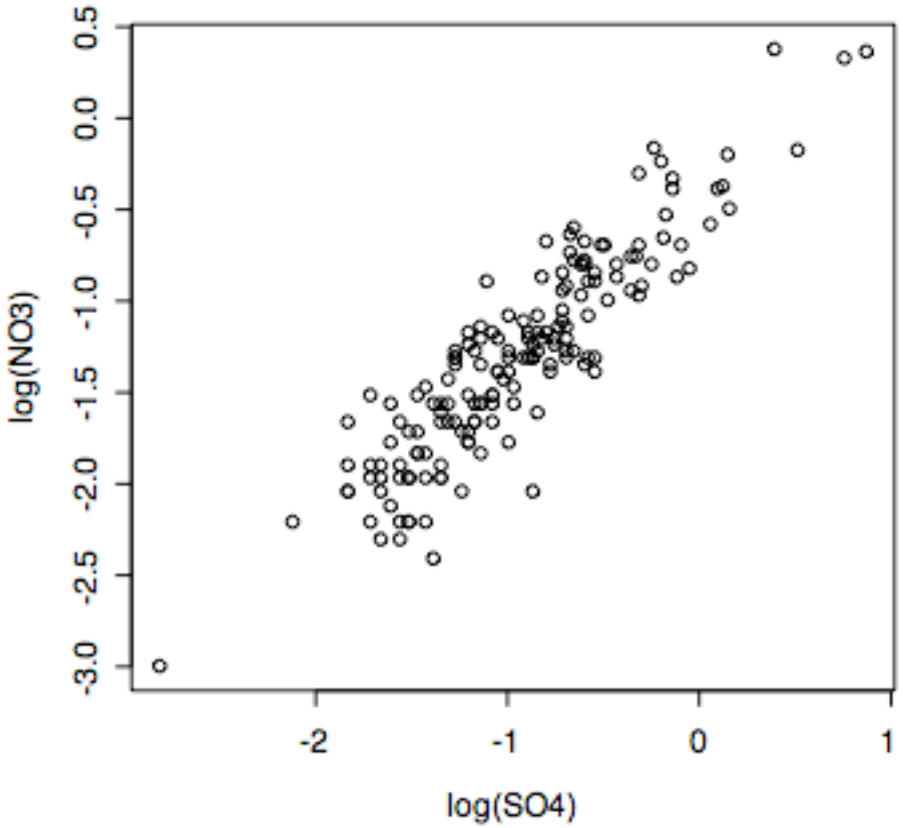
- Permet de ne pas avoir à calculer de seuil :
  - p-valeur $< 5\% \Rightarrow$ rejet au niveau 5%
  - p-valeur $< 1\% \Rightarrow$ rejet au niveau 1%
  - …

Test de signification : pratique

- En pratique, les logiciels donnent le tableau suivant :

coefficient	estimation	erreur d'estimation	t – ratio	p – valeur
β_i	$\hat{\beta}_{i,obs}$	$\hat{\sigma}_{\text{i,obs}}$	$T_{obs} = \frac{\hat{\beta}_{i,obs}}{\hat{\sigma}_{i,obs}}$	$P_{H_0}(\lvert T\rvert > \lvert T_{obs}\rvert)$

Exemple 1
Données de pollution (cf cours 1)



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(S

Residuals:
    Min   1Q Median   3Q
-0.80424 -0.14485 -0.01087 0.

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: 0.2417
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(SO

Residuals:
    Min   1Q Median   3Q
-0.79819 -0.14085 -0.01470 0.16

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 '**'

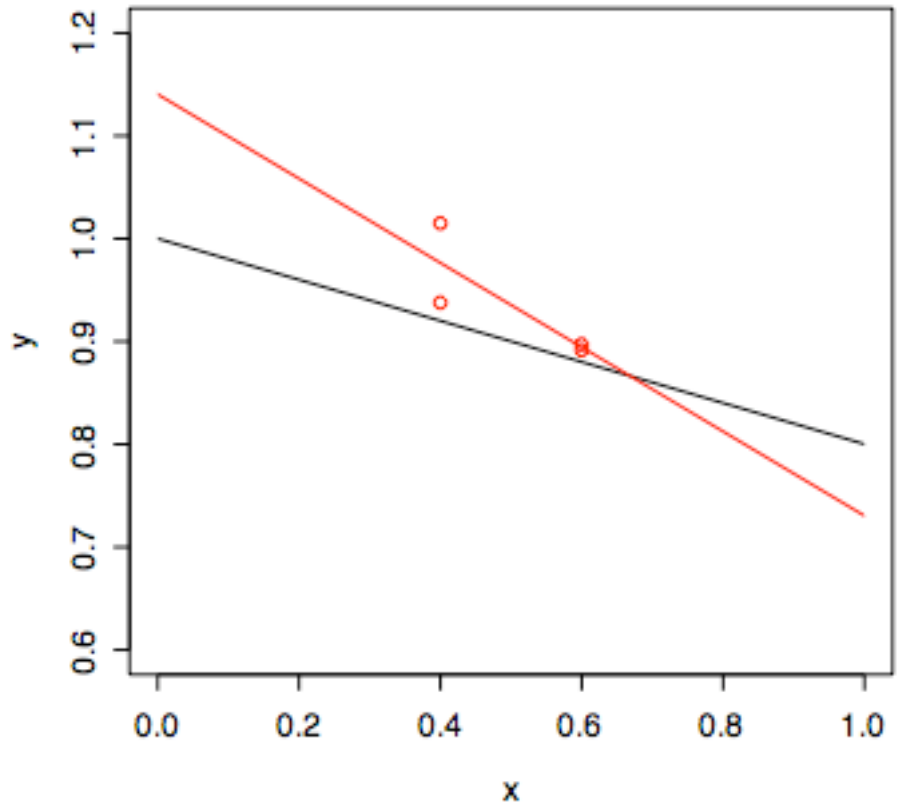
Residual standard error: 0.2423 o
Multiple R-Squared: 0.8208,
F-statistic: 375.7 on 2 and 164 DF

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
⇒ paramètre non significatif au niveau 5%

Exemple 2
Retour sur les simulations (cf transp. n°3)

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



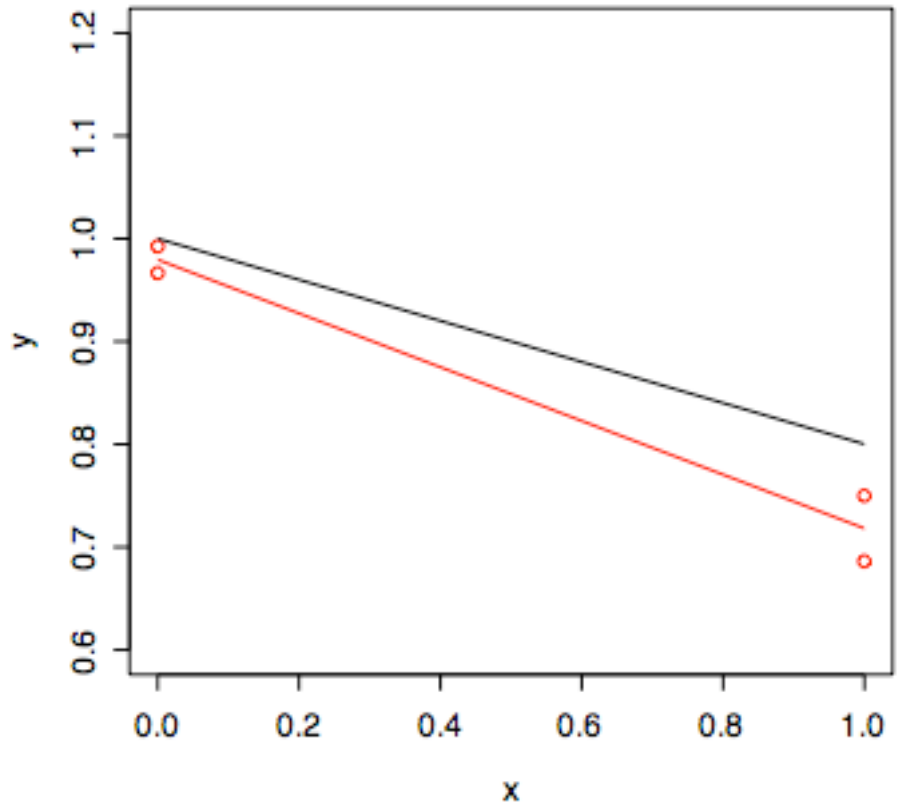
```
Call:
lm(formula = ysim ~ expe

Residuals:
      1       2       3       4
 0.038612 -0.038612  0.00

Coefficients:
             Estimate Std. Error  t valu  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: 0.038
Multiple R-Squared: 0.6916,
F-statistic: 4.485 on 1 and 2 D
```

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=4 << 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 !

Exemple 2 (suite)
Retour sur les simulations (cf transp. n°6)

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



```
Call:
lm(formula = ysim ~ experiences)

Residuals:
      1       2       3       4
 0.01300 -0.01300  0.0319(

Coefficients:
             Estimate Std. Error  t value Pr(
(Intercept)   0.97956    0.02436   40.22 0.00061  ***
experiences -0.26142    0.03444   -7.59 0.016921 *
---
Signif. co      0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Res     dard error: 0.03444 on 2 degrees of freedom
                    usted R-squared: 0.9497
                    e: 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

Exercice : planification des expériences en dimension 1

- Vérifier que, en dimension 1

$$X'X = n\begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \overline{x^2} \end{bmatrix} \quad puis \quad (X'X)^{-1} = \frac{1}{n}\frac{1}{\overline{x^2} - \bar{x}^2}\begin{bmatrix} \overline{x^2} & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}$$

$$puis \quad \mathrm{var}\left(\hat{\beta}_1\right) = \mathrm{cov}(\hat{\beta}_1, \hat{\beta}_1) = \left(\sigma^2 (X'X)^{-1}\right)_{11} = \frac{\sigma^2}{n}\frac{1}{\overline{x^2} - \bar{x}^2}$$

- En déduire que pour minimiser l'erreur d'estimation de la pente, il faut que la variance empirique des x_i soit la plus grande possible
  - Prenons n pair, et considérons le domaine expérimental [-1,1]. Montrer que le maximum est atteint lorsque la moitié des points est placée sur le bord gauche (en x = -1), et l'autre moitié sur le bord droit (x = 1)