1 Simple linear regression

Q1.1: Let $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ be n independent observations from a random vector (X, Y). Consider the regression equation $Y = \beta_0 + \beta_1 X$. Show that the the least squares estimators of β_0 and β_1 are, respectively:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}(x_i - \bar{x})^2}.$$

Hint: compute the critical point $(\hat{\beta}_0, \hat{\beta}_1)$ of RSS and prove that it is a minimum:

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2 \quad \rightarrow \quad min.$$

Q1.2: The "Old Faithful" geyser is located at Yellowstone National Park in Wyoming, USA. The name of the geyser comes from the fact that the eruptions follow a relatively stable pattern. The Park service tries to predict when the next eruptions occur and displays this information to the park visitors, so they can arrive at the geyser site in time. The main interest of the study is to understand the eruption patterns and predict the interval time until the next eruption.

The data set contains a sample of 272 inter-eruption times taken during August 1978 and August 1979. The variables are - the index of the date when the observation was taken, *eruptions* (the duration of an eruption of the geyser in minutes), and *waiting* (the time until the next eruption in minutes).

(a) Create a boxplot and a frequency histogram of the inter-eruption time. Can we assume that its distribution is approximately normal?

(b) Make a scatterplot of eruption duration (on Y-axis) vs waiting time (on X-axis). Which additional features of the data can you see on the plot?

(c) Why, to your opinion, shorter inter-eruption times are associated with shorter eruption times and longer inter-eruption times are associated with longer eruption times?

(d) Fit the simple linear regression between duration and waiting time and add the resulting regression line to the scatter plot above. Check the regression assumptions about the residuals and comment on the model fit.

2 Parametric bootstrap for a regression model

Consider a simple linear regression model: $Y = \beta_0 + \beta_1 X$, and let $(x_1, y_1), \dots, (x_n, y_n)$ be the observed data. The inference for regression coefficients β_0 and β_1 performed using confidence intervals is based on the assumption that the model residuals are identically normally distributed with mean zero. However, when the data do not (even approximately) satisfy this assumption, one may want to obtain a more robust version of confidence intervals. This can be achieved, for example, by using a bootstrap procedure. The parametric (model-based) bootstrap is usually applied when we believe that the considered model structure is true. There is a number of R packages which perform parametric bootstrap for various statistical models, but in this seminar, you will try to write your own R script implementing the parametric bootstrap for a simple linear regression model. The algorithm proceeds as follows:

1. Specify a (typically large) size N of bootstrap sequences for β_0 and β_1.

2. Fit the model $Y = \beta_0 + \beta_1 X$ to the given data and collect the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ (they will be the first elements of the corresponding bootstrap sequences). Under the model,

   $$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

   where $\epsilon_i = y_i - \hat{y}_i$ are the estimated residuals.

3. To carry out a bootstrap step, do:

   - Take a sample (with replacement) of n elements from the set of estimated residuals $\{\epsilon_1, \ldots, \epsilon_n\}$. Denote the sampled residuals, say, by ϵ_i^* and compute the new response values:

     $$y_i^* = \hat{y}_i + \epsilon_i^*.$$

   - Form the new data set, $(x_1, y_1^*), \ldots, (x_n, y_n^*)$, and fit the linear regression model to these data.
   - Collect the new estimates $\hat{\beta}_0^*$ and $\hat{\beta}_1^*$, and attach them to the corresponding bootstrap sequences.

4. Repeat the procedure N times. The quantiles resulting sampling distributions for β_0 and β_1 can be used to compute the desired confidence intervals.

Q2.1: Consider the subset of Old Faithful data set corresponding to the *eruption duration of less than 3 minutes*, fit the simple linear regression between duration and waiting time to these data, and implement parametric bootstrap to obtain a 95% confidence interval for each regression coefficient.

Hint: before performing the simulations, fix a random seed in R. It is useful if you want to get reproducible simulation results.