Stat 331 Applied Linear Models – Assignment 3

You need to use the cover sheet provided in Learn. Due on Nov 19 (Wednesday) 12pm to the drop boxes located across the hall from MC 4065/4066.

1. Assume that data are generated by the model

$$Y = X\beta + Z\gamma + \epsilon,$$

where Y is $n \times 1$, X is an $n \times (p_1 + 1)$ matrix of predictors, Z is an $n \times p_2$ matrix of predictors, β and γ are respectively $(p_1 + 1) \times 1$ and $p_2 \times 1$ parameter vectors, and ϵ is an $n \times 1$ vector of independent normally distributed random errors with mean 0 and variance σ^2. Notice that the intercept term is already included in X and β. Suppose a statistician is unaware of the Z predictors and fits the model $Y = X\beta + \epsilon$, calculating the standard lease squares estimator $\hat{\beta}$ under this assumption of ϵ. Let $H = X(X^TX)^{-1}X^T$.

   (a) Calculate the expectation of $\hat{\beta}$, and the expectation of the residual vector $r = Y - \hat{Y} = Y - X\hat{\beta}$.

   (b) Recall the definition of variance: $Var(r) = E\{r - E(r)\}\{r - E(r)\}^T$. Show that $E(rr^T) = Var(r) + E(r)E(r)^T$.

   (c) Using the conclusion in (a) and (b), show that the expected value of $SSE = r^Tr$ is $(n - p_1 - 1)\sigma^2 + \gamma^T Z^T(I - H)Z\gamma$.

2. The data set "math" explores the dependence of "income", the annual income of mathematics students' first job after graduation (response variable), on two explanatory variables: "GPA" (X_1), and an indicator variable "region" (X_2), where 0 stands for jobs in Europe and 1 for jobs in US. The GPA variable is rounded, hence we have repeated observations from the same predictors.

   (a) Suppose we are interested in the model

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i,$$

   where $x_{i,j}$ is the j^{th} predictor for the i^{th} observation. Perform a test for lack of fit for this model at 5% significance level. State the null and alternative hypotheses, calculate the F statistic and p-value, and make a conclusion.

   (b) Now we are interested to see if the interaction term "interaction" between "GPA" and "region" can contribute significantly in explaining the variation in the response variable. Construct a partial residual plot with respect to "interaction" and comment on the plot. Perform another test for lack of fit for the new model

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 (x_{i,1} * x_{i,2}) + \epsilon_i$$

   at 5% significance level. State the null and alternative hypotheses, calculate the F statistic and p-value, and make a conclusion.

   (c) Interpret the estimated parameter $\hat{\beta}_3$ in Step (b).

   (d) Based on the p-values for the t-tests in Step (b), can we drop the variable "GPA" in the model at 5% significance level? Explain.

   (e) Create a plot of leverages for the model in (b). Are there any outliers in x?

(f) Create a plot of Cook's D values for the model in (b). Do any observations have high influence? Compare to $F_{0.5}(p+1, n-p-1)$.

3. Consider the "Senic" data which you analyze in Homework 2, Q2. The variables we will use again for this analysis include:

y	InfctRisk, the risk of infection at a hospital
x_1	Stay, average length of stay at the hospital
x_2	Cultures, average number of bacterial cultures per day at the hospital
x_3	Age, average age of patients at hospital
x_4	Census, the average daily number of patients
x_5	Beds, the number of beds in the hospital

(a) Perform a best subsets regression in R to identify a possible model for predicting infection risk. Clearly state the final model and why you choose it.

(b) Perform a forward selection based on F test in R. Clearly state which variable is selected in each step.

(c) Perform a backward elimination based on AIC value. Clearly state which variable is removed in each step.