Homework # 8

STA 4210

Due: Monday, April 21

The submission must be typed. Show all work. Print all input and output code if using software to solve. Points will be deducted for insufficient work or missing steps to solving the problems.

The original data files for both problems can be found here:

`https://people.clas.ufl.edu/dlindberg/sta-4210/`

Problem 1. In any production process in which one or more workers are engaged in a variety of tasks, the total time spent in production varies as a function of the size of the workpool and the level of output of the various activities. In a large metropolitan department store, it is believed that the number of man-hours worked (Y) per day by the clerical staff depends on the following variables.

$$
\begin{aligned}
X_1 &= \text{Number of Pieces of Mail Processed} \\
X_2 &= \text{Number of Money Orders and Gift Certificates Sold} \\
X_3 &= \text{Number of Window Payments Transacted} \\
X_4 &= \text{Number of Change Order Transactions} \\
X_5 &= \text{Number of Checks Cashed} \\
X_6 &= \text{Number of Pieces of Miscellaneous Mail Processed} \\
X_7 &= \text{Number of Bus Tickets Sold}
\end{aligned}
$$

Data was recorded for each of these activities on 52 working days.

(a) Construct a multiple linear regression model with all 7 predictor variables and write down the fitted equation.

(b) Using the `stepAIC()` function from the `MASS` package, perform backward elimination on the model in part (a) based on AIC and identify the final model selected.

(c) Using the `stepAIC()` function, perform forward selection based on AIC and identify the final model selected.

(d) Manually calculate the AIC and BIC for the optimal model(s) in parts (b) and (c) using the formulas from Chapter 9. You may first need to construct the model using the `lm()` function, then use the `anova()` function to find values needed for the formulas.

(e) Fit the model selected in part (b). Write down the fitted equation. Calculate the $PRESS_p$ statistic for this model. Calculate $\frac{PRESS_p}{n}$ and compare to MSE_p. Is this an effective model in terms of model validation?

(f) Obtain Studentized deleted residuals, hat values, DFFITS, Cooks D, and DFBETAS for each observation, based on the model selected in part (b). Do any observations stand out as outliers or influential cases? Specifically give your "critical" cut-off value for each measure.

(g) Using the `vif()` function from the `car` package, calculate the Variance Inflation Factor (VIF) for each predictor and the average VIF for the model selected in part (b). Is there evidence of multicollinearity?

(h) Prepare an added variable plot for each of the predictors in the model selected in part (b). Comment on each plot in terms of whether the predictor should be in the model.

Problem 2. A small business is modeling the amount of revenue, Y, (in \$1,000's) with the number of total person hours worked on a particular day, X. A random sample of 22 days is collected.

(a) Fit a simple linear regression. Write down the fitted equation.

(b) Conduct the Modifed Levene Test to identify whether there is non-constant error variance. Split the data into two equally sized groups based on the value of X. Use a level of significance of $\alpha = 0.05$. State the null and alternative hypotheses, calculate a test statistic, compare the test statistic to a critical value, and state your conclusion.

(c) Create a simple linear regression equation based on the Weighted Least Squares method where $\sigma_i^2 = e_i^2$ for all observations. Write down the fitted equation.

(d) Produce a scatterplot with the fitted equations from parts (a) and (c) plotted over the data. Be sure to differentiate each line: use a different color (`col="red"`) or line type (`lty=2`).