Assignment 2 (R portion)

ST2137-2420

Introduction

This assignment covers topics 7 to 9. The questions on this pdf correspond to the R portion. The dataset can be found on Canvas.

For this R portion, you are not allowed to use any additional packages other than lattice.

Leukaemia Survival Times

The dataset `leuk_surv_times.csv` contains information on the survival times of 17 leukaemia patients. The columns are :

- `logWBC`: the base-10 logarithm of the patient's white blood cell count. Let this be X_1.
- `surv_time`: the number of weeks until the patient passed away. We denote this by Y, the response variable.

It is postulated that the following model is appropriate for the survival times, where $i = 1, 2, \ldots, 17$.

$$Y_i = \beta_0 \exp\{\beta_1(X_{1,i} - \bar{X_1})\}\epsilon_i$$

However, this is not linear in the coefficients. Taking *natural* logarithms, we obtain:

$$Y_i' = \ln Y_i = \beta_0' + \beta_1 W_{1,i} + \epsilon_i' \tag{1}$$

where:

- $\beta_0' = \ln \beta_0$.
- $W_{1,i} = X_{1,i} - \bar{X_1}$.

In this question, we are going to first fit the above model. Next, we shall deal with an influential point by creating an indicator column that identifies it. By adding this column to the model, we shall be able to estimate an "effect" for this outlier, while improving the model for the remaining points. We will not have to delete the outlier.

Answer the following questions in your R script:

1. Read the data into R as a dataframe named `leuk`, and create two new columns: `lnY` and `w`, that contain Y' and W_1 respectively.
2. Fit the model in equation (1) to the data. Extract the *adjusted* R^2 and store it in an R vector (of length 1) named `model_1_r2`.
3. Use `influence.measures()` to identify the point that has the greatest influence on the estimate of β_1. Suppose that this is point k. Add a column named `outlier` to the dataframe `leuk` that contains a 0 for all rows except row k. This row should have the value 1 for the column `outlier`.
4. Fit the following model to the data:

$$Y_i' = \beta_0' + \beta_1 W_{1,i} + \beta_2 I(W_{2,i} = 1) + \epsilon_i'$$

where W_2 corresponds to the column `outlier`. Extract the adjusted R^2 and store it in a vector (of length 1) named `model_2_r2`.

5. Create the following plot using the improved model, which depicts the 80% confidence interval for *non-outlier values*, but on the *original* Y and X-scale. Hint: The `lines()` function is useful for adding lines to an existing plot.