Uppsala Universitet
Matematiska Institutionen
I. Kaj, R. Müller, Y. Wang

Repetition and important formulas from modul 9

Linear regression describes a linear relation between different factors. One factor is the **explanatory/independent** variable x, the other the **response/dependent** variable y.

1 Model

Suppose we are given pairs of observations $(x_1, y_1), \dots, (x_n, y_n)$. The linear relationship between x and y (if it is linear!) is mathematically described as

$$Y_i = m + kx_i + \varepsilon_i$$

with model parameters m (intercept) and k (slope). The residuals $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, \dots, n$, are independent random variables and describe the deviation from a perfect linear line. With that definition the random variables have moments

$$\begin{aligned} \mathrm{E}[Y_i] &= \mathrm{E}[m + kx_i + \varepsilon_i] = m + kx_i + \mathrm{E}[\varepsilon_i] = m + kx_i \\ \mathrm{V}[Y_i] &= \mathrm{V}[m + kx_i + \varepsilon_i] = \mathrm{V}[\varepsilon_i] = \sigma^2, \end{aligned}$$

meaning that $Y_i \sim \mathcal{N}(m + kx_i, \sigma^2)$.
Note that an alternative formulation of the model is

$$Y_i = \tilde{m} + \tilde{k}(x_i - \bar{x}) + \varepsilon_i.$$

The two formulations are equivalent with $\tilde{k} = k$ and $\tilde{m} = m - k\bar{x}$.

2 Parameter estimation

In order to describe the linear relationship, the parameters m and k have to be fitted with the help of the observation pairs. This is done by minimizing the squared vertical difference between the data points $y_i's$ and the straight line ("minsta kvadratkriteriet"), i.e. one calculates

$$\min_{m,k} \sum_{i=1}^{n} (y_i - m - kx_i)^2$$

for example by deriving the above sum of squares. In doing so the estimators are

$$\hat{k} = \frac{S_{xy}}{S_{xx}}, \quad \hat{m} = \bar{y} - \hat{k}\bar{x}, \quad \hat{\sigma}^2 = s^2 = \frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right)$$

with

$$S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2, \quad S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

A measure of how well correlated the variables are is

$$R^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} \in (0,1).$$

Recall and compare it to the previously defined correlation coefficient $r = S_{xy}/\sqrt{S_{xx}S_{yy}}$, i.e. $R^2 = r^2$.

In order to check whether the model is meaningful, that means whether there is actually a linear relationship between the variables, you can ask the following questions:

- Is the variance σ^2 of the residuals ε_i constant?
- Are the residuals independent?
- Do they follow the normal distribution?

There are different ways to check these properties for example by investigating patterns in the QQ plot, scatterplot, ... (compare for example JR 8.3.2).

3 Confidence intervals for the estimators

One can show

$$\mathrm{E}[\hat{m}] = m, \quad \mathrm{E}[\hat{k}] = k, \quad \mathrm{V}[\hat{m}] = \frac{\sigma^2}{n}\frac{1}{S_{xx}}\sum_{i=1}^{n}x_i^2, \quad \mathrm{V}[\hat{k}] = \frac{\sigma^2}{S_{xx}}.$$

The $(1-\alpha)$ confidence interval for k is given by

$$I_k = \left[\hat{k} - t_{\alpha/2}(n-2)\frac{s}{\sqrt{S_{xx}}}, \hat{k} + t_{\alpha/2}(n-2)\frac{s}{\sqrt{S_{xx}}}\right]$$

where $t_{\alpha/2}(n-2)$ is the quantile of the t-distribution with $n-2$ degrees of freedom (you get the values from table 3 in JR for example). If $\alpha = 0.05$ for example and $n = 10$, then $t_{\alpha/2}(n-2) = t_{0.025}(8) = 2.31$.

Note that the interpretation of $k = 0$ is that there is no influence of X on Y.