MVE055 2018 Lecture 13

Marco Longfils

Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg

Monday 15th October, 2018

Regression

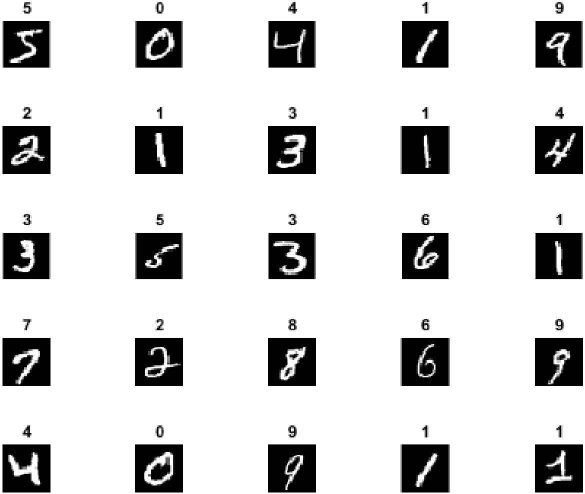
- Often we measure two or more quantitative variables on the same individuals/trials. Then, it could be useful to model some variables (dependent variables) in terms of other variables (independent variables) independent variables, for example to make predictions. This type of models is known as "Regression model".
- Independent variable are called **predictors** and they are **NOT** random variables. They are mathematical variables, i.e. they can assume different values but this is not determined by chance.
- Dependent variables are called **response variables** and they are random variables with a different probability distribution for different values of the predictors.
- Not always is clear what should be the dependent and the independent variables. We are treating here dependent and independent variables asymetrically, i.e. swapping their role will change the analysis.

Regression

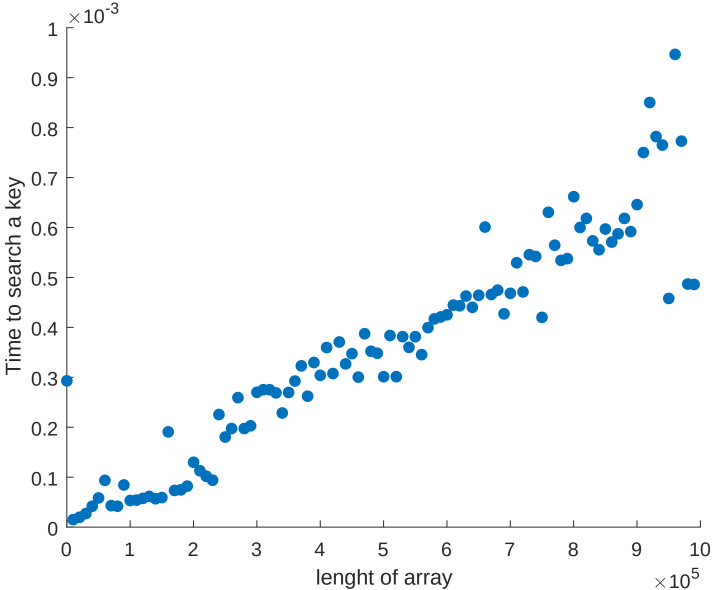
- Regression has a wide range of applications, mostly due to its simplicity;
- In some literature (e.g. Machine Learning), regression is divided into regression (for continuous outcomes) and classification (categorical outcomes).
- It is a form of supervised learning.

Regression

Regression



Regression


×10^-3
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Time to search a key
0
1
2
3
4
5
6
7
8
9
10
lenght of array
×10^5

Linear regression

- We will consider only the case where we have one predictor X and one response variable Y. This is called simple regression as opposed to multiple regression (i.e. more than one predictors).
- We restrict ourselves to case where the expected value for Y given x, denoted by $\mu_{Y|x}$, has the form

$$\mu_{Y|x} = \beta_0 + \beta_1 x. \tag{1}$$

- Before fitting the model, check that:
    - there is good reason to believe that X values could be used to predict Y values;
    - the scatterplot shows that a linear model could work well (if not, consider a transformation of variables);

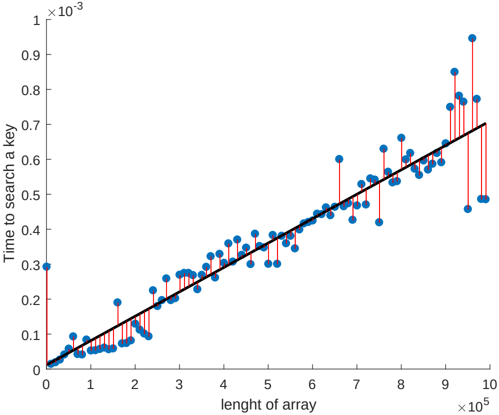
Model and parameter estimation

- Assume we have observed n pairs (x_i, y_i), where x_i is the value of X and y_i is the value of Y.
- The simple linear regression model can be expressed as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \tag{2}$$

where $\epsilon_1, ..., \epsilon_n$ are the residuals, i.e. the difference between y_i and its mean $\beta_0 + \beta_1 x_i$.

- Estimates for β_0 and β_1 can be obtained by minimizing $\text{SSE} = \sum_{i=1}^{n} \epsilon_i^2$ (i.e. Least square fitting).
- SSE can be minimized by computing the partial derivatives wrt β_0, β_1 and setting them to zero.



Estimates for β_0, β_1

Proposition (Least-squares estimates for β_0 and β_1)

$$b_1 = \frac{n\sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i\right)^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}.$$

- In practice/exercises: compute $\sum_{i=1}^n x_i, \sum_{i=1}^n x_i^2, \sum_{i=1}^n y_i, \sum_{i=1}^n x_i y_i$.
- Use the formulas above.

Properties of the least squares estimators

- The values of b_0, b_1 will vary from data set to data set. Thus, they are observed values of random variables denoted respectively by B_0, B_1.
- We are interested in the properties of B_0, B_1 to find confidence intervals and test hypothesis on β_0, β_1.
- We need the following assumptions
  1. The random variables $Y_1, \ldots, Y_n$ are independent and normally distributed.
  2. The mean of Y_i is $\beta_0 + \beta_1 x_i$.
  3. The variance of Y_i is σ^2.
- The above assumptions are equivalent to

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2).$$

Distribution of the estimator

- An unbiased estimator for σ^2 is

$$s^2 = \frac{\text{SSE}}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2} \tag{3}$$

- Under the previous assumptions we have

$$B_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}\right)$$

$$B_0 \sim N\left(\beta_0, \frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2} \sigma^2\right)$$

$$\frac{(n-2)s^2}{\sigma^2} \sim \chi^2(n-2)$$

Confidence interval for β_1

- Suppose we would like to test whether $H_0 : \beta_1 = \beta_1^0$ against $H_1 : \beta_1 \neq \beta_1^0$.
- The most common case is $\beta_1^0 = 0$, which tests if there is no linear relationship between X and Y
- It can be proved that

$$T_{n-2} = \frac{B_1 - \beta_1^0}{s/\sqrt{S_{xx}}},$$

  where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$, then T_{n-2} has a T distribution with $n-2$ degrees of freedom.
- $\rightarrow$ We reject the null hypothesis if the statistic T_{n-2} is outside the usual interval $\left[-t_{\frac{\alpha}{2}}, t_{\frac{\alpha}{2}}\right]$
- The confidence interval for β_1 has the form

$$B_1 \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{S_{xx}}}$$

Confidence interval for β_0

- Suppose we would like to test whether $H_0 : \beta_0 = \beta_0^0$ against $H_1 : \beta_0 \neq \beta_0^0$.
- It can be proved that

$$T_{n-2} = \frac{B_0 - \beta_0^0}{\frac{s\sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{nS_{xx}}}},$$

  has a T distribution with $n - 2$ degrees of freedom.
- → We reject the null hypothesis if the statistic T_{n-2} is outside the usual interval $\left[-t_{\frac{\alpha}{2}}, t_{\frac{\alpha}{2}}\right]$
- The confidence interval for β_0 has the form

$$B_0 \pm t_{\frac{\alpha}{2}} \frac{s\sqrt{\sum_{i=1}^n x_i^2}}{\sqrt{nS_{xx}}}$$

Confidence interval for $\mu_{Y|x}$ for a new observation

- Suppose we have a new observation x, i.e. we measured x for a new individual, and we want to make inference about $\mu_{Y|x}$;
- We know that an estimator for $\mu_{Y|x}$, i.e. the mean of Y given x, is $\hat{\mu}_{Y|x} = b_0 + b_1 x$
- The confidence interval for $\mu_{Y|x}$ has the form

$$\hat{\mu}_{Y|x} \pm t_{\frac{\alpha}{2}} s \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}$$

where $t_{\frac{\alpha}{2}}$ is the quantile of a T_{n-2} distribution.

Confidence interval for a new observation Y

- Suppose we have a new observation x, i.e. we measured x for a new individual, and we want to make inference about Y;
- We know that an estimator for $Y|x$, the value of Y when $X = x$, is $\hat{Y}|x = b_0 + b_1 x$ (as for $\mu_{Y|x}$)
- The confidence interval for $Y|x$ has the form

$$\hat{Y}|x \pm t_{\frac{\alpha}{2}} s \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}$$

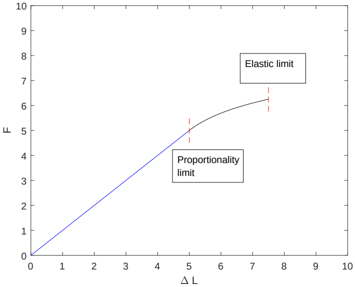
where $t_{\frac{\alpha}{2}}$ is the quantile of a T_{n-2} distribution.

Hooke's law

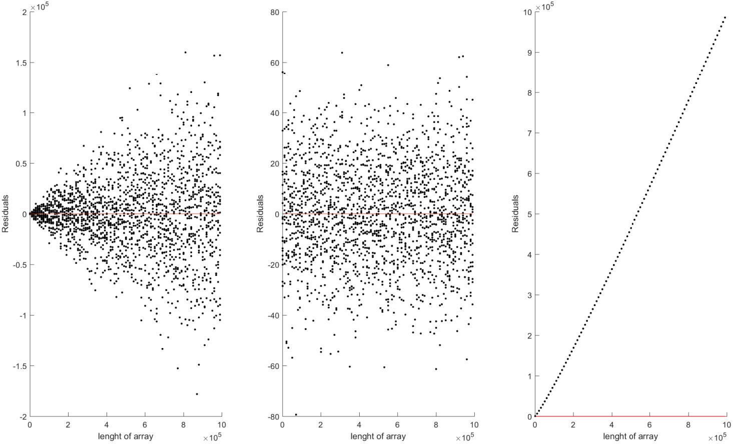
Proposition (Hooke's law)

The extension of a spring or wire is directly proportional to the force applied ***provided*** *the limit of proportionality is not exceeded*

$$F = k\Delta L$$



residual analysis



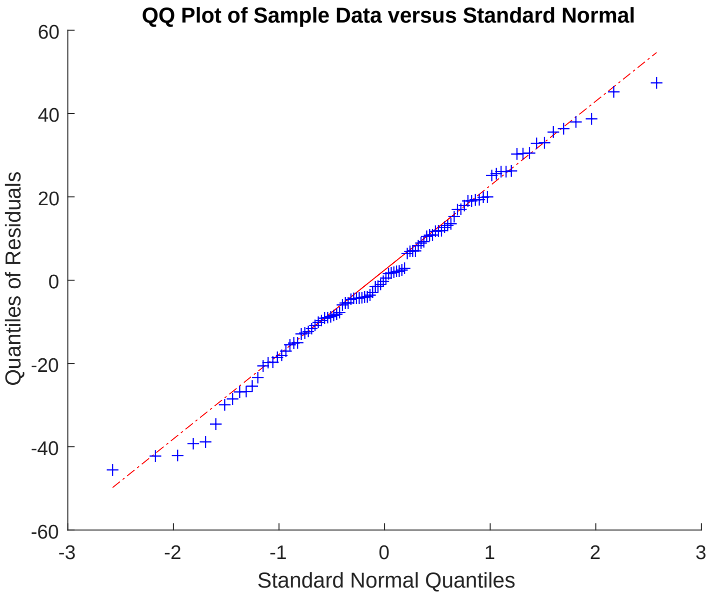
Coefficient of determination

- Let $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ and define

$$R^2 = \frac{S_{yy} - \mathrm{SSE}}{S_{yy}}$$

- R^2 is called coefficient of determination and representn the proportion of the total variation in the response values which is accounted for by the regression model, i.e. being caused by the variation in the independent variables.
- $R^2 \in [0, 1]$.

Probability plot


QQ Plot of Sample Data versus Standard Normal
Quantiles of Residuals
Standard Normal Quantiles