Probabilités et Statistiques

Année 2009/2010

laurent.carraro@telecom-st-etienne.fr

olivier.roustant@emse.fr

telecom Saint-Etienne

Cours n°13

Régression
Influence des prédicteurs et ANOVA

Rappel modèle linéaire

- Y vecteur des réponses
- X matrice du plan d'expériences
- β vecteur des paramètres
- ε vecteurs des écarts au modèle :

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

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

Ecriture développée :

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

Estimation de β

- **β est estimé par moindres carrés :**

  L'estimateur $\hat{\beta}$ vérifie :

  (Y-Xβ)'(Y-Xβ) minimum

  On montre que c'est équivalent à X'(Y-Xβ) = 0
  soit (X'X) β = X'Y

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

- **σ est estimé selon :**

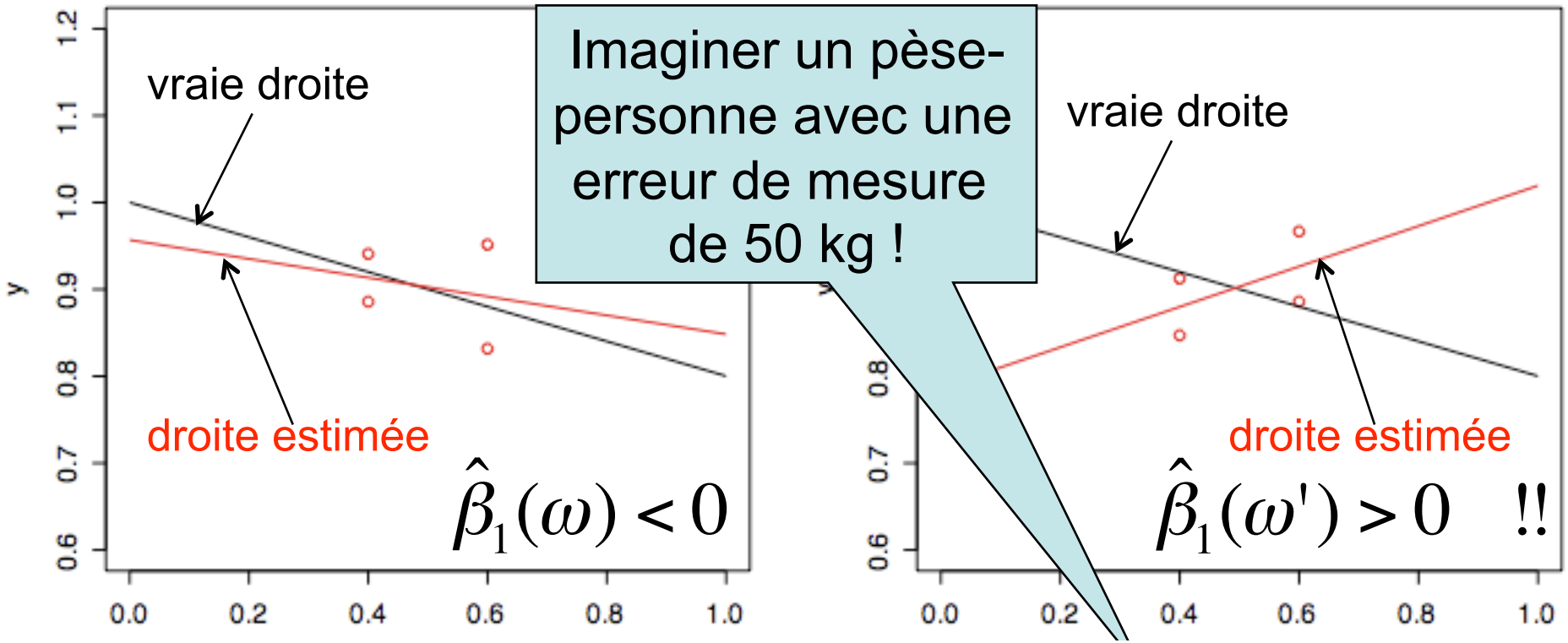
$$\hat{\sigma}^2 = \frac{1}{(n-p-1)} \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}$$

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 $M = (X'X)^{-1}$
  - $Var(\hat{\beta}_i) = \sigma^2 M_{i,i}$
  - $Cov(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 M_{i,j}$
- Les $\hat{\beta}_i$ sont tous de loi normale $N(\beta_i, \sigma^2 M_{i,i})$

Influent ou non influent ?

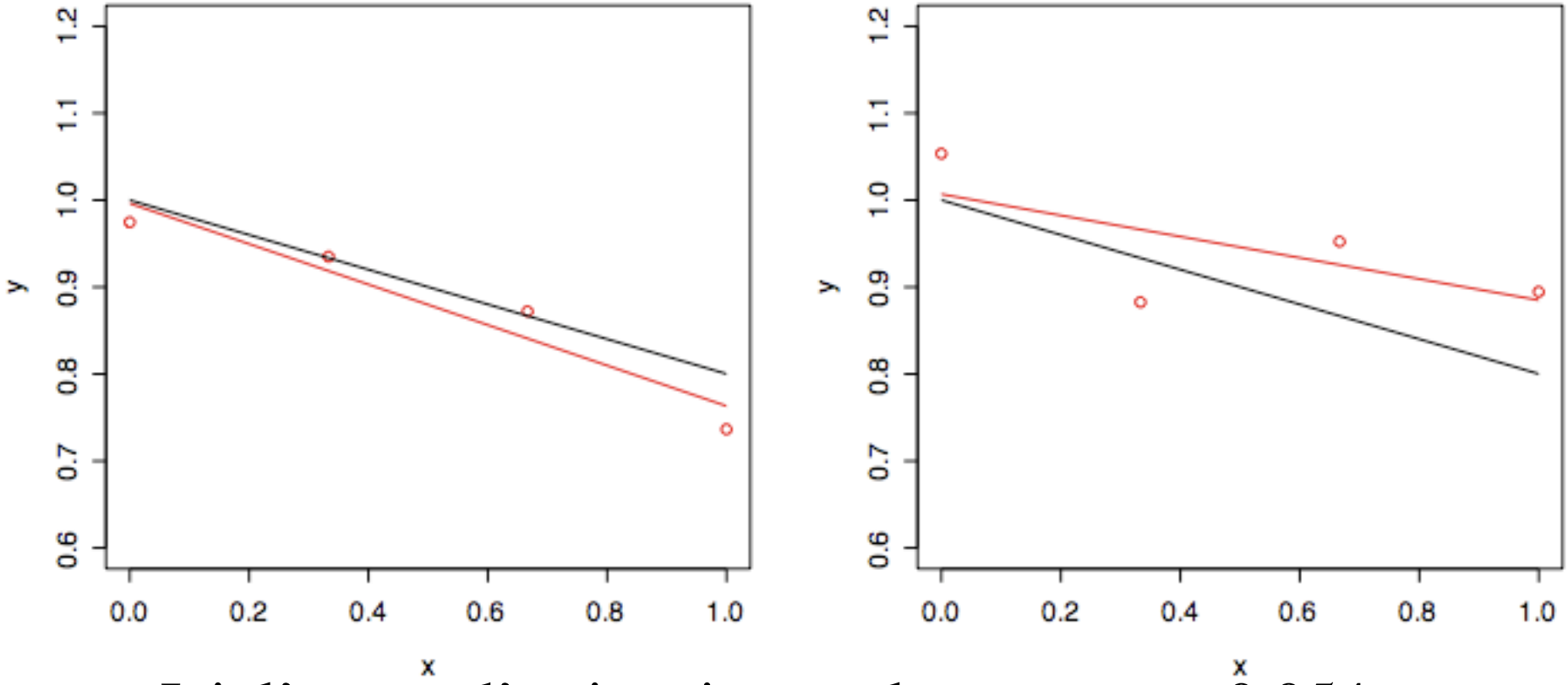
$$y_i = \beta_0 + \beta_1 x_i + e_i \quad \text{avec} \quad e_1, \ldots, e_4 \quad \text{i.i.d } N(0, 0.04^2)$$



Ici, l'erreur d'estimation sur β_1 vaut $|\beta_1|$! (=0.2)

Influent ou non influent (suite)

Le même ex., mais en planifiant mieux les expériences



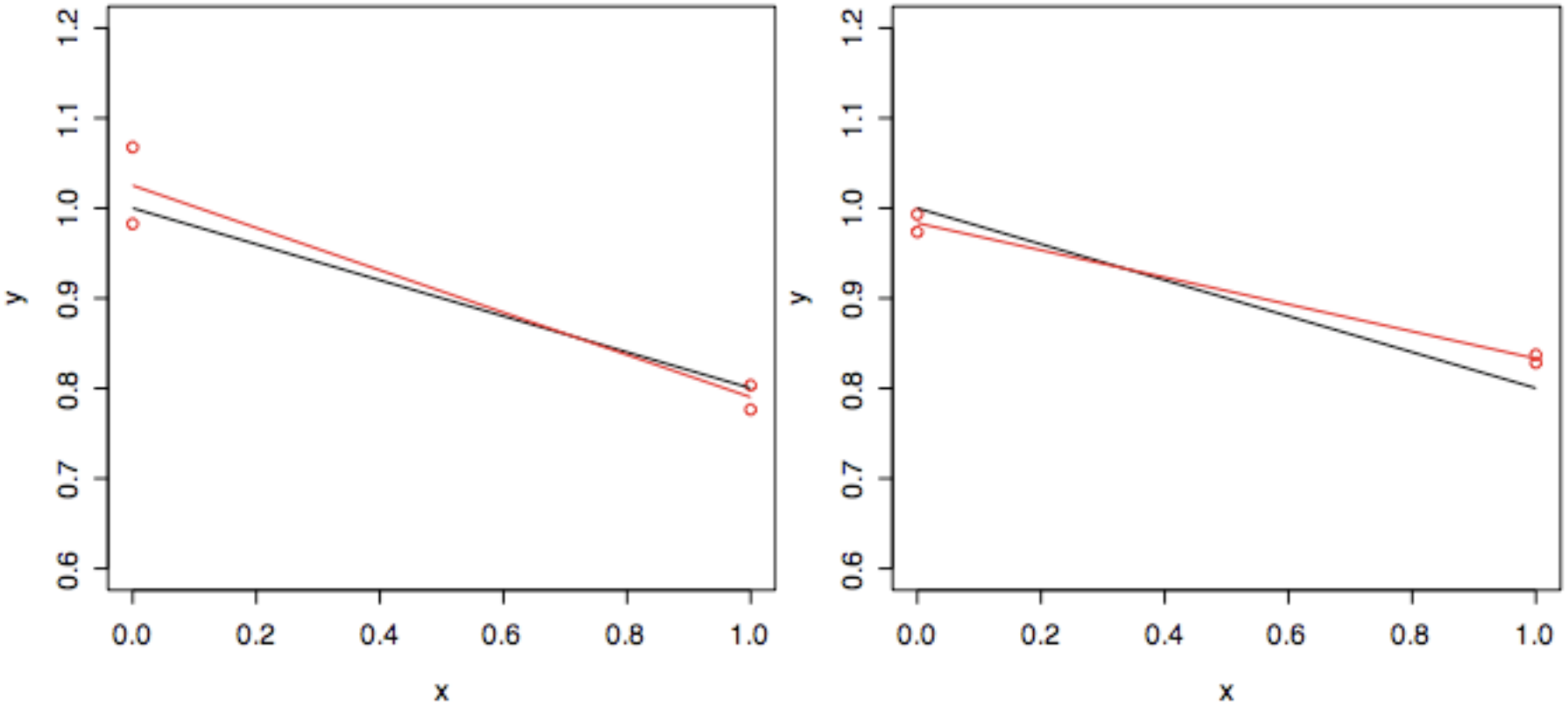
Ici, l'erreur d'estimation sur la pente vaut 0.054

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



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

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

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

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

- Le prédicteur x_i est influent si $\beta_i \neq 0$

$\Longrightarrow$ Test statistique opposant les hypothèses

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

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 ?

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

  - Raison inavouable : pouvoir faire les calculs…

Construction du test (suite)

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

et $\dfrac{\hat{\beta}_i}{\sigma\sqrt{M_{i,i}}}$ est de loi N(0,1)

  - On remplace σ par son estimation :

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

Construction du test (suite)

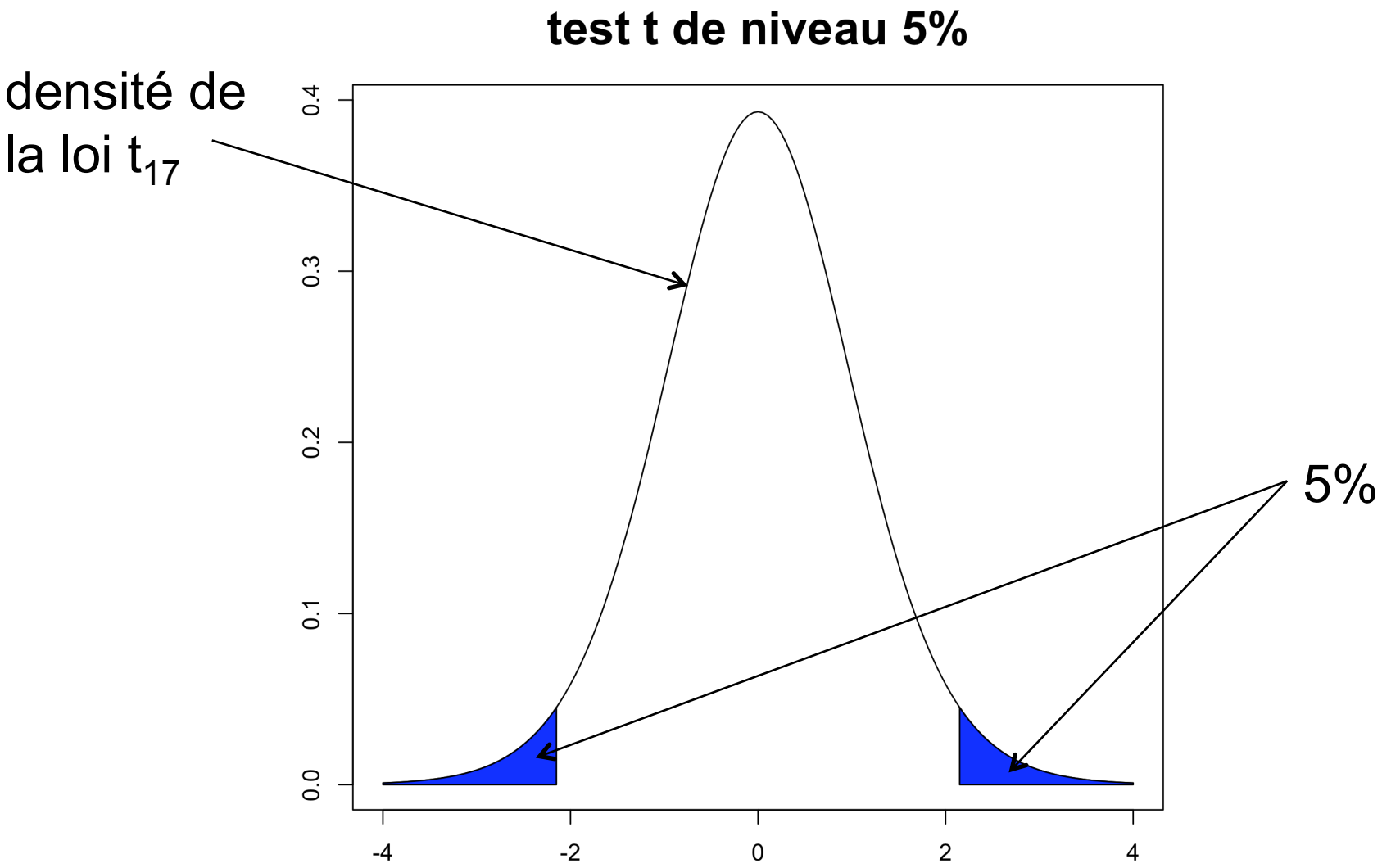
3ème étape : détermination d'un seuil

- Notation :

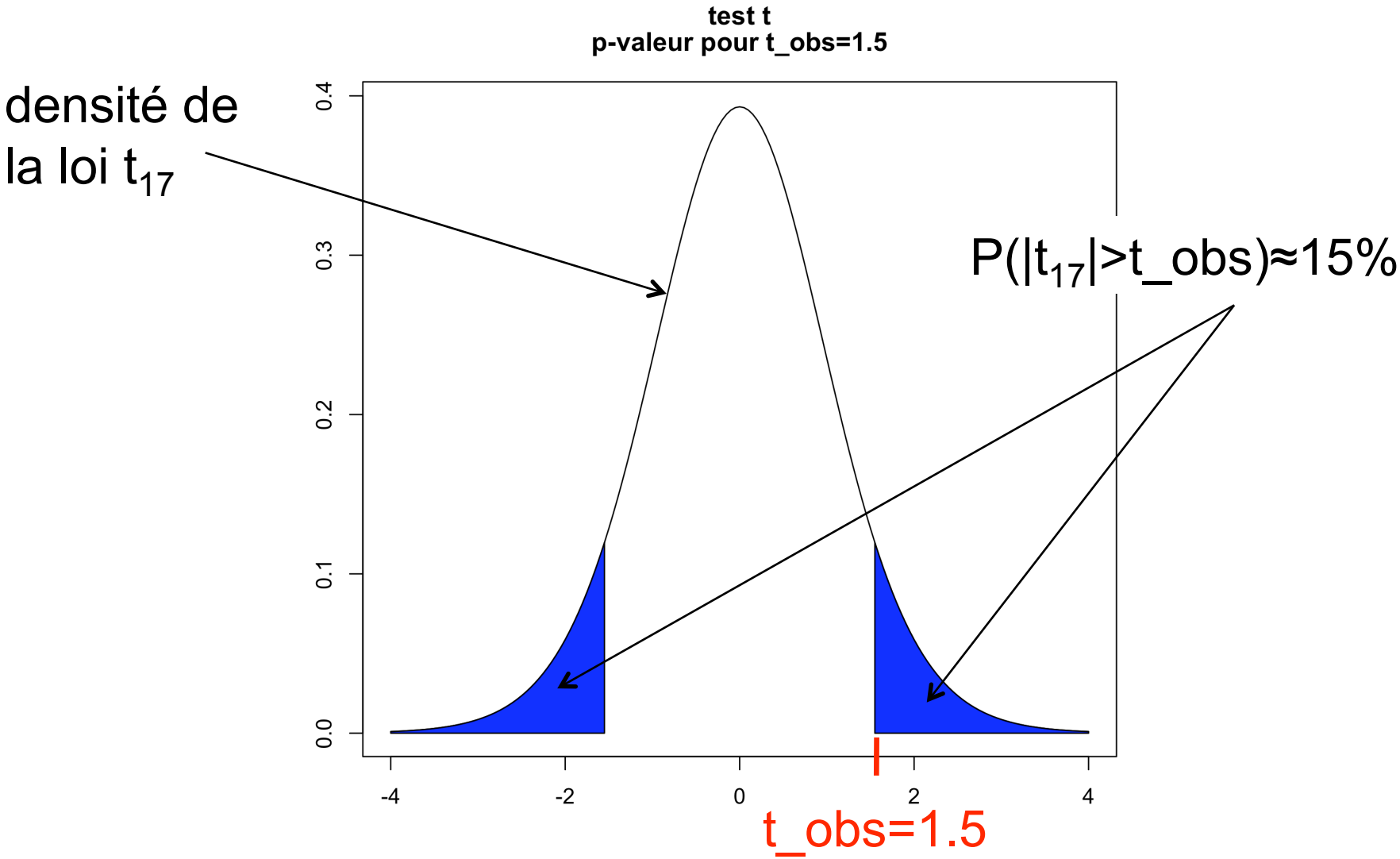
$$T_{obs} = \frac{\hat{\beta}_{i,obs}}{\hat{\sigma}_{obs}\sqrt{M_{i,i}}}$$

- Remarques :
  - T est appelé t-ratio (à cause de la loi de Student, notée t)
  - En pratique, pour n-(p+1)≥20, on approche $t_{n-(p+1)}$ par N(0,1)
- 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

Interprétation



Idem avec p-valeur



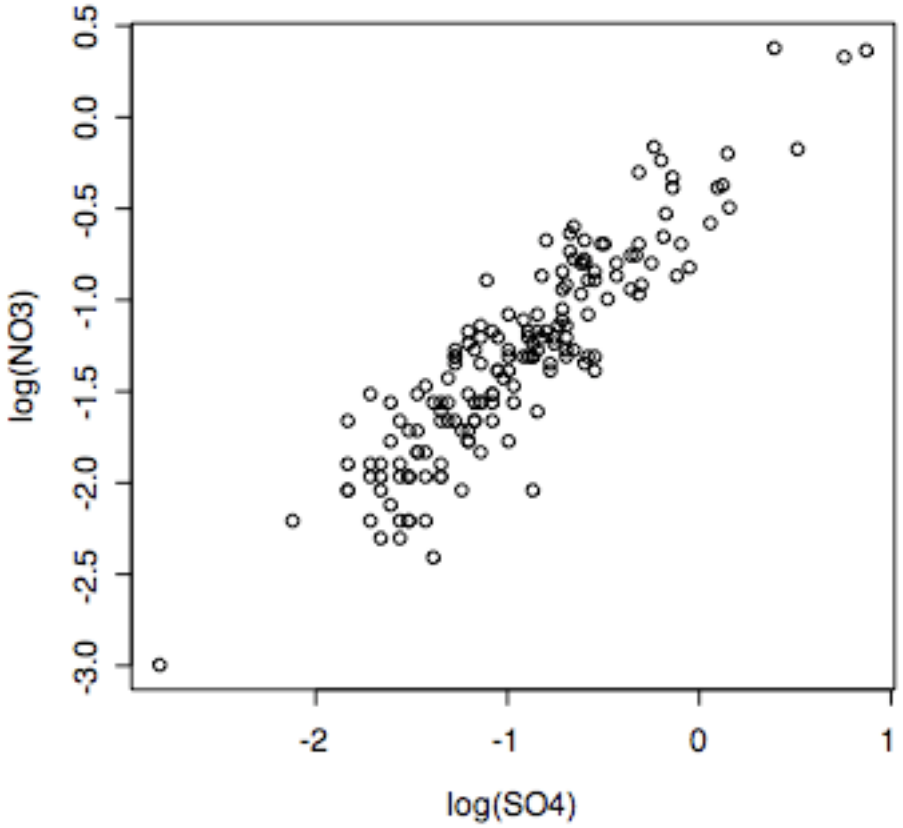
Test de signification : pratique

- En pratique, les logiciels donnent le tableau suivant :

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

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 '**'

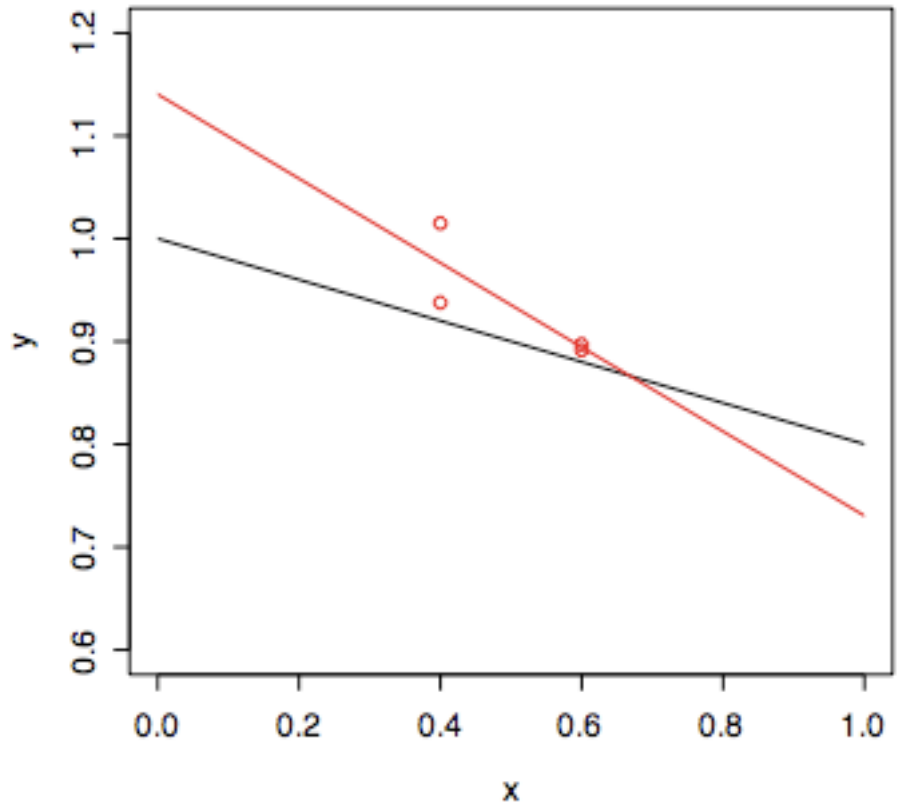
Residual standard error: 0.242
Multiple R-Squared: 0.8208,
F-statistic: 375.7 on 2 and 164

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%

Retour sur les simulations

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



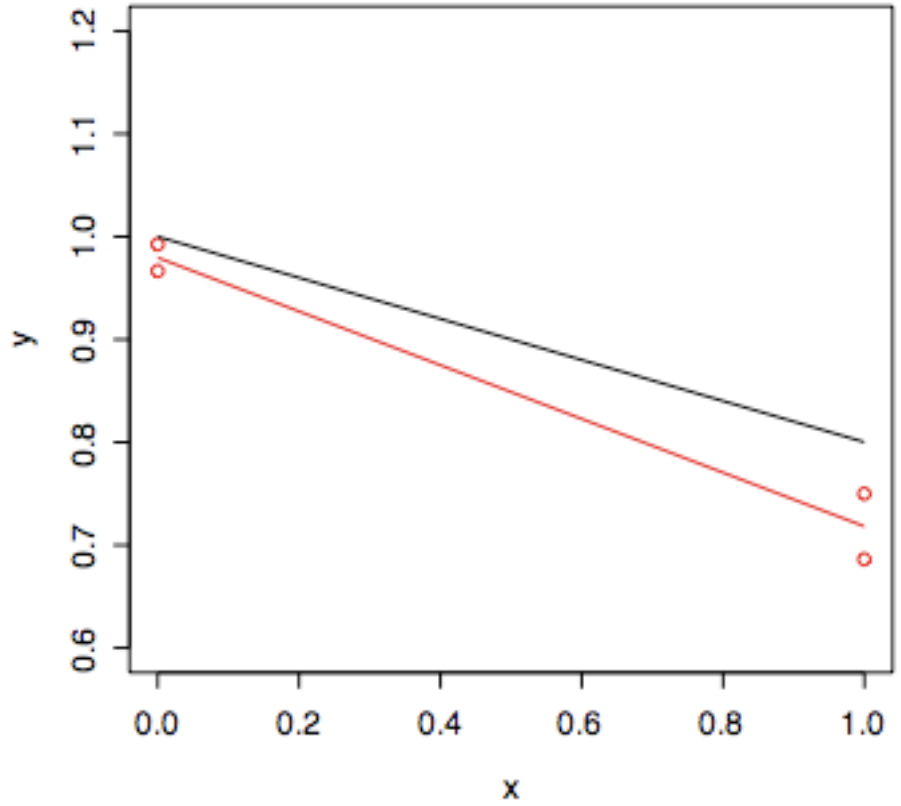
```
Call:
lm(formula = ysim ~ exp

Residuals:
       1        2        3
 0.038612 -0.038612  0

Coefficients:
             Estimate Std. Error  t    ue  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 0.00618 ***
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

Exercice planification d'expériences

Exemple précédent (n expériences et un facteur)

- Vérifier que l'on a

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

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

- 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
  - 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)