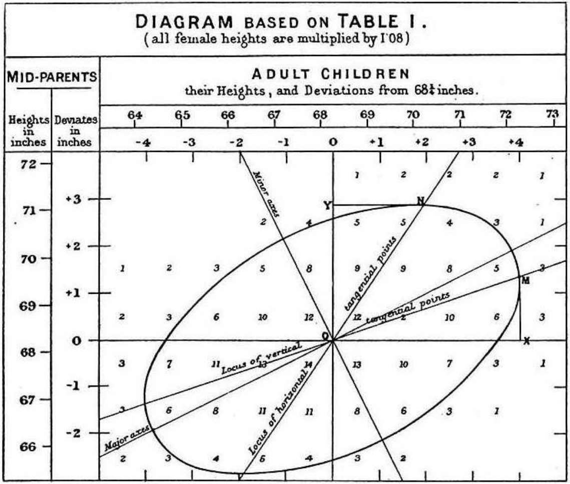
Introduction to data science & artificial intelligence (INF7100)

Arthur Charpentier

#321 (Simple) Regression

été 2020

Linear Regression

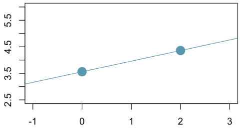


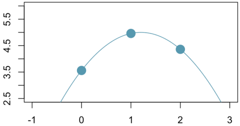
Galton regression towards mediocrity in hereditary stature, 1886.

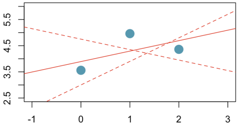
Regression

$\{(x_i, y_i)\}$ for $i = 1, \cdots, n$ $y_i = \alpha + \beta x_i + \varepsilon_i$

- y is the variable of interest
- x is the explanatory variable







For Ordinary Least Squares, solve

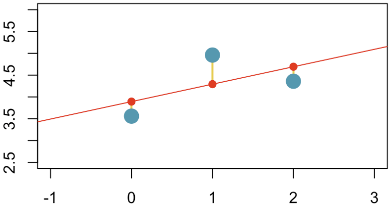
$$\min_{\alpha,\beta}\left\{\sum_{i=1}^{n}\varepsilon_i^2\right\} = \min_{\alpha,\beta}\left\{\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2\right\}$$

Regression

$\{(x_i, y_i)\}$ for $i = 1, \cdots, n$

$y_i = \alpha + \beta x_i + \varepsilon_i$

- y is the variable of interest
- x is the explanatory variable



For Ordinary Least Squares, solve

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

Solutions are (see #411)

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

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = r_{xy}\frac{s_y}{s_x}.$$

Significant Explanatory Variable

With centered and scaled variables, $\widehat{\beta} = r_{xy}$,

$$\frac{\widehat{y} - \bar{y}}{s_y} = r_{xy} \frac{x - \bar{x}}{s_x}.$$

$$t = \frac{\widehat{\beta} - \beta}{s_{\widehat{\beta}}} \simeq \mathcal{N}(0,1),$$

where

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

Significant: $H_0 : \beta = 0$? use $t = \dfrac{\widehat{\beta}}{s_{\widehat{\beta}}} \simeq \mathcal{N}(0,1)$,

Regression

```
> 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
```

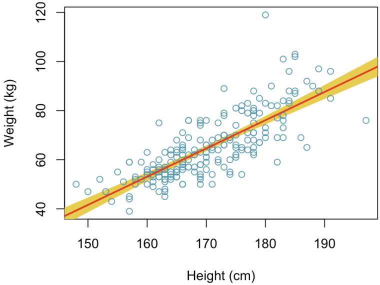
Regression

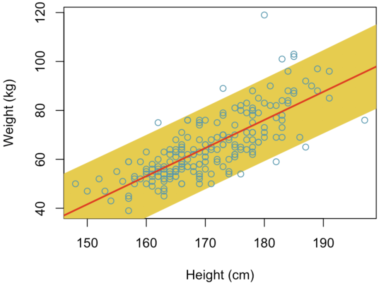
```
> df = data.frame(x=c(5, 15, 25, 35, 45, 55),
                  y=c(5, 20, 14, 32, 22, 38))
> model = lm(y~x, data=df)
> summary(model)

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
```

Confidence ?

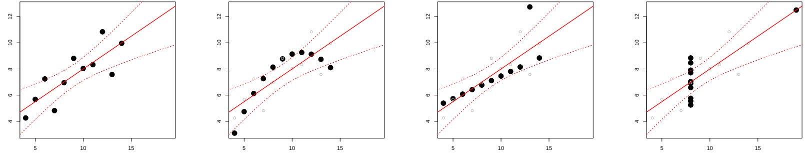
$$\widehat{y} = \widehat{\alpha} + \widehat{\beta}x \text{ and } y = \widehat{\alpha} + \widehat{\beta}x + \widehat{\varepsilon}$$

We can get the 95% confidence band for $\widehat{y}$ and y





Anscombe's Quartet



```
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
```