

# Probabilités et Statistiques

Année 2010/2011

[laurent.carraro@telecom-st-etienne.fr](mailto:laurent.carraro@telecom-st-etienne.fr)

[olivier.roustant@emse.fr](mailto:olivier.roustant@emse.fr)

# Cours n°13

Régression  
Influence des prédicteurs

# Rappel modèle linéaire

---

---

- $Y$  vecteur des réponses
- $X$  matrice du plan d'expériences
- $\beta$  vecteur des paramètres
- $\varepsilon$  vecteurs des écarts au modèle :

$$Y = X\beta + \varepsilon$$

$\varepsilon_1, \dots, \varepsilon_n$  indépendants et de même loi  $N(0, \sigma^2)$

Ecriture développée :

pour toute « expérience » n°i

$$y_i = \beta$$

$$0 + \beta_1 x_{1,i} + \dots + \beta_p x_{p,i} + \varepsilon_i$$

# Estimation de $\beta$

---

- $\beta$  est estimé par moindres carrés :  
L'estimateur  $\hat{\beta}$  vérifie :

$$(Y - X\beta)'(Y - X\beta) \text{ minimum}$$

On montre que c'est équivalent à  $X'(Y - X\beta) = 0$   
soit  $(X'X)\beta = X'Y$

D'où  $\hat{\beta} = (X'X)^{-1}X'Y$

- $\sigma$  est estimé selon :

$$\hat{\sigma}^2 = \frac{1}{(n - p - 1)} \sum_{i=1}^n \hat{\varepsilon}_i^2$$

# Propriétés de $\hat{\beta}$

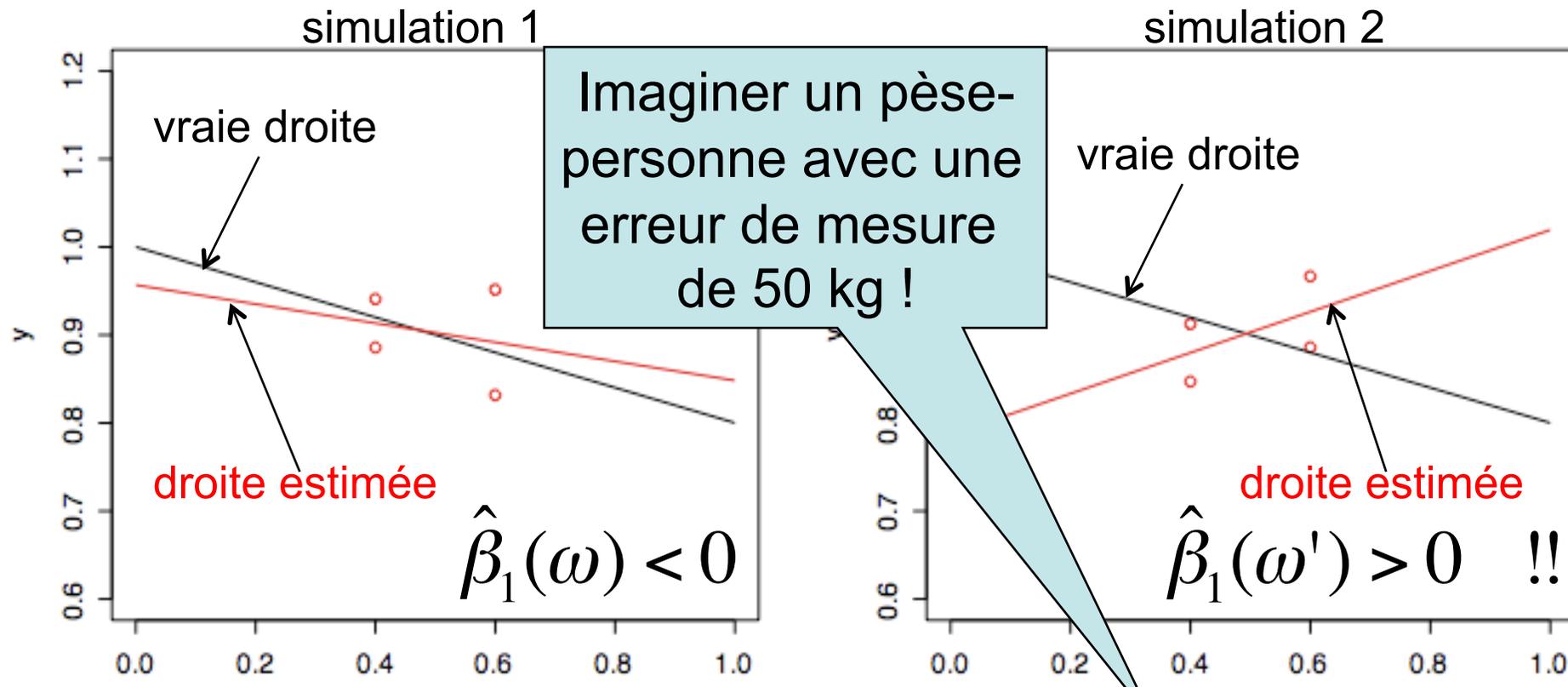
---

---

- $E(\hat{\beta}) = \beta$  : l'estimateur est sans biais.
- $\text{cov}(\hat{\beta}) = \sigma^2 (X' X)^{-1}$
- En d'autres termes, si  $M = (X' X)^{-1}$ 
  - $\text{Var}(\hat{\beta}_j) = \sigma^2 M_{j,j}$
  - $\text{Cov}(\hat{\beta}_j, \hat{\beta}_k) = \sigma^2 M_{j,k}$
- Les  $\hat{\beta}_j$  sont tous de loi normale  $N(\beta_j, \sigma^2 M_{j,j})$

# Influent ou non influent ?

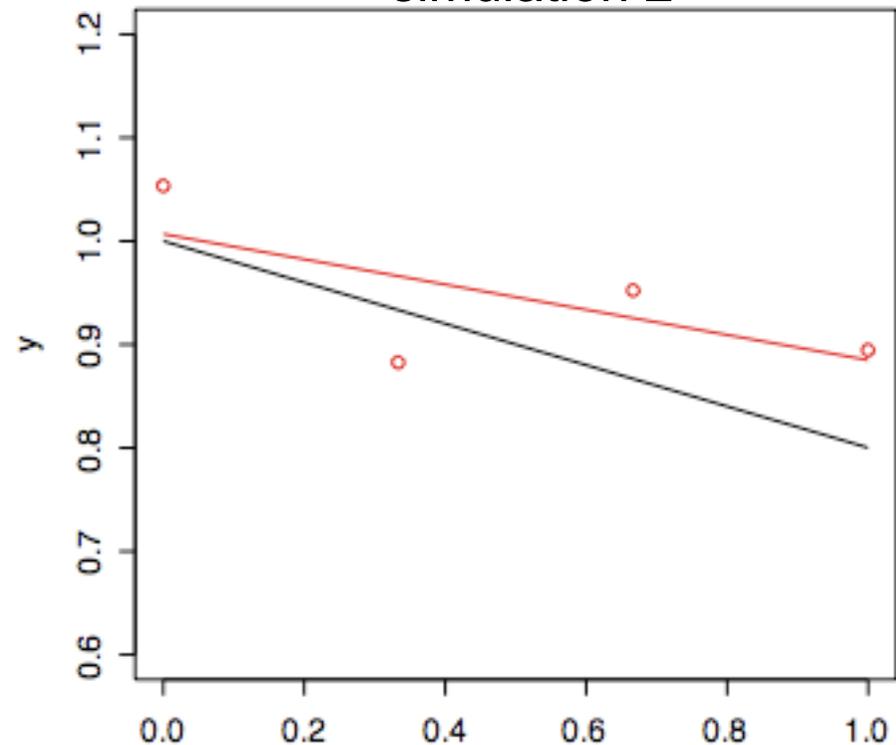
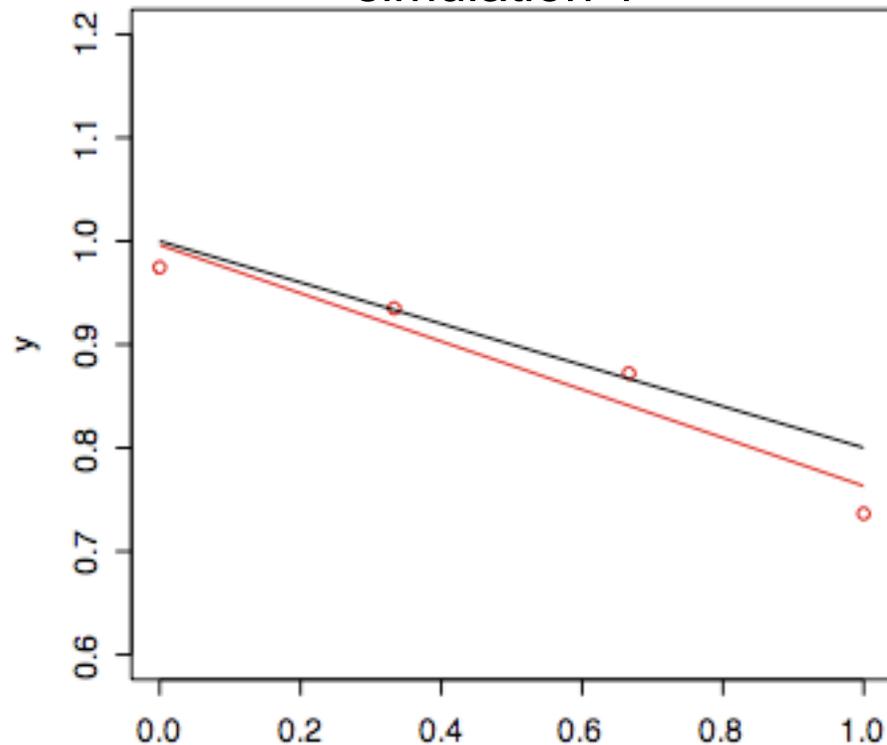
$$y_i = \beta_0 + \beta_1 x_i + e_i \text{ avec } e_1, \dots, e_4 \text{ i.i.d } N(0, 0.04^2)$$



Ici, l'erreur d'estimation sur  $\beta_1$  vaut  $|\beta_1|$  ! (=0.2)

# Influent ou non influent (suite)

Le même ex., mais en planifiant mieux les expériences



Ici, l'erreur d'estimation sur la pente vaut  $0.054$

# Question

---

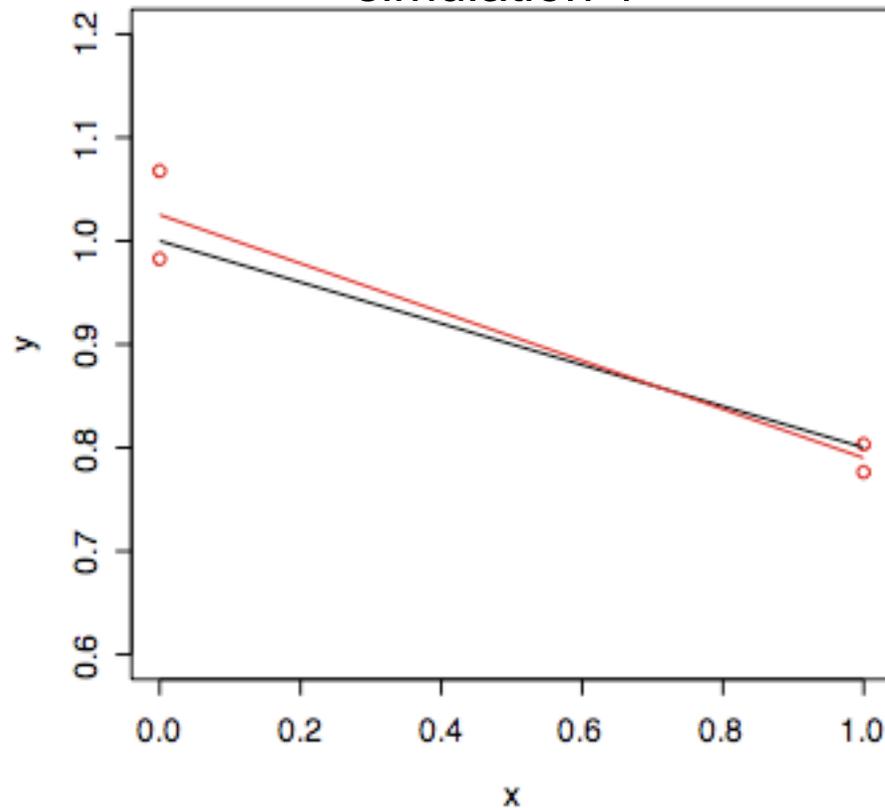
---

- Vous pouvez réaliser 4 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 ?

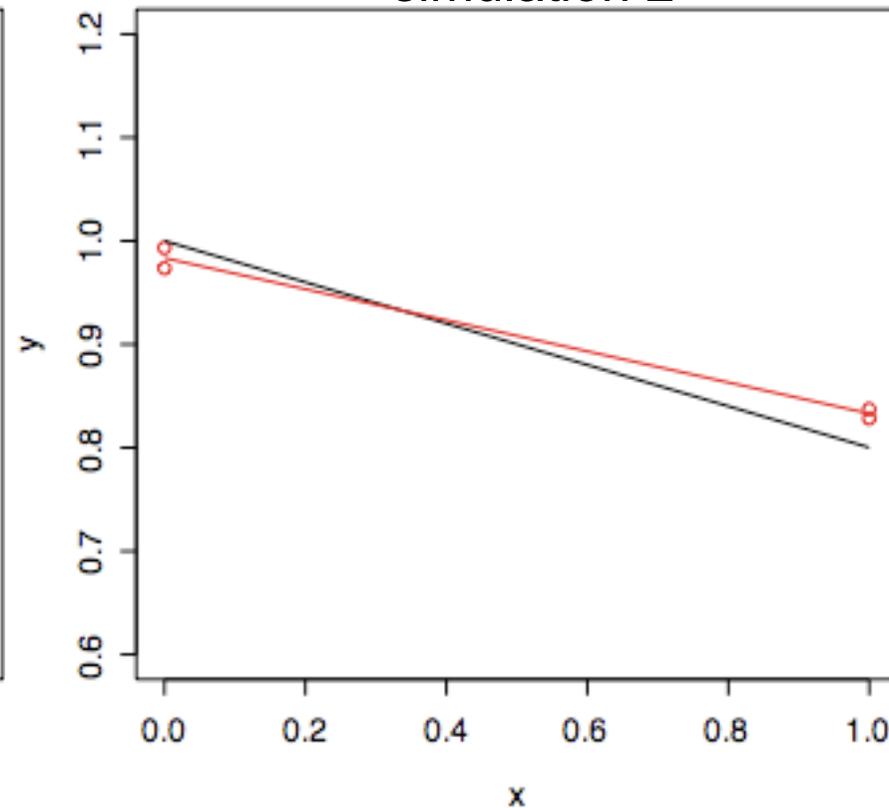
# Influent ou non influent (suite)

Le même exemple, planification optimale

simulation 1



simulation 2



L'erreur d'estimation sur la pente vaut ici 0.04

# Conclusion temporaire

---

---

- Prendre en compte l'erreur d'estimation d'un paramètre pour savoir s'il est important ou pas
  - Décision en milieu incertain : **test statistique**
- La difficulté à décider peut venir d'une mauvaise **planification des expériences**

# Formalisation

---

- Considérons le modèle linéaire

$$Y = X\beta + \varepsilon \text{ soit :}$$

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

avec  $\varepsilon_1, \dots, \varepsilon_n$  i.i.d  $N(0, \sigma^2)$

- Le prédicteur  $x_j$  est influent si  $\beta_j \neq 0$

⇒ Test statistique opposant les hypothèses  
 $\{\beta_j = 0\}$  et  $\{\beta_j \neq 0\}$

# Construction du test statistique

---

---

➤ 1ère étape : hypothèse  $H_0$

- On veut que les données nous montrent qu'un prédicteur est influent. Quelle est l'hypothèse  $H_0$  ?

$$H_0 = \left\{ \beta_j = 0 \right\}$$

- Raison inavouable : pouvoir faire les calculs...

# Construction du test (suite)

---

---

- 2ème étape : loi de  $\hat{\beta}_j$  sous  $H_0$  ?
- De façon générale,  $\hat{\beta}_j$  est de loi  $N(\beta_j, \sigma^2 M_{j,j})$
  - Mais sous  $H_0$  :  $\beta_j = 0$
  - donc sous  $H_0$   $\hat{\beta}_j$  est de loi  $N(0, \sigma^2 M_{j,j})$

et  $\frac{\hat{\beta}_j}{\sigma \sqrt{M_{j,j}}}$  est de loi  $N(0,1)$

- On remplace  $\sigma$  par son estimation :

$$T = \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{M_{j,j}}} \text{ est de loi de Student } t_{n-p-1}$$

# Construction du test (suite)

---

---

## ➤ 3ème étape : détermination d'un seuil

- Notation :

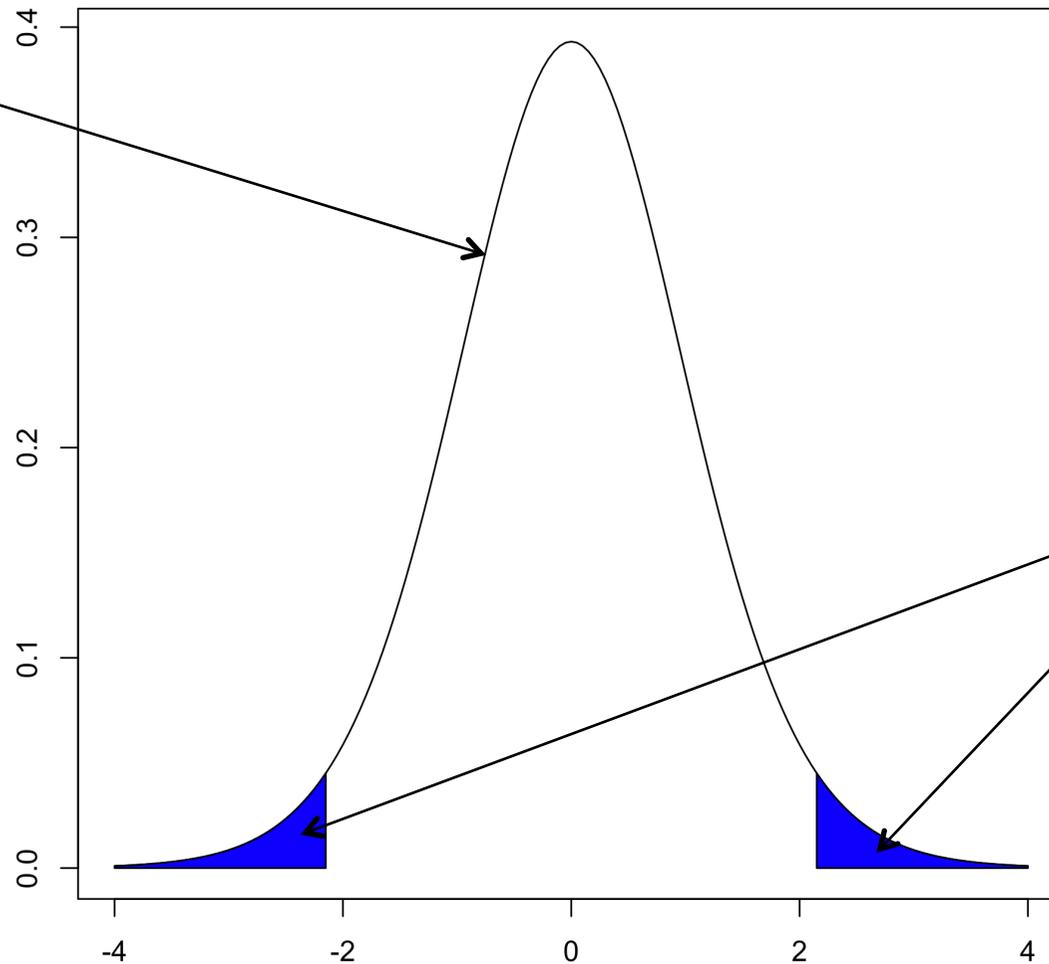
$$T_{obs} = \frac{\hat{\beta}_{j,obs}}{\hat{\sigma}_{obs} \sqrt{M_{j,j}}}$$

- Remarques :
  - T est appelé **t-ratio** (à cause de la loi de Student, notée **t**)
  - En pratique, pour  $n-(p+1) \geq 20$ , on approche  $t_{n-(p+1)}$  par  $N(0,1)$
- Au niveau 5%, on rejette  $H_0$ 
  - $n-(p+1) \geq 20$  : 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**

# Interprétation

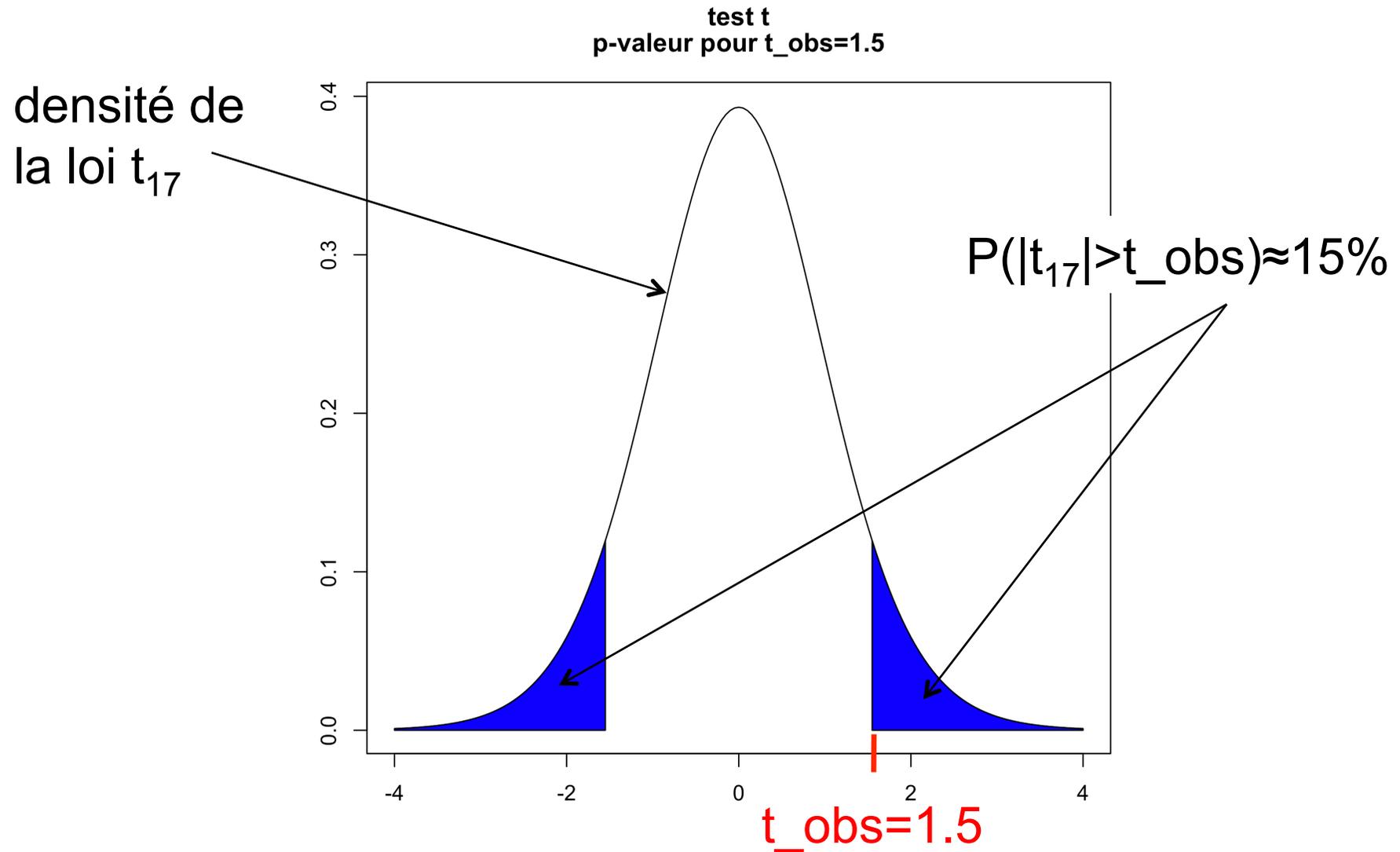
test t de niveau 5%

densité de  
la loi  $t_{17}$



5%

# Idem avec p-valeur



# Test de signification : pratique

- En pratique, les logiciels donnent le tableau suivant :

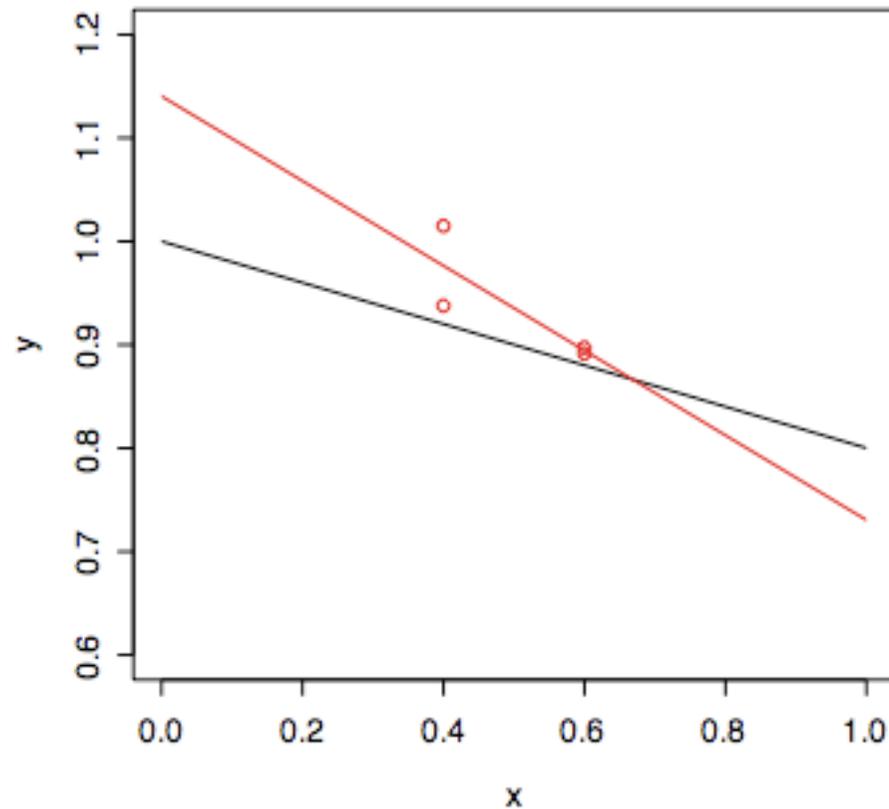
prédicteur	estimation	erreur d'estimation	t - ratio	p - valeur
$x_j$	$\hat{\beta}_{j,obs}$	$\hat{\sigma}_{j,obs}$	$T_{obs} = \frac{\hat{\beta}_{j,obs}}{\hat{\sigma}_{j,obs}}$	$P_{H_0} ( T  >  T_{obs} )$

# Retour sur les simulations

---

---

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



Call:

lm(formula = ysim ~ exp

Residuals:

1	2	3
0.038612	-0.038612	0

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  $n-2=2 \ll 20$ )  
 La p-valeur est calculée à partir de la loi de Student

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 2

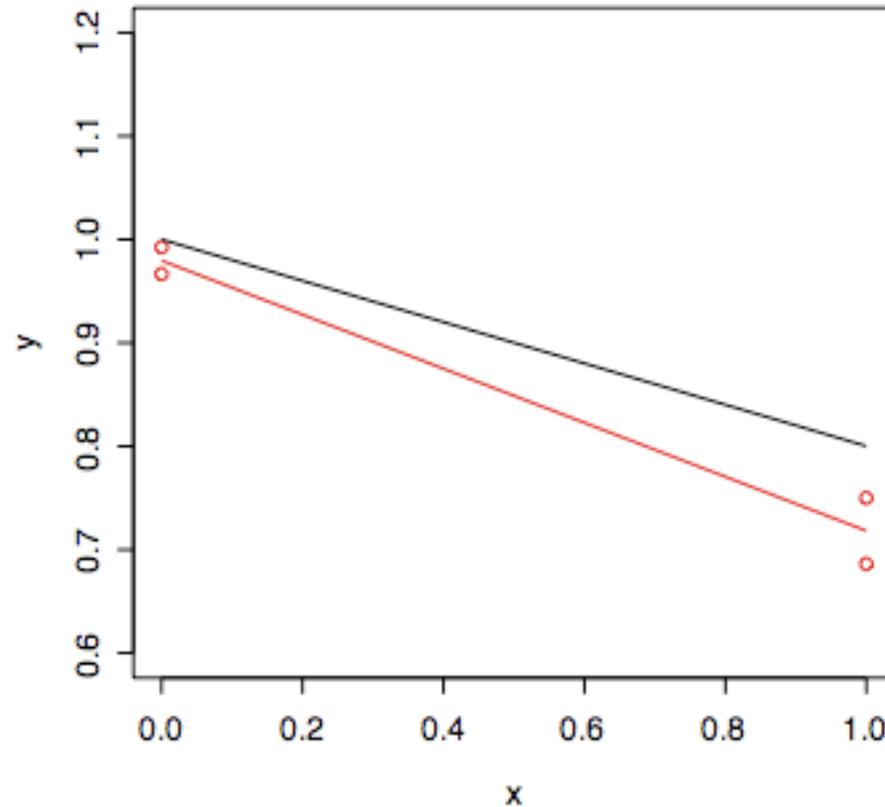
Moralité : la pente de la droite est négative,  
 Mais l'erreur d'estimation est trop importante  
 Et le paramètre est statistiquement non  
significatif au niveau 5% ...rassurant !

# Simulations (suite)

---

---

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



Call:

lm(formula = ysim ~ experiences)

Residuals:

1	2	3	4
0.01300	-0.01300	0.03	

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%)

Coefficients:

	Estimate	Std. Error	t value	t	
(Intercept)	0.97956	0.02436	40.22	0.000618	***
experiences	-0.26142	0.03444	-7.59	0.016921	*

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03444 on 2 degrees of freedom  
Adjusted R-squared: 0.9497  
value: 0.01692

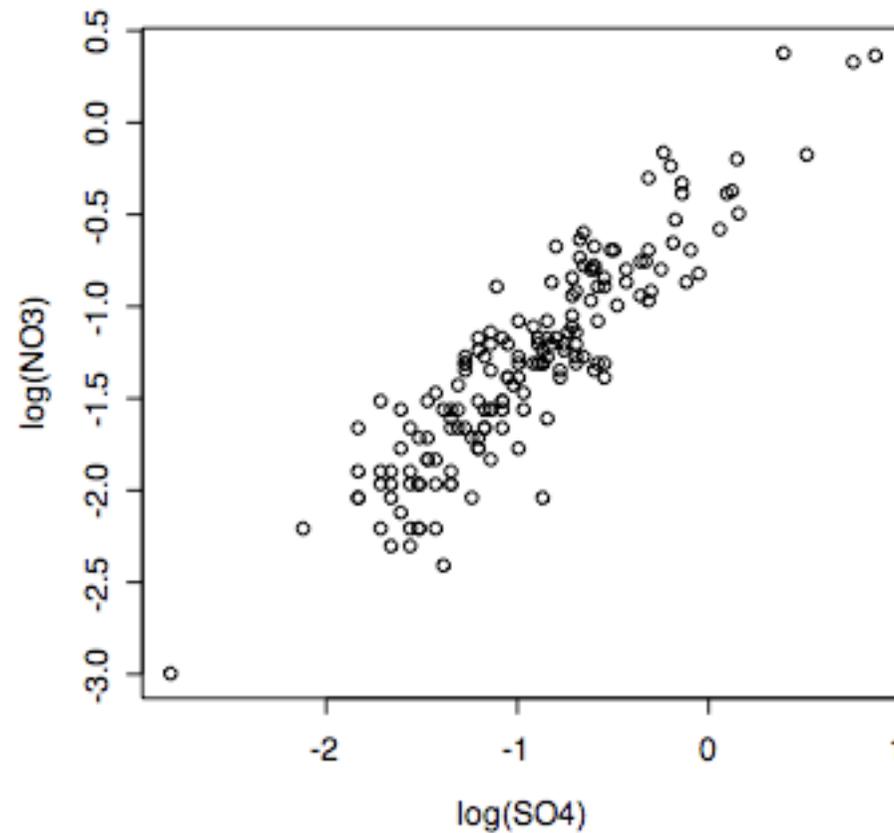
Remarque : la pente réelle (inconnue) est - 0.2

# Données de pollution

---

---

Rappel : on veut prévoir la teneur en  $\text{NO}_3$  par celle en  $\text{SO}_4$



# Régression avec R

Le fichier de données : NO3 SO4  
(format .txt) 0,45 0,78  
0,09 0,25  
Im : 1,44 2,39  
linear model ... ..

```
> pollution <- readtable("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(SO4), data = data)
   
 Comme  $n-p-1 > 20$ , on peut aussi se baser

sur le fait que  $|t\text{-ratio}| > 2$

ou

que l'erreur d'estimation est

< la moitié de l'estimation

Residuals:

Min	1Q	Median	3Q	Max
-0.80424	-0.14485	-0.01087	0.14485	0.80424

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 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

p-valeur < 0.05

⇒ paramètres significatifs au niveau 5%

(on est même très large :  $p=2e-16$  !)

Residual standard error: 0.24 on 165 degrees of 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

Call:

lm(formula = log(NO3) ~ log(SO4), data = data)
   
Somme n-p-1 > 20, on peut aussi se baser

sur le fait que  $|t\text{-ratio}| < 2$

ou

que l'erreur d'estimation est

> la moitié de l'estimation

Residuals:

Min	1Q	Median	3Q	0
-0.79819	-0.14085	-0.01470	0	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 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.242

p-valeur > 0.05

Multiple R-Squared: 0.8208,  $\Rightarrow$  paramètre non significatif au niveau 5%

F-statistic: 375.7 on 2 and 164

# Exercice

## planification d'expériences

Exemple précédent (n expériences et un facteur)

➤ Vérifier que l'on a

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

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

➤ 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

- Pour  $n=4$ , 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 ( $x = 1$ )