Introduction to linear regression

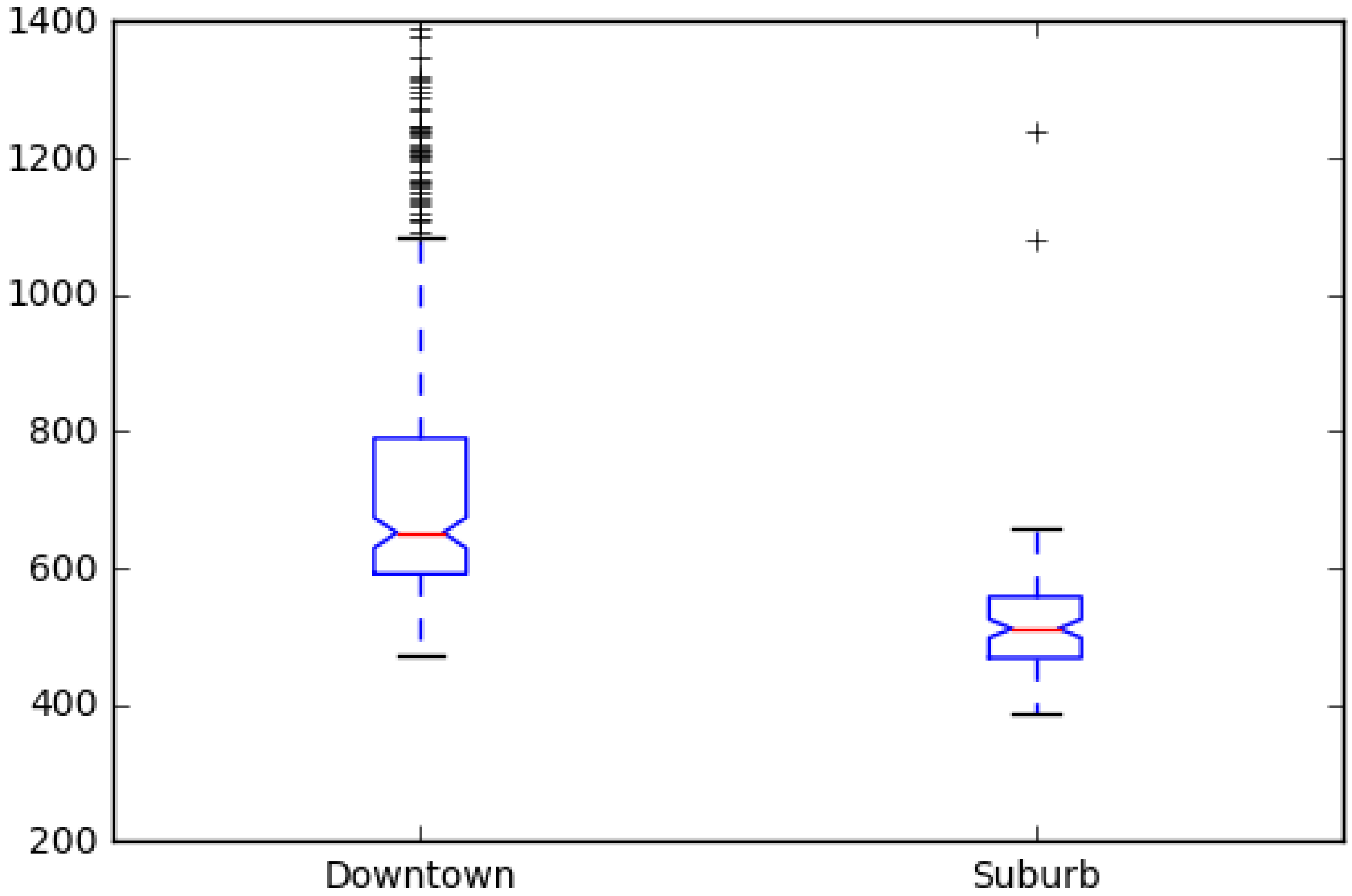
Agoston Torok

Multivariate statistics

Take an example

- House rent prices are higher in the centre of the city
- Is it linear?

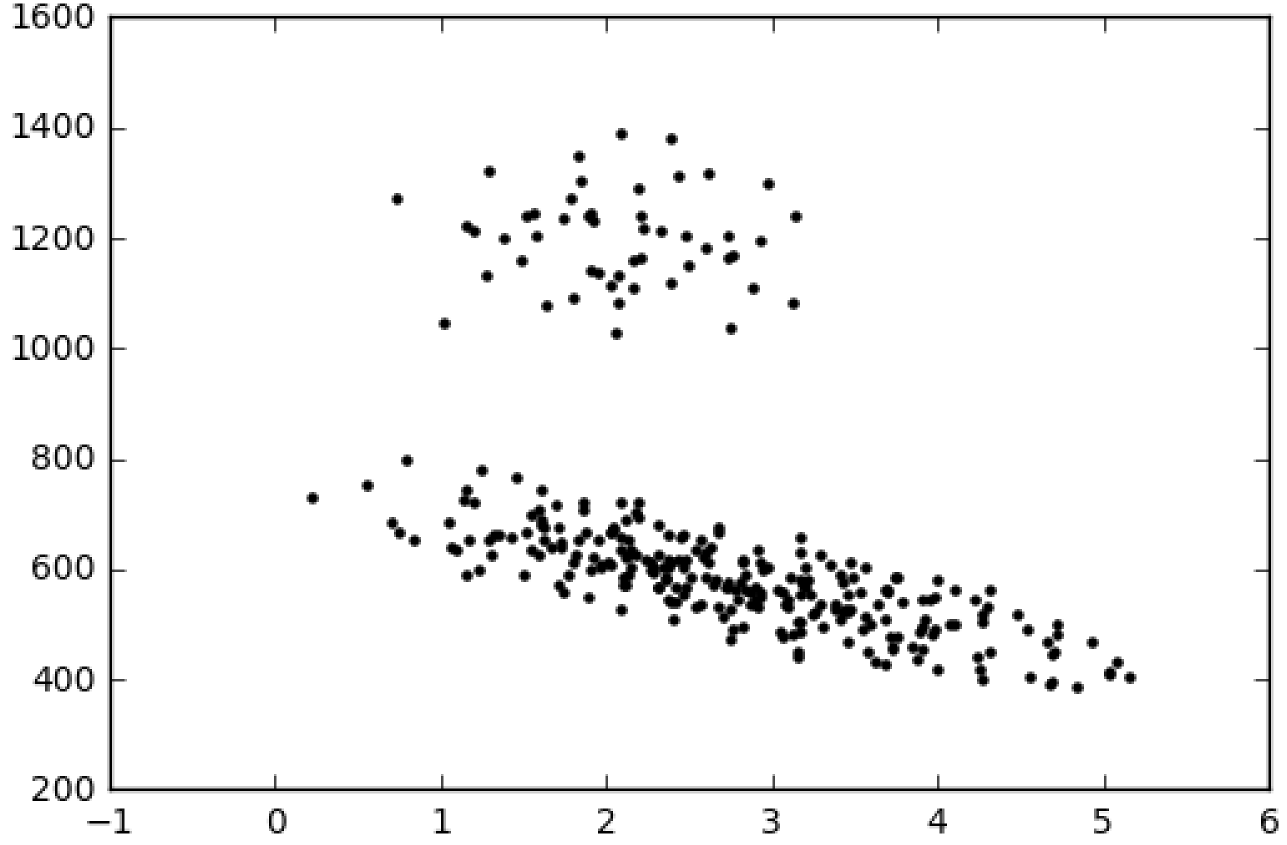
Show this on boxplots



The ANOVA would show a significant difference

Let’s see the highest resolution of the data

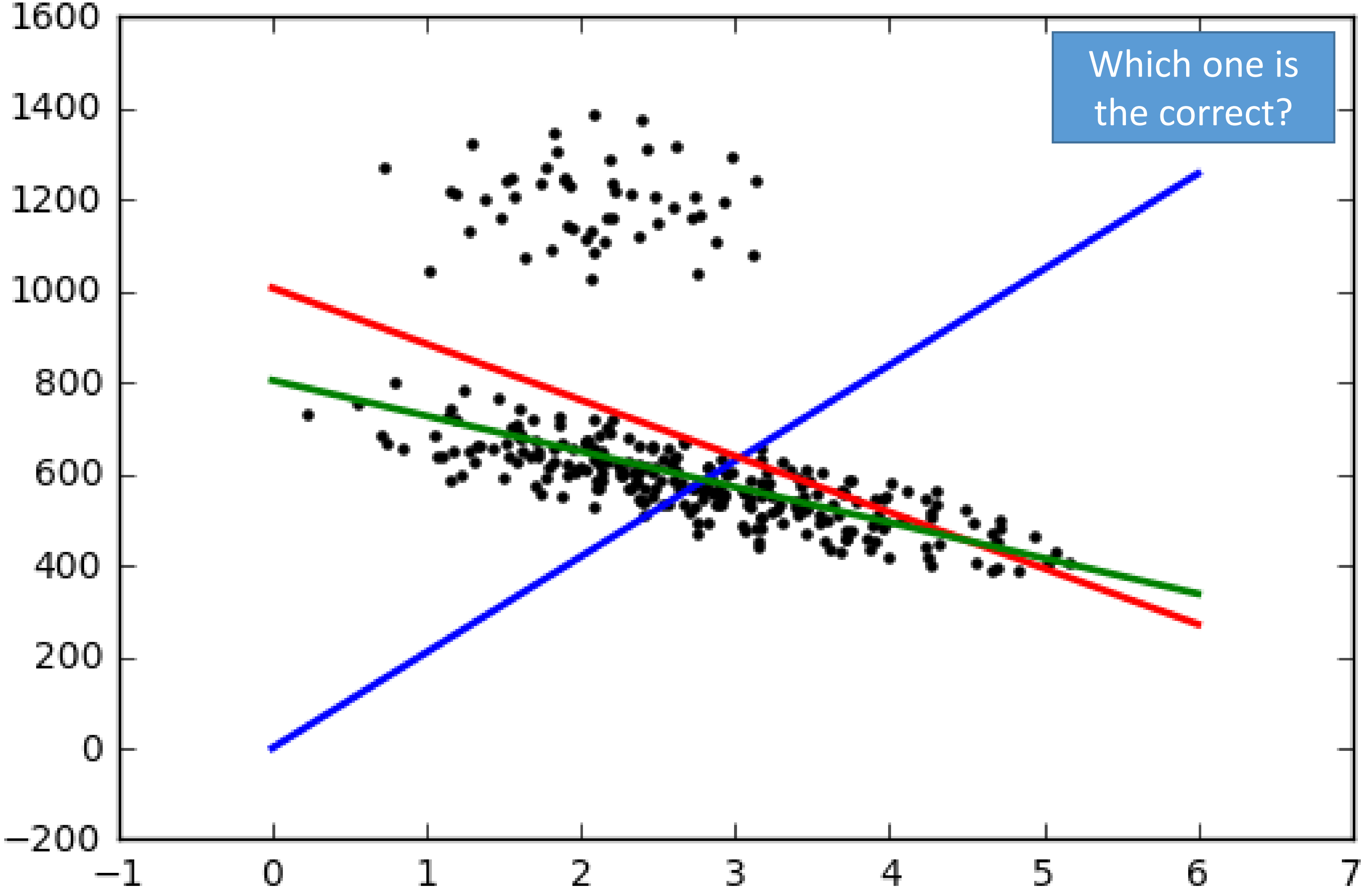
- Labeling values of a scalar variable → loss of information



Things to do before regression

- Centering the predictors
- Scale predictors if you want to compare there effects easily
- Transformation of predictors if needed
- Maybe centering the dependent variable (no intercept needed)
- Check data (outliers, informativeness)

Three solutions

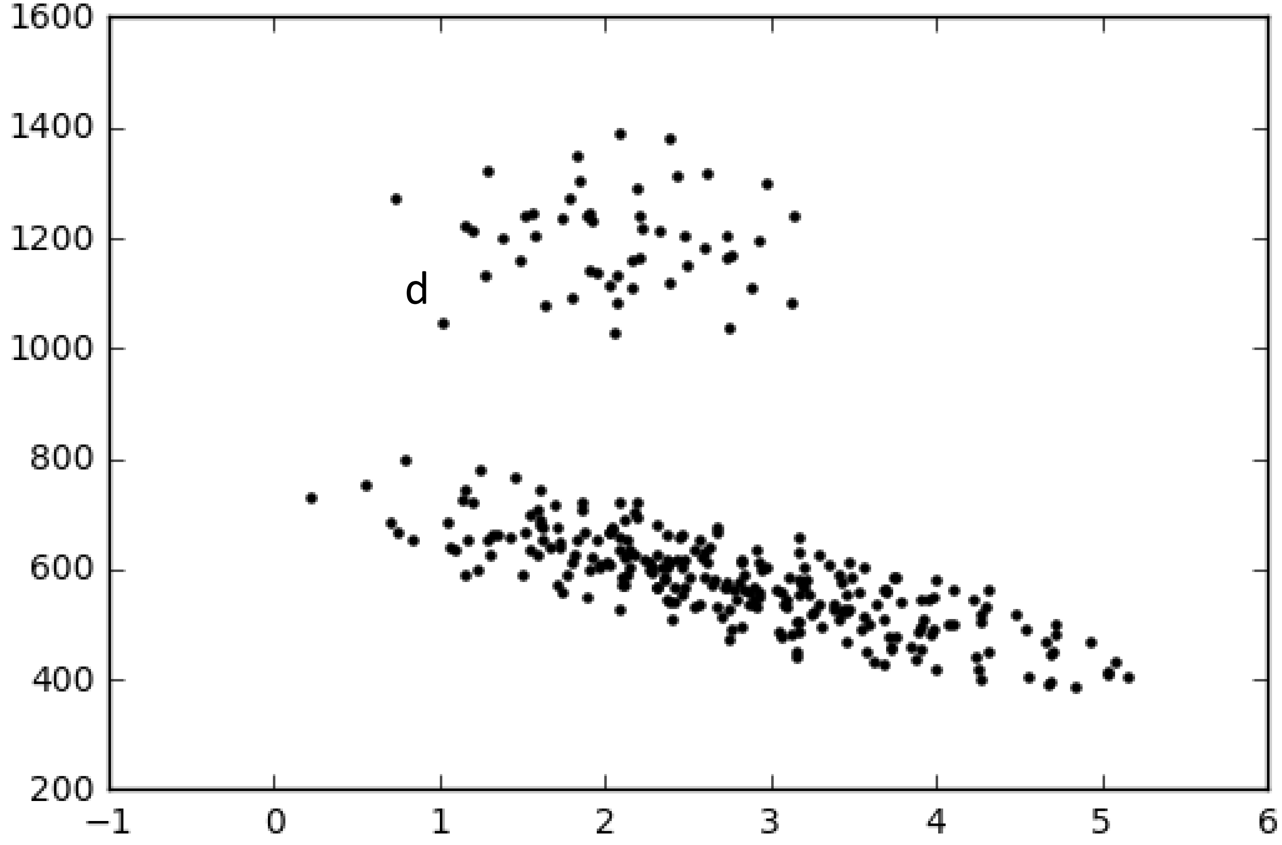


$y = \beta_1 x + \epsilon \mid O \text{ is } LS$

$y = \beta_0 + \beta_1 x + \epsilon \mid O \text{ is } LS$

$y = \beta_0 + \beta_1 x + \epsilon \mid O \text{ is } LS + r$

Calculation of the slope - OLS

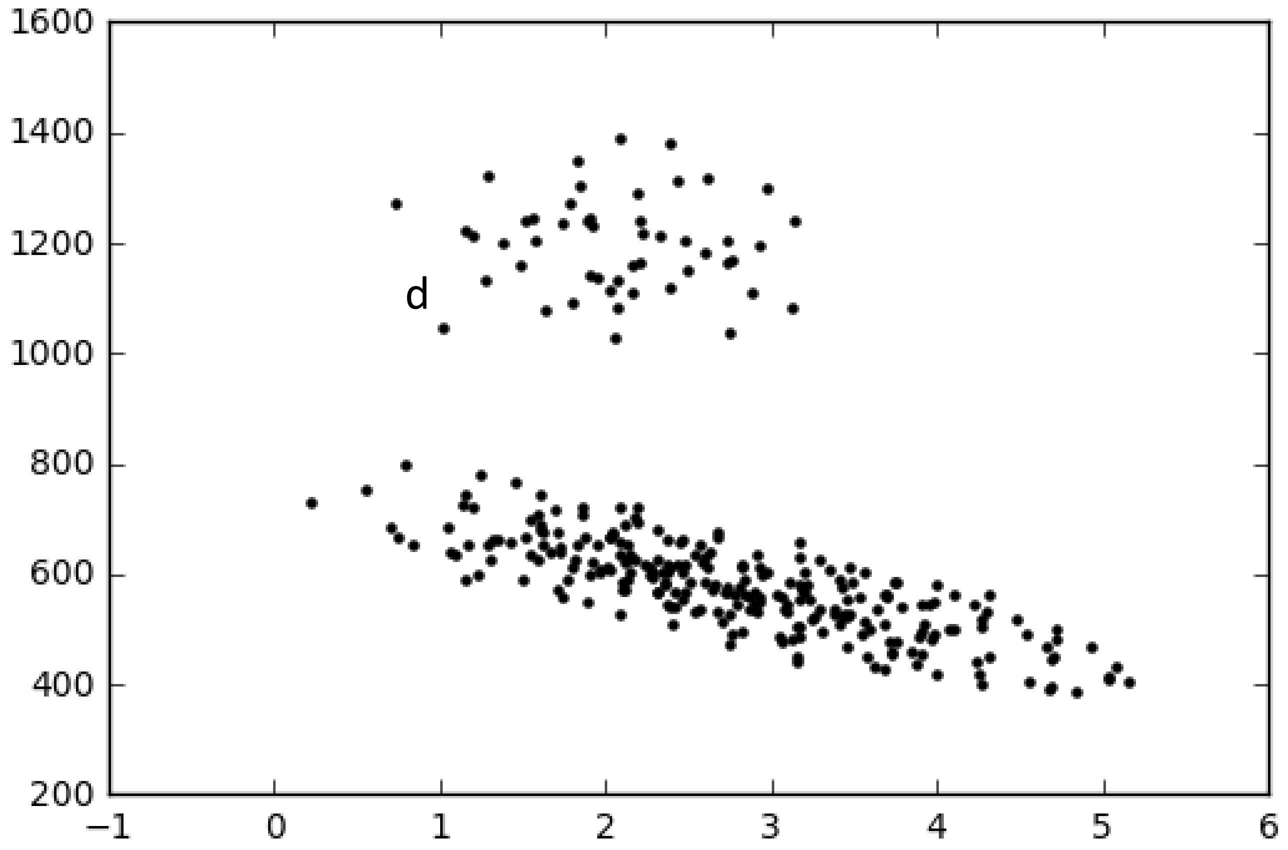


When the model is fitted each y hat will be one point on the regression line. The difference *d* between the real y and the predicted y (y hat) can be defined as:

$$d = \sqrt{(\hat{y} - y)^2}$$

Remember, we use the square root and the square only because we are interested in the magnitude of the difference

Calculation of the slope - OLS

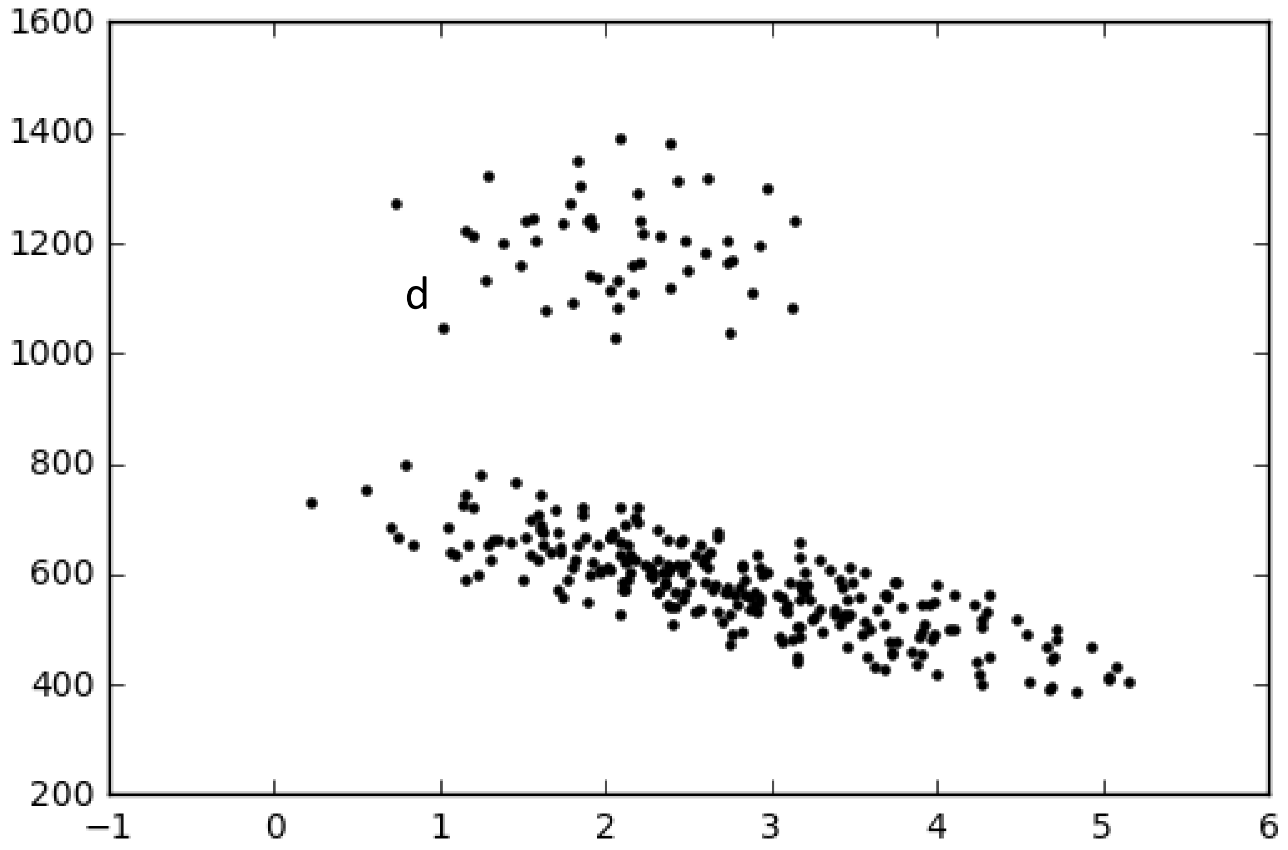


We can define the Sum of squares for every possible line we can draw by adding the differences between predicted and real values:

$$SS = \sum_{i=1}^{n} \sqrt{(\hat{y}_i - y_i)^2}$$

The goal is to find the one line that has the smallest SS, this is why we talk about Ordinary Least Squares (OLS) regression

Calculation of the slope - OLS

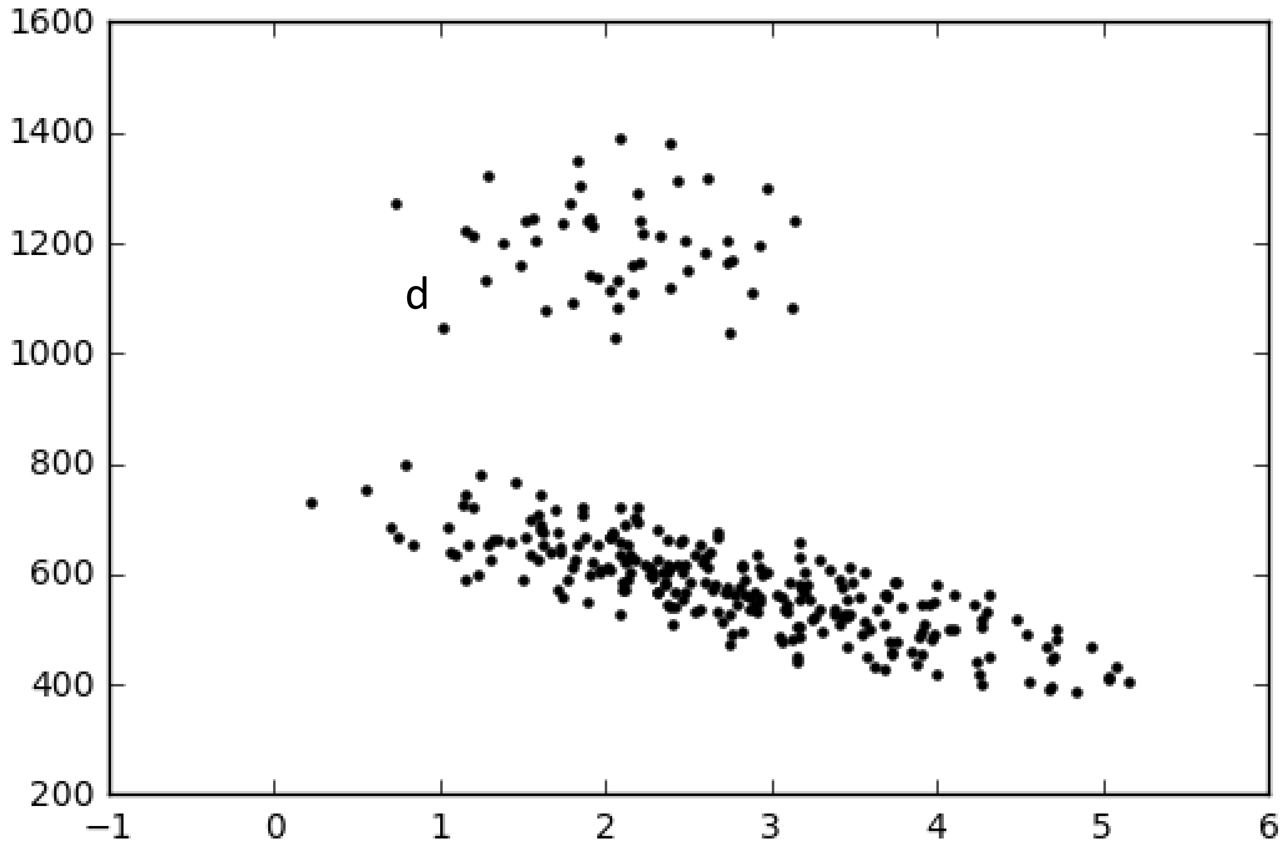


We calculate the predicted value by taking the Beta zero (the intercept) and adding the Beta one multiplied by predictor value:

$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

We are looking for the values of both Betas that minimize the SS.

Calculation of the slope - OLS



In the simple case we can do (the line over x and y means the average of the variable):

$$\beta_1 = \frac{Covariance\ (x,y)}{Variance\ (x)}$$

$$Covariance(x, y) = \sum (x - \overline{x})(y - \overline{y})$$

$$Variance(x) = \sum (x - \overline{x})^2$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

In more complex cases (multiple LR) we use e.g. gradient descent, MLE

Practice I

$$d = \sqrt{(\hat{y} - y)^2} \qquad \hat{y}_i = \beta_0 + \beta_1 x_i \qquad SS = \sum_{i=1}^{n} \sqrt{(\hat{y}_i - y_i)^2}$$

For example if $\beta_0 = 1,\ \beta_1 = 0.3, x = [\ 1,2,3]$, y = [2, 2.4, 3.4] then what is the SS?

$\hat{y}_1 = 1 + 0.3\ * 1 = 1.3$;
$\hat{y}_2$ = 1 + 0.3 * 2 = 1.6
$\hat{y}_3$ = 1 + 0.3 * 3 = 1.9

SS = $\sqrt{(1.3 - 2)^2}$ + $\sqrt{(1.6 - 2.4)^2}$ + $\sqrt{(\ 1.9\ - 3.4)^2}$ = 0.7 + 0.8 + 1.5 = 3

Practice II

$$\beta_1 = \frac{Covariance\ (x,y)}{Variance\ (x,y)}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$Covariance(x, y) = \sum (x - \overline{x})(y - \overline{y})$$

$$Variance(x) = \sum (x - \overline{x})^2$$

Was the latter the solution with the minimum square for the task? $x = [\ 1,2,3]$, y = [2, 2.4, 3.4]

$Covariance(x, y) =$ (1-2)(2- 2.6) + (2-2)(2.4- 2.6) + (3-2)(3.4- 2.6) = 0.6 + 0 + 0.8 = 1.4

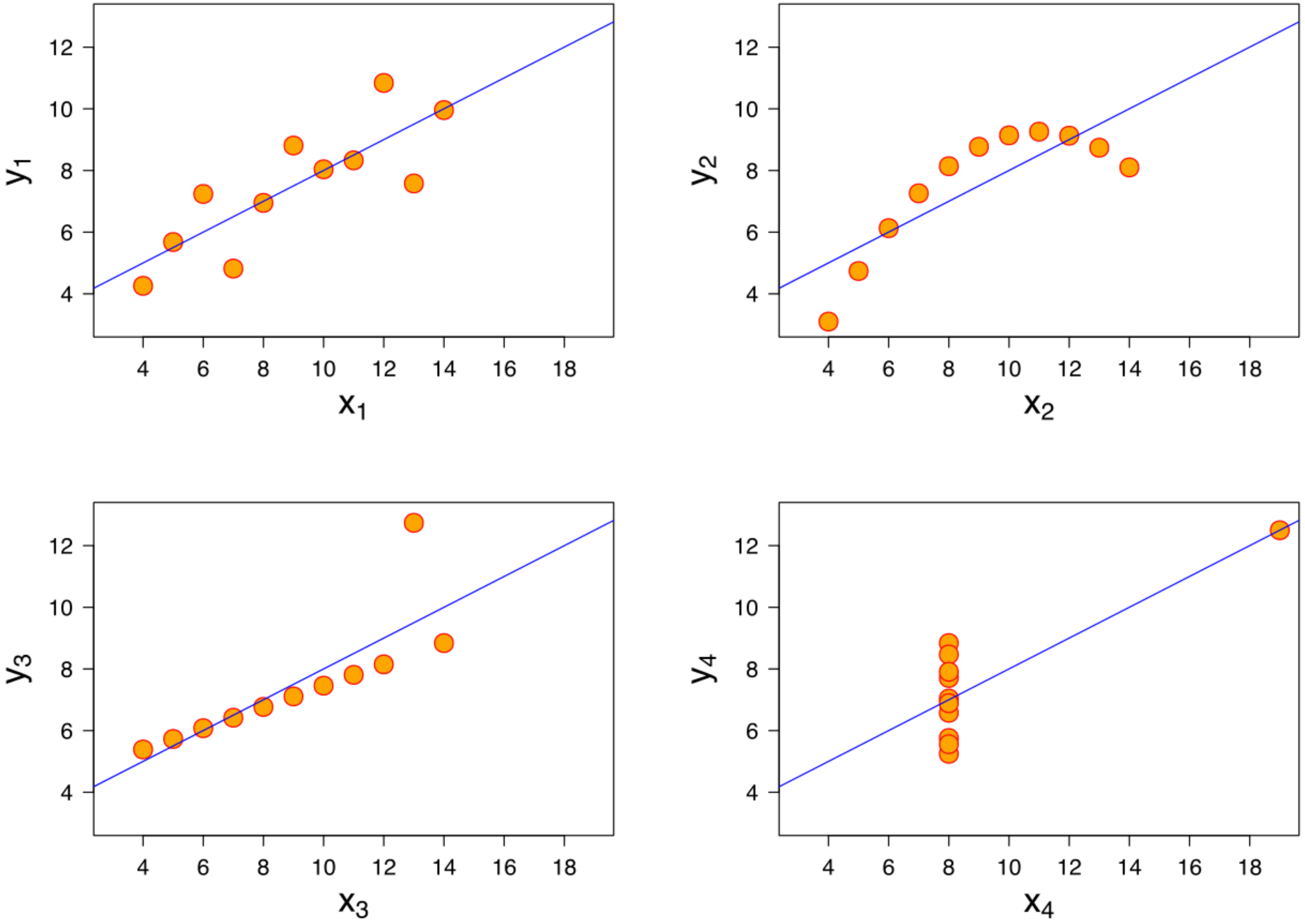
V$ariance(x, y) =$ (1-2)(1-2) + (2-2)(2-2) + (3-2)(3-2) = 1 + 0 + 1 = 2

$\beta_1 =$ 1.4 / 2 = 0.7

$\beta_0 = 2.6 - 0.7\ * 2 = 1.2$

Anscombe’s quartet

- Francis Anscombe:
  - Plotting before fitting
  - Outliers
  - Know your data



Next

- Assumptions
- Regularization
- Bayesian linear regression