MATH2831 Linear Models

Assignment

Note: This assignment is due by 11:59pm Monday 16 November (week 10)

Please follow the instructions below for completing the assignment, it's worth 20% of your final mark.

- This assignment must be completed individually.
- Your assignment **must be submitted as a pdf file**. It may be typed or handwritten, then converted into **one** pdf file. You must include the completed coversheet in your assignment.
- You must sign and date your submitted assignment, and include your name and zID below.

I declare that this assessment item is my own work, except where acknowledged, and has not been submitted for academic credit elsewhere, and acknowledge that the assessor of this item may, for the purpose of assessing this item:

- Reproduce this assessment item and provide a copy to another member of the University; and/or,
- Communicate a copy of this assessment item to a plagiarism checking service (which may then retain a copy of the assessment item on its database for the purpose of future plagiarism checking).

I certify that I have read and understood the University Rules in respect of Student Academic Misconduct.

_____

Student's full name and zID

Signed: _____ Date: _____

1. An experiment was conducted in order to study the size of squid eaten by sharks and tuna. The predictor variables are characteristic of the beak or mouth of the squid. The predictors and response considered for the study are:

$$\begin{aligned} x_1 &: \text{ Rostral length in inches} \\ x_2 &: \text{ Wing length in inches} \\ x_3 &: \text{ Rostral to notch length} \\ x_4 &: \text{ Notch to wing length} \\ x_5 &: \text{ Width in inches} \\ y &: \text{ Weight in pounds} \end{aligned}$$

The study involved measurements and weight taken on 22 specimen and is available in the `squid.txt` data set.

(I) Best subset selection. Carry out a best subset linear regression analysis using the `regsubsets()` function on the `squid` data set.

(a) Copy summary output in your report and briefly comment on the models identified in this output. From the summary ouput, identify the best model obtained with the four predictors.

(b) What is the best model based on adjusted R^2, PRESS and Cp from among the chosen models by the `regsubsets()` function? To provide evidence for your answer, include in your report a table showing the values of adjusted R^2, PRESS and Cp for the best subsets of each size. Include also two plots, one of adjusted R^2 and another of Cp for the best subsets of each size against the number of predictors.

(II) Sequential variable selection on the `squid` data set.

(a) Carry out forward model selection with the `stepAIC()` function, using all the available predictors and starting from the model with just an intercept.

(i) Copy the R output in your report and describe the selection procedure from the output. At each step state the 'current model', which predictor was added to the current model and why.

(ii) Clearly state the final model obtained, including the coefficient estimates of the fitted model. How does your answer compare to the results in (I)?

(b) Repeat a) using backward model selection, starting from the model with all the available predictors.

(i) Copy the R output in your report and describe the selection procedure from the output. At each step state the 'current model', which predictor was removed from the current model and why. What is the AIC for the model with just x_4?

(ii) Clearly state the final model obtained, including the coefficient estimates of the fitted model. Do you obtain the same model as in the forward selection above?

(c) Carry out a stepwise selection procedure with the `stepAIC()` function, using all the available predictors and starting from the model with just x_1. Do NOT include the R output, just state the final model in your answer, and compare to your findings in parts a) and b) above.

(III) Model criticism. Fit a linear model with y as the response and all the available predictors to the `squid` data.

(a) Include in your report diagnostic plots of residuals for the fitted model and comment on the appropriateness of the general linear model assumptions. In particular, comment on whether or not there appear to be any violation of model assumptions, such as incorrectly specified mean, failure of the constancy of error variance, departure from normality, outliers and observations that have a large influence on the model analysis.
Note, that you can plot all four residual plots using:

```
par(mfrow=c(2,2))
plot(model)
par(mfrow=c(1,1))
```

(b) Recommend the final optimal model based on you findings in (I) and (II). Give reasons to support your choice.
Fit chosen optimal model to the `squid` data and repeat part (a). Produce the summary output of the fitted model and include in your report. Are all the predictor variables "significant" in the final model according to the parial t-tests?

2. In this question we consider derivation of Mallows' C_p statistic for model selection discussed in lectures.

Suppose the experimenter proposes a model

$$y = X_1\beta_1 + \varepsilon^* \quad (p \text{ parameters})$$

where X_1 is $n \times p$ matrix and vector β_1 contains p parameters.

The "true" model however contains additional $m - p$ parameters described by vector β_2. So the "true" model is given by

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon \quad (m \text{ parameters}, m > p)$$

where X_2 is $n \times (m-p)$ matrix. Assume that errors ε are uncorrelated with mean zero and common variance σ^2.

Consider fitting the proposed general linear model to data and write $\widehat{y}_i$ for the fitted value at x_i and $MSE(\widehat{y}_i)$ for its mean squared error. Recall that if the error variance σ^2 is known, then an estimate of

$$\frac{\sum_{i=1}^n MSE(\widehat{y}_i)}{\sigma^2} = \frac{\sum_{i=1}^n Var(\widehat{y}_i)}{\sigma^2} + \frac{\sum_{i=1}^n Bias^2(\widehat{y}_i)}{\sigma^2}$$

is

$$p + \frac{(n-p)(\widehat{\sigma}^2 - \sigma^2)}{\sigma^2} \tag{1}$$

where $\widehat{\sigma}^2$ is the estimate of the error variance for the proposed model and p is the number of parameters.

You now have to provide a justification for (1).

(a) Writing $\widehat{y} = (\widehat{y}_1, ..., \widehat{y}_n)^\top$ for the vector of fitted values for the proposed model and observing that

$$\widehat{y} = X_1(X_1^\top X_1)^{-1}X_1^\top y = H_1 y,$$

show that

$$\sum_{i=1}^n Var(\widehat{y}_i) = \sigma^2 \text{tr}(H_1),$$

where $\text{tr}(A)$ denotes the trace of A and $H_1 = X_1(X_1^\top X_1)^{-1}X_1^\top$ denotes the hat matrix corresponding to the proposed model. By using the rules given in lectures about matrix traces, deduce that

$$\frac{\sum_{i=1}^n Var(\widehat{y}_i)}{\sigma^2} = p.$$

(b) Consider the estimate of σ^2 obtained in lectures for the proposed model,

$$\widehat{\sigma}^2 = \frac{y^\top (I - X_1(X_1^\top X_1)^{-1} X_1^\top) y}{n-p}.$$

By using the result stated in lectures about the expected value of a quadratic form $y^\top A y$, and noting that

$$\mathbb{E}(y) = X_1\beta_1 + X_2\beta_2,$$

show that

$$\mathbb{E}(\widehat{\sigma}^2) = \sigma^2 + \frac{1}{n-p}\beta_2^\top X_2^\top (I - H_1) X_2 \beta_2.$$

(c) Show that

$$\begin{aligned} \sum_{i=1}^{n} Bias^2(\widehat{y}_i) &= (\mathbb{E}(y) - \mathbb{E}(\widehat{y}))^\top (\mathbb{E}(y) - \mathbb{E}(\widehat{y})) \\ &= \beta_2^\top X_2^\top (I - H_1) X_2 \beta_2. \end{aligned}$$

(d) From b) and c), deduce that an unbiased estimator of

$$\frac{\sum_{i=1}^{n} Bias^2(\widehat{y}_i)}{\sigma^2}$$

is

$$\frac{(n-p)(\widehat{\sigma}^2 - \sigma^2)}{\sigma^2},$$

from which it follows that (1) is a sensible estimator of

$$\frac{\sum_{i=1}^{n} MSE(\widehat{y}_i)}{\sigma^2}.$$