5 Linear regression

5.1 Linear regression

We are interested in the association between the number of hours of learning per week and the score at the statistics exam. Say we know that the exam result is distributed conditionally normal: $Y|X \sim N(\beta_0 + \beta_1 X, \sigma^2)$.

- Read the parameter estimates from the output below. Interpret the results - which null hypotheses are tested and how?

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-20.683  -4.746   2.844   4.512  14.693

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  19.2049     7.5172   2.555 0.033921 *
x             3.6850     0.6217   5.927 0.000351 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.44 on 8 degrees of freedom
Multiple R-squared: 0.8145,     Adjusted R-squared: 0.7913
F-statistic: 35.13 on 1 and 8 DF,  p-value: 0.0003508
```

- How would the null hypothesis $H_0 : \beta_1 = 0$ be tested with the generalized likelihood ratio test?
  Hint: Whenever possible, use the results of the previous exercise

Understanding the ideas in R:

- Let X be a uniformly distributed variable (between 0 and 20, rounded dow), $\beta_0 = 15$, $\beta_1 = 4$, $\sigma = 10$. Generate a sample of size 10, plot the data using a scatterplot and add the estimated and the population line.

- Find the result of the generalized likelihood ratio test for the simulated example with R

5.2 Linear regression with matrices

The values of the independent variables are united into the matrix X (design matrix), the values of the outcome and the coefficient are represented by the vectors Y and β:

$$X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix}; \; Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}; \; \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

Matrix X is of dimension $n \times (p+1)$, where p is the number of variables. If constant is not included in the model, the first column of X can be omitted.

- Rewrite the sum $\sum_{i=1}^{n} Y_i^2$ in matrix form.
- What do we get with the matrix product $X\beta$?
- The estimate with the least squares method in matrix form is given by $\widehat{\beta} = (X^T X)^{-1} X^T Y$. Show that for $p = 1$ the estimates equal:

$$\widehat{\beta}_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$\widehat{\beta}_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

- Derive the least squares estimate in the matrix form.
  You will need the following matrix algebra formulas:

$$\begin{aligned} (A+B)^T &= A^T + B^T; \;\; (A^T)^T = A; \;\; (AB)^T = B^T A^T; \\ \frac{\partial \beta^T A}{\partial \beta} &= A; \;\; \frac{\partial \beta^T A^T A \beta}{\partial \beta} = 2A^T A \beta \end{aligned}$$

  Hint: Write the expression we are trying to minimize? How is the sum of squared residuals be written in matrix form?

- Derive the formula for the standard error of the estimated coefficients in matrix form. Intuitively explain what the standard error of the coefficient β_1 (for $p = 1$) depends on. Use the rule: $\text{var}(cY) = c\ \text{var}Y c^T$.
- Estimate the confidence interval for the predicted line in our example ($p = 1$)? What will such interval look like?
- Say we are interested in how the number of hours learning and gender (0=female, 1=male) are associated with the exam score. Assume a model that includes the interaction term. How would you check whether the number of hours learning for men is associated with the exam result?

Understanding the ideas in R:

- Write the data of the previous exercise in the matrix form, calculate the least squares estimates and compare them to the output of the R function `lm`.
- Estimate the standard error and compare it with R output
- Plot the predicted line and add a confidence interval.

5.3 Linear regression assumptions

The basic linear regression model has four assumptions:

(a) Residuals are normally distributed around the regression line

(b) The residual variance does not depend on the values of the independent variables (homoscedasticity)

(c) Residuals are independent

(d) The association between X and Y is linear

What happens with the estimates, there expected value, standard error and confidence intervals if any of the first three assumptions is not met?

- What changes in the derivations if the residuals are not normally distributed?
- What happens if the residual variance depends on x?
- What if residuals are not independent?