

# Modèles de prévision

## Partie 1 - régression

Arthur Charpentier

charpentier.arthur@uqam.ca

[http ://freakonometrics.blog.free.fr/](http://freakonometrics.blog.free.fr/)



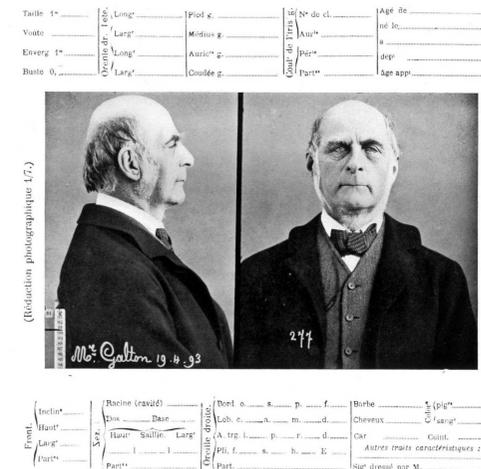
AUTOMNE 2012

## Plan du cours

- **Motivation et introduction aux modèles de régression**
- **Le modèle linéaire simple**
  - Résultats généraux
  - Approche matricielle
- **Le modèle linéaire multiple**
  - Résultats généraux
  - Tests, choix de modèle, diagnostique
- **Aller plus loin**
  - Les modèles non linéaires paramétriques
  - Les modèles non linéaires nonparamétriques

## Un peu de terminologie

L'étude de la transmission génétique de certaines caractéristiques a intéressé Galton en 1870 puis Pearson en 1896. Galton a proposé d'étudier la taille d'un enfant en fonction de la taille (moyenne) de ses parents, à partir de 928 observations.



*Table 8.1.* Galton's 1885 cross-tabulation of 928 adult children born of 205 midparents, by their height and their midparent's height.

Height of the midparent in inches	Height of the adult child														Total no. of adult children	Total no. of midparents	Medians
	<61.7	62.2	63.2	64.2	65.2	66.2	67.2	68.2	69.2	70.2	71.2	72.2	73.2	>73.7			
> 73.0	—	—	—	—	—	—	—	—	—	—	—	1	3	—	4	5	—
72.5	—	—	—	—	—	—	—	1	2	1	2	7	2	4	19	6	72.2
71.5	—	—	—	—	1	3	4	5	5	10	4	9	2	2	43	11	69.9
70.5	1	—	1	—	1	1	3	12	18	14	7	4	3	3	68	22	69.5
69.5	—	—	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68.9
68.5	1	—	7	11	16	25	31	34	48	21	18	4	3	—	219	49	68.2
67.5	—	3	5	14	15	36	38	28	38	19	11	4	—	—	211	33	67.6
66.5	—	3	3	5	2	17	17	14	13	4	—	—	—	—	78	20	67.2
65.5	1	—	9	5	7	11	11	7	7	5	2	1	—	—	66	12	66.7
64.5	1	1	4	4	1	5	5	—	2	—	—	—	—	—	23	5	65.8
<64.0	1	—	2	4	1	2	2	1	1	—	—	—	—	—	14	1	—
Totals	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	—
Medians	—	—	66.3	67.8	67.9	67.7	67.9	68.3	68.5	69.0	69.0	70.0	—	—	—	—	—

Source: Galton (1886a).

Note: All female heights were multiplied by 1.08 before tabulation. Galton added an explanatory footnote to the table: "In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62.2, 63.2, &c., instead of 62.5, 63.5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents." Galton republished these data in 1889, where they are referred to as the R.F.F. Data (Record of Family Faculties); he then noted that the first row must be in error (four children cannot have five sets of parents), but he claimed that "the bottom line, which looks suspicious, is correct" (p. 208).

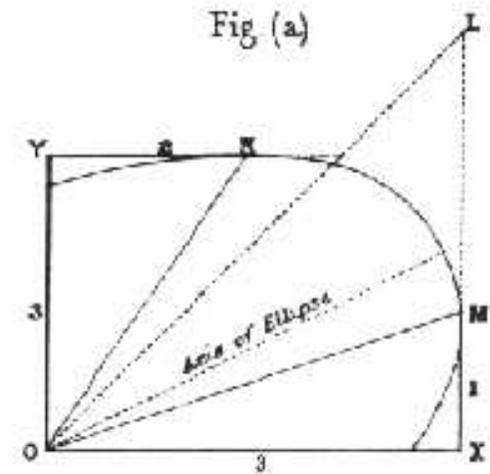
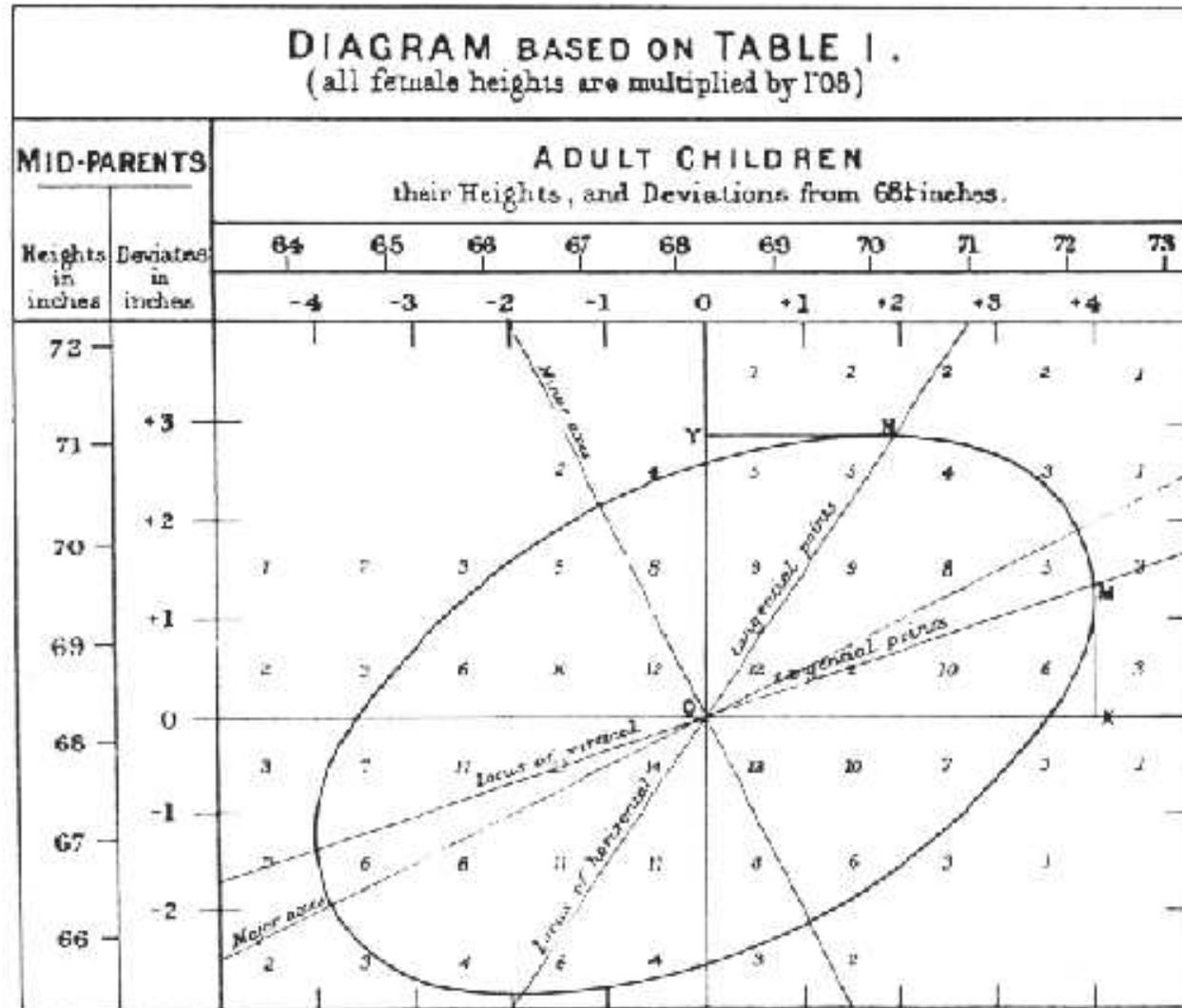
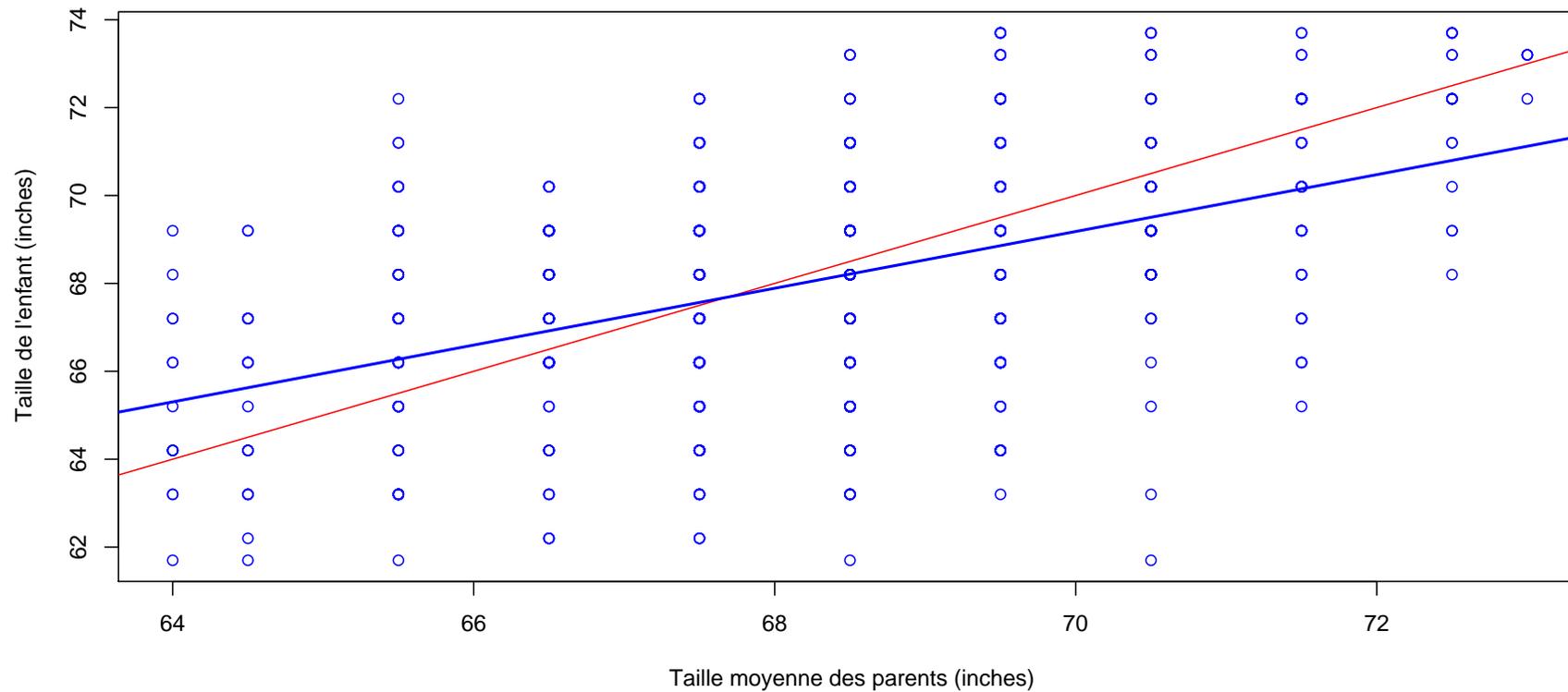


Figure 8.7. Galton's smoothed rendition of Table 8.1, with one of the "concentric and similar ellipses" drawn in. The geometric relationship of the two regression lines to the ellipse is also shown. (From Galton, 1886a.)



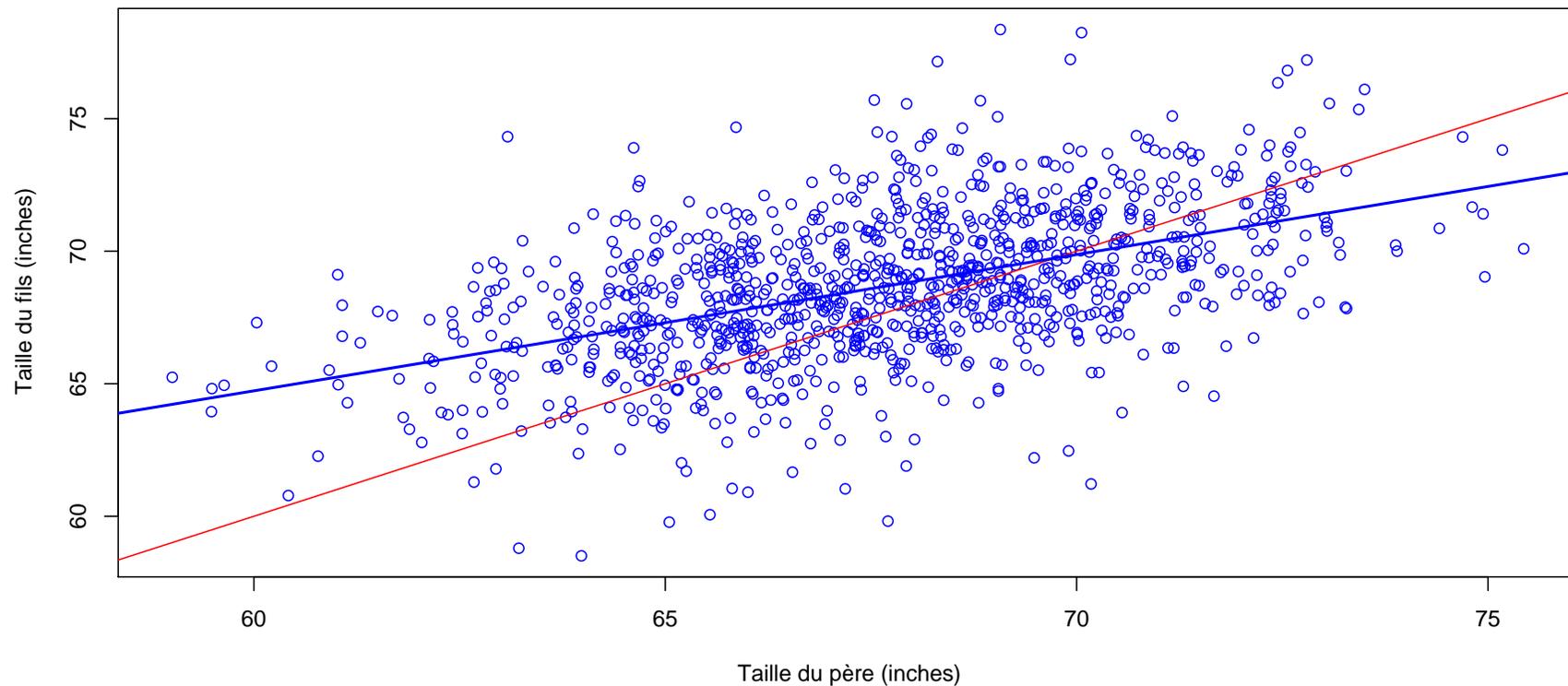
Un enfant de parents *grands* est en moyenne *grand*, mais moins que ses parents.

Un enfant de parents *petits* est en moyenne *petits*, mais moins que ses parents.

⇒ “*I have called this peculiarity by the name of regression*”, au sens **régression vers la moyenne**.

## Un peu de terminologie

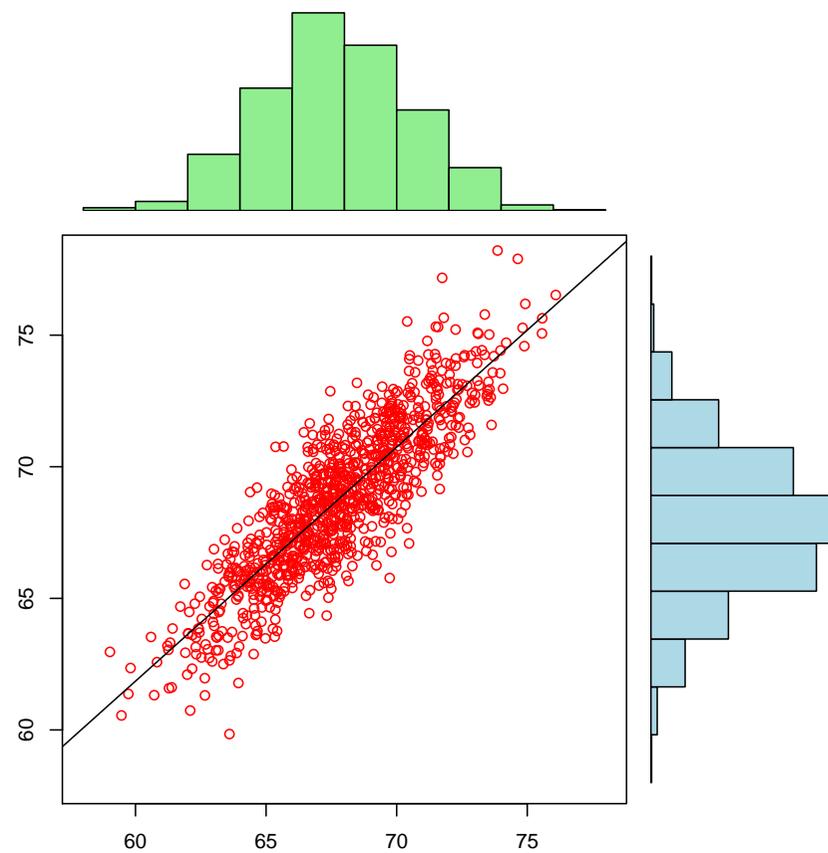
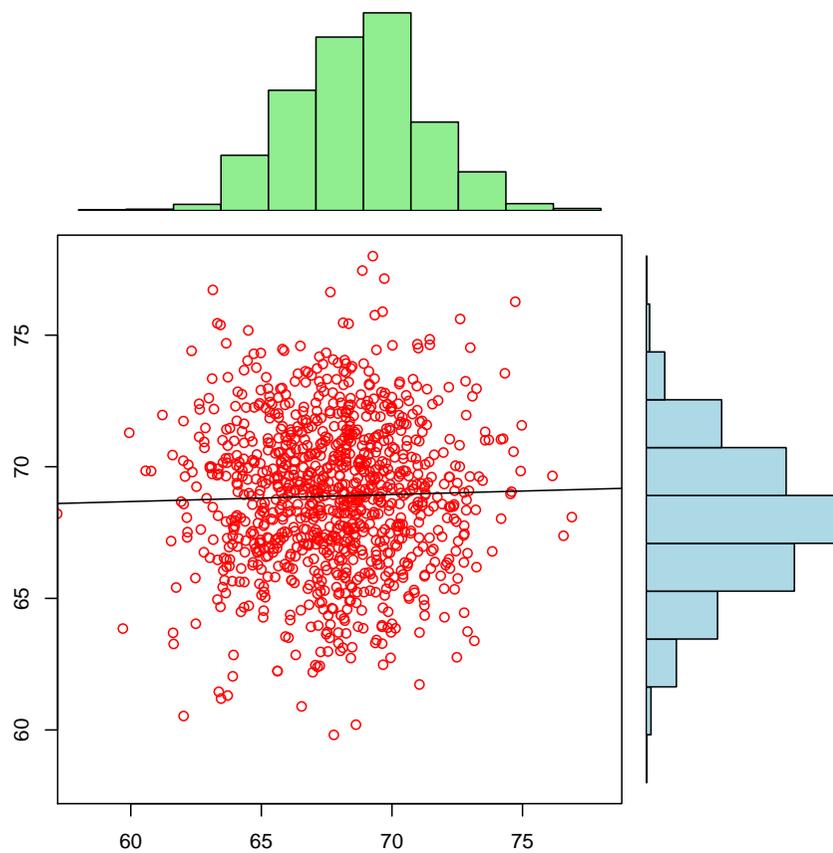
Pearson a proposé d'étudier le lien entre la taille d'un homme et celle de son père. La conclusion est la même, il y a **régression vers la moyenne**.



**régression = études de corrélations**

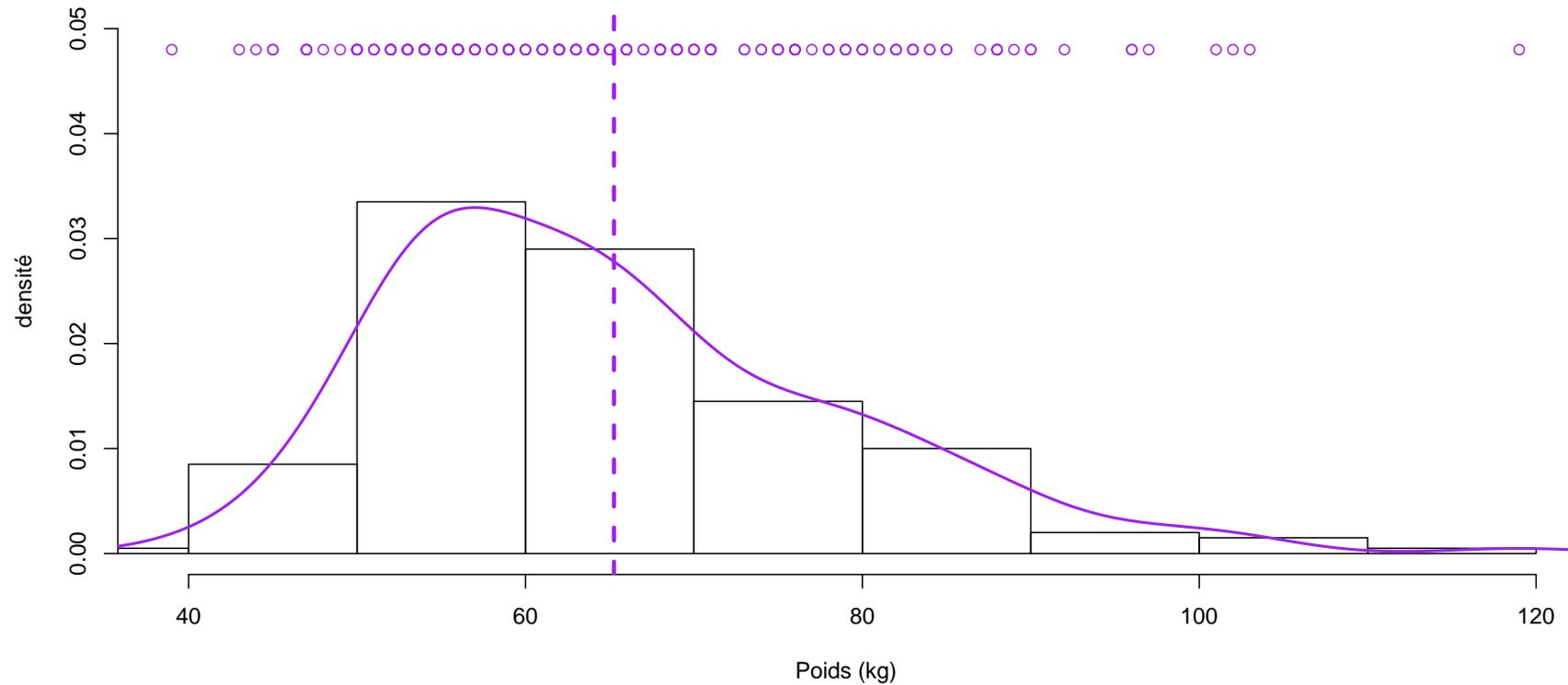
## Un peu de terminologie

**Remarque** cela ne signifie pas que les fils sont plus “*moyens*” que leurs pères : il ne s’agit que d’une notion de **corrélacion**, de **dépendance**, en aucun cas de lois marginales



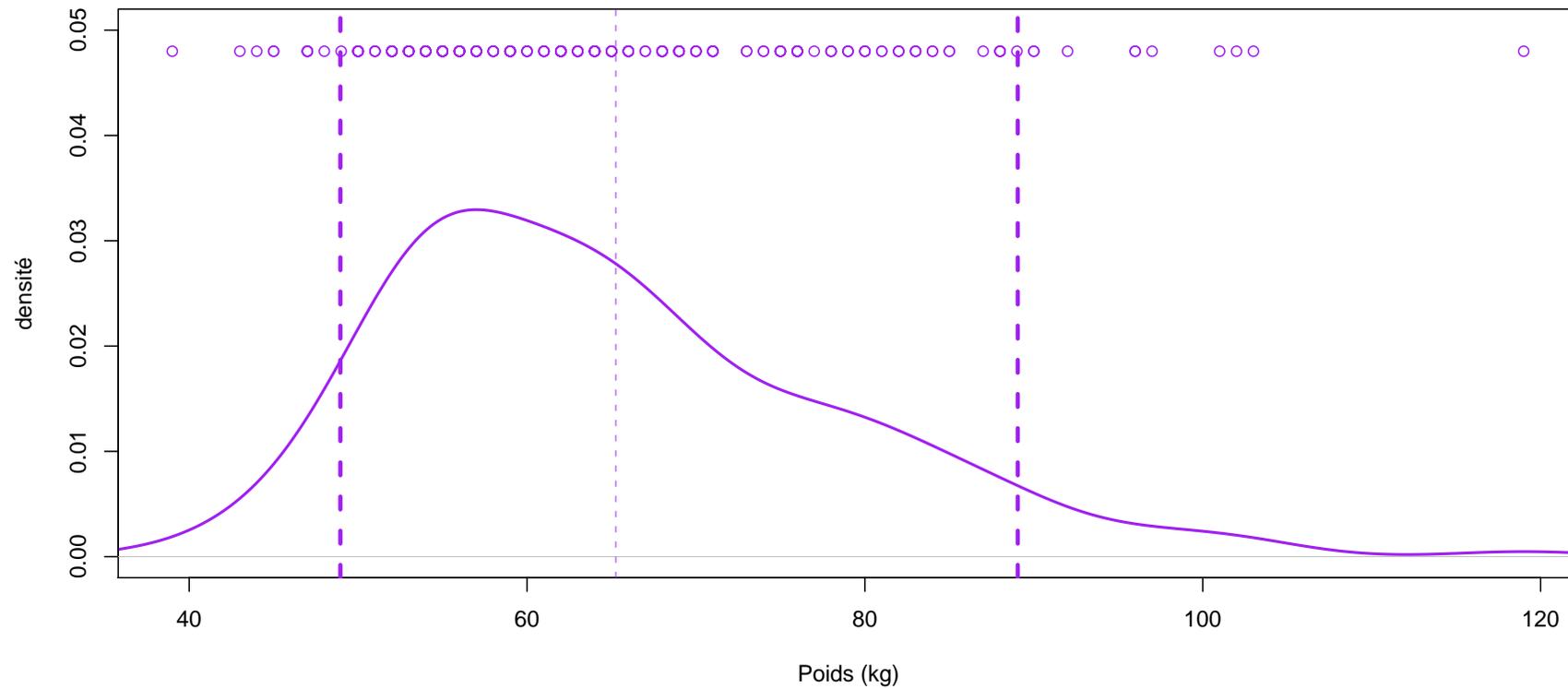
## de l'espérance à l'espérance conditionnelle

Prédire le poids  $Y$ , sans aucune autre information



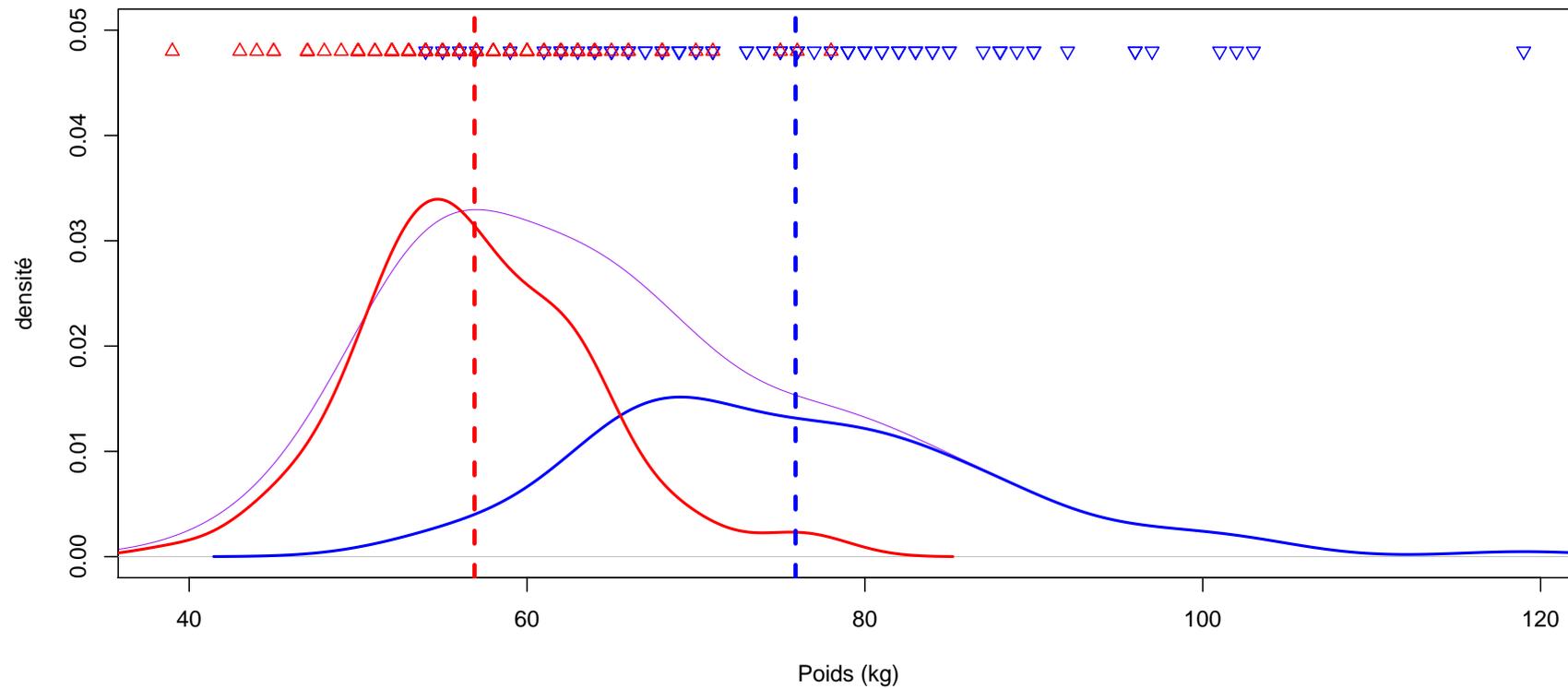
## de l'espérance à l'espérance conditionnelle

On peut aussi regarder la loi de  $Y$ , pour en déduire des quantiles (e.g. obésité)



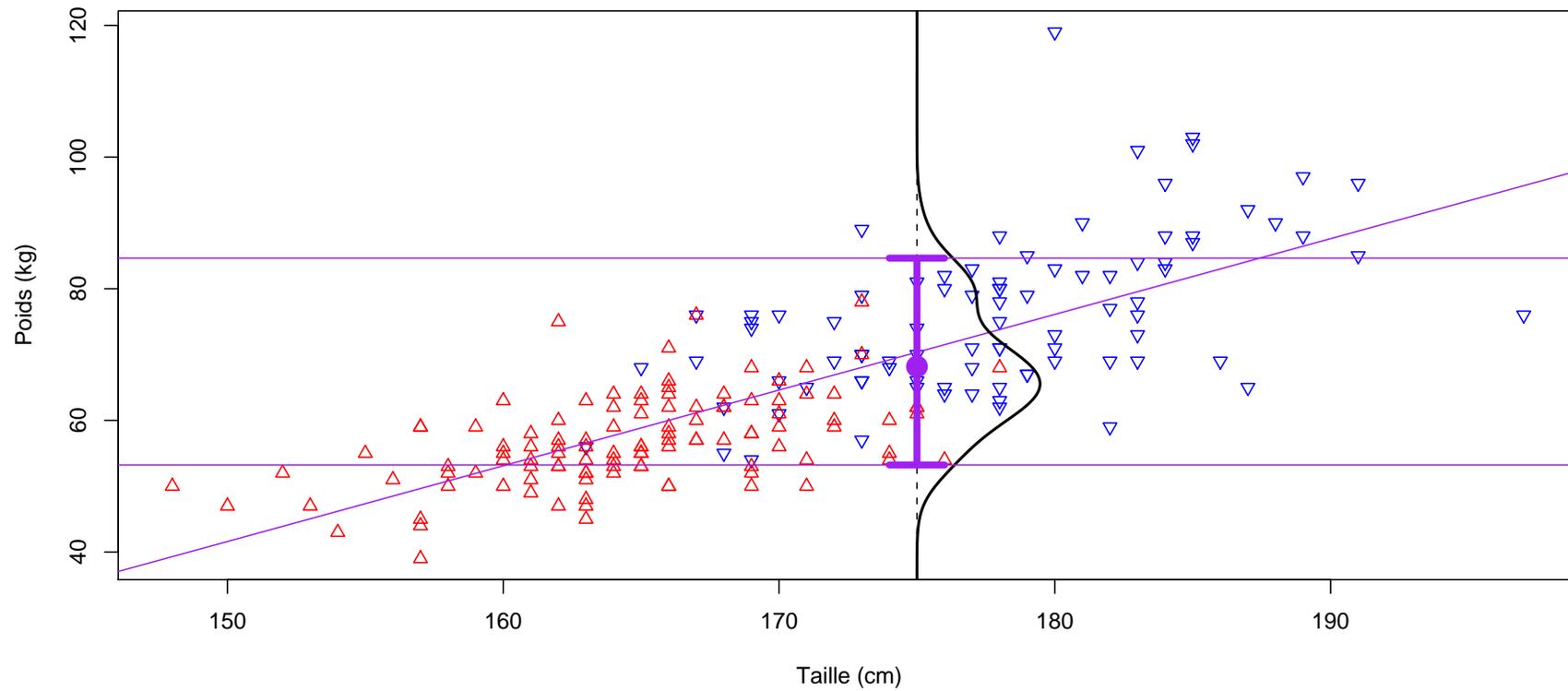
## de l'espérance à l'espérance conditionnelle

On peut aussi regarder la loi de  $Y$ , derrière se cache un mélange (par sexe)



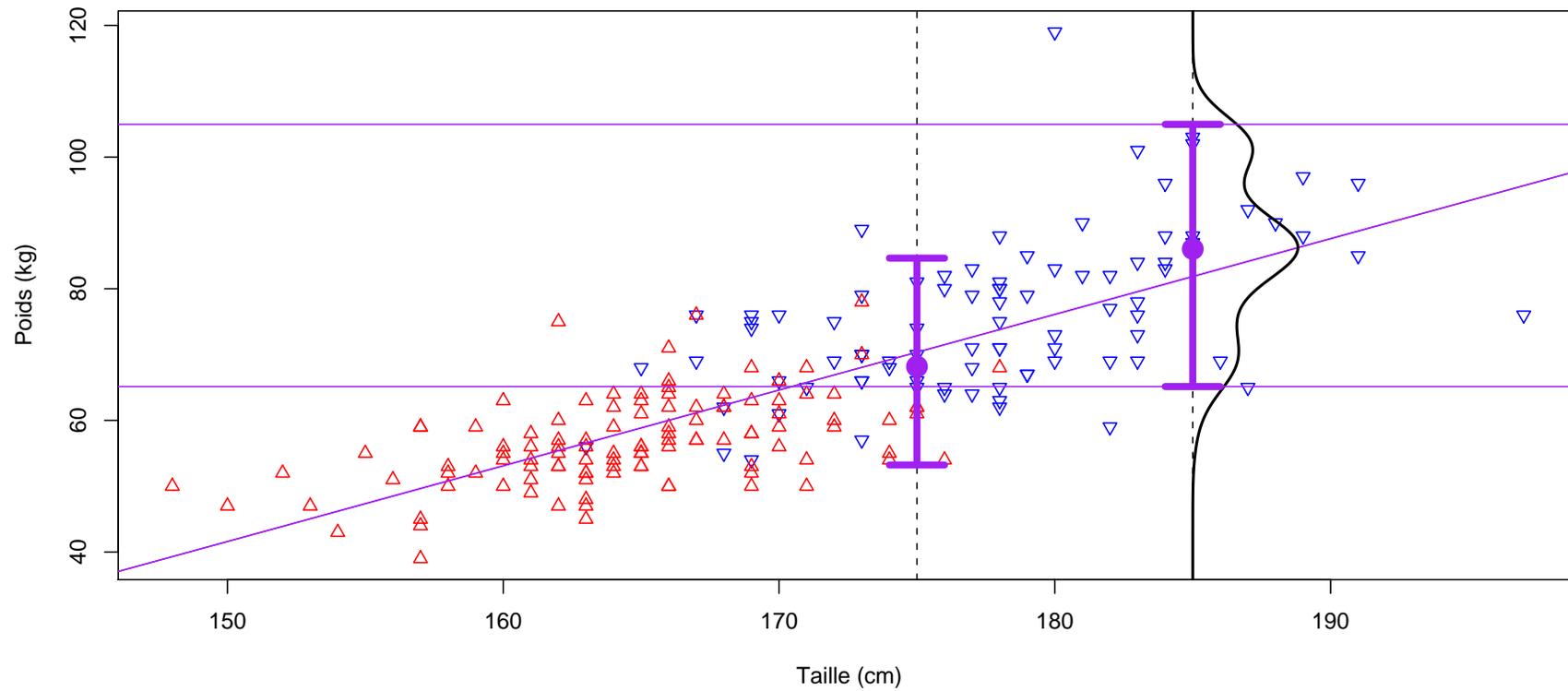
## de l'espérance à l'espérance conditionnelle

Estimation de la loi de  $Y|X = 175$



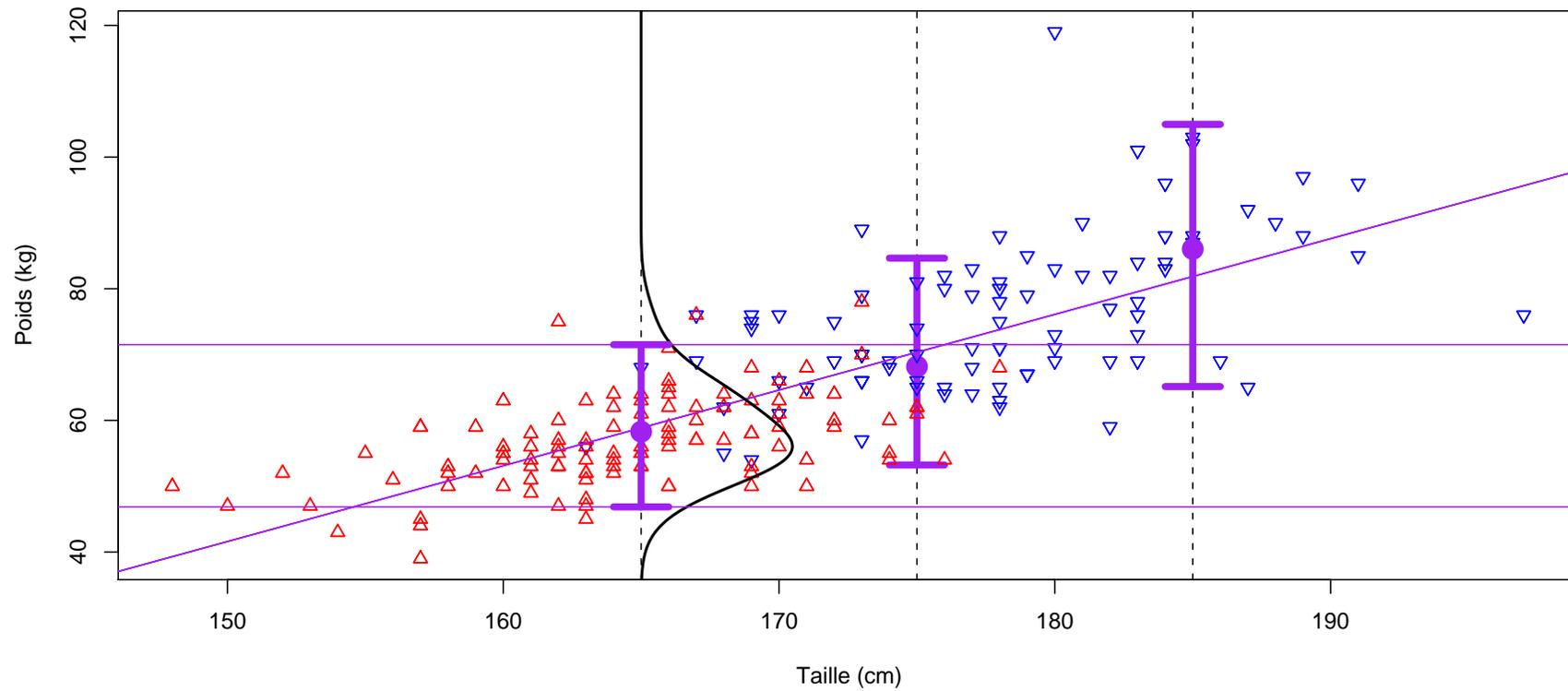
## de l'espérance à l'espérance conditionnelle

Estimation de la loi de  $Y|X = 185$



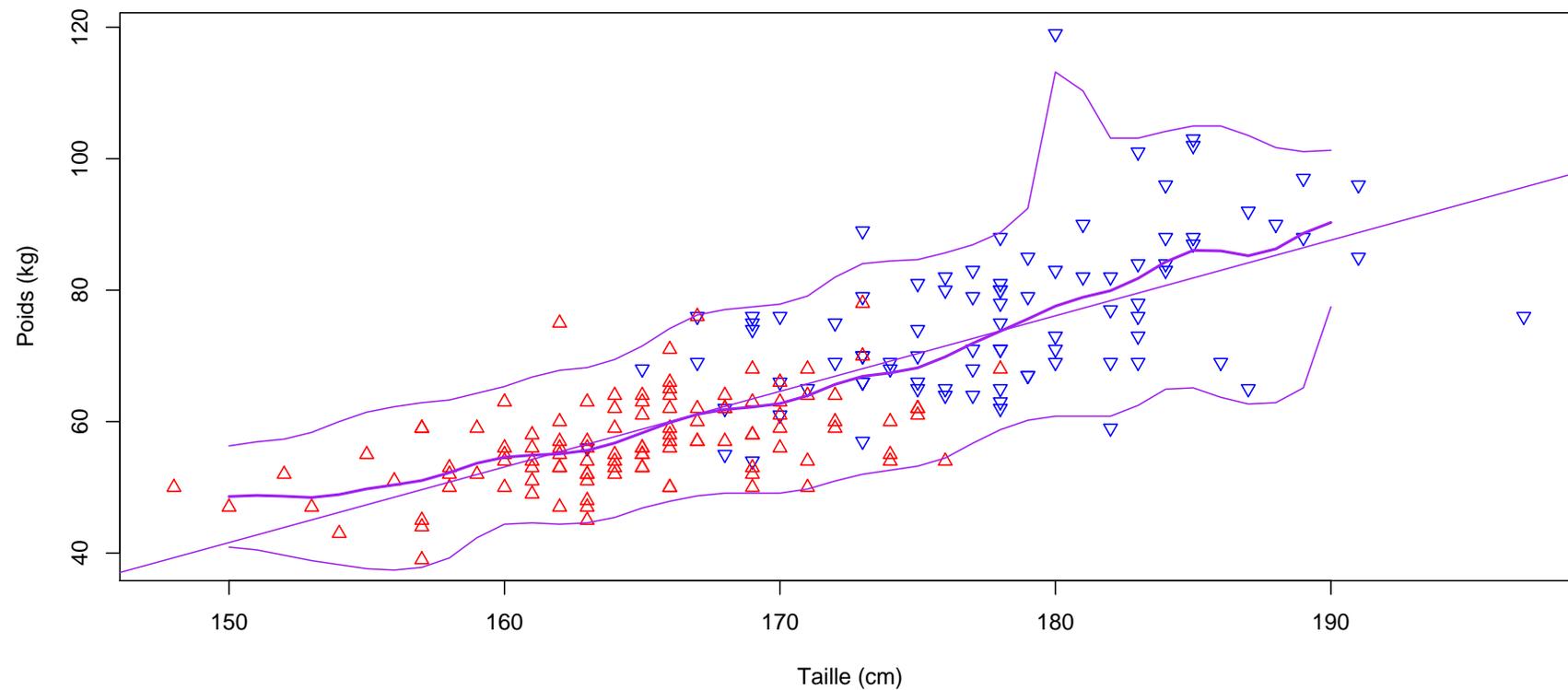
## de l'espérance à l'espérance conditionnelle

Estimation de la loi de  $Y|X = 165$



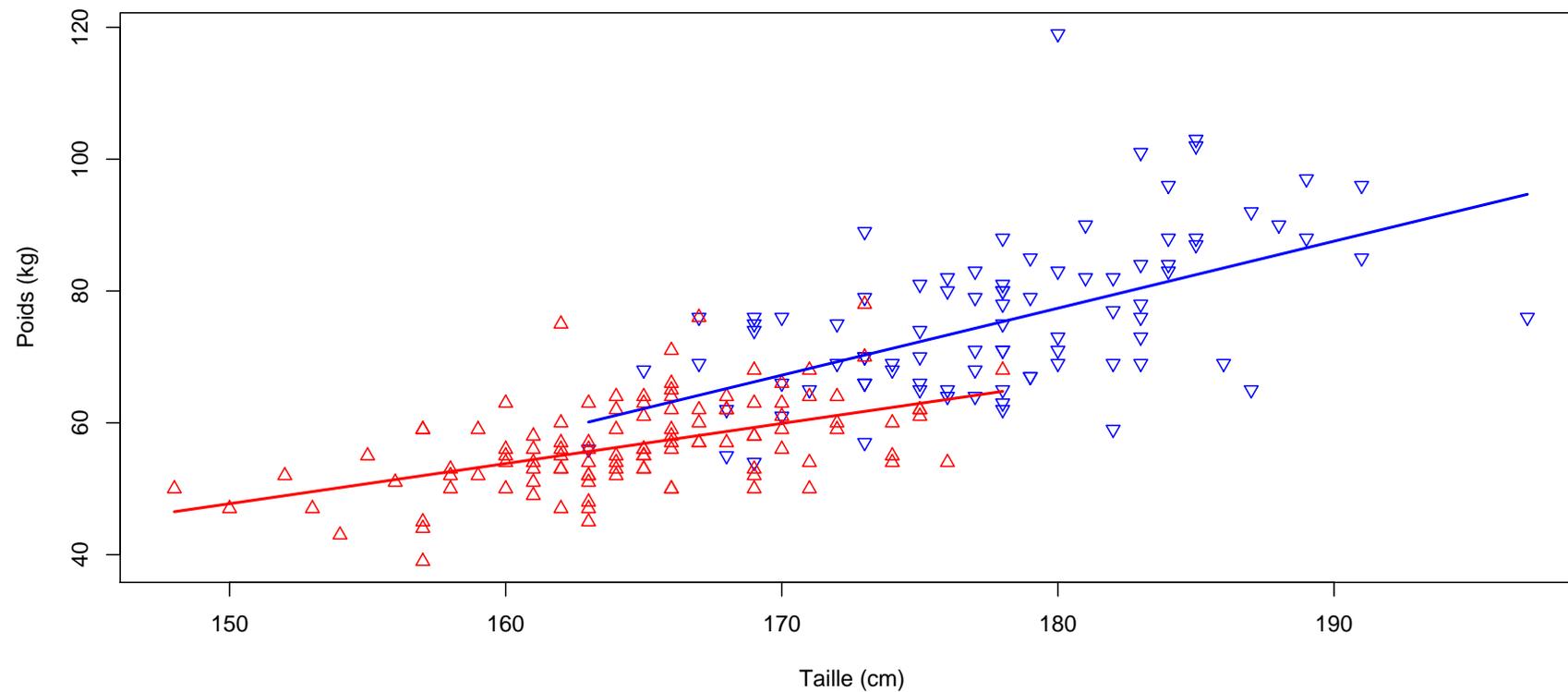
## de l'espérance à l'espérance conditionnelle

Pour plusieurs valeur de  $x$ , il possible d'estimer une loi de  $Y|X = x$



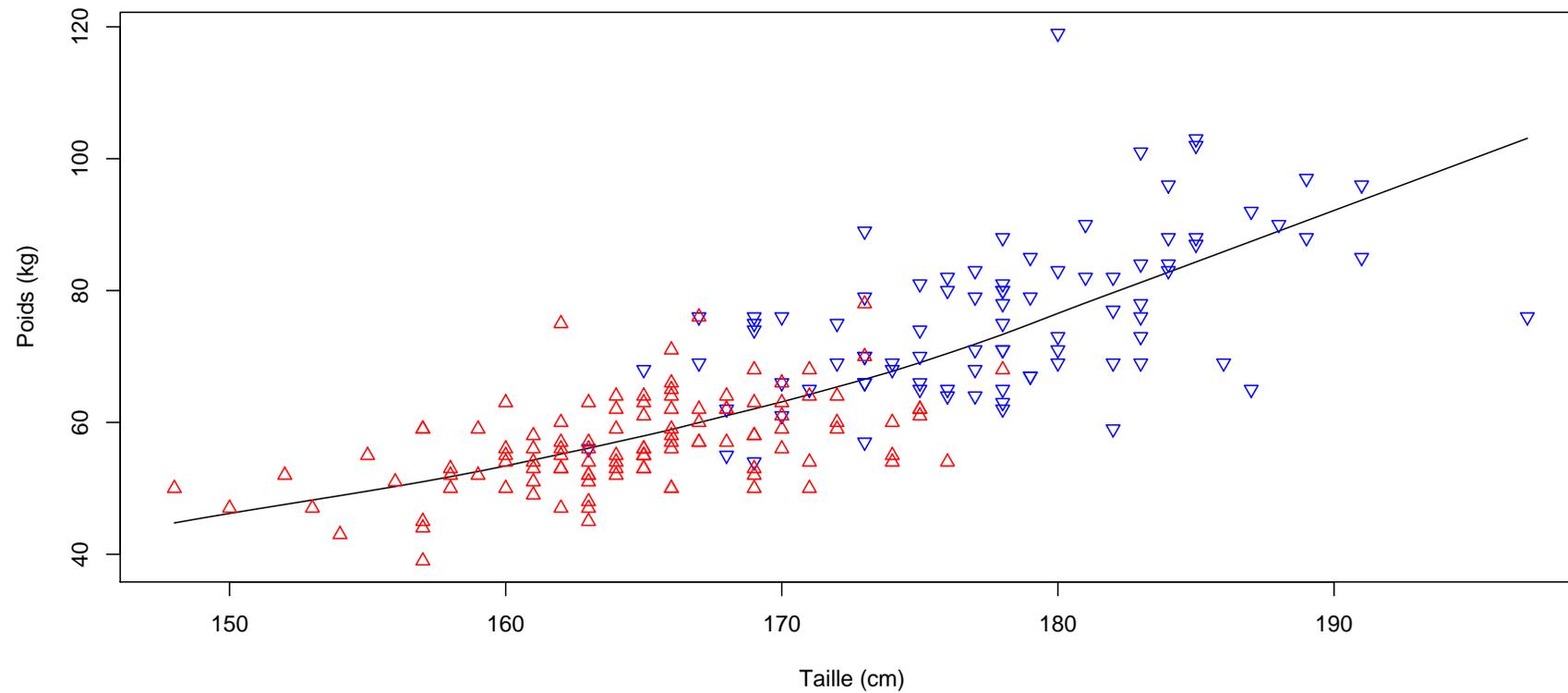
## de l'espérance à l'espérance conditionnelle

Une explication de la nonlinéarité, l'hétérogénéité hommes/femmes



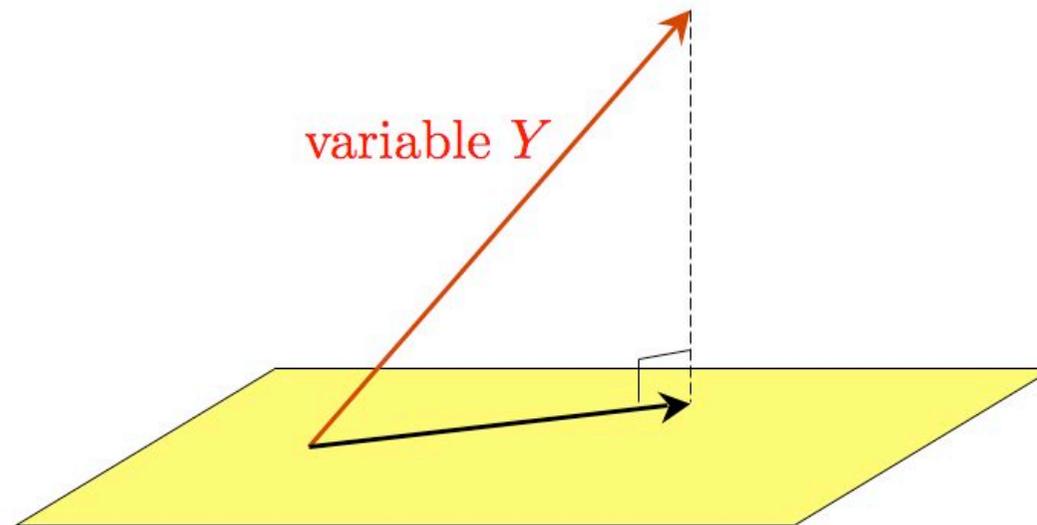
## de l'espérance à l'espérance conditionnelle

Une relation plus quadratique que linéaire ?



## Espérance conditionnelle et projection

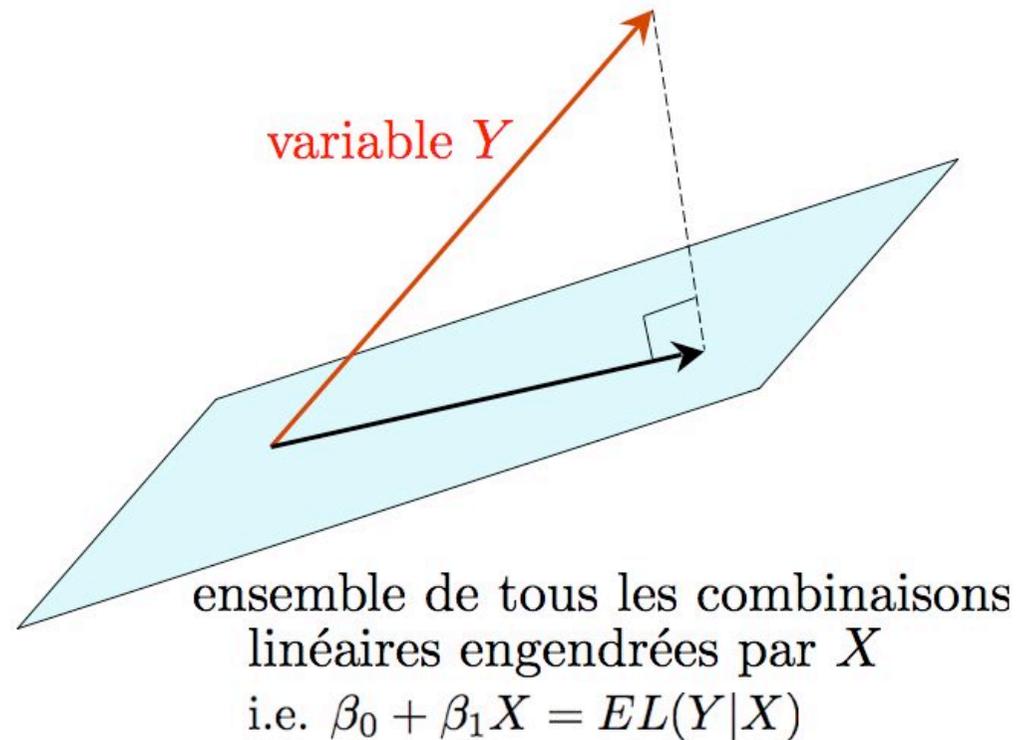
Faire une prédiction de  $Y$  à  $X$  fixé c'est projeter  $Y$  sur l'ensemble des variables aléatoires engendrées par  $X$  [...]



ensemble de toutes les variables engendrées par  $X$ , i.e.  $\varphi(X) = \mathbb{E}(Y|X)$

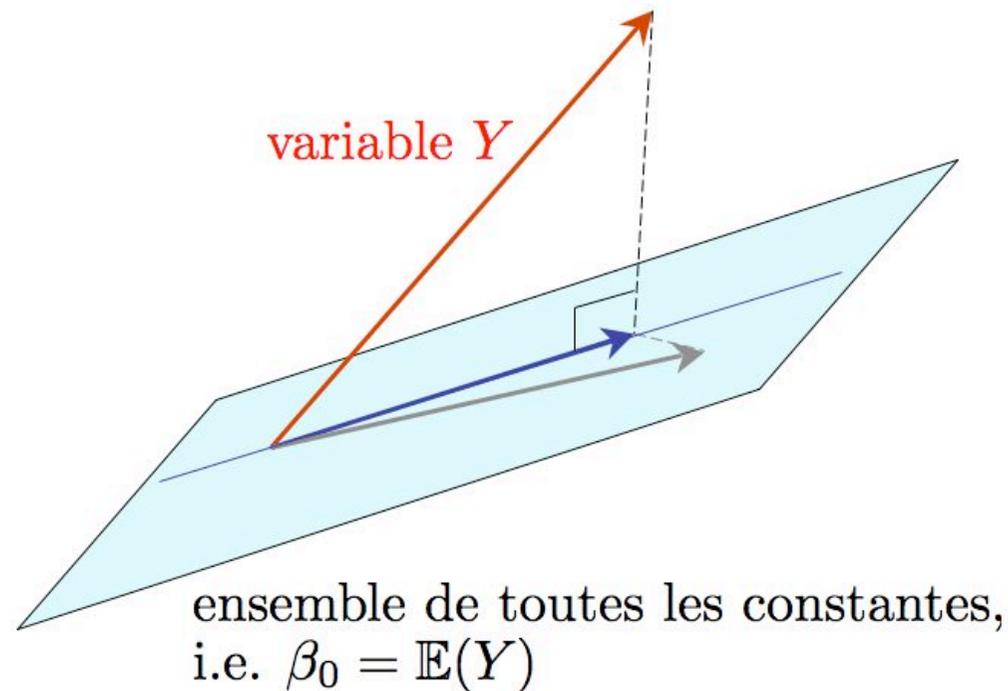
## Espérance conditionnelle et projection

[...] on peut se restreindre à un sous-ensemble, celui des transformations affines engendrées par  $X$  [...]



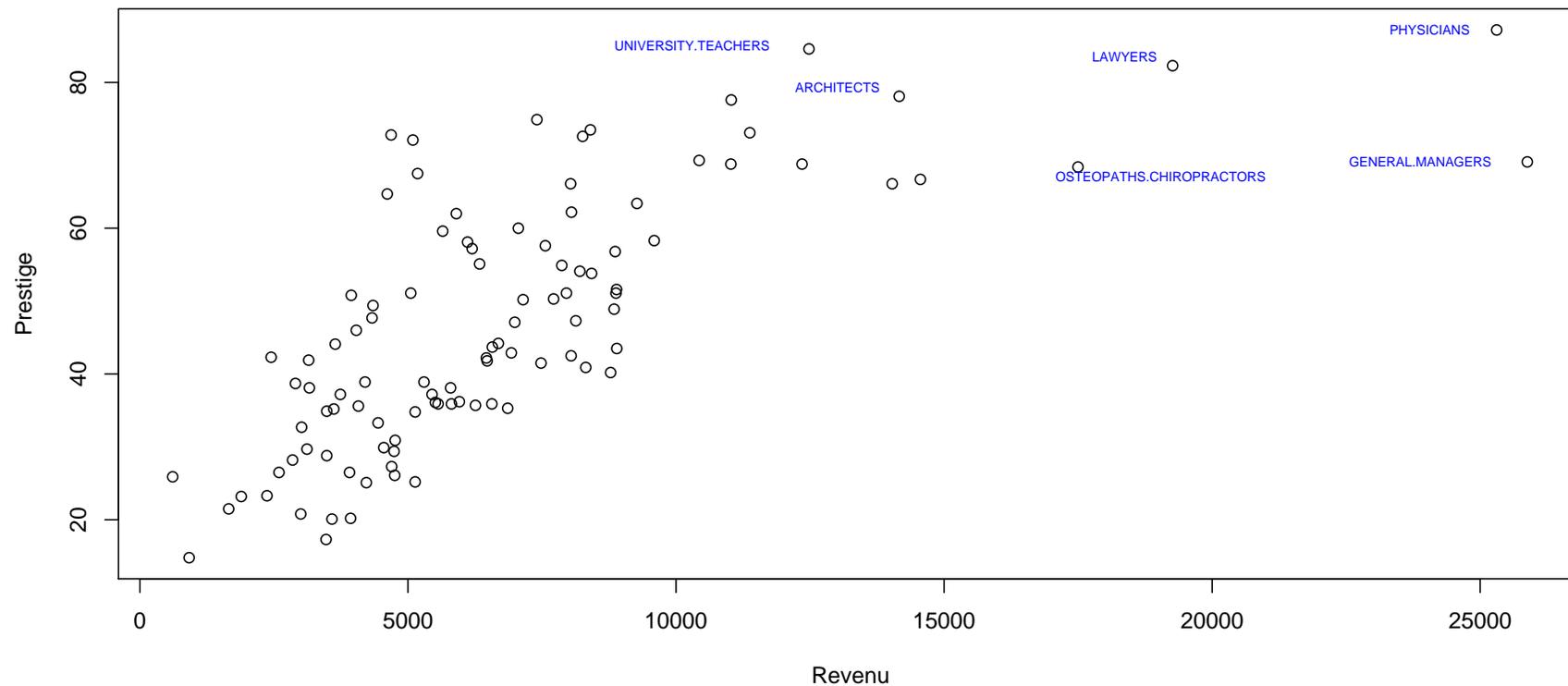
## Espérance conditionnelle et projection

[...] ou on se restreint à un sous-ensemble de ce sous-ensemble, les constantes.



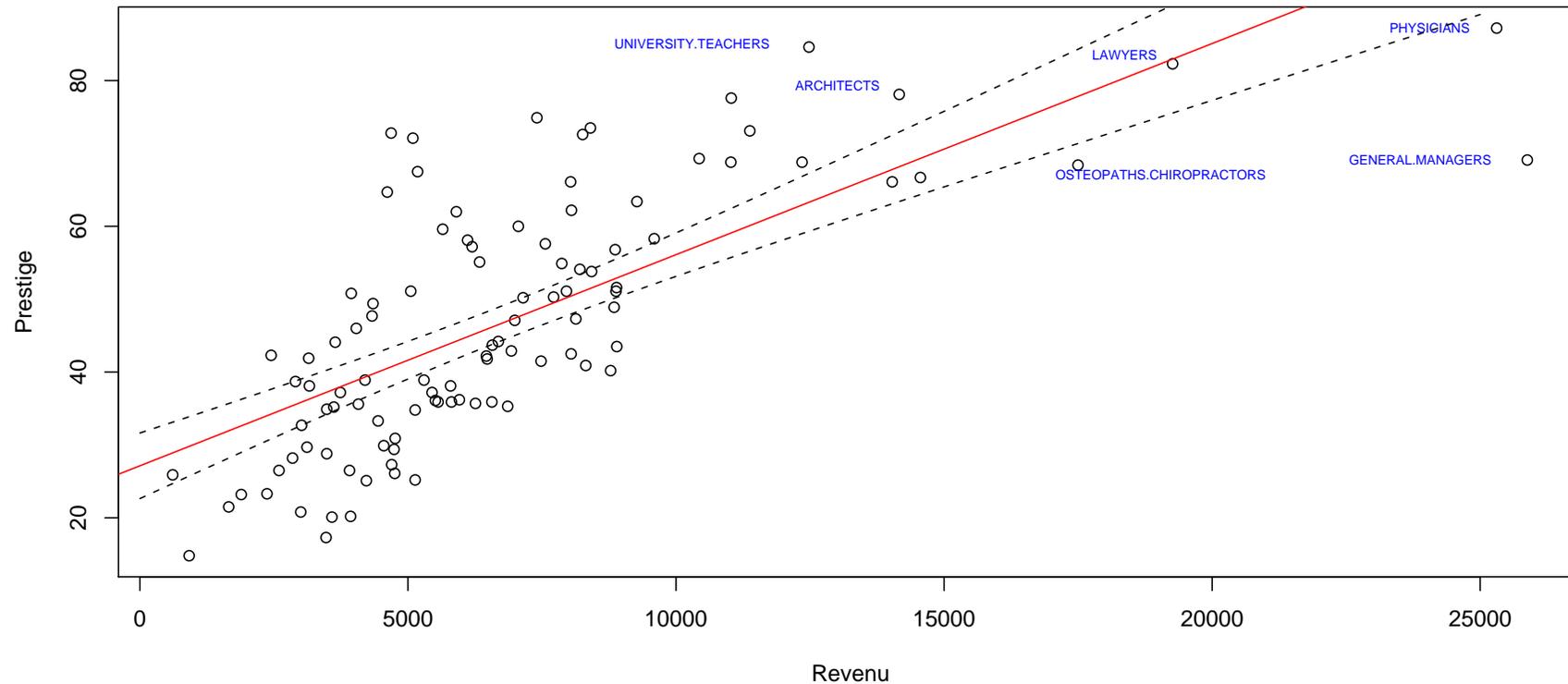
## Un peu de motivation ?

Le prestige ( $Y$ ) expliqué par le salaire ( $X$ ), cf. Blishen & McRoberts (1976), étude du “prestige” de 102 métiers au Canada.

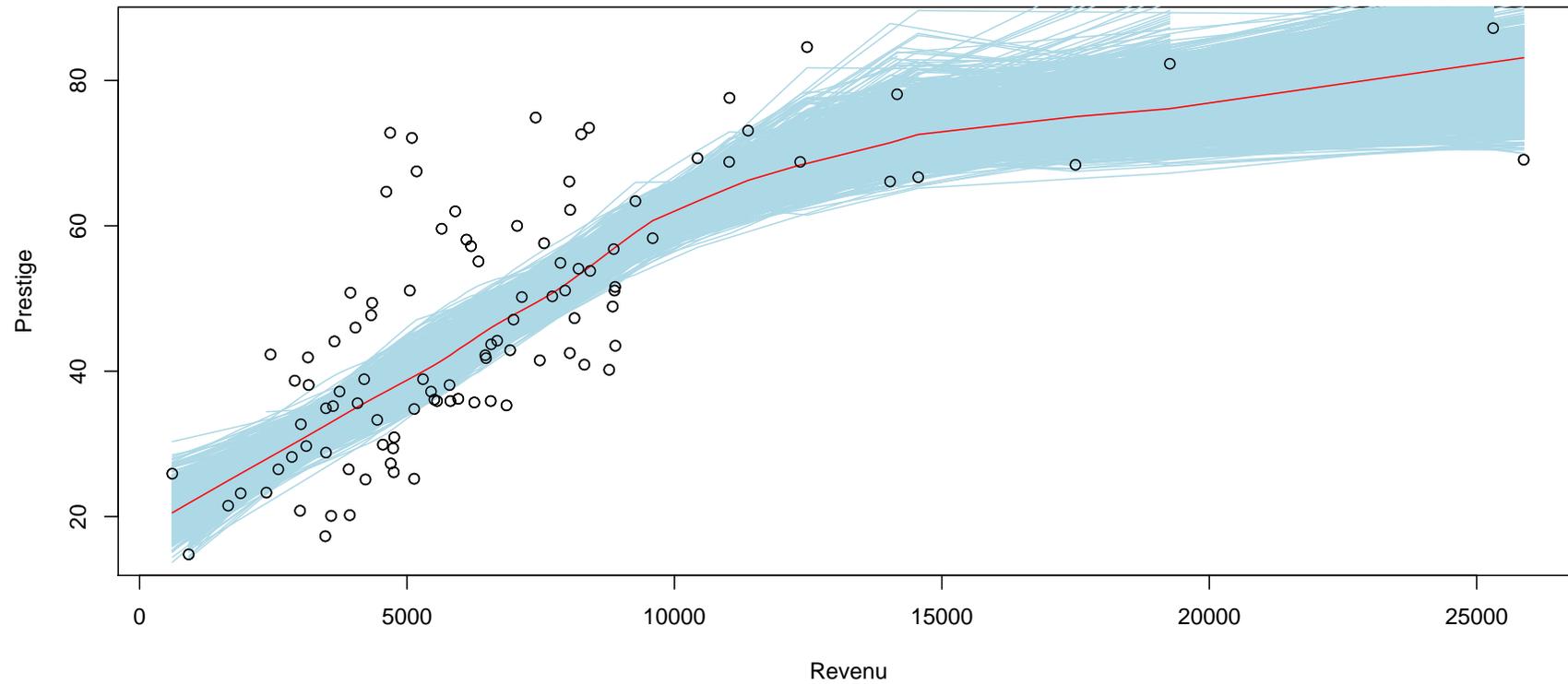


## Un peu de motivation ?

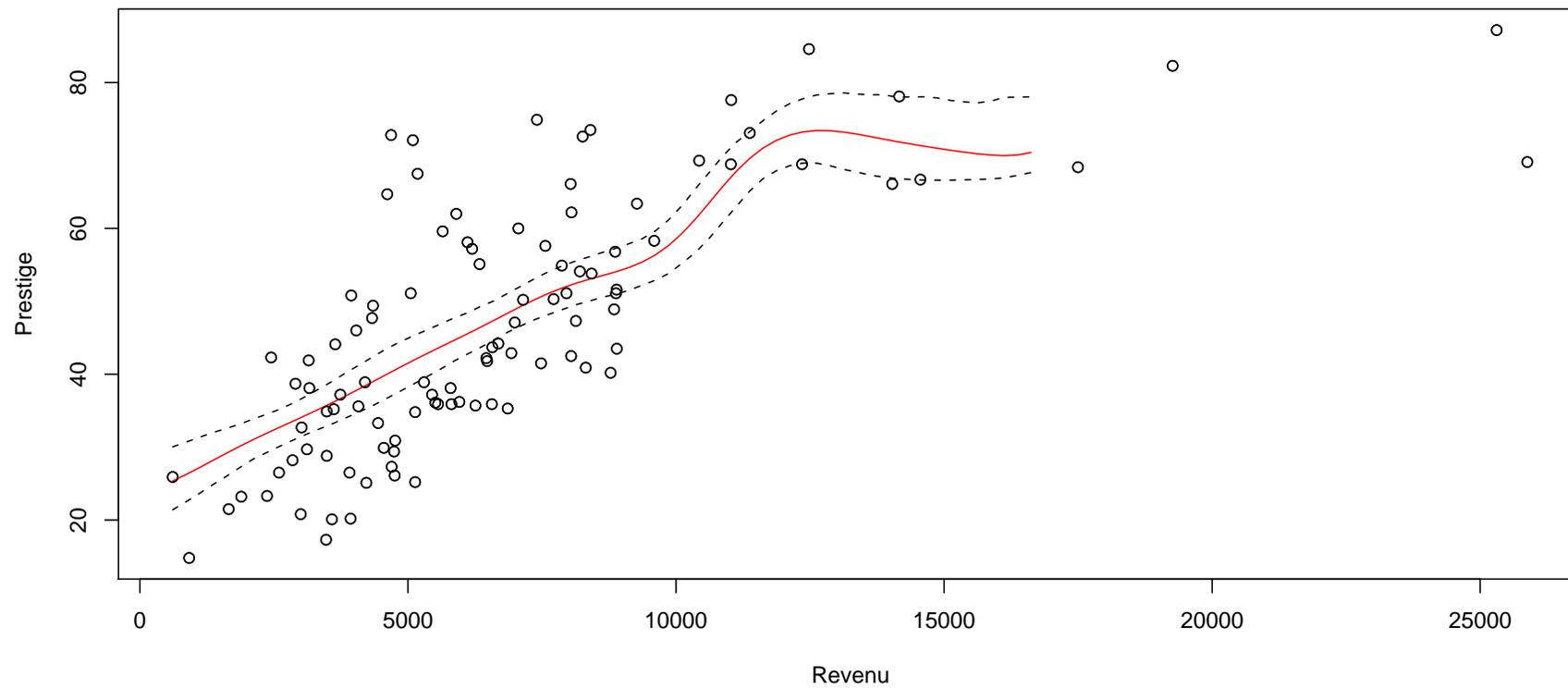
Le prestige ( $Y$ ) expliqué par le salaire ( $X$ ), cf. Blishen & McRoberts (1976), étude du “prestige” de 102 métiers au Canada.



## Un peu de motivation ?



## Un peu de motivation ?



## Lien avec le cours de statistique ?

- Cours de statistique ‘descriptive’

On dispose d'un échantillon  $\{y_1, \dots, y_n\}$ , de variables réelles,  $y_i \in \mathbb{R}$ .

On peut définir la moyenne, ou la variance (empirique)

- $$\bar{y} = \frac{y_1 + \dots + y_n}{n} = \frac{1}{n} \sum_{i=1}^n y_i$$

- $$s^2 = \frac{1}{n} \sum_{i=1}^n [y_i - \bar{y}]^2$$

## Lien avec le cours de statistique ?

- Cours de statistique mathématique

On dispose d'un échantillon  $\{y_1, \dots, y_n\}$ , vu comme des réalisations de variables aléatoires  $\{Y_1(\omega), \dots, Y_n(\omega)\}$ ,  $\omega \in \Omega$ , i.e.  $y_i = Y_i(\omega) \in \mathbb{R}$ . On a maintenant des variables aléatoires sous-jacentes,  $Y_i$ . Les moyennes et variances empiriques sont alors des réalisations des variables aléatoires

- $$\bar{Y} = \frac{Y_1 + \dots + Y_n}{n} = \frac{1}{n} \sum_{i=1}^n Y_i$$

- $$S^2 = \frac{1}{n} \sum_{i=1}^n [Y_i - \bar{Y}]^2$$

i.e.  $\bar{y} = \bar{Y}(\omega)$  et  $s^2 = S^2(\omega)$ . En statistique mathématique, on utilise des propriétés de ces variables aléatoires pour en déduire des propriétés sur telle ou telle statistique.

La **loi des grands nombres** garantit que  $\bar{Y} \xrightarrow{\mathbb{P}} \mathbb{E}(Y)$  si on suppose que les  $X_i$  sont des variables indépendantes, de même espérance et de même variance (finies, loi *faible* des grands nombres), i.e.  $\forall \varepsilon > 0$ .

$$\lim_{n \rightarrow +\infty} \mathbb{P} \left( \left| \frac{Y_1 + Y_2 + \cdots + Y_n}{n} - \mathbb{E}(Y) \right| \geq \varepsilon \right) = 0$$

On a aussi le **théorème central limite**, qui garantit que  $\sqrt{n}(\bar{Y} - \mathbb{E}(Y)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \text{Var}(Y))$  si on suppose que les  $X_i$  sont des variables indépendantes, de même espérance et de même variance (finies, loi *faible* des grands nombres), i.e.

$$\lim_{n \rightarrow \infty} \mathbb{P} \left( \sqrt{n} \frac{\bar{Y}_n - \mathbb{E}(Y)}{\sqrt{\text{Var}(Y)}} \leq z \right) = \Phi(z)$$

où  $\Phi(\cdot)$  est la fonction de répartition de la loi  $\mathcal{N}(0, 1)$ .

## Lien avec le cours de statistique ?

- Cours d'inférence (paramétrique)

On dispose d'un échantillon  $\{y_1, \dots, y_n\}$ , vu comme des réalisations de variables aléatoires  $\{Y_1(\omega), \dots, Y_n(\omega)\}$ , où les variables aléatoires sous-jacentes,  $Y_i$ , sont supposées indépendantes, et identiquement distribuées, de loi  $F \in \mathcal{F} = \{F_\theta, \theta \in \Theta\}$ . Aussi,  $F = F_{\theta_0}$ , mais  $\theta_0$  est inconnu.

**Remarque**  $\theta$  est généralement un paramètre dans  $\mathbb{R}^k$ , mais pour simplifier, on supposera  $\theta \in \Theta \subset \mathbb{R}$ .

Un estimateur  $\hat{\theta}$  est une fonction des observations. Attention, parfois

- $\hat{\theta} = s(y_1, \dots, y_n)$  est un réel, e.g.  $\theta = \bar{y}$
- $\hat{\theta} = s(Y_1, \dots, Y_n)$  est une variable aléatoire, e.g.  $\theta = \bar{Y}$

Pour estimer  $\theta$ , on dispose de deux méthodes standards

- la méthode des moments
- la méthode du maximum de vraisemblance

- **La méthode des moments**

On suppose que si  $Y \sim F_\theta$ ,  $\mathbb{E}(Y) = g(\theta)$  où  $g(\cdot)$  est bijective. Dans ce cas,  $\theta = g^{-1}(\mathbb{E}(Y))$ . Un estimateur naturel est alors

$$\hat{\theta} = g^{-1}(\bar{Y}).$$

On notera que l'on a (*a priori*) aucune information sur la qualité de l'estimateur, e.g. l'estimateur n'a aucune raison d'être sans biais,

$$\mathbb{E}(\hat{\theta}) = \mathbb{E}[g^{-1}(\bar{Y})] \neq g^{-1}(\mathbb{E}[\bar{Y}]) = g^{-1}(\mathbb{E}(Y)) = \theta$$

Par contre, si  $g$  est suffisamment régulière, on a des propriétés asymptotiques,

$$\mathbb{E}(\hat{\theta}) \rightarrow \theta \text{ lorsque } n \rightarrow \infty.$$

- **La méthode du maximum de vraisemblance**

Si les variables aléatoires sous-jacentes sont supposées indépendantes et identiquement distribuées, de loi  $F_\theta$ , de loi de probabilité  $f_\theta$

$$\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n) = \prod_{i=1}^n f_\theta(y_i) = \mathcal{L}(\theta)$$

(dans le cas discret, avec une expression similaire basée sur la densité dans le cas continu). On va chercher la valeur  $\hat{\theta}$  qui donne la plus grande probabilité  $\mathbb{P}(Y_1 = y_1, \dots, Y_n = y_n)$  (appelée **vraisemblance**). I.e.

$$\hat{\theta} \in \operatorname{argmax}_{\theta \in \Theta} \{\mathcal{L}(\theta)\}$$

**Remarque** On ne s'intéresse pas à la valeur de la vraisemblance, mais à la valeur  $\theta$  en laquelle la vraisemblance est maximale. Or

$$\hat{\theta} \in \operatorname{argmax}_{\theta \in \Theta} \{h(\mathcal{L}(\theta))\}$$

Comme  $\mathcal{L}(\theta) > 0$  (car  $\forall i, f_\theta(y_i) > 0$ ), on peut prendre  $h = \log$ , et donc

$$\hat{\theta} \in \operatorname{argmax}_{\theta \in \Theta} \{\log(\mathcal{L}(\theta))\}$$

car  $\log \mathcal{L}$  est une fonction beaucoup plus simple,

$$\log \mathcal{L}(\theta) = \sum_{i=1}^n \log f_\theta(y_i).$$

Si la fonction  $\log \mathcal{L}$  est différentiable, la **condition du premier ordre** s'écrit

$$\left. \frac{\partial \log \mathcal{L}(\theta)}{\partial \theta} \right|_{\theta=\hat{\theta}} = 0$$

avec une condition du second ordre pour assurer qu'en  $\hat{\theta}$ , on soit sur un maximum, et pas un minimum,

$$\left. \frac{\partial^2 \log \mathcal{L}(\theta)}{\partial^2 \theta} \right|_{\theta=\hat{\theta}} \leq 0$$

Là encore, on n'a aucune propriété distance finie. Mais quand  $n \rightarrow \infty$ ,  $\mathbb{P}(|\hat{\theta}_n - \theta| > \varepsilon) \rightarrow 0, \forall \varepsilon > 0$  (i.e.  $\hat{\theta} \xrightarrow{\mathbb{P}} \theta$ ). On a aussi une normalité asymptotique, et une efficacité asymptotique, i.e. la variance (asymptotique) est donnée par la borne de Cramér-Rao,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I^{-1}(\theta_0))$$

où  $I(\theta_0)$  est l'information de Fisher,

$$I(\theta) = \mathbb{E} \left[ \left( \frac{\partial \log f_{\theta}(Y)}{\partial \theta} \right)^2 \right], \text{ où } Y \sim F_{\theta}.$$

- **Construction de tests**

Considérons l'échantillon suivant, de pile ou face,  $\{y_1, \dots, y_n\}$

```
> set.seed(1)
> n=20
> (Y=sample(0:1,size=n,replace=TRUE))
[1] 0 0 1 1 0 1 1 1 1 0 0 0 1 0 1 0 1 1 0 1
```

Comme  $Y_i \sim \mathcal{B}(\theta)$ , avec  $\theta = \mathbb{E}(Y)$ , l'estimateur de la méthode des moments est  $\hat{\theta} = \bar{y}$ , soit ici

```
> mean(X)
[1] 0.55
```

On peut faire un test de  $H_0 : \theta = \theta_*$  contre  $H_1 : \theta \neq \theta_*$  (par exemple 50%)

On peut utiliser le test de Student,

$$T = \sqrt{n} \frac{\hat{\theta} - \theta_*}{\sqrt{\theta_*(1 - \theta_*)}}$$

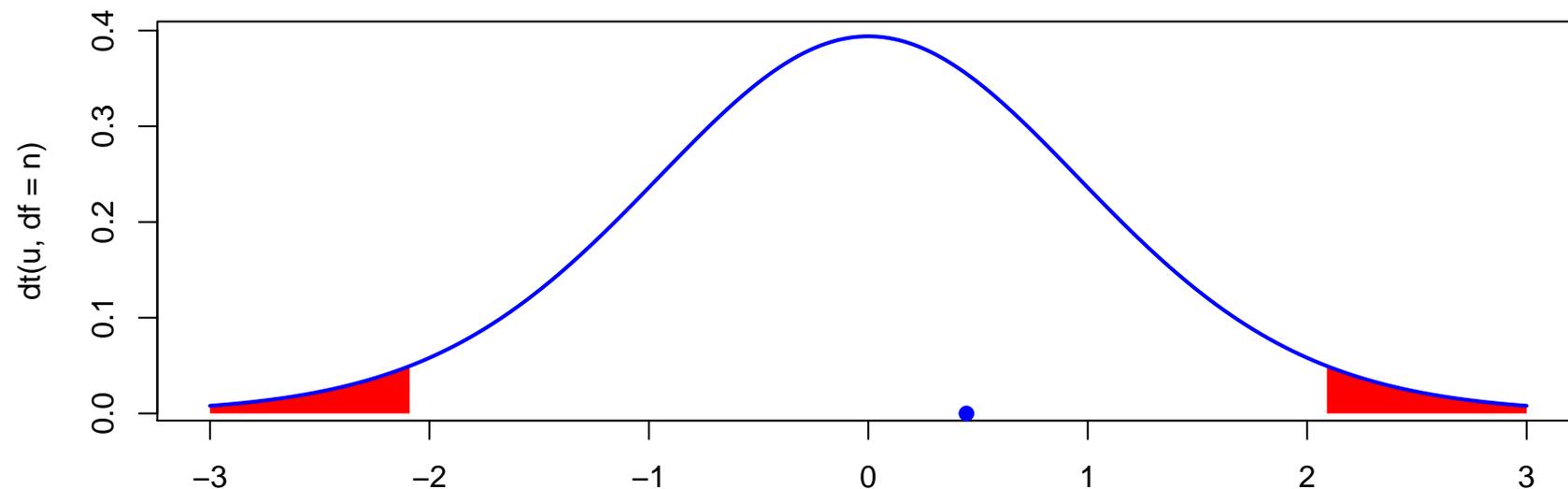
qui suit, sous  $H_0$  une loi de Student à  $n$  degrés de liberté.

```
> (T=sqrt(n)*(pn-p0)/(sqrt(p0*(1-p0))))
```

```
[1] 0.4472136
```

```
> abs(T)<qt(1-alpha/2,df=n)
```

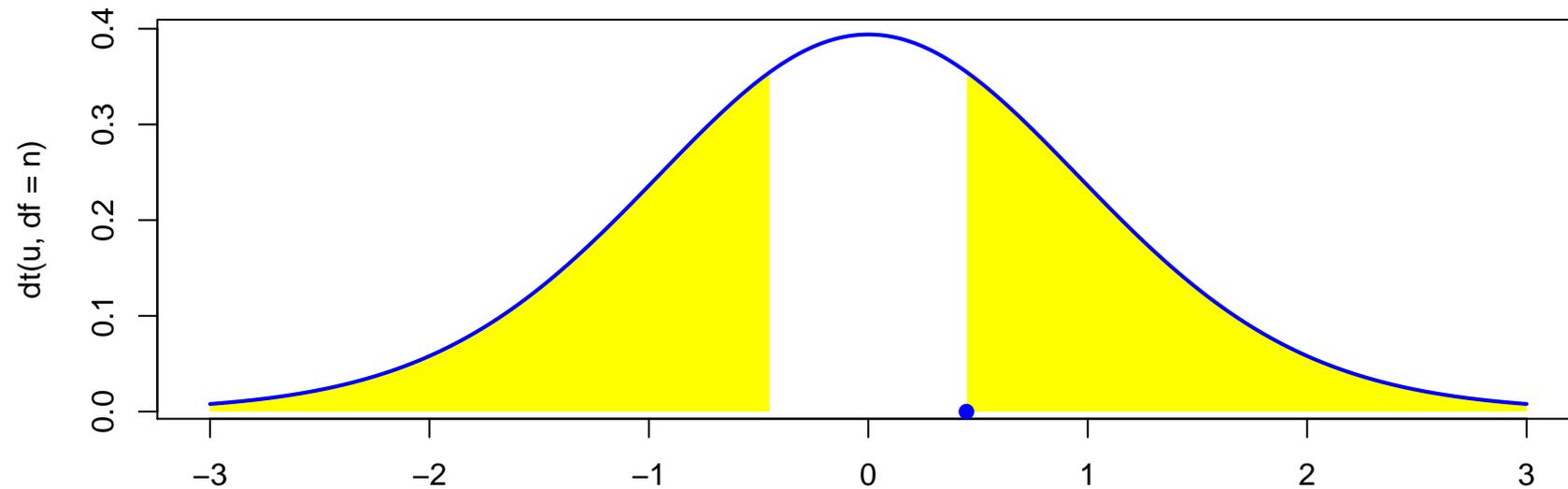
```
[1] TRUE
```



On est ici dans la région d'acceptation du test.

On peut aussi calculer la  $p$ -value,  $\mathbb{P}(|T| > |t_{obs}|)$ ,

```
> 2*(1-pt(abs(T),df=n))  
[1] 0.6595265
```



Le maximum de vraisemblance est important car il permet de faire des tests statistiques.

**Test de Wald** l'idée est d'étudier la différence entre  $\hat{\theta}$  et  $\theta_*$ . Sous  $H_0$ ,

$$T = n \frac{(\hat{\theta} - \theta_*)^2}{I^{-1}(\theta_*)} \xrightarrow{\mathcal{L}} \chi^2(1)$$

**Test du rapport de vraisemblance** l'idée est d'étudier la différence entre  $\log \mathcal{L}(\hat{\theta})$  et  $\log \mathcal{L}(\theta_*)$ . Sous  $H_0$ ,

$$T = 2 \log \left( \frac{\log \mathcal{L}(\theta_*)}{\log \mathcal{L}(\hat{\theta})} \right) \xrightarrow{\mathcal{L}} \chi^2(1)$$

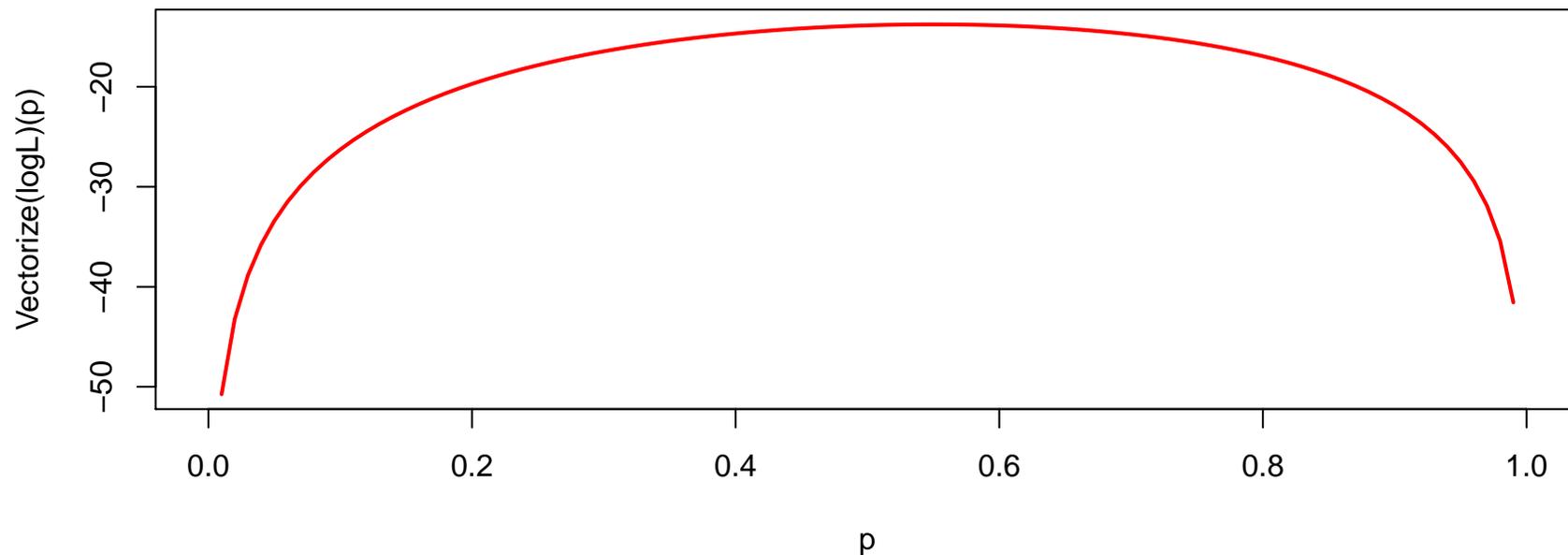
**Test du score** l'idée est d'étudier la différence entre  $\frac{\partial \log \mathcal{L}(\theta_*)}{\partial \theta}$  et 0. Sous  $H_0$ ,

$$T = \left( \frac{1}{n} \sum_{i=1}^n \frac{\partial \log f_{\theta_*}(x_i)}{\partial \theta} \right)^2 \xrightarrow{\mathcal{L}} \chi^2(1)$$

## Lien avec le cours de statistique ?

Pour l'estimateur de maximum de vraisemblance, on peut faire les calculs, ou le faire numériquement. On commence par tracer la fonction  $\theta \mapsto \log \mathcal{L}(\theta)$

```
> p=seq(0,1,by=.01)
> logL=function(p){sum(log(dbinom(X,size=1,prob=p)))}
> plot(p,Vectorize(logL)(p),type="l",col="red",lwd=2)
```



On peut trouver numériquement le maximum de  $\log \mathcal{L}$ ,

```
> neglogL=function(p){-sum(log(dbinom(X,size=1,prob=p)))}
> pml=optim(fn=neglogL,par=p0,method="BFGS")
> pml
$par
[1] 0.5499996

$value
[1] 13.76278
```

i.e. on retrouve ici  $\hat{\theta} = \bar{y}$ .

Maintenant, on peut tester  $H_0 : \theta = \theta_* = 50\%$  versus  $H_1 : \theta \neq 50\%$ . Pour le test de Wald, on peut commencer par calculer  $nI(\theta_*)$  ou ,

```
> nx=sum(X==1)
> f = expression(nx*log(p)+(n-nx)*log(1-p))
> Df = D(f, "p")
> Df2 = D(Df, "p")
> p=p0=0.5
```

```
> (IF==eval(Df2))
```

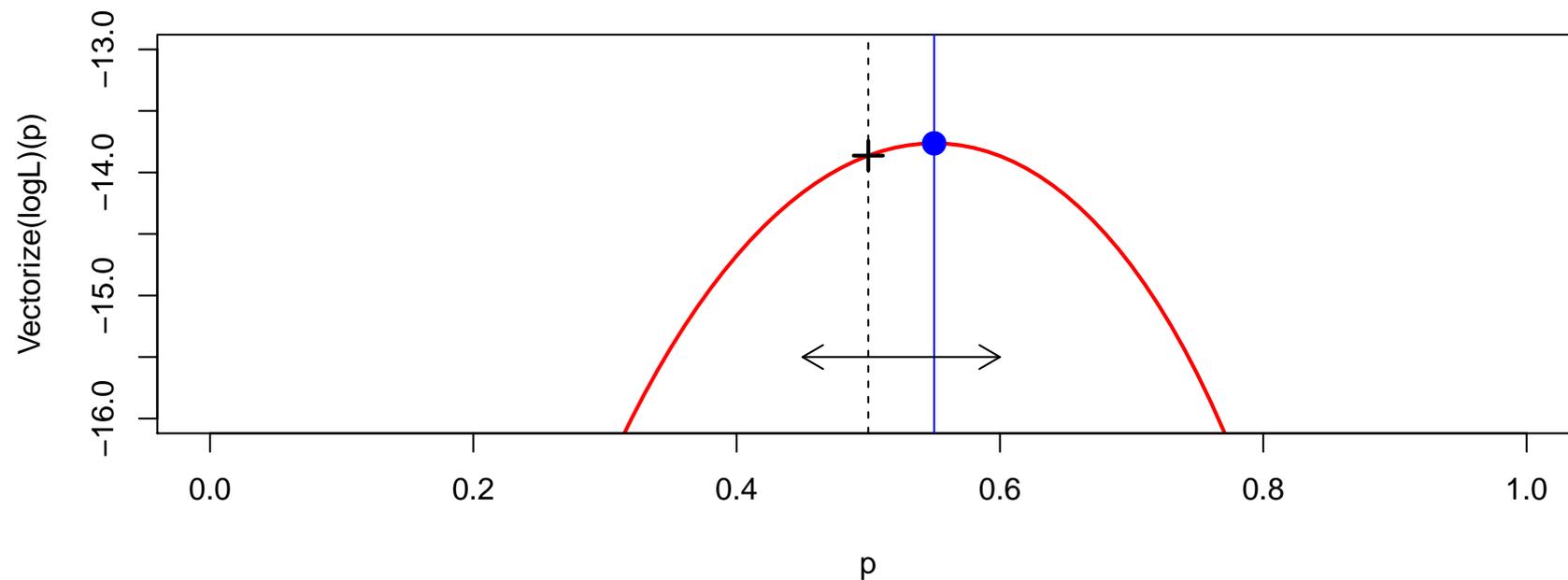
```
[1] 80
```

On peut d'ailleurs comparer cette valeur avec la valeur théorique, car

$$I(\theta)^{-1} = \theta(1 - \theta)$$

```
> 1/(p0*(1-p0)/n)
```

```
[1] 80
```



La statistique du test de Wald est ici

```
> pml=optim(fn=neglogL,par=p0,method="BFGS")$par  
> (T=(pml-p0)^2*IF)  
[1] 0.199997
```

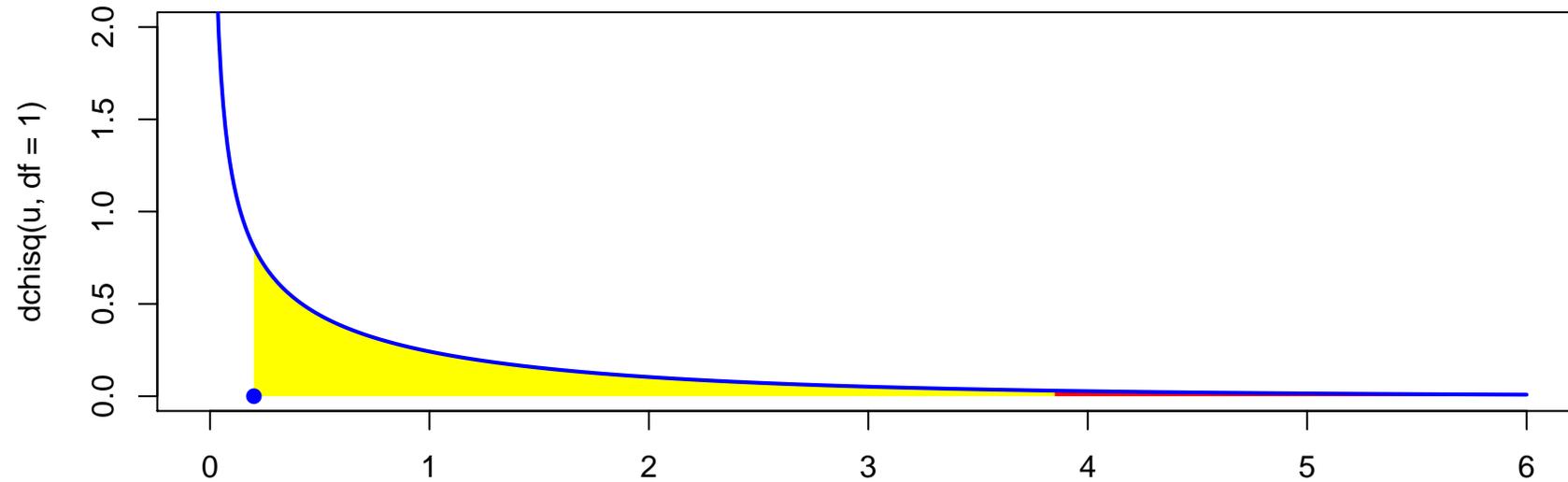
que l'on va chercher à comparer avec le quantile de la loi du  $\chi^2$ ,

```
> T<qchisq(1-alpha,df=1)  
[1] TRUE
```

i.e. on est dans la région d'acceptation du test. De manière duale, on peut calculer la  $p$ -value du test,

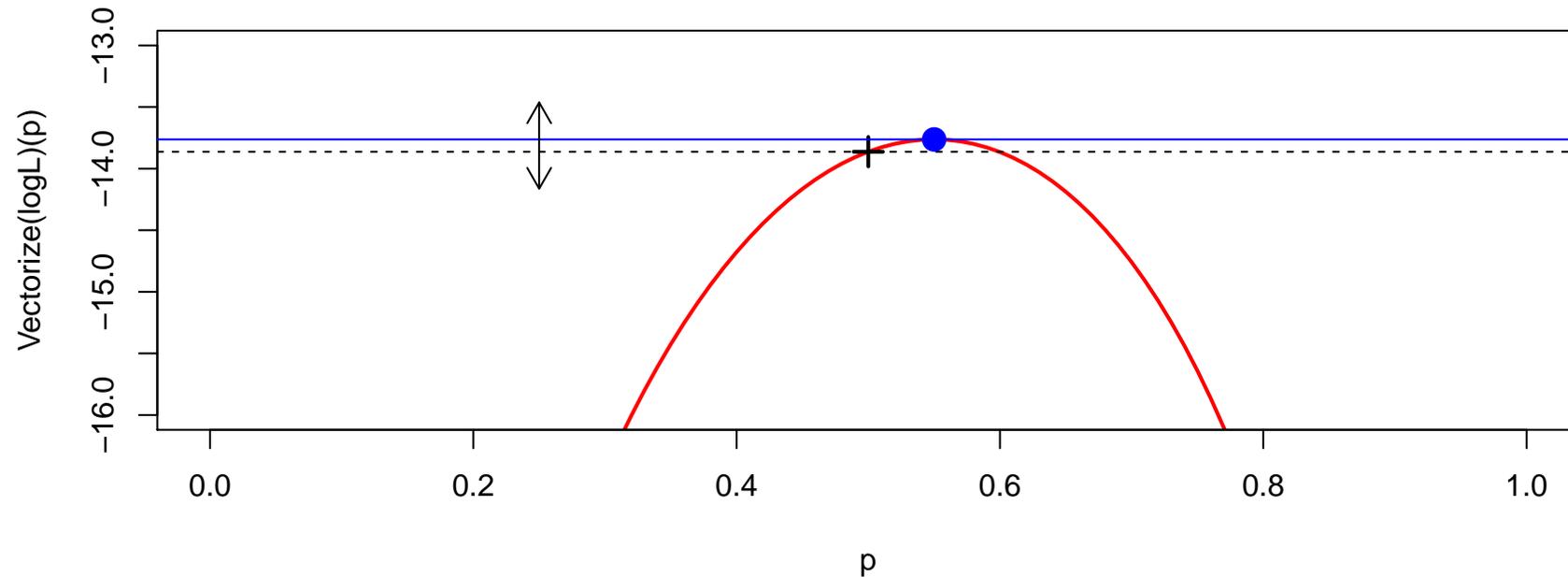
```
> 1-pchisq(T,df=1)  
[1] 0.6547233
```

Donc on va accepter  $H_0$ .



Pour le test du rapport de vraisemblance,  $T$  est ici

```
> (T=2*(logL(pml)-logL(p0)))  
[1] 0.2003347
```



Là encore, on est dans la région d'acceptation,

```
> T<qchisq(1-alpha,df=1)
[1] TRUE
```

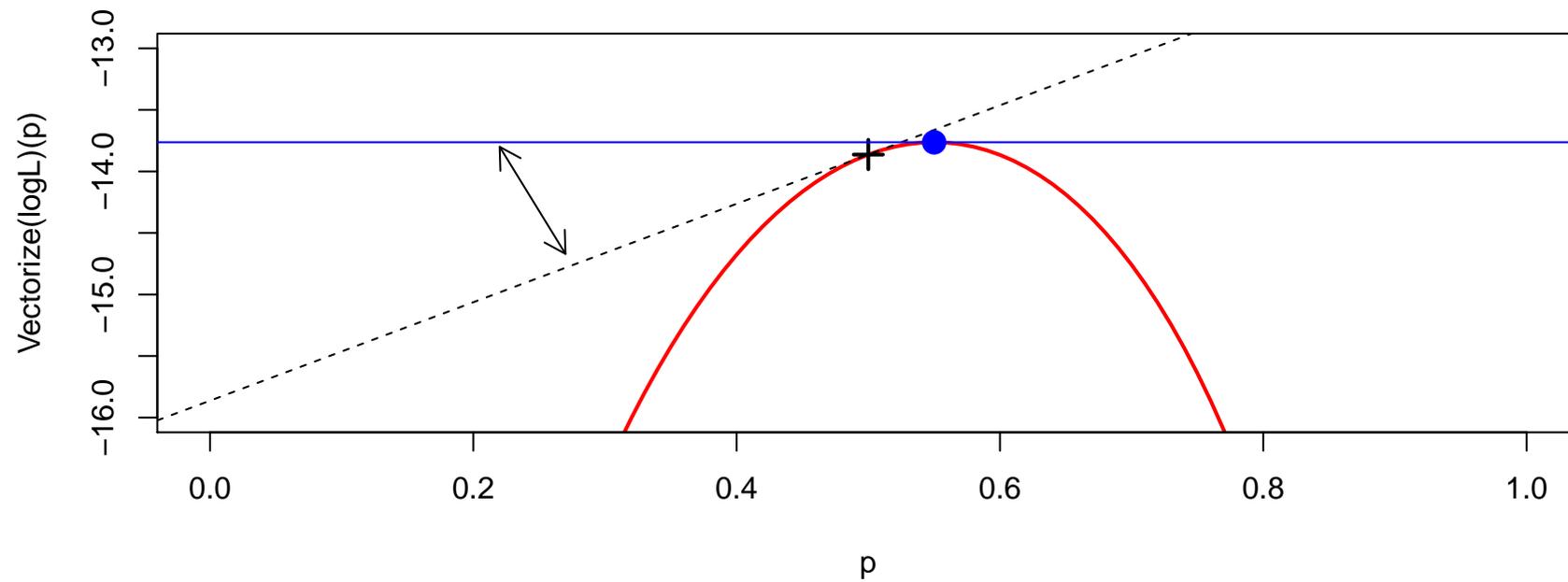
Enfin, on peut faire le test du score. Le score se calcule facilement,

```
> nx=sum(X==1)
```

```
> f = expression(nx*log(p)+(n-nx)*log(1-p))
> Df = D(f, "p")
> p=p0
> score=eval(Df)
```

La statistique de test est alors

```
> (T=score^2/IF)
[1] 0.2
```



Là encore, on est dans la région d'acceptation,

```
> T<qchisq(1-alpha,df=1)
[1] TRUE
```

## De la statistique mathématique au modèle de régression

Un modèle statistique standard supposerait que l'on dispose d'un échantillon  $\{Y_1, \dots, Y_n\}$  et de supposer que les observations  $Y$  sont i.i.d. de loi

$$Y \sim \mathcal{N}(\theta, \sigma^2)$$

Ici, on va disposer d'un échantillon  $\{(Y_1, X_1), \dots, (Y_n, X_n)\}$  et de supposer que les couples d'observations  $(X, Y)$  sont i.i.d.. de telle sorte que  $Y|X = x$  suive une loi

$$(Y|X = x) \sim \mathcal{N}(\theta_x, \sigma^2)$$

Aussi,  $\theta_x = \mathbb{E}(Y|X = x)$ . Dans un **modèle linéaire**

$$(Y|X = x) \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$$

## Le modèle linéaire simple, deux variables continues

Considérons - pour commencer un modèle linéaire simple,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \text{ ou } Y_i = \underbrace{(1, X_i)}_{\mathbf{X}'_i} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}}_{\boldsymbol{\beta}} + \varepsilon_i$$

où  $\varepsilon_i$  est un bruit. Si  $X$  est supposé “exogène”, déterministe, ou donné, on **suppose** que  $\varepsilon_i$  est

- centré  $\mathbb{E}(\varepsilon_i) = 0$ ,
- de variance constante  $\text{Var}(\varepsilon_i) = \sigma^2$ ,
- non autocorrélé,  $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$  pour  $i \neq j$ ,

Aussi,

- $\mathbb{E}(Y|X) = \beta_0 + \beta_1 X$
- $\text{Var}(Y|X) = \text{Var}(\varepsilon|X) = \sigma^2$  modèle **homoscédastique**.

## Estimation de $\beta = (\beta_0, \beta_1)'$

Par minimisation de la **somme des carrés des erreurs** (*ordinary least squares*)

$$SCR(\beta) = \sum_{i=1}^n \underbrace{[Y_i - (\beta_0 + \beta_1 X_i)]^2}_{\varepsilon_i^2}$$

i.e.

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^2}{\operatorname{argmin}} \{SCR(\beta)\}$$

Les conditions du premier ordre sont

$$\left. \frac{\partial SCR(\beta)}{\partial \beta_0} \right|_{\beta=\hat{\beta}} = 0 \text{ i.e. } \sum_{i=1}^n (-2)[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)] = 0$$

et

$$\left. \frac{\partial SCR(\beta)}{\partial \beta_1} \right|_{\beta=\hat{\beta}} = 0 \text{ i.e. } \sum_{i=1}^n (-2X_i)[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)] = 0$$

On obtient les équations normales

$$\sum_{i=1}^n Y_i = \sum_{i=1}^n \underbrace{\hat{\beta}_0 + \hat{\beta}_1 X_i}_{\hat{Y}_i} = \sum_{i=1}^n \hat{Y}_i$$

et

$$\sum_{i=1}^n X_i \underbrace{[Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)]}_{\varepsilon_i} = 0$$

**Proposition 1.** *Les estimateurs obtenus par minimisation de la somme des carrés des erreurs sont*

$$\hat{\beta}_1 = r \frac{\hat{s}_Y}{\hat{s}_X} \text{ et } \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

où  $\bar{X}$  et  $\bar{Y}$  sont les moyennes empiriques

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \text{ et } \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$$

où  $\hat{s}_X^2$  et  $\hat{s}_Y^2$  sont les variance empiriques

$$\hat{s}_Y^2 = \frac{1}{n-1} \sum_{i=1}^n [Y_i - \bar{Y}]^2 \text{ et } \hat{s}_X^2 = \frac{1}{n-1} \sum_{i=1}^n [X_i - \bar{X}]^2$$

et où  $r$  désigne la corrélation empirique

$$r = \frac{\frac{1}{n-1} \sum_{i=1}^n [X_i - \bar{X}][Y_i - \bar{Y}]}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n [X_i - \bar{X}]^2} \sqrt{\frac{1}{n-1} \sum_{i=1}^n [Y_i - \bar{Y}]^2}}$$

**Proposition 2.** *Les estimateurs obtenus par minimisation de la somme des carrés des erreurs sont*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i Y_i - n \bar{X} \bar{Y})}{\sum_{i=1}^n X_i^2 - n \bar{X}^2}$$

et

$$\hat{\beta}_0 = \frac{\sum_{i=1}^n X_i^2 - n \bar{X} \sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2 - n \bar{X}^2}$$

## Qualité du modèle et décomposition de la variance

On posera

- $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$  la prédiction de  $Y_i$ ,
- $\hat{\varepsilon}_i = Y_i - \hat{Y}_i$  l'erreur de prédiction

Notons que

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

aussi, en élevant à la puissance 2,

$$(Y_i - \bar{Y})^2 = (Y_i - \hat{Y}_i)^2 + (\hat{Y}_i - \bar{Y})^2 + 2(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)$$

et en sommant

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \underbrace{\sum_{i=1}^n 2(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)}_{=0}.$$

On notera que

$$\begin{aligned}
 \sum_{i=1}^n 2(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) &= \sum_{i=1}^n 2((\bar{Y} - \hat{\beta}_1 \bar{X} + \hat{\beta}_1 X_i) - \bar{Y})(Y_i - \hat{Y}_i) \\
 &= \sum_{i=1}^n 2((\bar{Y} + \hat{\beta}_1 (X_i - \bar{X})) - \bar{Y})(Y_i - \hat{Y}_i) \\
 &= \sum_{i=1}^n 2(\hat{\beta}_1 (X_i - \bar{X}))(Y_i - \hat{Y}_i) \\
 &= \sum_{i=1}^n 2\hat{\beta}_1 (X_i - \bar{X})(Y_i - (\bar{Y} + \hat{\beta}_1 (X_i - \bar{X}))) \\
 &= \sum_{i=1}^n 2\hat{\beta}_1 ((Y_i - \bar{Y})(X_i - \bar{X}) - \hat{\beta}_1 (X_i - \bar{X})^2).
 \end{aligned}$$

or

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i Y_i - n \bar{X} \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

donc ici

$$\begin{aligned}
 \sum_{i=1}^n 2(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) &= \sum_{i=1}^n 2\hat{\beta}_1 \left( (Y_i - \bar{Y})(X_i - \bar{X}) - \hat{\beta}_1 (X_i - \bar{X})^2 \right) \\
 &= 2\hat{\beta}_1 \left( \sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}) - \hat{\beta}_1 \sum_{i=1}^n (X_i - \bar{X})^2 \right) \\
 &= 2\hat{\beta}_1 \sum_{i=1}^n \left( (Y_i - \bar{Y})(X_i - \bar{X}) - (Y_i - \bar{Y})(X_i - \bar{X}) \right) \\
 &= 2\hat{\beta}_1 \cdot 0 = 0
 \end{aligned}$$

D'où la formule de décomposition de la variance

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

On notera  $R^2$  le coefficient d'ajustement

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

### Estimation de $\sigma^2$

Estimation de  $\sigma^2 = \text{Var}(\varepsilon)$ ,

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{\varepsilon}_i^2$$

**Remarque** Notons que  $\hat{\beta}_1 = \sum_{i=1}^n \omega_i Y_i$  où  $\omega_i = \frac{X_i - \bar{X}}{(n-1)\hat{S}_X^2}$  et  $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$

### Estimation de $\beta_1$

**Proposition 3.**  $\mathbb{E}(\hat{\beta}_1) = \beta_1$  et  $\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{(n-1)\hat{S}_X^2}$

$$\mathbb{E}(\hat{\beta}_1) = \mathbb{E} \left( \sum_{i=1}^n \omega_i Y_i \right) = \sum_{i=1}^n \omega_i \underbrace{\mathbb{E}(Y_i)}_{\beta_0 + \beta_1 X_i} = \beta_0 \underbrace{\sum_{i=1}^n \omega_i}_{=0} + \beta_1 \underbrace{\sum_{i=1}^n \omega_i X_i}_{=1} = \beta_1$$

et

$$\text{Var}(\hat{\beta}_1) = \text{Var} \left( \sum_{i=1}^n \omega_i Y_i \right) = \sum_{i=1}^n \omega_i^2 \text{Var}(Y_i)$$

### Estimation de $\beta_0$

De même, on peut dériver une expression similaire pour l'estimateur de la constante

**Proposition 4.**  $\mathbb{E}(\hat{\beta}_0) = \beta_0$  et  $\text{Var}(\hat{\beta}_0) = \frac{\sigma^2}{n} \left( 1 + \frac{\bar{X}^2}{\widehat{S}_X^2} \right)$

## Écriture matricielle du modèle

Posons

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_{n-1} \\ 1 & X_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}$$

Alors

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{n-1} \\ Y_n \end{pmatrix} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \begin{pmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_{n-1} \\ \beta_0 + \beta_1 X_n \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{n-1} \\ \varepsilon_n \end{pmatrix}$$

i.e.

$$\underset{n \times 1}{\mathbf{Y}} = \underset{n \times 2}{\mathbf{X}} \underset{2 \times 1}{\boldsymbol{\beta}} + \underset{n \times 1}{\boldsymbol{\varepsilon}}$$

Si on fait un peu de calcul matriciel, notons que

$$\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 \\ X_1 & X_2 & \cdots & X_{n-1} & X_n \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{n-1} \\ Y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n Y_i \\ \sum_{i=1}^n X_i Y_i \end{pmatrix}$$

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 \\ X_1 & X_2 & \cdots & X_{n-1} & X_n \end{pmatrix} \begin{pmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_{n-1} \\ 1 & X_n \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n X_i \\ \sum_{i=1}^n X_i & \sum_{i=1}^n X_i^2 \end{pmatrix}$$

Rappelons que

$$\text{si } \mathbf{A} = \begin{pmatrix} a & c \\ b & d \end{pmatrix} \text{ alors } \mathbf{A}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -c \\ -b & a \end{pmatrix}$$

Aussi

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n \sum_{i=1}^n X_i^2 - (\sum_{i=1}^n X_i)^2} \begin{pmatrix} \sum_{i=1}^n X_i^2 & -\sum_{i=1}^n X_i \\ -\sum_{i=1}^n X_i & n \end{pmatrix}$$

i.e.

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{\sum_{i=1}^n X_i^2 - n\bar{X}^2} \begin{pmatrix} \frac{1}{n} \sum_{i=1}^n X_i^2 & -\bar{X} \\ -\bar{X} & 1 \end{pmatrix}$$

Aussi

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \frac{1}{\sum_{i=1}^n X_i^2 - n\bar{X}^2} \begin{pmatrix} \sum_{i=1}^n X_i^2 - n\bar{X} \sum_{i=1}^n X_i Y_i \\ \sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y} \end{pmatrix} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix}$$

**Proposition 5.** *En posant  $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ , les estimateurs qui minimisent la somme des carrés des erreurs sont*

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Notons que  $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ . On notera

$$\hat{\mathbf{Y}} = \underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}_{\mathbf{H}}\mathbf{Y}$$

On a aussi

$$\hat{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}} = \underbrace{[\mathbb{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']}_{\mathbf{M}} \mathbf{Y}$$

où  $\mathbf{M} = (\mathbb{I} - \mathbf{H})$ .

Ici,  $\mathbf{M}$  est une matrice symétrique,  $\mathbf{M} = \mathbf{M}'$ , et  $\mathbf{M}^2 = \mathbf{M}$ , i.e.  $\mathbf{M}$  est une matrice de projection (orthogonale). Notons que  $\mathbf{M}\mathbf{X} = \mathbf{0}$ , i.e.  $\mathbf{M}$  est un matrice de projection orthogonale à  $\mathbf{X}$ .

En revanche,  $\mathbf{H}$  est une matrice symétrique,  $\mathbf{H} = \mathbf{H}'$ , et  $\mathbf{H}^2 = \mathbf{H}$ , i.e.  $\mathbf{H}$  est une matrice de projection (orthogonale). Notons que  $\mathbf{H}\mathbf{X} = \mathbf{X}$ , i.e.  $\mathbf{H}$  est un matrice de projection orthogonale sur  $\mathbf{X}$ .

## Le modèle linéaire multiple

On peut supposer que l'on régresse sur plusieurs variables explicatives

$$Y_i = \beta_0 + \beta_1 X_{1,i} + \cdots + \beta_k X_{k,i} + \varepsilon_i \text{ ou } Y_i = \underbrace{(1, X_{1,i}, \cdots, X_{k,i})}_{\mathbf{x}'_i} \underbrace{\begin{pmatrix} \beta_0 \\ \vdots \\ \beta_k \end{pmatrix}}_{\boldsymbol{\beta}} + \varepsilon_i$$

- $Y$  est la variable à expliquer, ou **réponse**, **output**, **variable dépendante**, ou **variable endogène**
- les  $X_j$  sont les variables explicatives, ou **prédicteurs**, **input**, **variables indépendantes**, ou **variables exogènes**
- $\varepsilon$  est un bruit, supposé non expliqué (ou orthogonal) par les variables explicatives.

## Le modèle linéaire multiple

Au niveau des notations, au lieu de supposer que pour  $i = 1, \dots, n$ ,

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \dots + \beta_k X_{i,k} + \varepsilon_i \text{ ou } Y_i = \mathbf{X}'_i \boldsymbol{\beta} + \varepsilon_i,$$

on peut utiliser l'écriture matricielle,

$$\underbrace{\begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}}_{\mathbf{Y}, n \times 1} = \underbrace{\begin{pmatrix} 1 & X_{1,1} & \cdots & X_{1,k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n,1} & \cdots & X_{n,k} \end{pmatrix}}_{\mathbf{X}, n \times (k+1)} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}_{\boldsymbol{\beta}, (k+1) \times 1} + \underbrace{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}_{\boldsymbol{\varepsilon}, n \times 1}.$$

## Le modèle linéaire

Aussi, pour les résidus, on **suppose** que l'on a

$$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbb{E} \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbb{I}_n = \sigma^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & 1 \end{pmatrix} = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & \cdots & \sigma^2 \end{pmatrix}$$

Ce sont des hypothèses (qu'il faudra tester un jour....)

## Le modèle linéaire

Si  $X$  est supposé aléatoire

- $\mathbb{E}(\boldsymbol{\varepsilon}|\mathbf{X}) = \mathbf{0}$ ,  
i.e.  $\mathbb{E}(Y|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$ , ou encore  $\mathbb{E}(Y|\mathbf{X})$  est linéaire en  $\mathbf{X}$
- $\mathbb{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|X) = \sigma^2\mathbb{I}$ ,  
i.e.  $\text{Var}(Y|\mathbf{X}) = \sigma^2$ ,  $\text{Var}(Y|\mathbf{X})$  est constante : **hypothèse d'homoscédasticité**.  
 $\text{Cov}(Y_i, Y_j|X) = 0$  pour  $i \neq j$ , ou encore  $\text{Cov}(Y_i, Y_j|X)$  est nulle
- on rajoute l'hypothèse  $\text{Cov}(X, \boldsymbol{\varepsilon}) = 0$  : le bruit est la composante qui ne peut être expliquée par  $X$ .

Pour estimer  $\boldsymbol{\beta}$ , il y a (au moins) trois méthodes classiques

- méthode du **maximum de vraisemblance** : besoin de spécifier la loi des résidus
- méthode des **moments**
- méthode des **moindres carrés** :  $\hat{\boldsymbol{\beta}} \in \text{argmin} \sum_{i=1}^n [Y_i - \mathbf{X}'_i\boldsymbol{\beta}]^2$

## Estimations à distance finie

Quelques rappels sur les propriétés (ou informations) *souhaitables* d'un estimateur

- **biais** :  $\hat{\theta}$  estime sans biais de  $\theta$  si  $\mathbb{E}(\hat{\theta}) = \theta$ , on note  $\text{biais}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$ .
- **“meilleur”** :  $\hat{\theta}_1$  est un meilleur estimateur que  $\hat{\theta}_2$  si  $\text{Var}(\hat{\theta}_1)$  est “plus petit” que  $\text{Var}(\hat{\theta}_2)$ , i.e. si  $k \geq 2$   $\text{Var}(\hat{\theta}_2) - \text{Var}(\hat{\theta}_1)$  est définie positive.
- **précision** : le MSE - *mean squared error* - d'un estimateur  $\hat{\theta}$  de  $\theta$  est

$$\text{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)(\hat{\theta} - \theta)'] = \text{Var}(\hat{\theta}) + \text{biais}(\hat{\theta})\text{biais}(\hat{\theta})'$$

- **efficience** : un estimateur est efficient si  $\text{Var}(\hat{\theta}) = nI(\hat{\theta})^{-1}$ , où  $I(\theta)$  désigne l'information de Fisher, i.e.  $I = [I_{i,j}]$  où

$$I_{i,j} = \mathbb{E} \left( \frac{\partial}{\partial \theta_i} \ln f_{\theta}(X) \cdot \frac{\partial}{\partial \theta_j} \ln f_{\theta}(X) \right).$$

## Estimations en grand échantillon (approche asymptotique)

$\hat{\theta}$  est un estimateur, construit à partir de  $\mathbf{X}_1, \dots, \mathbf{X}_n$ . On notera  $\hat{\theta}_n$  cet estimateur, l'étude du comportement **asymptotique** consistant à étudier le comportement de  $\hat{\theta}_n$  lorsque  $n \rightarrow \infty$ .

- **consistance** :  $\hat{\theta}_n$  est un estimateur consistant de  $\theta$  si  $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$
- **convergence en loi** :  $\hat{\theta}_n$  est un estimateur asymptotiquement normal de  $\theta$  si

$$\sqrt{n} \left( \hat{\theta}_n - \theta \right) \xrightarrow{\mathcal{L}} \mathcal{N}(\mu, \Sigma),$$

où  $\Sigma$  est une matrice connue, ou calculable, ou estimable

- **efficience asymptotique** :  $Var(\hat{\theta}_n) \sim nI(\theta)^{-1}$  si  $n \rightarrow \infty$ .

## Méthode du maximum de vraisemblance

Si on rajouter les hypothèses :  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$  et les  $\varepsilon_i$  sont indépendants, i.e.

$$f(\varepsilon_1, \dots, \varepsilon_n) = \prod_{i=1}^n f(\varepsilon_i) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right)$$

d'où la log-vraisemblance du modèle, en notant que  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$

$$\log \mathcal{L} = -\frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma^2 - \frac{1}{2\sigma^2} [Y - \mathbf{X}\beta]' [Y - \mathbf{X}\beta]$$

Maximiser  $\log \mathcal{L}$  est équivalent à minimiser  $[Y - \mathbf{X}\beta]' [Y - \mathbf{X}\beta]$ .

La condition du premier ordre est  $\frac{\partial [Y - \mathbf{X}\beta]' [Y - \mathbf{X}\beta]}{\partial \beta} = \mathbf{0}$ , soit

$$\frac{\partial [Y - \mathbf{X}\beta]' [Y - \mathbf{X}\beta]}{\partial \beta} = -2\mathbf{X}'Y + 2\mathbf{X}'\mathbf{X}\beta = \mathbf{0}.$$

## Méthode du maximum de vraisemblance

La condition du premier ordre,  $\mathbf{X}'Y = \mathbf{X}'\mathbf{X}\beta$  est parfois appelée **équation normale**.

La condition du second ordre est  $\frac{\partial^2[Y - \mathbf{X}\beta]'[Y - \mathbf{X}\beta]}{\partial\beta\partial\beta'}$  définie positive, soit  $\mathbf{X}'\mathbf{X}$  définie positive (et donc inversible).

Si  $\mathbf{X}'\mathbf{X}$  est inversible,  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$ .

**Remarque** En réalité, on cherche à estimer  $\theta = (\beta, \sigma^2)$ . Pour l'instant, on se préoccupe seulement de l'estimation de  $\beta$ .

**Proposition 6.** *Si  $\varepsilon_i$  est i.i.d. de loi  $\mathcal{N}(0, \sigma^2)$ , et si  $\mathbf{X}'\mathbf{X}$  est inversible, alors l'estimateur du maximum de vraisemblance  $\hat{\beta}$  est*

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

## Méthode du maximum de vraisemblance

Les propriétés du maximum de vraisemblance sont des propriétés **asymptotiques**, en particulier (moyennant quelques hypothèse de régularité)

- convergence,  $\hat{\beta} \xrightarrow{\mathbb{P}} \beta$
- normalité asymptotique,  $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \star)$
- efficacité asymptotique  $Var(\hat{\beta}) \sim nI(\beta)^{-1}$  si  $n \rightarrow \infty$ .

Dans le cas de la régression simple

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) \rightarrow \mathcal{N}\left(0, \frac{n \cdot \sigma^2}{\sum (x_i - \bar{x})^2}\right) \text{ et } \sqrt{n}(\hat{\beta}_0 - \beta_0) \rightarrow \mathcal{N}\left(0, \sigma^2 \frac{\sum x_i^2}{\sum (x_i - \bar{x})^2}\right).$$

Les variances asymptotiques de  $\hat{\beta}_0$  and  $\hat{\beta}_1$  peuvent être estimée par

$$\widehat{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum (x_i - \bar{x})^2} \text{ et } \widehat{Var}(\hat{\beta}_0) = \frac{\hat{\sigma}^2}{n} \left(1 + \frac{n(\sum x_i)^2}{\sum (x_i - \bar{x})^2}\right)$$

avec comme covariance entre ces deux estimateurs

$$\widehat{cov}(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\bar{x}\hat{\sigma}^2}{\sum (x_i - \bar{x})^2}.$$

Notons que

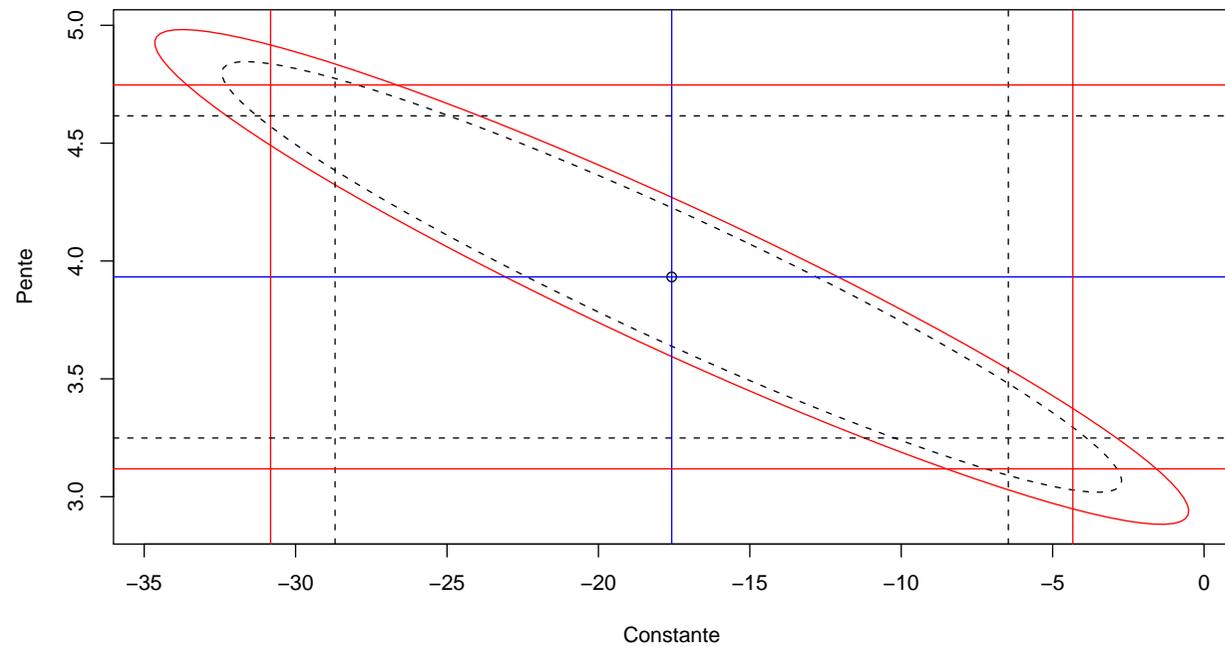
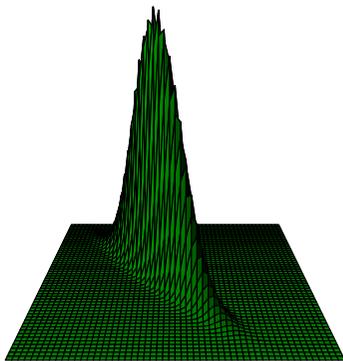
$$\hat{\beta}_1 = \frac{\text{cov}(X, Y)}{\text{Var}(X)}, \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}, \text{ et } \hat{\sigma}^2 = [1 - \text{cov}(X, Y)^2] \text{Var}(y),$$

ce qui se traduit également sous la forme

$$y_i = \bar{y} + \frac{\text{cov}(X, Y)}{\text{Var}(X)} [x_i - \bar{x}] + \varepsilon_i.$$

$\implies$  la régression est simplement une analyse de la corrélation entre des variables.

# Méthode du maximum de vraisemblance, loi asymptotique



## Méthode du maximum de vraisemblance

Mais le maximum de vraisemblance dépend explicitement de la loi des résidus  $\varepsilon$ .

**Remarque** En fait, compte tenu de la réécriture de la vraisemblance, cet estimateur correspond au maximum de vraisemblance dès lors que  $\varepsilon$  suit une loi sphérique (e.g. lois de Student indépendantes).

## Méthode des moments

La quatrième hypothèse stipulait que

$\text{Cov}(\mathbf{X}, \varepsilon) = \mathbb{E}(\mathbf{X}\varepsilon) = \mathbb{E}(\mathbf{X}[Y - \mathbf{X}\beta]) = \mathbf{0}$ , soit la version empirique

$$\frac{1}{n} \sum_{i=1}^n \mathbf{X}'_i (Y_i - \mathbf{X}_i \hat{\beta}) = 0.$$

Cette dernière condition correspond à la condition du premier ordre obtenue dans l'estimation du maximum de vraisemblance.

**Proposition 7.** *Si  $\mathbf{X}'\mathbf{X}$  est inversible, alors l'estimateur des moments  $\hat{\beta}$  est*

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

**Remarque** Cette méthode se généralise dans le cas de l'estimation nonlinéaire sous le nom GMM *Generalized Method of Moments*.

## Méthode des moindres carrés

Comme cela a été noté dans les rappels de statistiques, il existe une **interprétation géométrique** de l'espérance en terme de **projection orthogonale**,

$$\mathbb{E}(Y) = \operatorname{argmin}\{\|Y - y^*\|_{L^2}\} = \operatorname{argmin}\{\mathbb{E}(Y - y^*)^2\}$$

dont l'interprétation empirique est

$$\bar{y} = \operatorname{argmin}\left\{\sum_{i=1}^n (y_i - y^*)^2\right\}.$$

Si on suppose que  $Y = \beta_0 + \beta_1 X + \varepsilon$ , où  $\varepsilon$  est centré et indépendant de  $X$ ,  $\mathbb{E}(Y|X) = \beta_0 + \beta_1 X$ . De la même manière que précédemment, il est naturel de chercher

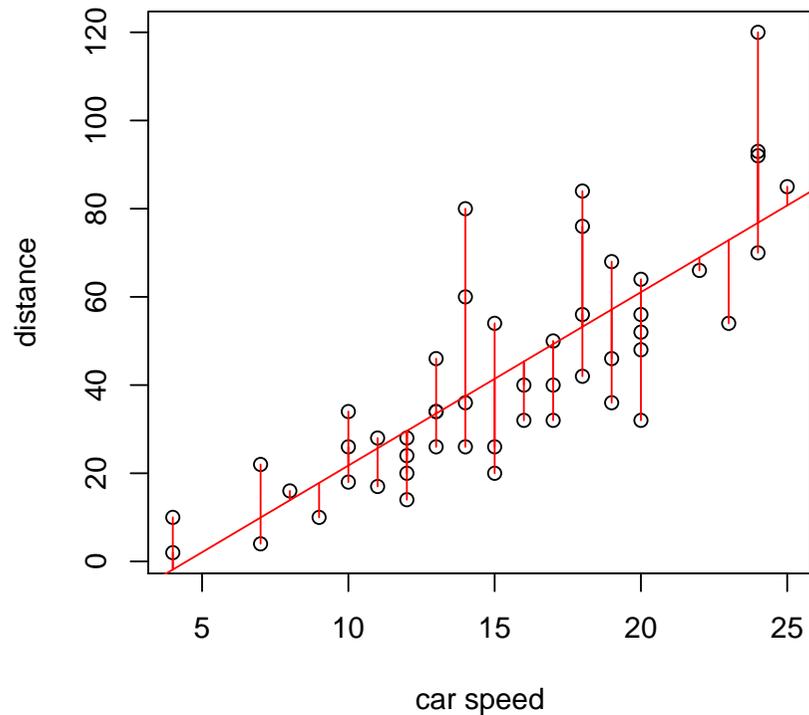
$$\operatorname{argmin}\left\{\sum_{i=1}^n (y_i - [\beta_0 + \beta_1 x_i])^2\right\}.$$

## Méthode des moindres carrés

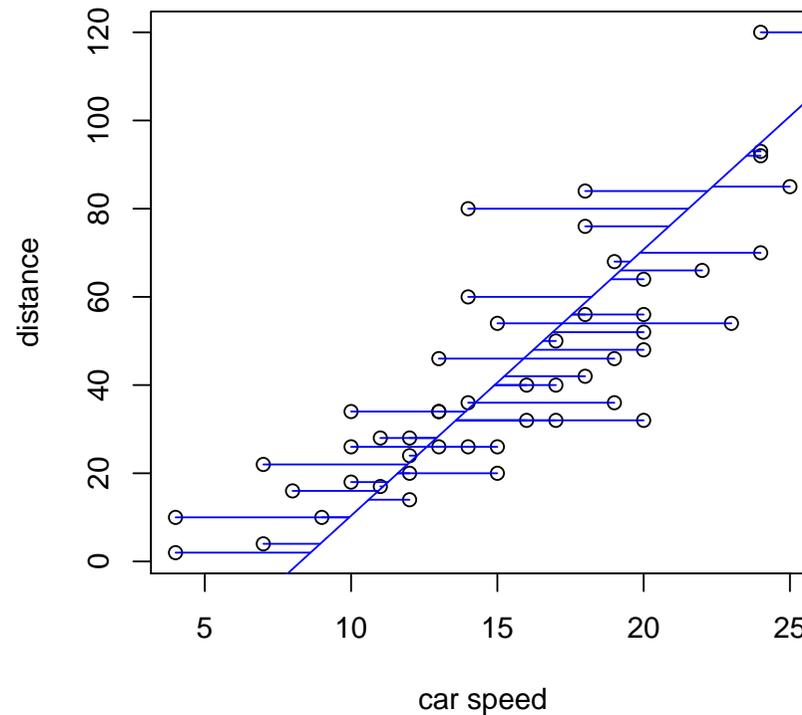
L'estimateur des moindres carrés  $\hat{\beta} \in \operatorname{argmin} \sum_{i=1}^n [Y_i - \mathbf{X}_i\beta]^2$ , soit

$$\hat{\beta} \in \operatorname{argmin}[Y - \mathbf{X}\beta]'[Y - \mathbf{X}\beta].$$

Linear regression, distance versus speed



Linear regression, speed versus distance



## Méthode des moindres carrés

La condition du premier ordre donne

$$\frac{\partial [Y - \mathbf{X}\beta]'[Y - \mathbf{X}\beta]}{\partial \beta} = -2\mathbf{X}'Y + 2\mathbf{X}'\mathbf{X}\beta = \mathbf{0}.$$

La condition du second ordre est équivalente à avoir  $\mathbf{X}'\mathbf{X}$  est inversible, et alors

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

**Proposition 8.** *Si  $\mathbf{X}'\mathbf{X}$  est inversible, alors l'estimateur des moments  $\hat{\beta}$  est*

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

## Un petit mot sur la variance des résidus $\sigma^2$

Pour l'instant nous n'avons parlé que de l'estimation de  $\beta$ , mais  $\sigma^2$  est également à estimer. La démarche est la suivante

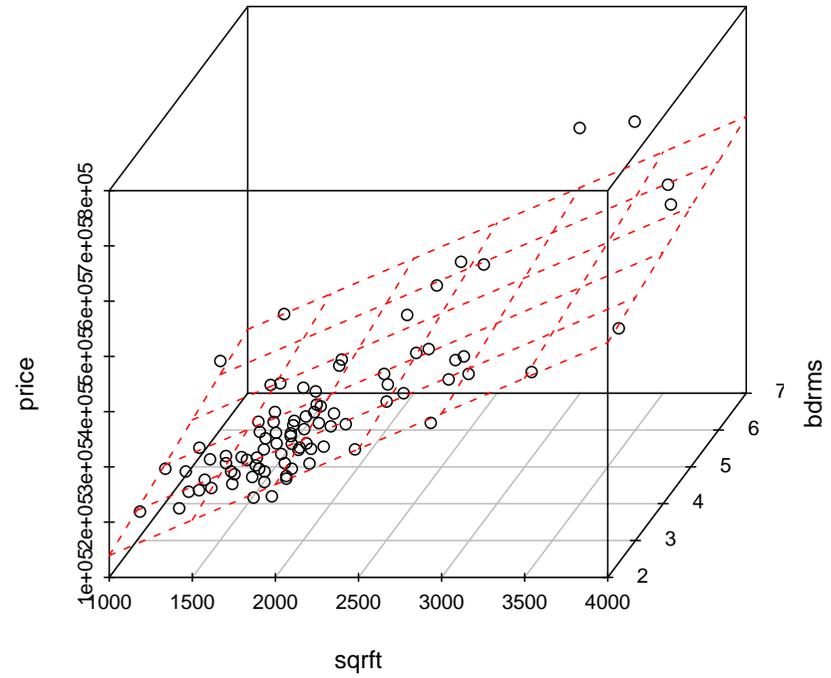
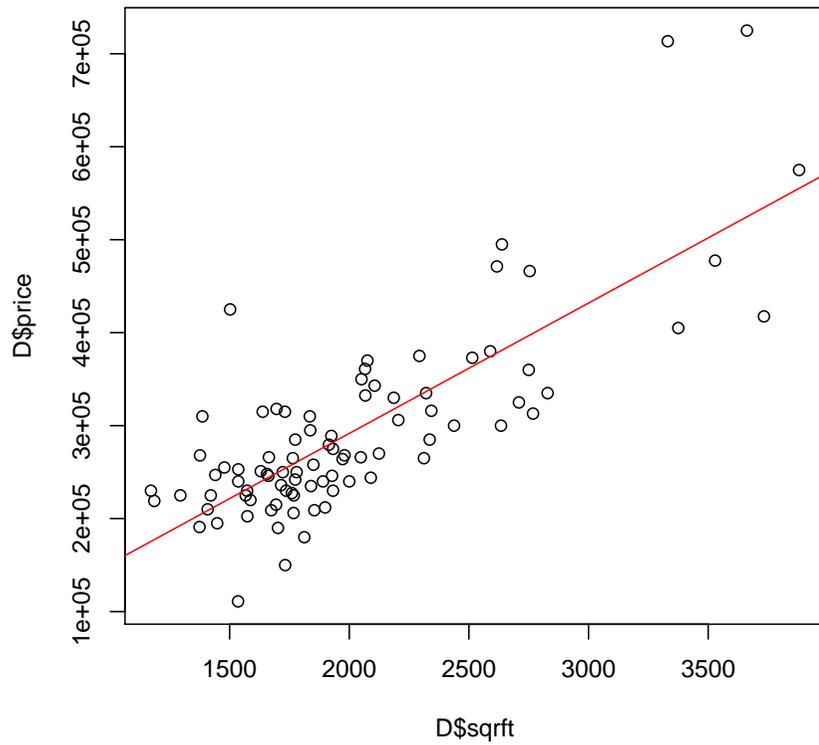
- on part d'un échantillon  $\{(Y_1, \mathbf{X}'_1), \dots, (Y_n, \mathbf{X}'_n)\}$ ,
- on construit l'estimateur par moindres carrés  $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$ ,
- on construit la série des *pseudo*-résidus  $\hat{\varepsilon}_i = Y_i - \mathbf{X}_i\hat{\beta}$ ,
- on estime la variance de la série des *pseudo*-résidus  $\hat{\sigma}^2 = \widehat{Var}(\hat{\varepsilon}_i)$ .

## Un peu de terminologie

Pour l'instant nous n'avons parlé que de l'estimation de  $\beta$ , mais  $\sigma^2$  est également à estimer. La démarche est la suivante

- $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$  est l'estimateur par moindres carrés (de  $\beta$ )
- $\hat{Y} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y = HY$  est le vecteur des **prédictions** de  $Y$ ,
- $\hat{\varepsilon} = Y - \hat{Y} = (\mathbf{I} - H)Y$  est le vecteur des **pseudo-résidus** associés à la régression,
- $\hat{\varepsilon}'\hat{\varepsilon} = Y'(\mathbf{I} - H)Y$  est la **somme des carrés des résidus**, ou RSS (*residual sum of squares*)

# La régression linéaire



## Propriétés de l'estimateur par mco

**Proposition 9.** Soit  $\hat{\beta}$  l'estimateur par moindres carrés,

- $\hat{\beta}$  estime sans biais  $\beta$ , i.e.  $\mathbb{E}(\hat{\beta}) = \beta$ ; la variance est  $\text{Var}(\hat{\beta}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ .
- Théorème de Gauss-Markov : parmi les estimateurs linéaires et sans biais de  $\beta$ , l'estimateur par moindres carrés est de variance minimale, i.e. BLUE
- $\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^n (Y_i - \mathbf{X}_i \hat{\beta})^2$  estime sans-biais  $\sigma^2$ .

Si  $\varepsilon \sim \mathcal{N}(0, \sigma^2)$ ,

- $\hat{\beta}$  est également l'estimateur du maximum de vraisemblance
- parmi les estimateurs sans biais de  $\beta$  est l'estimateur de variance minimale, i.e. BUE,
- $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$  et  $(n-k) \frac{\hat{\sigma}^2}{\sigma} \sim \chi^2(n-k)$ .

## Propriétés de l'estimateur par mco

Sous l'hypothèse où  $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I})$ , alors

- $Y = \mathbf{X}\boldsymbol{\beta} + \varepsilon$  est gaussien,  $Y \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbb{I})$
- $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'Y$  est gaussien,  $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I})$
- la somme des carrés des erreurs suit une loi du chi-deux,  $\hat{\varepsilon}'\hat{\varepsilon} \sim \sigma^2 \chi^2(n - k)$
- $\hat{\boldsymbol{\beta}}$  est indépendant de  $\hat{\sigma}^2$
- $\hat{Y}$  est indépendant de  $\hat{\varepsilon}$

## Intervalle de confiance pour $\beta_j$

Sous l'hypothèse où  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ ,  $Y_i \sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2)$ , et comme

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &\sim \mathcal{N}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}, [(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']'[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\sigma^2) \\ &\sim \mathcal{N}(\boldsymbol{\beta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2),\end{aligned}$$

soit, pour tout  $j$ ,

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\text{Var}(\widehat{\beta}_j)}} = \frac{\widehat{\beta}_j - \beta_j}{\sqrt{[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}\sigma^2}} \sim \mathcal{N}(0, 1),$$

or comme  $\sigma$  est inconnue, on la remplace par un estimateur indépendant de l'estimateur des  $\beta_j$ .

## Intervalle de confiance pour $\beta_j$

Aussi

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{\widehat{Var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j - \beta_j}{\sqrt{[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}\hat{\sigma}^2}} \sim St(n - k),$$

où  $k$  est le nombre de variables explicatives. Aussi,

$$\beta_j \in \left[ \hat{\beta}_j \pm t_{1-\alpha/2} \sqrt{[(\mathbf{X}'\mathbf{X})^{-1}]_{jj}\hat{\sigma}^2} \right],$$

sous l'hypothèse où  $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ .

## Test de nullité de $\beta_j$

Les résultats précédents permettent de proposer un test simple de

$$H_0 : \beta_j = 0 \text{ contre l'hypothèse } H_1 : \beta_j \neq 0.$$

La statistique de test

$$T_j = \frac{\hat{\beta}_j}{\sqrt{\widehat{Var}(\hat{\beta}_j)}} \sim St(n - k) \text{ sous } H_0.$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	-17.5791	6.7584	-2.601	0.0123	*
speed	3.9324	0.4155	9.464	1.49e-12	***

## Continuous sur les propriétés asymptotiques

**Proposition 10.** *Si les résidus sont centrés, de variance finie  $\sigma^2$ , alors  $\hat{\beta} \xrightarrow{\mathbb{P}} \beta$  si  $(\mathbf{X}'\mathbf{X})^{-1} \rightarrow \mathbf{0}$ .*

**Proposition 11.** *Si les résidus sont i.i.d. centrés, de variance finie  $\sigma^2$ , alors  $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \sigma Q^{-1})$  si  $\frac{1}{n}(\mathbf{X}'\mathbf{X})^{-1} \rightarrow Q$ , où  $Q$  est définie positive.*

**Proposition 12.** *Si les résidus sont i.i.d. centrés, de variance finie  $\sigma^2$ , alors  $\hat{\sigma}^2 \xrightarrow{\mathbb{P}} \sigma^2$  et  $\sqrt{n}(\hat{\sigma}^2 - \sigma^2) \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, \sigma \text{Var}(\varepsilon^2))$  si  $\frac{1}{n}(\mathbf{X}'\mathbf{X})^{-1} \rightarrow Q$ , où  $Q$  est définie positive.*

Dans le modèle linéaire gaussien, l'estimateur par moindres carrés est asymptotiquement efficace.

## Intervalles de confiance pour $Y$ et $\hat{Y}$

Supposons que l'on dispose d'une nouvelle observation  $\mathbf{X}_0$  et que l'on souhaite prédire la valeur  $Y_0$  associée.

L'incertitude sur la prédiction  $\hat{Y}_0$  vient de l'erreur d'estimation sur les paramètres  $\beta$ .

L'incertitude sur la réalisation  $Y_0$  vient de l'erreur d'estimation sur les paramètres  $\beta$  et de l'erreur associée au modèle linéaire, i.e.  $\varepsilon_0$ .

Dans le cas de la régression simple,  $Y_0 = \hat{Y}_0 + \varepsilon_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0 + \varepsilon_0$ . Aussi,

- $Var(\hat{Y}_0) = Var(\hat{\beta}_0) + 2X_0 cov(\hat{\beta}_0, \hat{\beta}_1) + X_0^2 Var(\hat{\beta}_1)$
- $Var(Y_0) = Var(\hat{Y}_0) + Var(\varepsilon_0)$ , si l'on suppose que le bruit est la partie non expliquée. Aussi  $Var(Y_0) = Var(\hat{\beta}_0) + 2X_0 cov(\hat{\beta}_0, \hat{\beta}_1) + X_0^2 Var(\hat{\beta}_1) + \sigma^2$

## Intervalle de confiance pour $\hat{Y}$

L'incertitude sur la prédiction  $\hat{Y}_j$  vient de l'erreur d'estimation sur les paramètres  $\beta$ .

Dans ce cas, si l'on dispose d'une nouvelle observation  $\mathbf{X}_0$ , l'intervalle de confiance pour  $\hat{Y}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y_0$  est

$$\left[ \hat{Y}_0 \pm t_{n-k}(1 - \alpha/2)\hat{\sigma}\sqrt{\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0} \right]$$

où  $t_{n-k}(1 - \alpha/2)$  est le quantile d'ordre  $(1 - \alpha/2)$  de la loi de Student à  $n - k$  degrés de liberté.

## Intervalle de confiance pour $Y$

L'incertitude sur la réalisation  $Y_j$  vient de l'erreur d'estimation sur les paramètres  $\beta$  et de l'erreur associée au modèle linéaire, i.e.  $\varepsilon_i$ .

Dans ce cas, si l'on dispose d'une nouvelle observation  $\mathbf{X}_0$ , l'intervalle de confiance pour  $Y_0 = \hat{Y}_0 + \varepsilon_0$  est

$$\left[ \hat{Y}_0 \pm t_{n-k}(1 - \alpha/2)\hat{\sigma}\sqrt{1 + \mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0} \right]$$

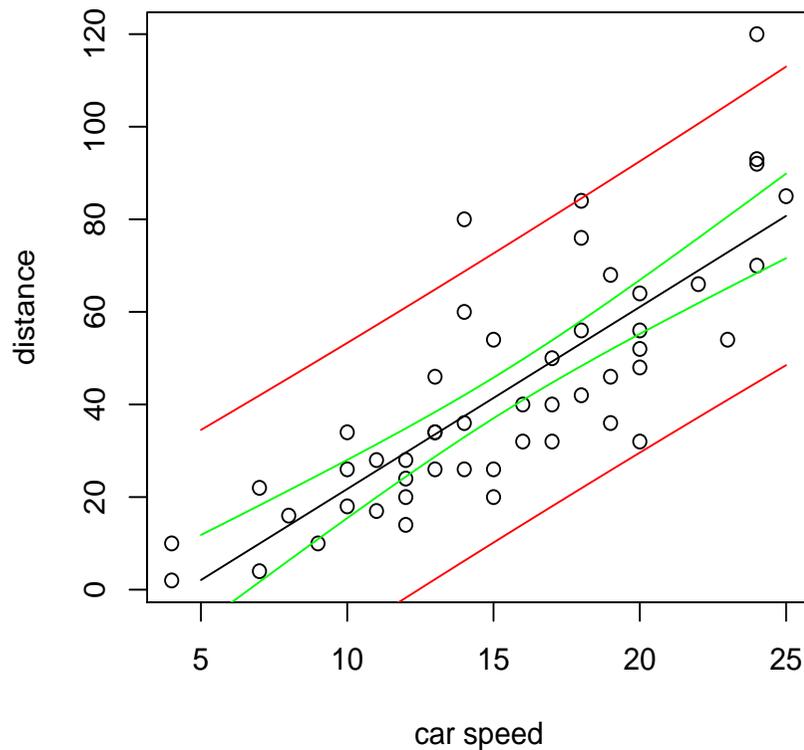
En fait, sous  $\mathbf{R}$ , l'intervalle de confiance pour  $Y$  n'inclue pas l'erreur associée à l'erreur d'estimation, aussi, dans ce cas, si l'on dispose d'une nouvelle observation  $\mathbf{X}_0$ , l'intervalle de confiance pour  $Y_0 = \hat{Y}_0 + \varepsilon_0$  est

$$\left[ \hat{Y}_0 \pm t_{n-k}(1 - \alpha/2)\hat{\sigma} \right].$$

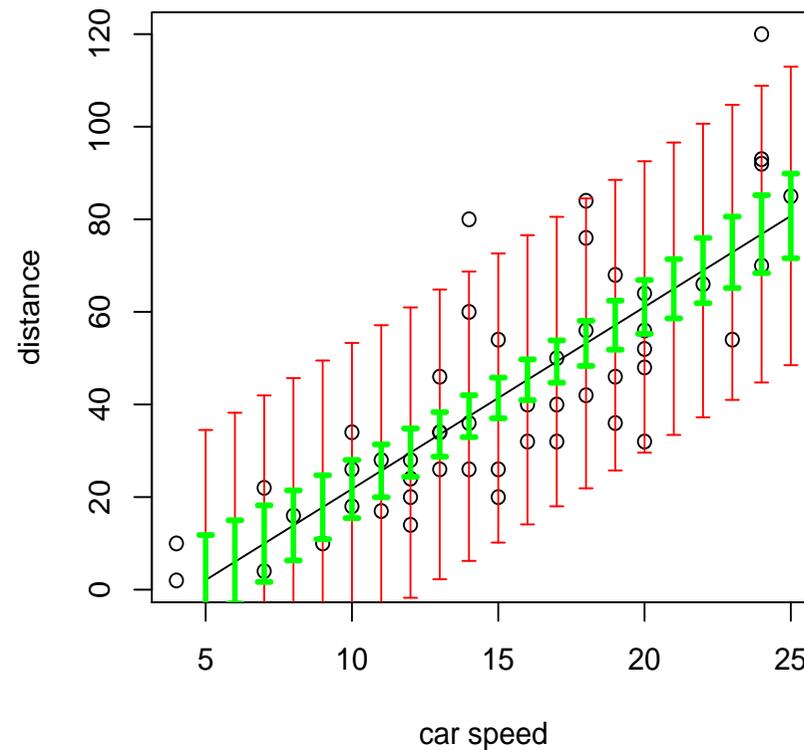
## Intervalles de confiance pour $Y$ et $\hat{Y}$

```
pp <- predict(lm(y~x,data=D), new=data.frame(x=seq(0,30)), interval='prediction')  
pc <- predict(lm(y~x,data=D), new=data.frame(x=seq(0,30)), interval='confidence')
```

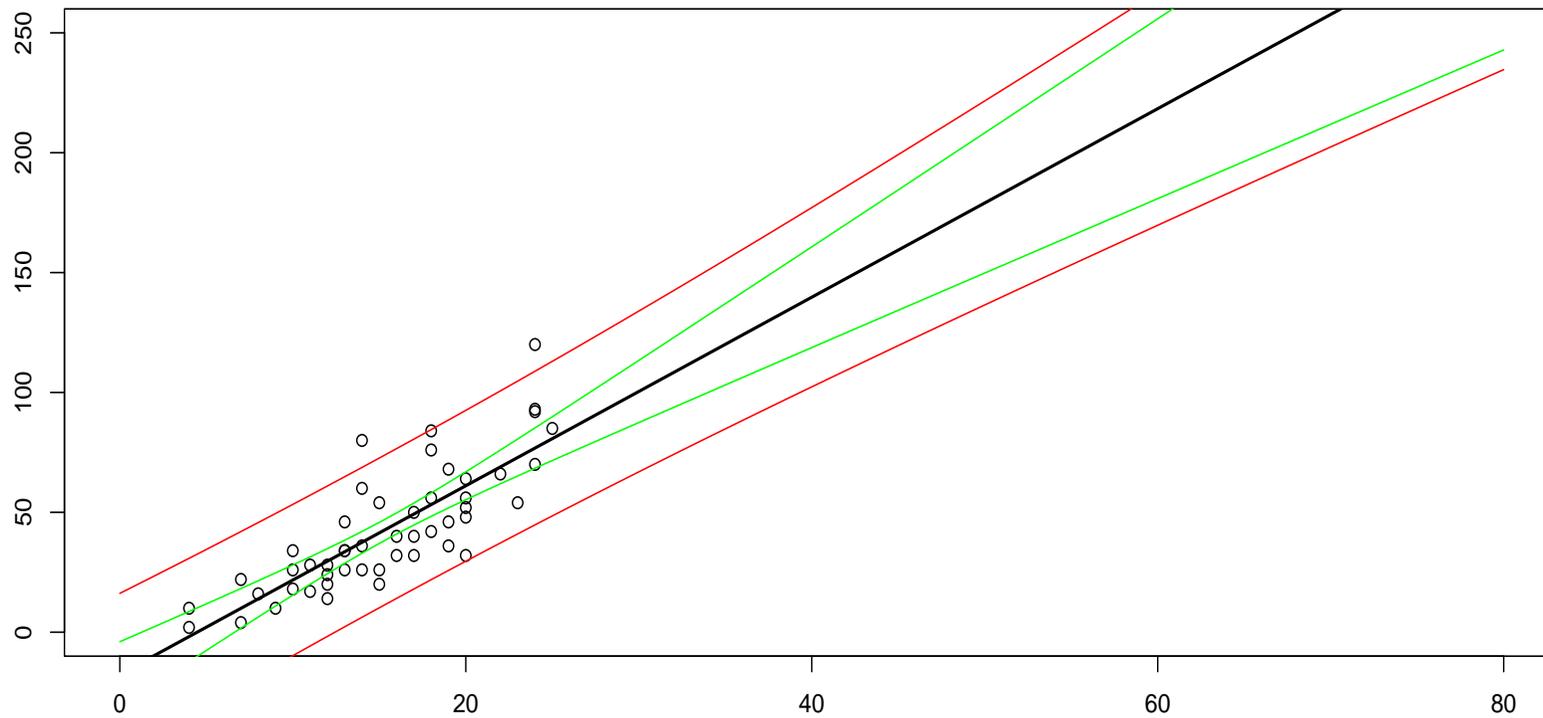
Confidence and prediction bands



Confidence and prediction bands



## Intervalles de confiance pour $Y$ et $\hat{Y}$



## Un mot sur les extensions du modèle linéaire

Nous avons données deux interprétations de la régression par moindres carrés,

- $\mathbb{E}(Y|\mathbf{X} = \mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$ ,
- $Y|\mathbf{X} = \mathbf{x} \sim \mathcal{N}(\mathbf{x}'\boldsymbol{\beta}, \sigma^2)$ .

Des extensions de ces interprétations sont les suivantes,

- médiane( $Y|\mathbf{X} = \mathbf{x}$ ) =  $\mathbf{x}'\boldsymbol{\beta}$ , dans le cas de la régression  $L^1$ , ou plus généralement, n'importe quel quantile,  $Q_\alpha(Y|\mathbf{X} = \mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$  : **régression robuste** et **régression quantile**,
- $\mathbb{E}(Y|\mathbf{X} = \mathbf{x}) = \varphi(\mathbf{x})$  : **régression nonlinéaire** ou **régression nonparamétrique**,
- $Y|\mathbf{X} = \mathbf{x} \sim \mathcal{L}(g(\mathbf{x}'\boldsymbol{\beta}))$ ; où  $\mathcal{L}$  est une loi “quelconque”, e.g. binomiale dans le cas de la régression probit/logit.

## Les variables explicatives, $\mathbf{X} = (X_1, \dots, X_k)$

- oublie d'une variable (importante) dans la régression : sous-identification

Supposons que le vrai modèle est de la forme

$$Y_i = \mathbf{X}_{1,i}\boldsymbol{\beta}_1 + \mathbf{X}_{2,i}\boldsymbol{\beta}_2 + \varepsilon_i, \quad (1)$$

où les  $\varepsilon_i$  vérifient les conditions usuelles du modèle linéaire.

Supposons que l'on régresse seulement  $Y$  sur  $\mathbf{X}_1$ ,

$$Y_i = \mathbf{X}_{1,i}\boldsymbol{\alpha}_1 + \eta_i. \quad (2)$$

L'estimation de  $\boldsymbol{\alpha}_1$  dans le modèle 2 donne

$$\begin{aligned} \hat{\boldsymbol{\alpha}}_1 &= (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1Y \\ &= (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1[\mathbf{X}_{1,i}\boldsymbol{\beta}_1 + \mathbf{X}_{2,i}\boldsymbol{\beta}_2 + \varepsilon] \\ &= (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{X}_1\boldsymbol{\beta}_1 + (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{X}_2\boldsymbol{\beta}_2 + (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\varepsilon \\ &= \boldsymbol{\beta}_1 + \underbrace{(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{X}_2\boldsymbol{\beta}_2}_{\boldsymbol{\beta}_{12}} + \underbrace{(\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\varepsilon_i}_{\nu_i} \end{aligned}$$

i.e.  $\mathbb{E}(\hat{\alpha}_1) = \beta_1 + \beta_{12}$ .

Si  $\mathbf{X}'_1 \mathbf{X}_2 = \mathbf{0}$  (variables non corrélées),  $\hat{\alpha}_1$  estime sans biais  $\beta_1$

Si on oublie une variable  $\hat{\alpha}_1$  est un estimateur biaisé de  $\beta_1$ , mais possède une variance plus faible.

- rajout d'une variable dans la régression : sur-identification

De manière duale, si on rajoute une variable, i.e. le vrai modèle est

$$Y = \mathbf{X}_1 \beta_1 + \varepsilon$$

mais que l'on estime

$$Y = \mathbf{X}_1 \alpha_1 + \mathbf{X}_2 \alpha_2 + \eta$$

L'estimation est sans biais, i.e.  $\mathbb{E}(\alpha_1) = \beta_1$ , mais l'estimateur n'est pas efficient.

## Un (petit) mot sur l'interprétation des paramètres

On interprète souvent le signe de  $\hat{\beta}_1$  comme le sens de la relation entre la variable explicative  $X_1$  et la réponse  $Y$ .

Une lecture naïve serait *“un changement d'une unité de  $X_1$  va conduire à un changement - en moyenne - de  $\hat{\beta}_1$  sur  $Y$ ”*

Mais il s'agit d'une interprétation **causale** d'une simple corrélation, et surtout, ceci suppose que *“ $\hat{\beta}_1$  mesure l'effet de  $X_1$  sur  $Y$ , toutes choses étant égales par ailleurs”*. Mais peut-on supposer que  $X_1$  et  $X_2$  soient indépendantes (et que l'on puisse faire varier  $X_1$  sans faire varier  $X_2$  ?).

## Un (petit) mot sur la multicolinéarité

On parlera de (multi)colinéarité entre deux variables exogènes lorsque la corrélation entre ces variables est (très) élevée.

- colinéarité parfait, i.e.  $\text{rang}(\mathbf{X}'\mathbf{X}) < k + 1$ , alors  $(\mathbf{X}'\mathbf{X})$  n'est pas inversible
- colinéarité très forte, alors le calcul de  $(\mathbf{X}'\mathbf{X})^{-1}$  est instable.

L'idée de la régression ridge (proposée par Hoerl & Kennard (1970)) est de considérer

$$\hat{\beta} = (\mathbf{X}'\mathbf{X} + \lambda\mathbb{I})^{-1} \mathbf{X}'\mathbf{Y}.$$

**Attention** La multicolinéarité biaise les interprétations dans les contextes de prédiction, c.f. [paradoxe de Simpson](#)

Considérons le cas où

- $Y$  désigne le fait d'être guéri ou décédé
- $X_1$  désigne l'hôpital (noté ici  $A$  ou  $B$ )
- $X_2$  désigne le fait d'être entré (ou pas) en bonne santé

## Un (petit) mot sur la multicolinéarité

Un peu de statistique descriptive donne, pour  $\mathbb{E}(Y|X_1)$

hôpital	total	survivants	décès	taux de survie
hôpital A	1,000	800	200	80%
hôpital B	1,000	900	100	90% ★

Personnes en bonne santé, i.e.  $\mathbb{E}(Y|X_1, X_2 = \text{bonne santé})$

hôpital	total	survivants	décès	taux de survie
hôpital A	600	590	10	98% ★
hôpital B	900	870	30	97%

Personnes malades, i.e.  $\mathbb{E}(Y|X_1, X_2 = \text{malade})$

hôpital	total	survivants	décès	taux de survie
hôpital A	400	210	190	53% ★
hôpital B	100	30	70	30%

## Un (petit) mot sur la multicolinéarité

Aussi,  $\mathbb{E}(Y|X_1)$  et  $\mathbb{E}(Y|X_1, X_2 = x_2)$  peuvent être sensiblement différents pour tout  $x_2$ .

L'explication du paradoxe est que le choix de l'hôpital et la condition médicale à l'entrée sont très liés ( $B$  ne soignant pas des personnes malades)

Pour détecter la multicolinéarité, on étudie les **corrélations croisées**. Il existe généralement deux règles pour détecter de la multicolinéarité entre  $X_1$  et  $X_2$ , i.e.

- si  $|r_{1,2}| > 0.8$ , mais cela ne prend pas en compte les particularités de la régression,
- si  $r_{1,2}^2 > R^2$ , règle dite de **Klein**.

## Un (petit) mot sur la multicolinéarité

Une autre interprétation est la suivante : il est possible d'avoir

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \text{ avec } \beta_1 > 0,$$

mais

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \eta \text{ avec } \alpha_1 < 0.$$

⇒ Attention aux interprétations fallacieuses des valeurs de coefficients.

Autre exemple, considérons le cas où

- $Y$  désigne le pourcentage d'admission à Berkeley
- $X_1$  désigne le sexe de la personne ( $H$  ou  $F$ )
- $X_2$  désigne la majeure

## Un (petit) mot sur la multicolinéarité

Un peu de statistique descriptive donne, pour  $\mathbb{E}(Y|X_1)$

sexe	postulants	admis	(%)
hommes $H$	2,074	1,018	48%
femmes $F$	849	261	30%

Personnes en bonne santé, i.e.  $\mathbb{E}(Y|X_1, X_2)$

Hommes				Femmes			
majeure	post.	admis	(%)	majeure	post.	admis	(%)
$A$	825	511	62%	$A$	108	89	82%
$B$	560	353	63%	$B$	25	17	68%
$C$	417	138	33%	$C$	375	131	35%
$D$	272	16	6%	$D$	341	24	7%

## Un (petit) mot sur la multicollinéarité

On peut retrouver des dizaines d'exemples de ce type,

**Example** : confrontation de deux joueurs de base-base

joueur	1995		1996		Total	
Derek Jeter	12/48	25.0%	183/582	31.4%	195/630	31.0%
David Justice	104/411	25.3%	45/140	32.1%	149/551	27.0%

**Example** : validation clinique d'un médicament

traitement	petits caillots		gros caillots		Total	
A 234/270	87%	55/80	69%	289/350	83%	
B 81/87	93%	192/263	73%	273/350	78%	