Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2020

OLS #4 (régression sur une variable continue - 3)

Poids & Taille

```
> str(Davis)
'data.frame': 200 obs. of  5 variables:
 $ sex       : Factor w/ 2 levels "F","M": 2 1  ...
 $ weight    : int  77 58 53 68 59 76 76 69 71  ...
 $ height    : int  182 161 161 177 157 170 167 ...
> X = Davis$height
> Y = Davis$weight
> (B1hat = cor(X,Y) * sd(Y)/sd(X))
[1] 1.150092
> (B0hat = mean(Y) - B1hat*mean(X))
[1] -130.9104
```

$$y_i = \underbrace{-130.91}_{\hat{\beta}_0} + \underbrace{1.15}_{\hat{\beta}_0} x_i + \widehat{\varepsilon}_i$$

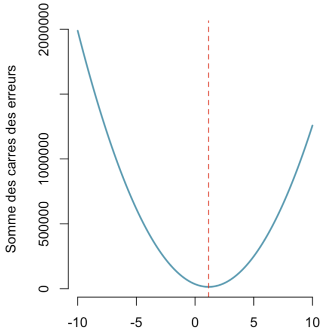
```
> plot(X,Y)
> abline(B0hat, B1hat)
```

Poids & Taille

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} = \text{corr}(\mathbf{x}, \mathbf{y}) \cdot \frac{s_y}{s_x}$$

and $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

```
> (B1hat = cor(X,Y) * sd(Y)/sd(X))
[1] 1.150092
> (B0hat = mean(Y) - B1hat*mean(X))
[1] -130.9104
> SCR = function(B){sum(((Y-mean(Y)) -B*(X-mean(X)))^2)}
> optim(0,SCR)$par
[1] 1.15
> x = seq(-10,10,length=251)
> y = Vectorize(SCR)(x)
> plot(x,y)
```



Note: $\hat{\beta}_1 = \text{argmin}\left\{\sum_{i=1}^{n}\Big((y_i - \bar{y}) - \beta_1(x_i - \bar{x})\Big)^2\right\}$

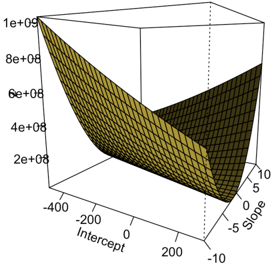
Poids & Taille

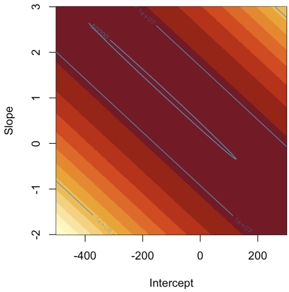
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \text{corr}(\mathbf{x}, \mathbf{y}) \cdot \frac{s_y}{s_x}$$

and $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$

```
> (B1hat = cor(X,Y) * sd(Y)/sd(X))
[1] 1.150092
> (B0hat = mean(Y) - B1hat*mean(X))
[1] -130.9104
> SCR2 = function(B){sum((Y-B[1] -B
    [2]*X)^2)}
> optim(c(0,0),SCR2)
$par
[1] -130.96193     1.15037

$value
[1] 14321.11
```





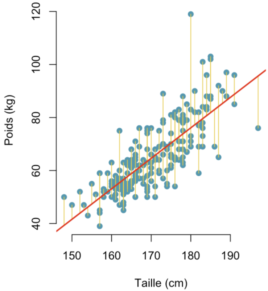
Poids & Taille

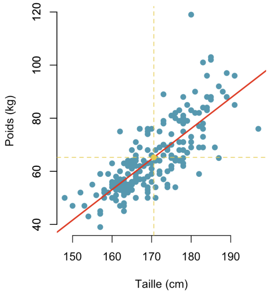
$$\widehat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i \text{ and } \widehat{\varepsilon_i} = y_i - \widehat{y_i}$$

```
> Y_hat = B0hat + B1hat*X
> E_hat = Y-Y_hat
> sum(E_hat)
[1] -2.046363e-12
> sum(E_hat*X)
[1] -3.517471e-10
> B0hat + B1hat*mean(X)
[1] 65.255
> mean(Y)
[1] 65.255
```

i.e. $\overline{y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{x}$

$$\sum_{i=1}^{n} \widehat{\varepsilon_i} = 0 \text{ and } \sum_{i=1}^{n} x_i \widehat{\varepsilon_i} = 0$$





$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\substack{\text{total sum of squares}\\ TSS}} = \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\substack{\text{residual sum of squares}\\ RSS}} + \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}_{\substack{\text{explained sum of squares}\\ ESS}}$$

```
> (TSS=sum( (Y-mean(Y))^2 ))
[1] 35322
> RSS=sum( E_hat^2 )
> ESS=sum( (Y_hat-mean(Y))^2 )
> ESS+RSS
[1] 35322
```

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = \widehat{\beta}_1^2 \frac{s_x}{s_y} = \frac{s_{xy}^2}{s_x s_y} = \text{corr}(\mathbf{x},\mathbf{y})^2$$

```
> (R2 = 1-RSS/TSS)
[1] 0.5945555
```

Poids & Taille

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

```
> (s = sqrt(sum(E_hat^2)/(length(Y)-2)))
[1] 8.504635
```

$$\widehat{\text{Var}}[\widehat{\beta}_0] = \widehat{\sigma}^2\left(\frac{1}{n} + \frac{\overline{x}^2}{s_x^2}\right) \text{ and } \widehat{\text{Var}}[\widehat{\beta}_1] = \frac{\widehat{\sigma}^2}{s_x^2}$$

```
> (s1 = sqrt(s^2/sum((X-mean(X))^2)))
[1] 0.06749465
> (s0 = sqrt(s^2*(1/length(Y)+mean(X)^2/sum((X-mean(X))^2))))
[1] 11.52792
```

To test $H_0 : \beta_1 = 0$ against $H_1 : \beta_0 \neq 0$

$$T = \frac{\widehat{\beta}_1 - 0}{\sqrt{\widehat{\text{Var}}[\widehat{\beta}_1]}} \sim \mathcal{S}td(n-2) \text{ if } H_0 \text{ is true}$$

```
> (t0 = B0hat/s0)
[1] -11.35594
> (t1 = B1hat/s1)
[1] 17.03975
```

or

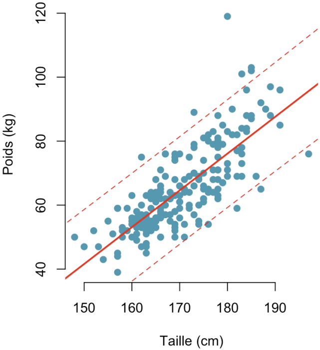
$$F = \frac{(TSS - RSS)/1}{RSS/(n-2)} = \frac{ESS}{RSS/(n-2)} \sim \mathcal{F}_{1,n-2} \text{ if } H_0 \text{ is true}$$

```
> (F = ESS*(length(Y)-2)/RSS)
[1] 290.353
```

For prediction, $\widehat{y}_x = \widehat{\beta}_0 + \widehat{\beta}_1 x$ and

$$\sqrt{\widehat{\mathrm{Var}}[y_x - \widehat{y}_x]} = \widehat{\sigma} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x - \overline{x})^2}{s_x^2}}$$

```
> pred = function(x){
+   yx = B0hat + B1hat*x
+   sd_yx = s*sqrt(1+1/length(Y)+(x-
    mean(X))^2/sum((X-mean(X))^2))
+   c(yx+qt(.025,length(Y)-2)*sd_yx,
+     yx,
+     yx+qt(.975,length(Y)-2)*sd_yx)
    }
> pred(170)
   lower     pred    upper
47.79186 64.60520 81.41853
```



Using R Functions

To fit a model, use

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

Coefficients:
              Estimate Std. Error t value   Pr(>t)
(Intercept) -130.91040   11.52792  -11.36   <2e-16 ***
height         1.15009    0.06749   17.04   <2e-16 ***

Residual standard error: 8.505 on 198 DF
Multiple R-squared: 0.5946, Adjusted R-squared: 0.5925
F-statistic: 290.4 on 1 and 198 DF, p-value: < 2.2e-16
```

and for the prediction

```
> predict(model,newdata=data.frame(height=170),
    interval = "prediction",level = .95)
      fit      lwr      upr
1 64.6052 47.79186 81.41853
```