MAST30025: Linear Statistical Models

Assignment 2, 2017

Due: 5pm Friday April 28 (week 8)

This assignment is worth 7% of your total mark.

You may use R for this assignment, including the `lm` function unless otherwise specified. Include your R commands and output.

1. Prove Theorem 4.8: show that the maximum likelihood estimator of the error variance σ^2 is
$$\hat{\sigma^2} = \frac{SS_{Res}}{n}.$$

2. 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 (\$$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

For this question, you may NOT use the `lm` function in R.

(a) Fit a linear model to the data and estimate the parameters and variance.

(b) Which two of the parameters have the highest (in magnitude) covariance in their estimators?

(c) Find a 99% confidence interval for the average number of \$8,000 cars sold in a year which has unemployment rate 9% and interest rate 5%.

(d) A prediction interval for the number of cars sold in such a year is calculated to be $(4012, 7087)$. Find the confidence level used.

(e) Test for model relevance using a corrected sum of squares.

3. Consider two full rank linear models $\mathbf{y} = X_1\boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon}_1$ and $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}_2$, where all predictors in the first model are also contained the second model. Show that the SS_{Res} for the first model is at least the SS_{Res} for the second model.

4. In this question, we study a dataset of 50 US states. This dataset contains the variables:

- `Population`: population estimate as of July 1, 1975
- `Income`: per capita income (1974)
- `Illiteracy`: illiteracy (1970, percent of population)
- `Life.Exp`: life expectancy in years (1969–71)
- `Murder`: murder and non-negligent manslaughter rate per 100,000 population (1976)
- `HS.Grad`: percentage of high-school graduates (1970)

- `Frost`: mean number of days with minimum temperature below freezing (1931–1960) in capital or large city
- `Area`: land area in square miles

The dataset is distributed with R. Open it with the following commands:

```
> data(state)
> statedata <- data.frame(state.x77, row.names=state.abb, check.names=TRUE)
```

We wish to use a linear model to model the murder rate in terms of the other variables.

(a) Plot the data and comment. Should we consider any variable transformations?

(b) Perform model selection using forward selection, using all variable transformations which may be relevant.

(c) Starting from the full model, perform model selection using stepwise selection with the AIC.

(d) Write down your final fitted model (including any variable transformations used).

(e) Produce diagnostic plots for your final model and comment.

5. For ridge regression, we choose parameter estimators $\mathbf{b}$ which minimise

$$\sum_{i=1}^{n} e_i^2 + \lambda \sum_{j=0}^{k} b_j^2,$$

where λ is a constant penalty parameter.

(a) Show that these estimators are given by

$$\mathbf{b} = (X^T X + \lambda I)^{-1} X^T \mathbf{y}.$$

(b) Calculate the ridge regression estimates for the data from Q2 with penalty parameter $\lambda = 0.5$. In order to avoid penalising some parameters unfairly, we must first scale every predictor variable so that it is standardised (mean 0, variance 1), and centre the response variable (mean 0), in which case an intercept parameter is not used. (*Hint:* This can be done with the `scale` function).

(c) One way to calculate the optimal value for the penalty parameter is to minimise the AIC. Since the number of parameters p does not change, we use a slightly modified version:

$$AIC = n \ln \frac{SS_{Res}}{n} + 2\,df,$$

where df is the "effective degrees of freedom" defined by

$$df = tr(H) = tr(X(X^T X + \lambda I)^{-1} X^T).$$

For the data from Q2, construct a plot of λ against AIC. Thereby find the optimal value for λ.