Tutorial 10:

Introduction to Data Science

Fit and Overfit

Ernst C. Wit

November 25, 2024

Exercises

1. **Fitting Mayer's Lunar Data**. Over the period from April 1748 to March 1749, Mayer made numerous observations of the positions of several prominent lunar features; and in his 1750 memoir he showed how these data could be used to determine various characteristics of the moon's orbit, and ultimately to the longitude.

   His analysis let to the equation

   $$\beta - (90 - h) = \alpha \sin(g - k) - \gamma \cos(g - k),$$

   where h, g and k can be measured via observations, and $\gamma = \alpha \sin \theta$. Consider the data `mayer.csv` — also available as `Mayer` in the R-package `HistData`, which contains for 27 observations, $i = 1, \ldots, 27$:

   - `Y`: $-(90 - h_i)$
   - `X2`: $\sin(g_i - k_i)$
   - `X3`: $\cos(g_i - k_i)$

   (a) Last week we used the function for least squares (e.g., `lm` in R) to obtain the estimated values for β, α and γ. Reconsider the output from this regression and interpret the p-values associated with the parameters and the R-squared value.

   (b) Perform a diagnostic analysis to evaluate the following aspects:

      i. Linearity
      ii. Normality
      iii. Constant variance
      iv. Outlying or influential observations

   (c) Perform model selection using the stepwise AIC criterion.

2. **Overfitting Galton's height data.** Last week we saw how Galton in 1886 studied the the question of the relation between heights of parents and their offspring. We present the original dataset involving the heights and gender from 934 individuals as well as the heights of their parents. However, we also added 20 noise variables to the dataset. The dataset is available as `galton2.csv` on iCorsi.

   (a) Consider the above data and perform an exploratory analysis.

(b) Let's assume that Galton had no idea a priori which of the 23 variables was important. Consider the following model:

$$\text{childHeight} = \beta_0 + \beta_1\text{father} + \beta_2\text{mother} + \beta_3\text{gender} + \sum_{j=1}^{20} \gamma_j x_j + E.$$

Use *least squares* to fit the model to the data.

(c) Consider the output and evaluate according to the p-values which covariates seem to be important in predicting the height of a child.

(d) Use stepwise BIC criterion to select the most likely true model.