Stat 601 Homework 5 Due 11/20/2015

1. Suppose that the regression model postulated is

$$E(Y) = \beta_0 + \beta_1 x$$

when, in fact, the true model is

$$E(Y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3.$$

If we use observations of Y at $x = -3, -2, -1, 0, 1, 2, 3$ to estimate β_0 and β_1 in the postulated model, what bias will be introduced in these estimates.

2. Consider a linear function $d'\beta$ of β. Show that the change in the estimate $d'\hat{\beta}$ when the ith observation is deleted is

$$d'\hat{\beta} - d'\hat{\beta}(i) = (C'd)_i \hat{\epsilon}_i/(1 - h_i),$$

where C is the catcher matrix $(X'X)^{-1}X'$.

3. An avid fan of the PGA tour with limited background in statistics has sought your help in answering one of the age-old questions in golf, namely, *what is the relative importance of each different aspect of the game on average prize money in professional golf ?*

The following data on the top 196 tour players in 2006 can be found in the file pgatour2006.csv:

Y, PrizeMoney = average prize money per tournament

x_1, Driving Accuracy is the percent of time a player is able to hit the fairway with his tee shot.

x_2, GIR, Greens in Regulation is the percent of time a player was able to hit the green in regulation. A green is considered hit in regulation if any part of the ball is touching the putting surface and the number of strokes taken is two or less than par.

x_3, Putting Average measures putting performance on those holes where the green is hit in regulation (GIR). By using greens hit in regulation the effects of chipping close and one putting are eliminated.

x_4, Birdie Conversion% is the percent of time a player makes birdie or better after hitting the green in regulation.

x_5, SandSaves% is the percent of time a player was able to get "up and down" once in a greenside sand bunker.

x_6, Scrambling% is the percent of time that a player misses the green in regulation, but still makes par or better.

x_7, PuttsPerRound is the average total number of putts per round.(http://www.pgatour.com/r/stats/; accessed March 13, 2007)

(a) A statistician from Australia has recommended to the analyst that they not transform any of the predictor variables but that they transform Y using the log transformation. Do you agree with this recommendation? Give reasons to support your answer.

(b) Develop a valid full regression model containing all seven potential predictor variables listed above. Ensure that you provide justification for your choice of full model, which includes scatter plots of the data, plots of standardized residuals, and any other relevant diagnostic plots.

(c) Identify any points that should be investigated. Give one or more reasons to support each point chosen.

(d) Describe any weaknesses in your model.

(e) The golf fan wants to remove all predictors with insignificant t -values from the full model in a single step. Explain why you would not recommend this approach.

4. Show that the first step in the forward selection is equivalent to selecting the variable most highly correlated with the response.

5. The AIC criterion for a model M for which the mle's provide a log-likelihood of l and the total number of parameters is q is

$$AIC = -2l + 2q$$

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.

6. 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 $n = 500$.

   - For this to be a simulation study, you should generate a matrix of covariates with at least 10000 columns.
   - Report both the bias and the variance of your estimates.

7. Consider the linear model

$$Y \sim N(X\beta + Z\gamma, \sigma^2 I)$$

where X is a known $n \times p$ matrix of rank $p < n$, z is a known $n \times 1$ vector that is linearly independent of the columns X, and β, γ and σ are unknown parameters.

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

(b) Consider a full least squares fit of the model shown above. Let $M = X(X'X)^{-1}X'$. Show that

$$\hat{\gamma} = \frac{z'(I-M)Y}{z'(I-M)z}.$$

Hint: First rewrite $X\beta + z\gamma$ as $X\delta + (I-M)z\gamma$.

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

8. **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).

```
> job <- read.table("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.

9. This is a continuation of Problem 3. The golf fan was so impressed with your answers to part 1 that your advice has been sought re the next stage in the data analysis, namely using model selection to remove the redundancy in full the model developed in part 1.

$$\log(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \epsilon$$

Interest centers on using variable selection to choose a subset of the predictors to model the transformed version of Y . Throughout this question we shall assume the above model is a valid model for the data.

(a) Identify the optimal model or models based on R^2_{adj} , AIC, BIC from the approach based on all possible subsets.

(b) Identify the optimal model or models based on AIC and BIC from the approach based on backward selection.

(c) Identify the optimal model or models based on AIC and BIC from the approach based on forward selection.

(d) Carefully explain why the models chosen in (a) & (c) are not the same while those in (a) and (b) are the same.

(e) Recommend a final model. Give detailed reasons to support your choice.

(f) Interpret the regression coefficients in the final model. Is it necessary to be cautious about taking these results to literally?