Stat 191 **Assignment 2** **Winter 2018**

The assignment is due on February 2. 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. Suppose that in the model

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, \dots, n,$$

the errors have mean 0 and are independent, but $\mathrm{Var}(\varepsilon_i) = \rho_i^2 \sigma^2$, where the ρ_i are known constants so that the errors do not have equal variance. This situation arises when the Y_i are averages of several observations at X_i; in this case, if Y_i is an average of n_i independent observations, $\rho_i^2 = 1/n_i$. Since the variances are not equal the theory developed in lectures does not apply; intuitively, it seems that observations with large variability should influence the estimates of β_0 and β_1 less than the observations with small variability.

The problem may be transformed as follows:

$$\rho_i^{-1} Y_i = \rho_i^{-1} \beta_0 + \rho_i^{-1} \beta_1 X_i + \rho_i^{-1} \varepsilon_i, \qquad i = 1, \dots, n,$$

or,

$$Z_i = u_i \beta_0 + v_i \beta_1 + \delta_i, \qquad i = 1, \dots, n,$$

where

$$u_i = \rho_i^{-1}, \qquad v_i = \rho_i^{-1} X_i, \qquad \delta_i = \rho_i^{-1} \varepsilon_i.$$

(a) Show that the new model satisfies the assumptions of the standard statistical model. (Assume that the X's are fixed in the above model.)

(b) Find the least squares estimates of β_0 and β_1.

(c) Show that performing a least squares analysis on the new model, as was done in part (b), is equivalent to minimizing

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \rho_i^{-2}.$$

This is a weighted least squares criterion; the observations with large variances are weighted less.

2. A researcher in a scientific foundation wishes to evaluate the relation between intermediate and senior level annual salaries of research mathematicians (Y, thousands of dollars) and an index of publication quality (X_1), number of years experience (X_2), and an index of success in obtaining grant support (X_3). The data for a sample of 24 intermediate and senior level research mathematicians are in the file 'math.txt', with Y being in the last column.

(a) Fit a regression model with all three independent variables to the data. What is the estimated regression function?

(b) Obtain the fitted values and residuals from the fit in (a). Construct a box-plot of the residuals. What do you conclude?

(c) Construct appropriate diagnostic plots to assess the validity of normality and homoscedasticity assumptions on the residuals. Also, plot the standardized residuals against each of the independent variables and each two-factor interaction (i.e. product of two X predictors) on separate graphs. Analyse your plots and summarize your findings.

(d) Assume that the regression model with three independent variables and independent normal errors is appropriate. Test the significance of the regression (using $\alpha = 0.05$). Carefully state the null and alterniative hypotheses and your conclusions.

(e) Test whether $\beta_1 = \beta_3$; use $\alpha = 0.01$. State the alternatives, full and reduced models, decision rule and conclusion.

(f) Prepare a partial regression plot (added-variable plot) for each of the independent variables. Do these plots suggest that the regression relationships in the fitted regression function in question (a) are inappropriate for any of the independent variables? Explain.

(g) Identify any outlying X values.

(h) Obtain the externally studentized residuals and identify any outlying Y observations.

(i) Data point 19 (4.0,35,6.0,38.0) appears to be a borderline outlying Y observation. Obtain the DFFITS, DFBETAS, and Cook's distance values for this point to assess its influence. What do you conclude?

(j) Calculate Cook's distance for the estimated regression coefficients for each point. Are any points influential by this measure?