Homework 5 (67 pts)

Data Background

The primary objective of the study on the Efficacy of Nosocomial Infection Control (**SENIC**) was to determine whether infection surveillance and control programs have reduced the rates of nosocomial (hospital-acquired) infection in United States hospitals. The data set consists of a random sample of 113 hospitals selected from the original 338 hospitals surveyed.

Each line of the data set has an identification number and provides information on 11 other variables for a single hospital. The 12 variables and descriptions are the following. Note that in this data, the region has been recoded such that 1=NC, 2=NE, 3=S, 4=W, and the default baseline level is NC or 1.

Variable number	Variable name	Description
1	Id	1-113
(Y) 2	Length of stay	Average length of stay of all patients in hospital (in days)
3	Age	Average age of patients (in years)
(Infection) 4	Infection risk	Average estimated probability of acquiring infection in hospital (in percent)
5	Routine culturing ratio	Ratio of number of cultures performed to number of patients without signs or symptoms of pneumonia, times 100
6	Routine chest X-ray ratio	Ratio of X-rays performed to number of patient without signs or symptoms of pneumonia, times 100
7	Number of beds	Average number of beds in hospital during study period
8	Medical school affiliation	1=yes, 2=no
(Region) 9	Region	Geographic region, where 1=NE, 2=NC, 3=S, 4=W
10	Average daily census	Average number of patients in hospital per day during study period
(Nurse) 11	Number of nurses	Average number of full-time equivalent registered and licensed practical nurses during study period
12	Available facilities and services	Percent of 35 potential facilities and services that are provided by the hospital

Use R for the homework. For each question, attach code, R output and interpret the relevant output clearly.

Problem 1 (4 pts)

The length of stay (Y) is to be regressed on X_4 (infection) and $X_{11(\text{new})}$ (the availability of nurses). Note that the **categorical variable, $X_{11(\text{new})}$, is created from the original continuous variable, nurse in the data**. Specifically, $X_{11(\text{new})} = 0$ if the number of nurses is no less than the average value (173), and $X_{11(\text{new})} = 1$ otherwise. That is $\text{X}_{11(\text{new})} \leftarrow \text{ifelse}(\text{X}_{11} \geq 173, 0, 1)$.

(a) [2 pts] Consider the MLR model: $Y = \beta_0 + \beta_4 X_4 + \beta_{11(\text{new})} X_{11(\text{new})} + \beta_{4,11(\text{new})} X_4 X_{11(\text{new})} + \epsilon$. Use R to complete the model summary.

(b) [1 pt] The general form of the mean response functions for Y at the baseline level when $X_{11(\text{new})} = 0$ is $Y = \beta_0 + \beta_4 X_4 + \beta_{11(\text{new})} * 0 + \beta_{4*11(\text{new})} X_4 * 0 = \beta_0 + Y = \beta_0 + \beta_4 X_4$. Find the estimated mean response function by finding the value of the parameter estimate, i.e., b_0 and b_4 from the model summary_____.

(c) [1 pt] The general form of the mean response functions at the level where $X_{11(\text{new})} = 1$ is $Y = \beta_0 + \beta_4 X_4 + \beta_{11(\text{new})} * 1 + \beta_{4*11(\text{new})} X_4 * 1 = (\beta_0 + \beta_{11(\text{new})}) + (\beta_4 + \beta_{11(\text{new})}) X_4$.
Find the estimated form _____.

Problem 2 (10 pts)

Refer to Problem 1, complete the following hypothesis with a GLT F-test. Specifically, define the full model, reduced model, compute the test statistic, critical value, p-value, and state the conclusion.

(a) [5 pts] Infection has **no** impact on Y, whether the number of nurses is no less than the average or not (i.e.,$X_{11(\text{new})} = 0$ or $X_{11(\text{new})} = 1$).

(b) [5 pts] Infection has the same impact on Y when both when the number of nurses is no less than the average and when the number of nurses is greater than the average. In other words, there is no interaction impact between X_4 and $X_{11(\text{new})}$ on Y.

Problem 3 (8 pts)

Consider the multiple linear regression (MLR) model that regresses the response on infection, region, and the interaction between infection and region. Region has four levels and can be modeled with 3 dummy variables: $X_{9,1}$, $X_{9,2}$, and $X_{9,3}$. The default baseline chosen in R is NC based on alphabetical order, with $X_{9,1} = X_{9,2} = X_{9,3} = 0$, if the region is NC, $X_{9,1} = 1$, $X_{9,2} = 0$, and $X_{9,3} = 0$ if the region is NE, $X_{9,1} = 0$, $X_{9,2} = 1$, and $X_{9,3} = 0$ if the region is S, and $X_{9,1} = 0$, $X_{9,2} = 0$, and $X_{9,3} = 1$ if the region is W.

The MLR full model can be written as following:

$$Y = \beta_0 + \beta_4 X_4 + \beta_{9,1} X_{9,1} + \beta_{9,2} X_{9,2} + \beta_{9,3} X_{9,3} + \beta_{4*9,1} X_4 * X_{9,1} + \beta_{4*9,2} X_4 * X_{9,2} + \beta_{4*9,3} X_4 * X_{9,3}$$

The response function can be specifically defined based on the 4 levels of the categorical variable, region. For the baseline (NC): $Y = \beta_0 + \beta_4 X_4$ which is estimated as $b_0 + b_4 X_4$

For Level 2 (NE): $Y = \beta_0 + \beta_{9,1} + (\beta_4 + \beta_{4*9,1})X_4$
For Level 3 (S): $Y = \beta_0 + \beta_{9,2} + (\beta_4 + \beta_{4*9,2})X_4$
For Level 4 (W): $Y = \beta_0 + \beta_{9,3} + (\beta_4 + \beta_{4*9,3})X_4$

(a) [4 pts] Obtain the model summary and ANOVA table using R. Then, using the model summary, write the estimated response function for each level of the four regions (1 pts each).

- Hint1: There are two ways to do this in R. You can manually create dummy variable columns; or use the factor(region) in the lm function: lm(Y~infection+factor(Region)+infection*factor(Region)). Although the factor() method can generate the summary for the full model quicky, it is not as flexible as the dummy variable method to evaluate the parameters in the model.
- Hint2: You may use the relevel() function in R to change the baseline category. For example, if you want to use the second, or the NE category as the baseline, do data$region<-relevel(data$region,2). This is useful when you are not comparing the mean response value of Y in a level to that in the baseline level.

(b) [4 pts] Consider the MLR model $\mathrm{lm}(Y \sim \mathrm{infection} * \mathrm{Region})$ instead of $\mathrm{lm}(Y \sim \mathrm{infection} * \mathrm{as.factor}(\mathrm{Region}))$ in R, which causes R to treat "region" as a continuous variable. Compare the model summary and ANOVA output for this incorrect model to those in the previous questions. Be sure to specify in detail the differences between the two models, particularly in terms of the beta coefficients, MSE and dfE.

Problem 4 (24 pts)

What is the implication of centering X variable to simply the mean response prediction? If we center the X variable,

$$X' = (X - \bar{X}), \text{ then } X' = 0 \text{ when } X = \bar{X}$$

We can then fit the model $Y \sim \beta_0 + \beta_1 X'$. The mean response prediction at the mean level ($X = \bar{X}$) simplifies to $Y_h = \beta_0$, because there is only the intercept value in the equation.

Now center the infection variable and denote the centered variable as $X_{4(\mathrm{center})}$, where $X_{4(\mathrm{center})} = X_4 - \bar{X}_4$ (infection minus the mean of infection). Then $X_{4(\mathrm{center})} = 0$ whenever $X_4 = \bar{X}_4$ (whenever it takes on the average value).

Refit the MLR model with $X_{4(\mathrm{center})}$ (centered infection), X_9 (region, represented with three dummy variables $X_{9,1}, X_{9,2}$, and $X_{9,3}$), and the interaction between them.

(a) [6 pts] Predict the mean response value, Y_h when infection takes the average value and fill in the blanks. The purpose of this question is to interpret the meaning of $\beta_0, \beta_4, \beta_{9,1}, \beta_{9,2}, \beta_{9,3}, \beta_{4*9,1}, \beta_{4*9,2}, \beta_{4*9,3}$ when the predictor takes the average value.

(i) The average Y for NC is represented by β_0 and has a predicted value of _____.

(ii) The average Y for NE is different from NC by $\beta_{9,1}$ and has a predicted value of _____.

(iii) The average Y for S is different from NC by _____ (which beta?) and has a predicted value of _____.

(iv) The average Y for W is different from NC by _____ (which beta?) and has a predicted value of _____.

(b) [14 pts] Now consider the impact of infection on Y.

(i) The impact of infection on Y is $\beta_{4(\text{center})}$ for NC. This impact is _____(significant/insignificant), because_____.

(ii) The impact of infection on Y for NE is different from NC by $\beta_{4*9,1}$ and has a predicted value of _____, the difference is _____(significance/insignificance), because_____.

(iii) The impact of infection on Y for S is different from NC by _____ (which beta?) and has a predicted value of _____. the difference is _____(significance/insignificance), because_____.

The impact of infection on Y for W is different from NC by _____ (which beta?) and has a predicted value of _____. the difference is _____(significance/insignificance), because_____.

(c) [4 pts] Based on the result in part (b), can you judge whether the impact of infection on Y for NE is significantly different from for S? If not, find a way to compare and then conclude.

Problem 5 (6 pts)

Perform a hypothesis test to see whether the interaction effect between infection and region on Y is significant. Define the hypothesis with appropriate notation. Then complete the hypothesis with a GLT F-test. Specifically, define the full model, reduced model, compute the test statistic, critical value, p-value, and state the conclusion.

SEE LAST PAGE FOR PROBLEM 6

Problem 6 (15 pts)

Consider the life expectancy data, and the MLR model given by $Y \sim X_1 + X_2 + X_3$.

Implement the best subsets regression algorithm. Look at all 7 possible models (all but the one with just an intercept). Then answer each of the following questions, providing R output to support your answers:

(a) [2 pts] Which model is best according to R^2_{adj}?

(b) [2 pts] Which model is best according to Cp?

(c) [2 pts] Which model is best according to AIC_p?

(d) [2 pts] Which model is best according to BIC_p?

(e) [2 pts] Which model is best according to $PRESS$?

(f) [5 pts] Based on your answers in parts (a) through (e), which model do you believe is best? Explain the reasoning behind your answer.