Biometry 755 Homework 2 KEY

Use the SENIC data to conduct a multiple linear regression of nosocomial infection risk on hospital average length of stay (LOS), routine culturing ratio (CULT), routine chest X-ray ratio (XRAY), and the percent of available facilities and services (FACS). Then answer the following questions.

1. *Use the fitted model to estimate the risk of nosocomial infection for an 'average' hospital, that is, for a hospital with values of LOS, CULT, XRAY and FACS equal to the average as estimated from the data.*

   First, we use the following code to generate a multiple linear regression model of infection risk given the predictors of interest:

```
proc reg data = one;
    model infrisk = los cult xray facs;
run;
quit;
```

   Our resulting regression model is:

   INFRISK = -0.0636 + 0.188(LOS)+0.0465(CULT)+0.0121(XRAY)+0.0205(FACS)

   Then, we use the following code to generate the mean values for these predictor variables:

```
proc means data = one;
     var los cult xray facs;
run;
```

   We get the following means:

   LOS = 9.648
   CULT = 15.793
   XRAY = 81.628
   FACS = 43.159

   Substituting in the mean values to the regression equation yields:

   INFRISK = -0.0636 + 0.188(9.648) + 0.0465(15.793) + 0.0121(81.628) + 0.0205(43.159)
   = 4.355

   Thus, for the average hospital, the risk of nosocomial infection is 4.355%.

2. *What is the value of the model R^2 and what is its interpretation?*

   For this regression model, the value for $R^2 = 0.5161$. This means that 51.61% of the total variability in the risk of nosocomial infection is explained by its linear association with length of stay, culturing ratio, routine chest x-ray ratio, and percent of available facilities and services.

3. *What are the full and reduced models for the overall F-test of the regression? What can we conclude based on the results of this test?*

The full regression model includes all predictor variables:

$$INFRISK = \beta_0 + \beta_1 (LOS) + \beta_2 (CULT) + \beta_3 (XRAY) + \beta_4 (FACS) + \varepsilon$$

The reduced model includes only an intercept:

$$INFRISK = \beta_0 + \varepsilon$$

Our hypotheses for the overall F-test are as follows:

H_0: $\beta_{LOS}, \beta_{CULT}, \beta_{XRAY}, \beta_{FACS} = 0$

H_1: at least one of $\beta_{LOS}, \beta_{CULT}, \beta_{XRAY}, \beta_{FACS} \neq 0$

Our p-value for the overall F-test for our regression model is < 0.0001, which is significant at the 5% level of significance. Thus, we reject the null hypothesis (reduced model) in favor of the full model. We conclude that AT LEAST ONE of the predictors in the full model is a significant predictor.

4. *Does LOS contribute significantly to a model already containing CULT, XRAY and FACS? Justify your answer.*

H_0: $\beta_{CULT} = 0$, given that CULT, XRAY, and FACS are in the model

H_1: $\beta_{CULT} \neq 0$, given that CULT, XRAY, and FACS are in the model

We refer to the regression model with all four of these predictors, and look at the p-value associated with LOS. Since this p-value is significant at the 5% level (p-value = 0.0008), we reject the null hypothesis and conclude that LOS does significantly contribute to a model already containing CULT, XRAY and FACS.

5. *Interpret the slope parameter for LOS.*

For every unit increase in length of stay, the risk of nosocomial infection increases by 0.18%, adjusting for routine culturing ratio, routine chest X-ray ratio, and the percent of available facilities and services.

6. *Use the 'test' statement in PROC REG to conduct a multiple partial F-test of the variables XRAY and FACS in the presence of LOS and CULT. Then answer the following questions.*
   a. *State the null and alternative hypothesis for this test in terms of a reduced and full model.*

H_0: $INFRISK = \beta_0 + \beta_1 (LOS) + \beta_2 (CULT) + \varepsilon$

H_1: $INFRISK = \beta_0 + \beta_1 (LOS) + \beta_2 (CULT) + \beta_3 (XRAY) + \beta_4 (FACS) + \varepsilon$

An alternative way to express these hypotheses is:

H_0: $\beta_{XRAY}, \beta_{FACS} = 0$ (while $\beta_{LOS}, \beta_{CULT} \neq 0$)

H_1: at least one of $\beta_{XRAY}, \beta_{FACS} \neq 0$ (while $\beta_{LOS}, \beta_{CULT} \neq 0$)

b. *Use PROC REG to calculate SSR for the reduced and full models stated in part a.*

First we look at the full model, and determine that SSR(full) = 103.938.
Then, we run the reduced model with the following code:

```
proc reg data = one;
    model infrisk = los cult;
run;
quit;
```

From this reduced model, we determine SSR(reduced) = 90.702.

c. *Using the results from part b, demonstrate how the numerator and denominator for the multiple partial F test are generated.*

$$F = \frac{\frac{SSR(full) - SSR(reduced)}{deg\ freedom}}{MSE(full)} = \frac{\frac{103.938 - 90.702}{2}}{0.902} = 7.337$$

Note that the # deg of freedom for the F-test is equivalent to the number of variables eliminated from the full model to form the reduced model.

d. *State your conclusion based on the results of the multiple partial F test.*

The p-value for the partial F-test is 0.0010, which is significant at the 5% level, so we reject the null hypothesis. Thus, we conclude that the full model is a better fit than the reduced model, meaning that AT LEAST ONE of XRAY and FACS is a significant predictor when added to a model containing LOS and CULT.