HW Question Week 1

STA2101F 2022

Due September 21 2022 11.59 pm

Homework to be submitted through Quercus

You can submit this HW in Word, Latex, or R Markdown, but in future please use R Markdown. If you are using Word or Latex with a R script for the computational work, then this R script should be provided as an Appendix. In the document itself you would just include properly formatted output.

You are welcome to discuss questions with others, but the solutions and code must be written independently. Any R output that is included in a solution should be formatted as part of the discussion (i.e. not cut and pasted from the Console).

The dataset `wafer` concerns a study on semiconductors. You can get more information about the data with `?wafer`; you will first need `library(faraway);data(wafer)`, and possibly `install.packages("faraway")`. The questions below are adapted from LM Ch.3.

(a) Fit the linear model `resist ~ x1 + x2 + x3 + x4`. Extract the X matrix using the `model.matrix` function. How have the levels of the factors been coded?

(b) Compute the correlation between the columns of the X matrix. Why are there some missing values?

(c) What difference in resistance is expected when moving from the low to the high level of `x1`?

(d) Refit the model without `x4` and examine the regression coefficients and standard errors. What stayed the same and what changed? How is this related to the correlation matrix of X?

(e) (PhD only) An alternative to least squares regression is a penalized version called ridge regression. This estimate is defined as:

$$\hat{\beta}_\lambda = \arg\min_\beta \{(y - X\beta)^T(y - X\beta)\} + \lambda\beta^T\beta\}$$

Give an explicit expression for $\hat{\beta}_\lambda$ and show that it is equivalently defined by

$$\arg\min_\beta \{(y - X\beta)^T(y - X\beta)\}, \text{ subject to} \sum \beta_j^2 \leq t.$$

Show that the two solutions are the same, when $t = \hat{\beta}_\lambda^T\hat{\beta}_\lambda$. Compare the least-squares estimates and the ridge regression estimates of β for the `prostate` data.