Stat 191 **Assignment 3** **Winter 2018**

The assignment is due on February 21 at 9 am. You need to show your work and answers to each problem as well as plots and **R** code and output to substantiate your answers. You don't have to typeset your solution; handwritten answers together with a printout of code and output are fine.

1. Most of the regression data we have seen in class is easily described by linear regression equations. Exceptions certainly occur, but often other types of models can be easily transformed into a linear model. A particularly common type of nonlinear model is the model

$$Y = \gamma_0 X^{\gamma_1}.$$

(a) Show that if $Y' = \log Y$ and $X' = \log X$, the above model is equivalent to a simple linear model. Express the usual parameters of the simple linear model in terms of the parameters γ_0 and γ_1.
(b) Among mammals, the relationship between the age at which an animal develops locomotion and the age at which it begins to play has been widely studied. Listed in the following table are typical onset times for locomotion and for play in 11 different species (including humans).

Species	Locomotion begins X_i (days)	Play begins Y_i (days)
Homo sapiens	360	90
Gorilla gorilla	165	105
Felis catus	21	21
Canis familaris	23	26
Rattus norvegicus	11	14
Tudus merula	18	28
Macaca mulatta	18	21
Pan troglodytes	150	105
Saimiri sciurens	45	68
Cercocebus alb.	45	75
Tamiasciureus hud.	18	46

Analyze the data. In particular, your analysis should address the following questions: Is a linear model relating Y and X appropriate? Do the data appear normal? Is a transformation along the lines of that studied in (a) appropriate? Do the transformed variables appear normal? Do any of the data points look suspicious (in either model)? Fit both a linear and nonlinear model (i.e. using transformations) and carry out appropriate tests of significance in answering this question. Be careful to check all assumptions. Data are in Locomotion.txt and Play.txt.

2. 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 purposes 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 (last column) for the 25 employees are given in the file `Job.txt`.

(a) Prepare stem-and-leaf plots for each of the four sets of newly-developed aptitude test scores. Are there any noteworthy features in these plots? Comment.

(b) Prepare scatterplots of job proficiency scores against each of the four independent variables. What do the scatterplots suggest about the nature of the functional relationship between the dependent variable Y and each of the independent variables?

(c) Obtain the correlation matrix of the X variables. Are any serious multicollinearity problems evident from this matrix? Explain.

(d) Fit the multiple regression model containing all four independent variables.

(e) Obtain the variance inflation factors for the regression model fitted in part (d). Are there indications that serious multicollinearity problems exist here? Explain.

(f) Obtain the residuals and plot them separately against $\hat{Y}$ and each of the independent variables. Also prepare a normal probability plot of the residuals. On the basis of these plots, should any modifications be made to the regression model?

(g) Find the four best subset regression models according to the adjusted R^2 criterion. According to the C_p criterion? Now, using the unadjusted R^2 criterion, what do you think the four best subset regression models are? Comment on your choices and compare them with those you found using the adjusted R^2 criterion.

(h) Since there is relatively little difference in the adjusted R^2 values for the four best subset models, what other criteria would you use to help in the selection of the best model? Discuss.

(i) Using forward stepwise regression, find the best subset of independent variables to predict job proficiency.

(j) How does the best subset according to forward stepwise regression compare with the best subset according to the adjusted R^2 criterion obtained in (g)?

The subset model containing only X_1 and X_3 is to be evaluated in detail. **The remainder of the questions concern this model.**

(k) To assess internally the predictive ability of this regression model, compute the PRESS statistic and compare it to SSE. What does this comparison suggest about the validity of MSE as an indicator of the predictive ability of the fitted model?

To assess externally the validity of this regression model, 25 additional applicants for entry-level clerical positions in the agency were similarly tested and hired irrespective of their test scores. Their data are given in the file `Jobval.txt`.

(l) Obtain the correlation matrix of the X variables for the validation set and compare it with that obtained for the model-building data set. Comment on your findings.

(m) Fit the regression model specified above (using X_1 and X_3) to the validation data set. Compare the estimated regression coefficients and their estimated standard deviations to those obtained for the model-building data set. Also compare the error mean squares and coefficients of multiple determination. Do the estimates for the validation data set appear to be reasonably similar to those obtained for the model-building data set?

(n) Calculate the mean squared prediction error (MSPR) and compare it to MSE obtained from the model-building data set. Is there evidence of a substantial bias problem in MSE here? Is this conclusion consistent with your finding in (k)?

(o) Combine the model-building data set given in `Job.txt` with the validation data set and fit the selected regression model to the combined data. Are the estimated standard deviations of the estimated regression coefficients appreciably reduced now from those obtained for the model-building data set?