FALL 2010
Stat 849: Homework Assignment 4
Due: November 24, 2010
Total points = 250

1. (30) Show that the first step in the forward selection is equivalent to selecting the variable most highly correlated with the response.

2. (20) The AIC criterion for a model $\mathcal{M}$ for which the mle's provide a log-likelihood of ℓ and the total number of parameters is q is
$$\mathrm{AIC} = -2\ell + 2q$$
Find an expression for the AIC in terms of residual sum of squares in the Gaussian linear model and simplify it as much as you can.

3. (75) Design a simulation study to investigate the effects of over- and under-fitting in linear regression models. Summarize and report your conclusions at two sample sizes, $n = 50$ and $n = 500$.

   Hint: For example, let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_5$ be normally distribution random vectors of length n. Let the true regression model be
$$\mathcal{Y} \sim \mathcal{N}(\beta_0 + \beta_1\boldsymbol{x}_1 + \beta_2\boldsymbol{x}_2 + \beta_3\boldsymbol{x}_3, \sigma^2\boldsymbol{I}_n)$$
   You will need to generate $n = 50$ (later $n = 500$) observations according to this mechanism. For studying the effect of under-fitting, fit
```
> underfit <- lm(Ysim ~ 1 + x1 + x2)
```
   and summarize the estimates of β_0, β_1, β_2 and σ^2. Similarly, for studying the effect of over-fitting, fit
```
> overfit <- lm(Ysim ~ 1 + x1 + x2 + x3 + x4)
```
   and report the estimates of the coefficients in the true model. For this to be a simulation study, you should generate a matrix of responses with at least 10000 columns. Report both the bias and the variance of your estimates.

4. (75) Consider the linear model
$$\mathcal{Y} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z}\gamma, \sigma^2\boldsymbol{I}_n)$$
   where $\boldsymbol{X}$ is a known $n \times p$ matrix of rank $p < n$, $\boldsymbol{z}$ is a known $n \times 1$ vector that is linearly independent of the columns $\boldsymbol{X}$, and $\boldsymbol{\beta}$, γ and σ^2 are unknown parameters.

   (a) Consider fitting the model shown above by ignoring the $\boldsymbol{z}\gamma$ term. The corresponding ordinary least squares estimator of $\boldsymbol{\beta}$ is obtained as $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. Let $\boldsymbol{r}$ be the vector residuals, with $r_i = y_i - \widehat{y}_i$, from the fitted model. Derive $E(\boldsymbol{r})$ and $\mathrm{cov}(\boldsymbol{r})$.

   (b) Consider a full least squares fit of the model shown above. Let $\boldsymbol{M} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Show that
$$\widehat{\gamma} = \frac{\boldsymbol{z}^T(\boldsymbol{I} - \boldsymbol{M})\boldsymbol{y}}{\boldsymbol{z}^T(\boldsymbol{I} - \boldsymbol{M})\boldsymbol{z}}.$$
   Hint: First rewrite $\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z}\gamma$ as $\boldsymbol{X}\boldsymbol{\delta} + (\boldsymbol{I} - \boldsymbol{M})\boldsymbol{z}\gamma$, where $\boldsymbol{\delta}$ can involve both parameters and elements of $\boldsymbol{X}$ and/or $\boldsymbol{z}$.

(c) Argue whether or not the following claim is correct: "If the plot of $\boldsymbol{r}$ versus $\boldsymbol{z}$ represents the influence of $\boldsymbol{z}$ after accounting for other variables, then the slope from fitting a simple linear regression of $\boldsymbol{r}$ on $\boldsymbol{z}$ will be equal to the γ estimate that I would get from fitting the model shown above."

5. (50) **Job proficiency data (`jobdata.txt`).** A personnel officer in a governmental agency administered four newly developed aptitude tests to each of 25 applicants for entry-level clerical positions in the agency. For purpose of the study, all 25 applicants were accepted for positions irrespective of their test scores. After a probationary period, each applicant was rated for proficiency on the job. The scores on the four tests (X_1, X_2, X_3, X_4) and the job proficiency score (Y) for the 25 employees are given in `jobdata.txt`, where the first column represents Y and the rest represent the test scores (X_1, X_2, X_3, X_4).

```
> DataURL <- "http://www.stat.wisc.edu/~st849-1/data/"
> job <- read.table(paste(DataURL, "jobdata.txt", sep=''),
+                   col.names=c("Y","X1","X2","X3","X4"))
> str(job)
'data.frame':        25 obs. of  5 variables:
 $ Y : num  88 80 96 76 80 73 58 116 104 99 ...
 $ X1: num  86 62 110 101 100 78 120 105 112 120 ...
 $ X2: num  110 97 107 117 101 85 77 122 119 89 ...
 $ X3: num  100 99 103 93 95 95 80 116 106 105 ...
 $ X4: num  87 100 103 95 88 84 74 102 105 97 ...
```

Using forward selection, backward deletion, and forward selection & backward deletion stepwise methods to find the best subset of predictor variables to predict job proficiency as a linear function of test scores. Compare models obtained by each stepwise algorithm, choose a final model and discuss how and why you chose it.