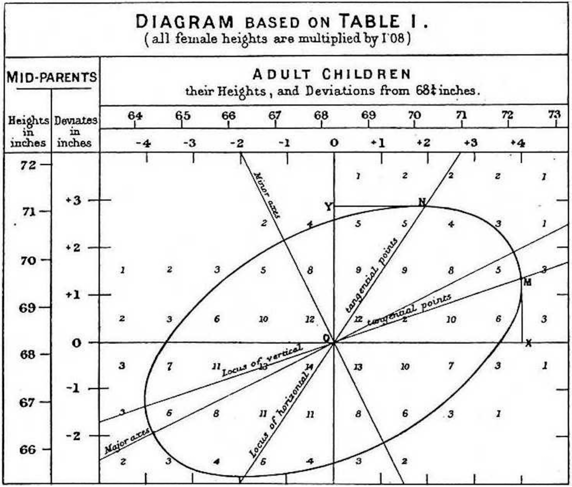
Statistiques pour les sciences (MAT-4681)

Arthur Charpentier

15 - Régression simple

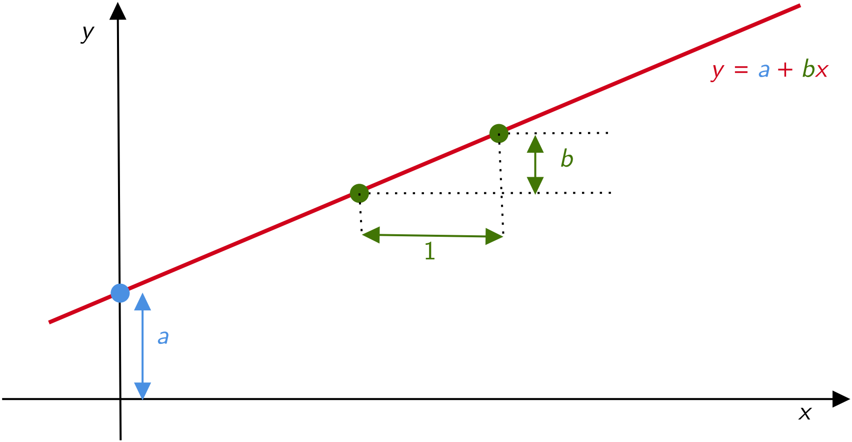
été 2022

Linear Regression



Galton regression towards mediocrity in hereditary stature, 1886.

Droite (dans le plan)



Covariance et corrélation I

Covariance

On appelle covariance d'un couple de variable aléatoire

$$\text{Cov}(X,Y)=\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]=\mathbb{E}[XY]-\mathbb{E}[X]\mathbb{E}[Y]$$

(à condition que $\mathbb{E}[X^2]<\infty$ et $\mathbb{E}[Y^2]<\infty$).

Corrélation

On appelle corrélation d'un couple de variable aléatoire

$$\text{Cor}(X,Y)=\frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)\cdot\text{Var}(Y)}}$$

(à condition que $\mathbb{E}[X^2]<\infty$ et $\mathbb{E}[Y^2]<\infty$).

Moindrs carrés

Pour rappel (partie 4)

Moyenne (empirique) / empirical mean / average

$\overline{y}$ est la solution de $\overline{y} = \underset{m\in\mathbb{R}}{\text{argmin}} \left\{ \sum_{i=1}^{n} (y_i - m)^2 \right\}$

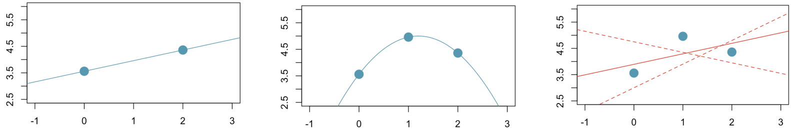
"*De tous les principes qu'on peut proposer pour cet objet, je pense qu'il n'en est pas de plus général, de plus exact, ni d'une application plus facile que celui dont nous avons fait usage dans les recherches précédentes, et qui consiste à rendre minimum la somme des quarrés des erreurs. Par ce moyen, il s'établit entre les erreurs une sorte d'équilibre qui empêchant les extrêmes de prévaloir, est très-propre à faire connoître l'état du système le plus proche de la vérité*", Legendre (1806)

Moindres carrés

Soit $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ un échantillon. On suppose

$$y_i = \alpha + \beta x_i + \varepsilon_i, \text{ pour } i = 1, \cdots, n,$$

- y est la variable d'intérêt (que l'on veut prédire)
- x est une variable explicative (possible)
- ε est une variable résiduelle (non observée et imprévisible)



Si $n \geq 3$ et que les points ne sont pas alignés, il y a une infinité de droites de régression possibles. On va chercher

$$\min_{\alpha,\beta}\left\{\sum_{i=1}^{n}\varepsilon_i^2\right\} = \min_{\alpha,\beta}\left\{\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right\}$$

Moindres carrés ★★★

Formellement, on va avoir besoin de quelques hypothèses pour aller plus loin,

$\mathcal{H}_1$: La variable explicative X n'est pas constante

$\mathcal{H}_2$: Les erreurs sont centrées, de même variance et non corrélées $\Leftrightarrow \mathbb{E}(\varepsilon_i) = 0$ et $\text{Var}(\varepsilon_i) = \sigma^2$.

$\mathcal{H}_2^{\mathcal{N}}$: Les erreurs sont indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$

$\mathcal{H}_3$: Lorsque $n \to \infty$, $\displaystyle\frac{1}{n}\sum_{i=1}^{n} x_i^2 \to Q > 0$

- $\mathcal{H}_1$: permet de démontrer l'existence de $\widehat{\alpha}$ et $\widehat{\beta}$.
- $\mathcal{H}_2$: permet de démontrer des propriétés pour $\widehat{\alpha}$ et $\widehat{\beta}$ (sans biais, calcul de variance).
- $\mathcal{H}_2^{\mathcal{N}}$: permet de faire des tests.
- $\mathcal{H}_3$: permet de démontrer la convergence de $\widehat{\alpha}$ et $\widehat{\beta}$.

Moindres carrés

Droite de régression, moindres carrées (OLS)

Soit $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ un échantillon. La droite de régression qui minimise la somme des carrés des erreurs est $y = \widehat{\alpha} + \widehat{\beta}x$ où

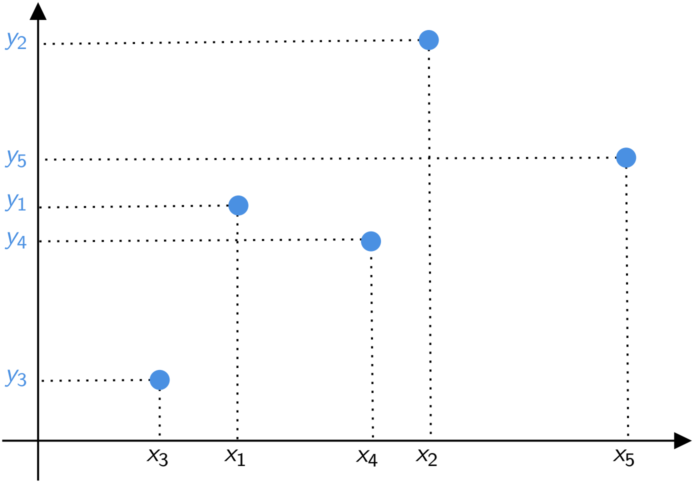
$$(\widehat{\alpha}, \widehat{\beta}) = \underset{\alpha,\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} \varepsilon_i^2 \right\} = \underset{\alpha,\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \right\}$$

Note il est possible de considérer d'autre critère, comme la somme des valeurs absolue des erreurs

$$\underset{\alpha,\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} |\varepsilon_i| \right\} = \underset{\alpha,\beta}{\text{argmin}} \left\{ \sum_{i=1}^{n} |y_i - \alpha - \beta x_i)| \right\}$$

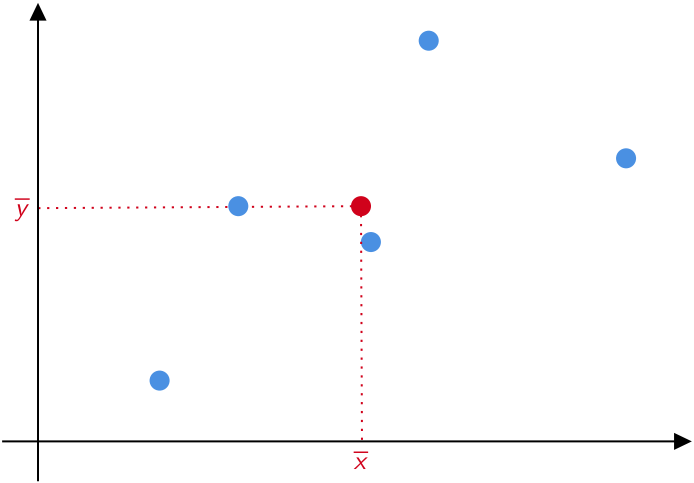
Régression : notations I

On considère l'échantillon $\{(x_1, y_1), \cdots, (x_n, y_n)\}$



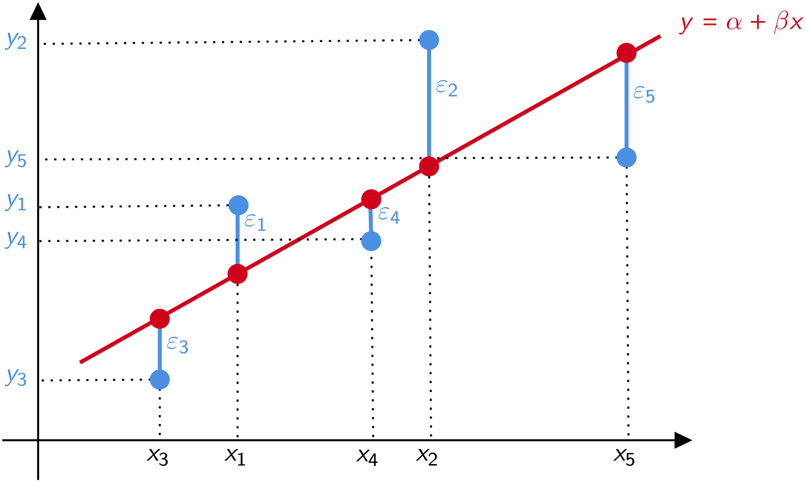
Régression : notations II

On note $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ et $\overline{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$ les moyennes empiriques



Régression : notations III

Les résidus sont $\varepsilon_i = y_i - (\alpha + \beta x_i)$



Régression

Droite de régression, moindres carrées (OLS)

Soit $(\boldsymbol{x},\boldsymbol{y})=\{(x_1,y_1),\cdots,(x_n,y_n)\}$ un échantillon. La droite de régression qui minimise la somme des carrés des erreurs est $y=\widehat{\alpha}+\widehat{\beta}x$, où

$$\begin{cases}\widehat{\alpha} & =\overline{y}-\widehat{\beta}\,\overline{x},\\[1em] \widehat{\beta} & =\dfrac{\displaystyle\sum_{i=1}^n(x_i-\overline{x})(y_i-\overline{y})}{\displaystyle\sum_{i=1}^n(x_i-\overline{x})^2}=r_{xy}\dfrac{s_y}{s_x}.\end{cases}$$

(admis)

Régression

Pour obtenir (numériquement) $\widehat{\alpha}$ et $\widehat{\beta}$,

```
> model = lm(weight~height, data=Davis)
> model

Coefficients:
(Intercept)      height
    -130.91        1.15
```

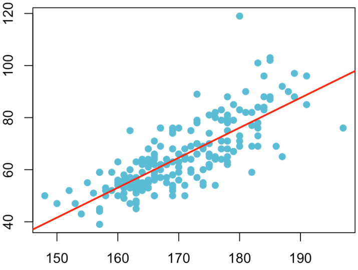
On peut vérifier que $\widehat{\beta} = r_{xy}\dfrac{s_y}{s_x}$

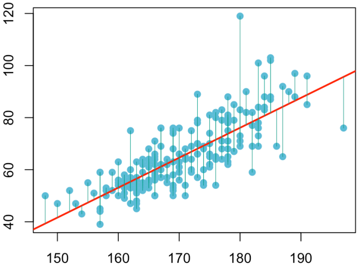
```
> (b = cor(Davis$weight,Davis$height)*sd(Davis$weight)
    /sd(Davis$height))
[1] 1.150092
```

et que $\widehat{\alpha} = \overline{y} - \widehat{\beta}\,\overline{x}$

```
> (a = mean(Davis$weight)-b*mean(Davis$height))
[1] -130.9104
```

Régressions



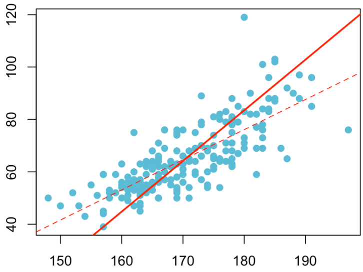


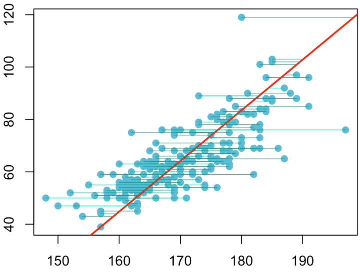
Régresser y sur x et régresser x sur y ne sont pas équivalent...

```
> model = lm(height~weight, data=Davis)
> model

Coefficients:
(Intercept)        weight
    136.831         0.517
```

Régressions

```
> model = lm(height~weight, data=Davis)
> model

Coefficients:
(Intercept)       weight
    136.831        0.517
```





Prévision et Résidus

Prévision et résidus

Soit $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ un échantillon. La droite de régression qui minimise la somme des carrés des erreurs est $y = \widehat{\alpha} + \widehat{\beta}x$. La différence entre la valeur observée y_i et la valeur prédite $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}x_i$ est le résidu empirique $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$.

Étant donnée une nouvelle observation x_{n+1}, la prévision est

$$\widehat{y}_{n+1} = \widehat{\alpha} + \widehat{\beta}x_{n+1} = \overline{y} + r_{xy}\frac{s_y}{s_x}(x_{n+1} - \overline{x})$$

et la nouvelle observation sera

$$y_{n+1} = \overline{y} + r_{xy}\frac{s_y}{s_x}(x_{n+1} - \overline{x}) + \widehat{\varepsilon}_{n+1}$$

Estimation et prévision

$\widehat{\alpha}$, $\widehat{\beta}$ et $\widehat{y}_{n+1}$ estiment sans biais α, β et $\mathbb{E}[Y|X = x_{n+1}]$.

Prévision et Résidus

Résidus

Soient $\widehat{\varepsilon}_i = y_i - \widehat{y}_i$ les résidus estimés. Les résidus sont centrés

$$\sum_{i=1}^{n} \widehat{\varepsilon}_i = 0,$$

et leur variance σ^2 est estimée par s^2 où

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

s^2 estime sans biais σ^2.

Test de significativité

Test $H_0 : \beta = 0$ contre $H_1 : \beta \neq 0$

Soit $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ un échantillon. Si la droite de régression est $y = \alpha + \beta y$, pour tester $H_0 : \beta = 0$ contre $H_1 : \beta \neq 0$, la statistique de test est

$$T = \frac{\widehat{\beta}}{s_{\widehat{\beta}}} \text{ où } s_{\widehat{\beta}} = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Si $H_0 : \beta = 0$ est vraie, et si $\varepsilon \sim \mathcal{N}(0, \sigma^2)$, $T \sim \mathcal{S}td(n-2)$. Et donc

- on rejette H_0 si $|t| > T_{n-2}^{-1}(1 - \alpha/2)$

où T_ν est la fonction de répartition de la loi de Student $\mathcal{S}td(\nu)$

Test de significativité ★★★

Test $H_0 : \alpha = 0$ contre $H_1 : \alpha \neq 0$

Soit $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ un échantillon. Si la droite de régression est $y = \alpha + \beta y$, pour tester $H_0 : \beta = 0$ contre $H_1 : \beta \neq 0$, la statistique de test est

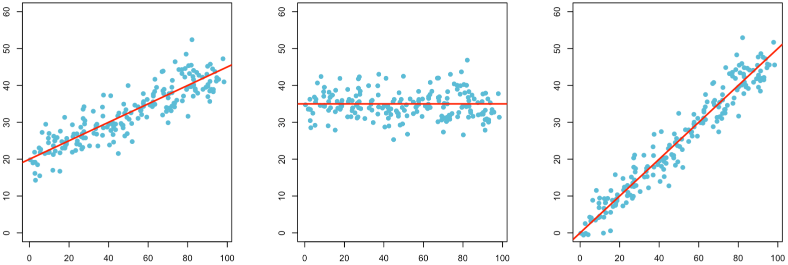
$$T = \frac{\widehat{\alpha}}{s_{\widehat{\alpha}}} \text{ où } s_{\widehat{\alpha}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)}$$

Si $H_0 : \alpha = 0$ est vraie, et si $\varepsilon = Y - (\alpha + \beta X) \sim \mathcal{N}(0, \sigma^2)$, $T \sim \mathcal{S}td(n-2)$. Et donc

- on rejette H_0 si $|t| > T_{n-2}^{-1}(1 - a/2)$

où T_ν est la fonction de répartition de la loi de Student $\mathcal{S}td(\nu)$

Test de significativité



Si $H_0 : \beta = 0$ est accepté, la pente est nulle, et x n'influence pas y

Si $H_0 : \alpha = 0$ est accepté, $y = \beta x$, autrement dit, il y a une relation de proportionalité entre x et y

Regression avec Python

```
> import numpy as np
> import statsmodels.api as sm
> x = np.array([5, 15, 25, 35, 45, 55])
> x = x.reshape((-1, 1))
> x = sm.add_constant(x)
> y = np.array([5, 20, 14, 32, 22, 38])
> model = sm.OLS(y, x)
> results = model.fit()
> print(results.summary())
=====================================================
          coef.  std err       t   P>|t|    [0.025   0.975]
-----------------------------------------------------
const     5.6333   5.872   0.959   0.392  -10.670   21.936
x1        0.5400   0.170   3.175   0.034    0.068    1.012
-----------------------------------------------------
Dep. Variable:   y       R-squared:          0.716
Model:           OLS     Adj. R-squared:     0.645
                         F-statistic:        10.08
```

Regression avec R

```
> df = data.frame(x=c(5, 15, 25, 35, 45, 55),
                  y=c(5, 20, 14, 32, 22, 38))
> model = lm(y~x, data=df)
> summary(model)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.6333     5.8719   0.959   0.3917
x             0.5400     0.1701   3.175   0.0337 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

Residual standard error: 7.116 on 4 degrees of freedom
Multiple R-squared: 0.7159, Adjusted R-squared: 0.6448
F-statistic: 10.08 on 1 and 4 DF,  p-value: 0.03371
```

Intervalle de confiance

Intervalle de confiance pour β et α

Soit $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ un échantillon. Si la droite de régression est $y = \alpha + \beta y$, et si $\varepsilon = Y - (\alpha + \beta X) \sim \mathcal{N}(0, \sigma^2)$, l'intervalle de confiance pour β est

$$\left[\widehat{\beta} \pm t_{n-2,1-a/2} s_{\widehat{\beta}}\right]$$

et l'intervalle de confiance pour α est

$$\left[\widehat{\alpha} \pm t_{n-2,1-a/2} s_{\widehat{\alpha}}\right]$$

```
> confint(model)
                    2.5 %    97.5 %
(Intercept) -10.66959886 21.936266
x             0.06773221  1.012268
```

Décomposition de la variance

Décomposition de la variance

Soit $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ un échantillon, et $\widehat{y}_i$ la prévision par régression linéaire. Alors

$$\underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{variance totale}} = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{variance résiduelle}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{variance expliquée}}$$

(admis)

On décompose aussi la somme des carrés totaux en la somme des carrés des résidus et la somme des carrés expliqués

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{\text{SCT}} = \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{\text{SCR}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{\text{SCE}}$$

R^2

R^2

Soit $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ un échantillon, et $\widehat{y}_i$ la prévision par régression linéaire. Alors

$$R^2 = \frac{\text{variance expliquée}}{\text{variance totale}} \in [0, 1]$$

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

R^2

$$R^2 = \text{Cor}(x, y)^2 = \text{Cor}(\widehat{y}, y)^2$$

(admis)

Test de significativité ★★★

Test $H_0 : \beta = 0$ contre $H_1 : \beta \neq 0$

Soit $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ un échantillon. Si la droite de régression est $y = \alpha + \beta y$, pour tester $H_0 : \beta = 0$ contre $H_1 : \beta \neq 0$, la statistique de test est

$$F = (n-2)\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = (n-2)\frac{r_{xy}^2}{1 - r_{xy}^2}$$

Si $H_0 : \beta = 0$ est vraie, $F \sim \mathcal{F}(1; n-2)$. Et donc

- on rejette H_0 si $f > F_{1,n-2}^{-1}(1 - a/2)$

où F_{ν_1,ν_2} est la fonction de répartition de la loi de Fisher $\mathcal{F}(\nu_1, \nu_2)$

Note on fait un test de type $H_0 : \text{Corr}(X, Y) = 0$, cf partie 11

Test de significativité ★★★

On regroupe souvent ces informations dans une tableau dit d'analyse de la variance

source de variation	somme des carrés	degrés de liberté	moyenne des carrés	F
régression	SCE	1	MCE$=\dfrac{\text{SCE}}{1}$	$\dfrac{\text{MCE}}{\text{MCR}}$
résidus	SCR	$n-2$	MCR$=\dfrac{\text{SCR}}{n-2}$	
total	SCT	$n-1$		

$$F=\frac{\text{MCE}}{\text{MCR}}=\frac{\text{SCE}/1}{\text{SCR}/(n-2)}=(n-2)\cdot\frac{\text{SCE}}{\text{SCR}}=(n-2)\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

Test de significativité ★★★

source de variation	somme des carrés	degrés de liberté	moyenne des carrés	F
régression	SCE	1	$\text{MCE}=\dfrac{\text{SCE}}{1}$	$\dfrac{\text{MCE}}{\text{MCR}}$
résidus	SCR	$n-2$	$\text{MCR}=\dfrac{\text{SCR}}{n-2}$	
total	SCT	$n-1$		

```
> anova(model)
Analysis of Variance Table

Response: weight
           Df Sum Sq Mean Sq F value    Pr(>F)
height      1  21001 21000.9  290.35 < 2.2e-16 ***
Residuals 198  14321    72.3
---
Signif.: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Test de significativité ★★★

On peut noter que la statistique de test F

$$F = \frac{\text{MCE}}{\text{MCR}} = \frac{\text{SCE}/1}{\text{SCR}/(n-2)} = (n-2)\cdot\frac{\text{SCE}}{\text{SCR}} = (n-2)\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

```
> anova(model)$F
[1] 290.353      NA
```

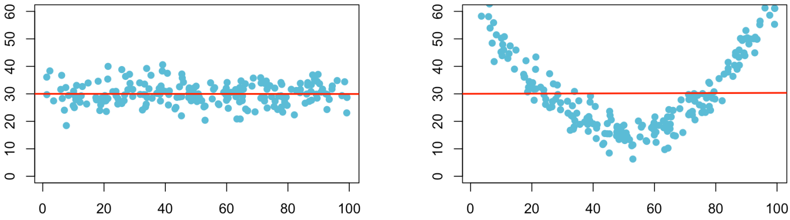
est le carré de la statistique T

$$T = \frac{\widehat{\beta}}{s_{\widehat{\beta}}} \text{ où } s_{\widehat{\beta}} = \sqrt{\frac{\frac{1}{n-2}\sum_{i=1}^n \widehat{\varepsilon}_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

```
> summary(model)$coefficients["height",3]^2
[1] 290.353
```

Test de significativité ★★★

La notion de significativité est à comprendre au sens où il existe une relation linéaire significative entre X et Y



Pour ces deux modèles, x n'est pas significatif

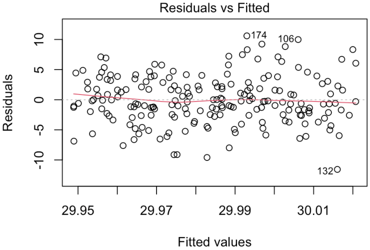
```
Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) 30.017474   2.213545  13.561   <2e-16 ***
x            0.003498   0.037963   0.092    0.927
```

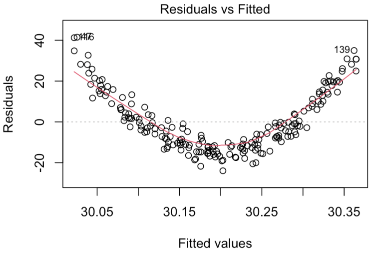
(même sortie de régression dans les deux cas)

Visualisation des résidus

On peut visualiser $(x_i, \widehat{\varepsilon}_i)$ ou $(\widehat{y}_i, \widehat{\varepsilon}_i)$ (comme le propose R)

```
> plot(model)
```



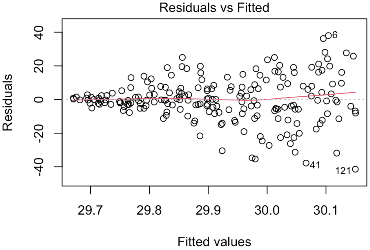


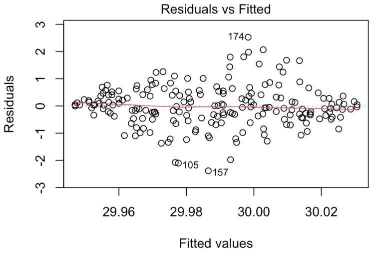
- avec un graphique convenable (conforme aux hypothèses du modèle linéaire)
- avec un graphique qui montre que le modèle linéaire n'est pas approprié (ici quadratique, ou linéaire par morceaux)

Visualisation des résidus

On peut visualiser $(x_i, \widehat{\varepsilon}_i)$ ou $(\widehat{y}_i, \widehat{\varepsilon}_i)$ (comme le propose R)

```
> plot(model)
```





- avec un graphique qui montre que la variance des $\widehat{\varepsilon}_i$ croît avec $\widehat{y}_i$ (ou x_i)
- avec un graphique qui montre que la variance des $\widehat{\varepsilon}_i$ n'est pas constante (hétéroscédastique)

Confidence et prédiction

Intervalle de prédiction

Soit $(\boldsymbol{x}, \boldsymbol{y}) = \{(x_1, y_1), \cdots, (x_n, y_n)\}$ un échantillon. On dispose d'une nouvelle observation x_{n+1}. L'intervalle de confiance de la valeur moyenne prédite est:

$$\hat{y}_{n+1} \pm t_{1-a/2,n-2} \sqrt{s^2 \left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right]}$$

L'intervalle de confiance pour une valeur particulière est:

$$\hat{y}_{n+1} \pm t_{1-a/2,n-2} \sqrt{s^2 \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right]}$$

Confidence et prédiction

$$\hat{y}_{n+1} \pm t_{1-a/2,n-2} \sqrt{s^2 \left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]}$$

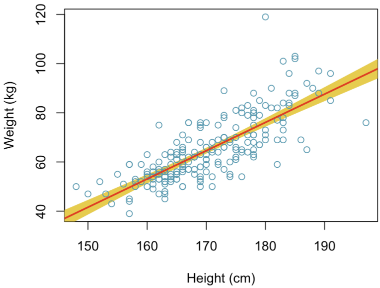
```
> model = lm(weight~height, data=Davis)
> nouvdonnee = data.frame(height=c(170,192))
> predict(model, newdata = nouvdonnee , interval = ’confidence’)
       fit      lwr      upr
1 64.60520 63.41691 65.79349
2 89.90722 86.81755 92.99689
```

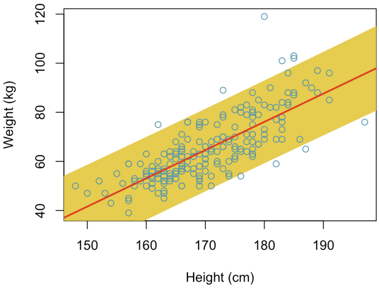
$$\hat{y}_{n+1} \pm t_{1-a/2,n-2} \sqrt{s^2 \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right]}$$

```
> predict(model, newdata = nouvdonnee , interval = ’prediction’)
       fit      lwr       upr
1 64.60520 47.79186  81.41853
2 89.90722 72.85371 106.96073
```

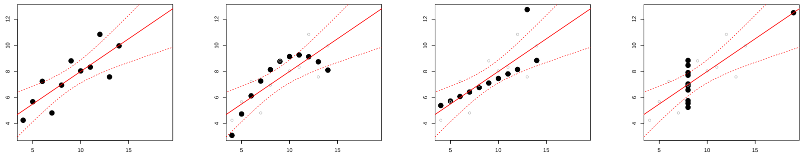
Confidence et prédiction

Les intervalles à 95% pour $\widehat{y}$ et y sont respectivement





Anscombe's Quartet



```
> library(datasets)
> summary(lm(y1~x1,data=anscombe))
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0000     1.1247   2.667  0.02573 *
x1            0.5000     0.1179   4.241  0.00217 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'

Residual standard error: 1.237 on 9 degrees of freedom
Multiple R-squared: 0.6665, Adjusted R-squared: 0.6295
F-statistic: 17.99 on 1 and 9 DF,  p-value: 0.00217
```