

MAST90104: A First Course in Statistical Learning

Assignment 2, 2022

Due: 11:59pm Sunday 4 September. Please submit a scanned or other electronic copy of your work via the Learning Management System

This assignment is worth 5% of your total mark. The total point is 40. You must fill in the online plagiarism declaration form prior to submitting your assignment. You may use R for this assignment for calculations. If you do, include your R commands and output.

1. (4 pt) Consider this linear regression model

$$y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \epsilon_i$$

where \bar{x} is the sample mean of x . The error terms ϵ_i have mean 0 and variance σ^2 .

Show that the least square estimator b_0, b_1 of β_0 and β_1 are uncorrelated.

2. (8 pt) The file income.csv contains the income at age 30 and the number of years of formal education of 15 individuals. We would like to model the effect of education on income via a linear model

$$y = X\beta + \epsilon,$$

where $\beta = [\beta_0 \ \beta_1]^T$.

Solve the following questions using matrix calculations as detailed in the lectures. Do not use R's `lm()` function.

- (a) Write down the linear model in matrix form.
 - (b) Obtain the least squares estimates of the parameters using the formula in lecture notes.
 - (c) Calculate the sample variance s^2 .
 - (d) Estimate the average income of a person who has had 18 years of formal education.
3. (8 pt) Consider the dataset from question 2.
- (a) A confidence interval for β_1 is calculated to be (2.21929, 3.017580). Find the confidence level used.
 - (b) Define the joint 95% confidence region of β_0, β_1 .
 - (c) Calculate a 99% confidence interval for the income of a single person who has had 18 years of formal education.
4. (20 pt) An experiment is conducted to estimate the annual demand for cars, based on their cost, the current unemployment rate, and the current interest rate. A survey is conducted and the following measurements obtained:

Cars sold ($\times 10^3$)	Cost ($\times \$k$)	Unemployment rate (%)	Interest rate (%)
5.5	7.2	8.7	5.5
5.9	10.0	9.4	4.4
6.5	9.0	10.0	4.0
5.9	5.5	9.0	7.0
8.0	9.0	12.0	5.0
9.0	9.8	11.0	6.2
10.0	14.5	12.0	5.8
10.8	8.0	13.7	3.9

- (a) Fit a linear model to the data and estimate the parameters and variance σ^2 , using the formulas in the lecture notes.
- (b) Fit the model using the function `lm()` in R. Plot the diagnostics plots of the model. Discuss the plots.
- (c) Calculate the standardised residual, leverage, and Cook's distance for the 5th observation.
- (d) Calculate a 90% prediction interval for the number of \$7,000 cars sold in a year which has unemployment rate 8.6% and interest rate 5%.
- (e) Test for model relevance using a corrected sum of squares. Do NOT use **R** functions `lm()` and `anova()`
- (f) Test the hypothesis that $\beta_2 = 1$, using both t test and F test.
- (g) Test the hypothesis that $\beta_1 = \beta_3 = 0$, using both the formulas in lecture notes and R's function(s).