Data Science

Modelling with Regression

Alexandra Posekany

Summerschool 2023

Linear Regression - simple univariate model

(linear) regression models the dependence between

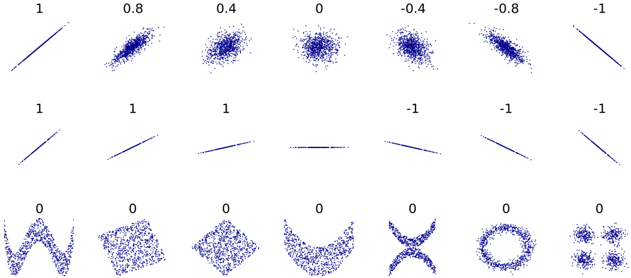
- a **dependent** numeric variable, **regressand** Y, and
- one or more **independent** explanatory numeric variables, **regressor(s)** X, $\boldsymbol{X}$

Mathematically, the simple linear regression model is

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

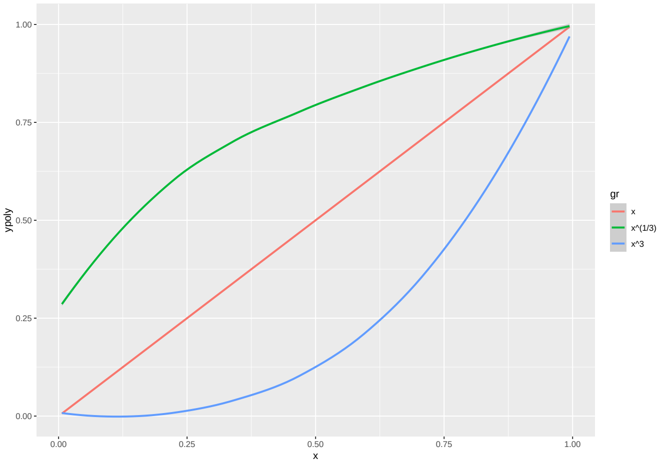
- α and β are unknown parameters of the population
- ε_i are iid errors with mean 0 and a common unknown variance σ^2 (no heteroscedasticity).

Visualising Correlation



Root and Polynomials

```
ggplot(dfpoly,aes(x=x,y=ypoly,col=gr))   + geom_smooth(meth
## `geom_smooth()` using formula = 'y ~ x'
```



Non-linear Transformations

- E_i are a **exponential** transformation of data X_i, if

$$E_i = exp(X_i)$$

- L_i are a **logarithmic** transformation of data X_i, if

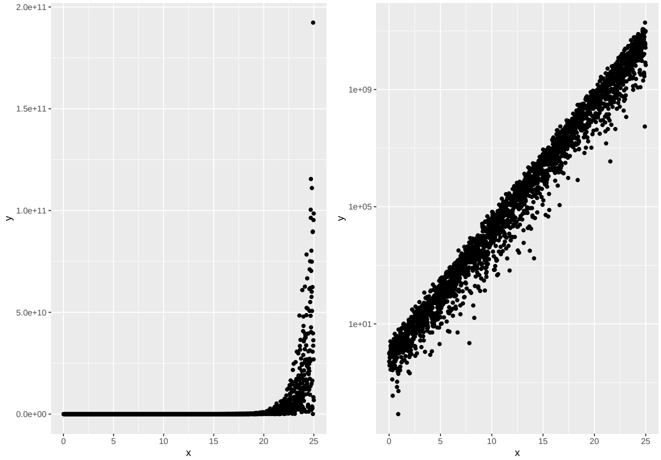
$$L_i = log(X_i)$$

These two transformations form the bridge between the class of exponential models

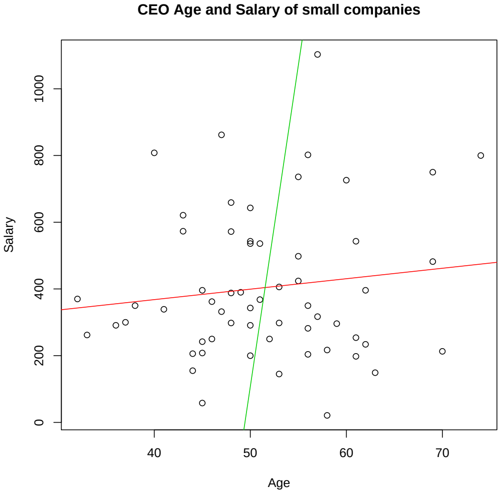
$$Y_i = C \cdot exp(\beta \boldsymbol{X}_i) \cdot \epsilon_i$$

and linear models

Non-linear Transformations of Exponential function to linear

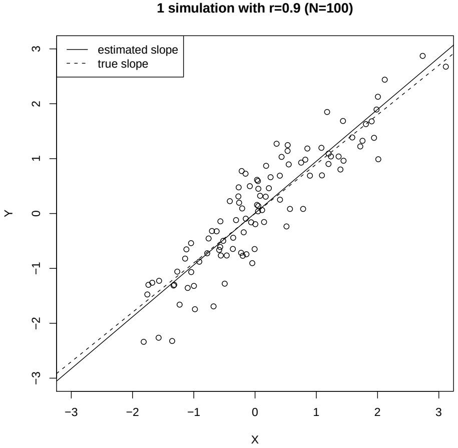


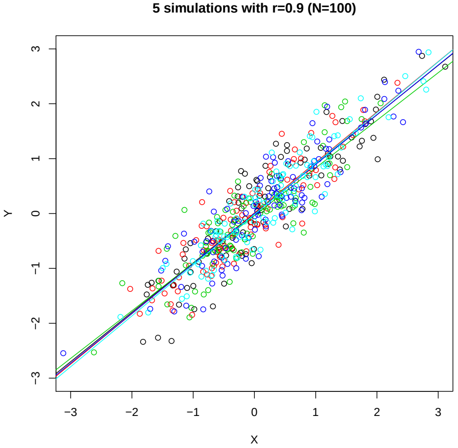
Which line is the “right’ ’ line?



Different Regression of Y onto X (red) and X onto Y (green).

Let's talk about the precision of $\hat{\beta}$

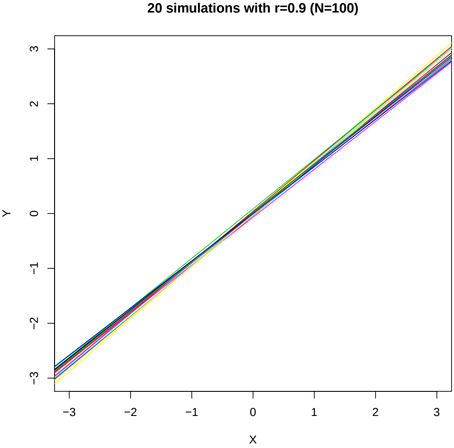


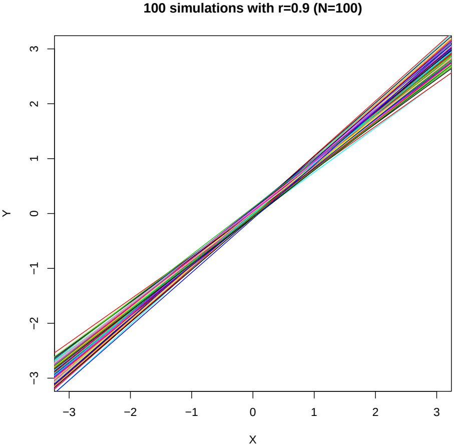


$$0.9448 \leq \hat{\beta} \leq 0.9448$$

$$0.8677 \leq \hat{\beta} \leq 0.9448$$

Let's talk about the precision of $\hat{\beta}$

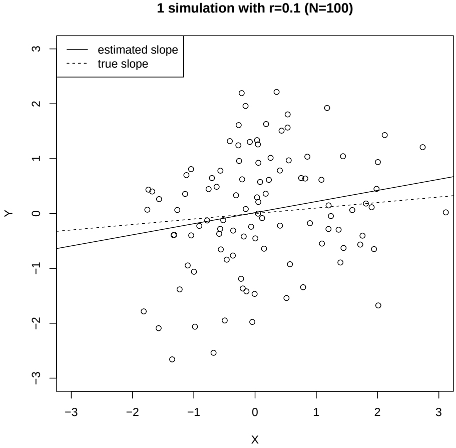


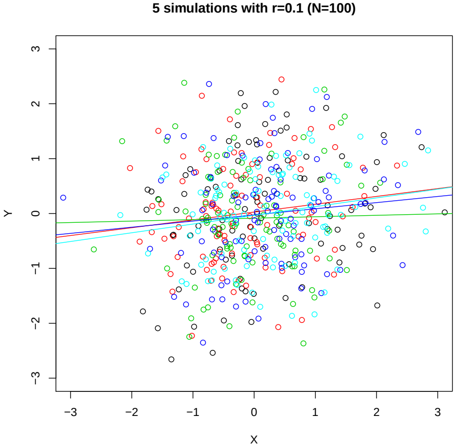


$0.8588 \leq \hat{\beta} \leq 0.9582$

$0.7874 \leq \hat{\beta} \leq 1.0089$

Let's talk about the precision of $\hat{\beta}$

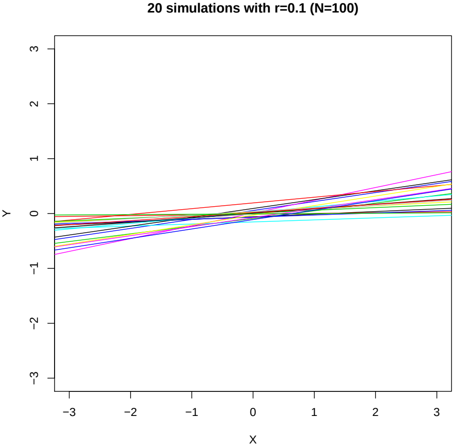


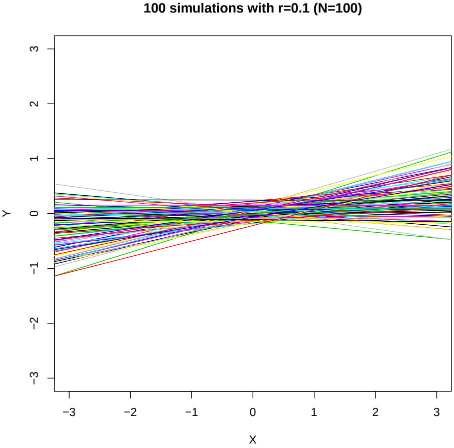


$$0.2022 \leq \hat{\beta} \leq 0.2022$$

$$0.0263 \leq \hat{\beta} \leq 0.2022$$

Let's talk about the precision of $\hat{\beta}$

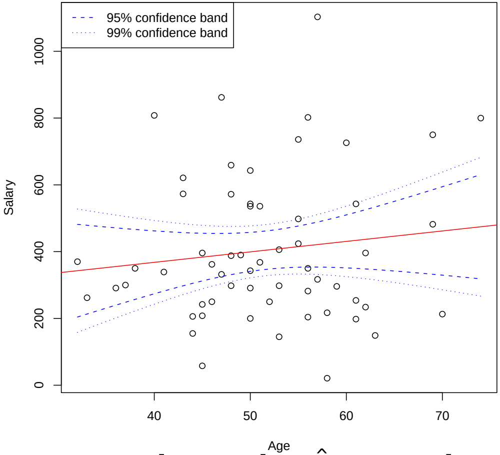




$$0.0061 \leq \hat{\beta} \leq 0.2329$$

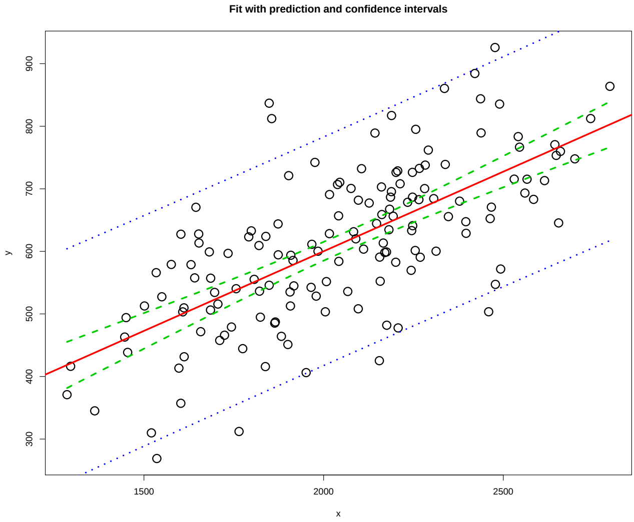
$$-0.1569 \leq \hat{\beta} \leq 0.3486$$

CEO regression with confidence bands



$\hat{\alpha} = 242.702$ [168.760], $\hat{\beta} = 3.133$ [3.226].

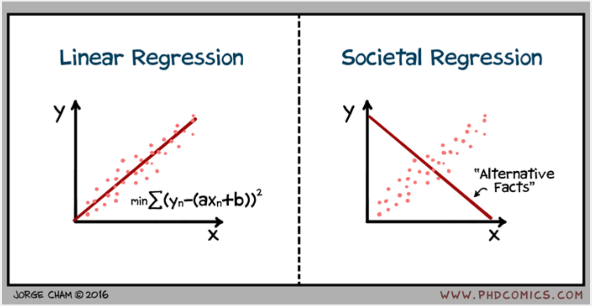
Confidence and Prediction bands



Leverage

Leverage

Leverage points are observations made at extreme or outlying values of the independent variables $\boldsymbol{X}$ which therefore have large influence on the slope of the regression line β.



Linear Regression - multiple model

Mathematically, the simple linear regression model is

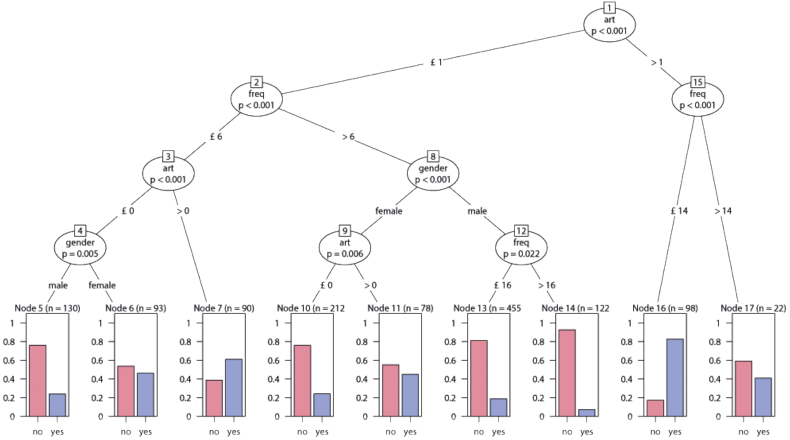
$$y_i = \alpha + \beta_1 x_{1,i} + \ldots + \beta_k x_{k,i} + \varepsilon_i$$

in the notation of vectors and matrices this model corresponds to

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

- μ and the α_i are unknown parameters of the population
- ϵ_{ij} are iid errors with mean 0 and a common unknown variance σ^2 (no heteroscedasticity).
- in case of multivariate X, the columns of $x_{k,.}$ have to be stochastically independent

Tree based method



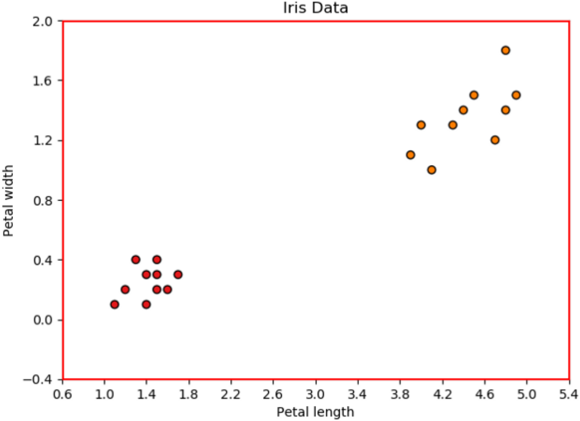
Machine Learning and types of Learning

Creating models based on data has two main goals: - learning relations between the variables in the models and their structure - predicting future data based on previous one

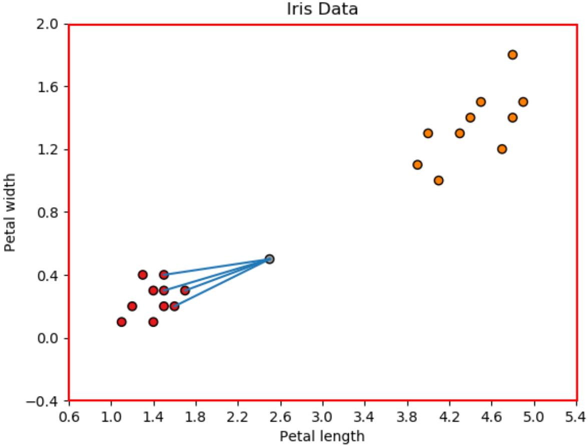
In **Machine Learning** these approaches are often split dependent on the amount of knowledge and data available on the process you wish to learn about or predict:

- *supervised learning* (all outcomes are already know for training the algorithm)
- *semi-supervised learning* (some outcomes are already know for training the algorithm, other training data or validation data have no outcomes known in advance)
- *unsupervised learning* (what is to be learned be the algorithm is not available as previous data, because it is unknown, unmeasurable etc.)

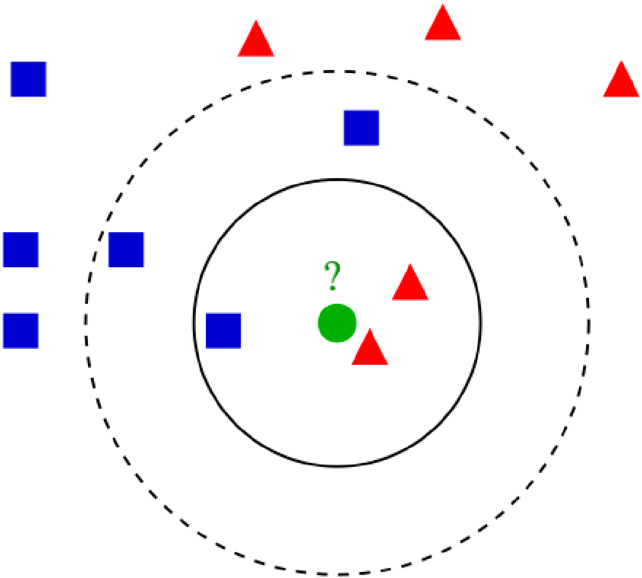
Concepts of Classification



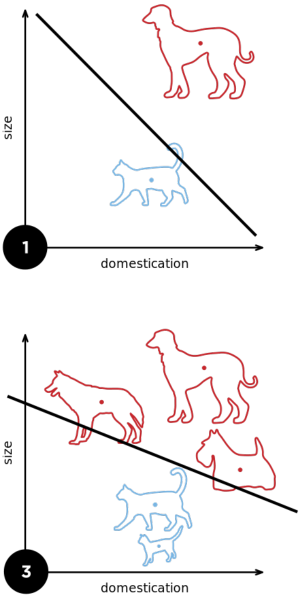
Concepts of Classification



k Nearest Neighbours



Perceptron = Single-Layer aNN



k means Clustering