Stat 331 Applied Linear Models – Assignment 2

You need to use the cover sheet provided in Learn. Due on Oct 29 (Wednesday) 12pm to the drop boxes located across the hall from MC 4065/4066.

1. In multiple linear regression, recall that H is the hat matrix, $\hat{Y} = HY$. Assume that the normality assumption of Y holds.

   (a) Prove that $SSE = Y^T(I - H)^T(I - H)Y$.

   (b) Prove that SSE and $\hat{Y}$ are independent.

   (c) Show that $SSR = Y^T HY - n\bar{y}^2$.

2. Consider the "Senic" data. You can access the data on Learn. The variables we will use 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) Plot the `InfctRisk` against each of the five predictors. Based on the plots, which predictors appear to be related to `InfctRisk`?

   (b) Fit the following model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$$

   Find the value of $\hat{\beta}_2$ and interpret its meaning.

   (c) Predict the risk of infection (%) for a new hospital with $x_1 = 7, x_2 = 9, x_3 = 56, x_4 = 200, x_5 = 250$. Construct a 95% prediction interval for the infection risk.

   (d) In the five-predictor model, test whether age, census, and beds can be removed from the model at the same time. State the null and alternative hypotheses, the value of the test statistic and draw the conclusion ($\alpha = 0.05$).

   (e) Re-fit the model with only two predictors: Stay and Cultures. What is the value of R^2 and interpret its meaning. Compare it with the R^2 value for the five-predictor model and comment.

   (f) Based on the two-predictor model, plot the studentized residuals versus the fitted values and make a Q-Q plot of the studentized residuals. Comment on the two plots.

3. Consider the "nuclear" data. You can access the data in R by using R commands:

```
> library(boot)
> data(nuclear)
```

   The data set `nuclear` contains data collected on 32 light-water reactor (LWR) plants constructed in the U.S.A in the late 1960′s and early 1970′s. The variables include:

`cost`	Cost in millions of dollars, adjusted to 1976 base
`date`	Date construction permit issued. The data are measured in years since Jan. 1, 1990 to the nearest month
`t1`	Time between application for and issue of permit
`t2`	Time between issue of operating license and construction permit
`cap`	Power plant net capacity (MWe)
`pr`	Prior existence of an LWR on same site (1=yes)
`ne`	Plant constructed in north-east region of USA (1=yes)
`ct`	Use of cooling tower (1= yes)
`bw`	Nuclear steam supply system manufactured by Babcock-Wilcox (1=yes)
`cum.n`	Cumulative number of power plants constructed by each architect-engineer
`pt`	Partial turnkey plant (1=yes)

The data was collected with the aim of predicting the cost of construction of further LWR plants. We shall use `cost` as the response variable and all the others as the explanatory variables.

(a) Plot the `cost` against each of the predictors (except $1/0$ indicator variables such as `pr, ne ct, bw, pt`). Based on the plots, which predictors appear to be related to `cost`?

(b) Apply the Box-Cox transformation analysis to the response variable. What does this analysis suggest about transforming the cost?

(c) Apply the transformation you picked in (b). Write down the fitted equation for the model based on the transformation. Give a careful explanation of the estimated β coefficient for the variable `cap` and of the estimated β coeofficient for the variable `ne`.

(d) Do the estimated coefficients in your model make sense? In other words, do the $\hat{\beta}$'s have the right sighs given your knowledge of nuclear engineering?

(e) Based on the transformed model, plot the studentized residuals versus the fitted values and make a Q-Q plot of the studentized residuals. Comment on the two plots.