Tutorial 10

ST2137-2420

Material

This tutorial covers concepts on linear regression, corresponding to chapter 9 from the textbook.

Crabs dataset

The file `crab.txt` contains data on female horseshoe crabs. The columns in the data are :

- Colour, recorded as an integer (values from 1 – 4),
- Spine condition (1 = both good, 2 = one worn or broken, 3 = both worn or broken).
- Width of the carapace in cm,
- Number of satellites a female crab has,
- Weight of the crab in in grams, and
- Whether a female crab has a satellite (1=yes, 0=no).

Here are the questions for this data:

1. Fit the following model to the dataset.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 I(X_2 = 2) + \beta_3 I(X_2 = 3) + \beta_4 X_1 I(X_2 = 2) + \beta_5 X_1 I(X_2 = 3) + e$$

   where

   - Y: weight
   - X_1: width
   - X_2: spine

2. Write code that will extract the following quantities:

   - The p-value for the t-test for $H_0 : \beta_0 = 0$.
   - The estimate of β_4.
   - The residual sum of squares.
   - The estimate of σ^2.
   - The adjusted R^2 value.

3. Compute the prediction for width = 27, spine = 1.

Cars dataset

The `Cars93` dataset from the `MASS` package in R contains information on 93 cars. The description of the 27 columns can be found in `?Cars93`. Consider the two columns `MPG.city` and `Cylinders`. Both of these variables are numeric. Please skim through the help page to understand the dataset before proceeding.

4. Create a boxplot of `MPG.city` (y-axis) vs `Cylinders` (x-axis), comment on the distribution of values.
5. First, fit an ANOVA model to this data to assess if there is any significant difference in mileage between the three groups. Then, fit a simple linear regression model of `MPG.city` versus `Cylinders`

(as integers). Inspect the two outputs. What is the difference between these two models? Ignoring the assumptions that need to be met, which is the "correct" one to use?

Population dataset

The dataset in `sg_population.csv` was obtained from the Singapore Department of Statistics. It contains information on the resident population in Singapore since 1950. Population growth is sometimes modeled with the following equation:

$$P = \frac{K}{1 + ae^{-bX}}$$

where:

- P is the population in year X
- a and b are constants to be estimated.
- K is a constant that defines the maximum possible population.

6. Read the data into R and Python and plot it.
7. Prove that the equation can be re-written as

$$\underbrace{\ln\left(\frac{K}{P} - 1\right)}_{Y} = \ln a - bX$$

8. Using $K = 7 \times 10^6$, fit a simple linear regression model and report the estimates of a and b for this data.
9. Use the model to estimate the mean population until year 2050 and plot it, along with the original data.