BIS 505b – Problem Set 3 Due 11:59 PM Eastern, 3/28/2021

Instructions: Follow the homework instructions outlined in the syllabus. Round your answers to 2 decimal places. Perform all tests at the α = 0.05-level and follow the steps of hypothesis testing.

Assignment

Question 1: The authors of a paper looking at the impact of weight-bearing activity during youth on bone mass used a multiple linear regression model to describe the relationship between:

y = Bone mineral density (g/cm^3)
x_1 = Body weight (kg)
x_2 = A measure of weight-bearing activity, with higher values indicating greater activity

a. [5] The authors concluded that both body weight and weight-bearing activity were important predictors of bone mineral density and that there is no significant interaction between body weight and weight-bearing activity. Write the general form of the multiple regression function that is consistent with this description. (i.e., $y = \cdots + \epsilon$)

$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

b. [5] The value of the coefficient of body weight in the multiple regression function given in the paper is 0.587. Interpret this value.

$\beta_1 = 0.587$: Holding weight bearing activity constant, a one-kg increase in body weight is associated with a 0.587 g/cm^3 increase in mean bone mineral density.

c. [5] Use the coefficient reported in part (b) to determine the impact of a 2-kg increase in body weight on bone mineral density, controlling for the measure of weight-bearing activity.

Holding weight bearing activity constant, a 2-kg increase in body weight is associated with a 1.174 g/cm^3 increase in mean bone mineral density.

Question 2: A study was conducted to determine whether infection surveillance and control programs have reduced the rates of hospital-acquired infection in U.S. hospitals. The data analyzed were a random sample of n = 113 hospitals. A multiple linear regression model was used to describe the relationship between:

y = Average length of stay of all patients in hospital (days)
x_1 = Average age of patients (years)
x_2 = Average daily census: average number of patients in hospital per day during study period (hundreds of patients)
x_3 = Affiliation of hospital with medical school (0 = No, 1 = Yes)

x_4, x_5, x_6 = Geographic region (x_4 = 1 if NE and 0 otherwise; x_5 = 1 if NC and 0 otherwise; x_6 = 1 if S and 0 otherwise). In this coding, W is the reference category.

Results are shown below.

Table 1. ANOVA Table [$x_1, x_2, x_3, x_4, x_5, x_6$ included in model of average length of stay]

	SS	df	MS	F	p-value
Model	185.67494	6			
Error	223.53544	106			

Table 2. Least Squares Results [$x_1, x_2, x_3, x_4, x_5, x_6$ included in model of average length of stay]

	Estimate	Standard Error	t	p-value
Intercept	3.08827	1.73997	1.775	
AGE x_1	0.08187	0.03177		
DAILYCENSUS x_2	0.49310	0.11561		
AFFILIATION x_3	0.23772	0.49671		
NE x_4	2.47809	0.46339		
NC x_5	1.27053	0.45343		
S x_6	0.75886	0.44322		

a. [5] Using Table 2, report the fitted least squares regression equation? (i.e., $\hat{y} = \cdots$)

$\hat{y} = 3.09 + 0.08x_1 + 0.49x_2 + 0.24x_3 + 2.48x_4 + 1.27x_5 + 0.76x_6$

b. [20] Interpret the values of b_1, b_2, b_3, and b_4 given in Table 2.

$b_1 = 0.08$: Adjusted for daily census, affiliation and geography, a one-year increase in average age of patients is associated with 0.08 days increase in average length of stay of all patients in hospital.

$b_2 = 0.49$: Adjusted for age, affiliation and geography a one-hundreds of patients increase is associated with 0.49 days increase in average length of stay of all patients in hospital.

$b_3 = 0.24$: Adjusted for age, daily census and geography, average length of stay of all patients in hospital with Affiliation of medical school is 0.24 days higher than that in hospital without Affiliation of medical school.

$b_4 = 2.48$: Adjusted for age, daily census and affiliation, average length of stay of all patients in hospital in NE geographic region is 2.48 days higher than that in W geographic region.

c. [5] Just as in simple linear regression, we can use the fitted model to estimate the mean response for given values x. What is the expected average length of stay in a hospital where mean age is 55, average daily census is 200 patients, there is no university affiliation, and the hospital is in the West? Assume these values are in the range of the data used to fit the model.
Plug in $x_1 = 55, x_2 = 200, x_3 = 0, x_4 = 0, x_5 = 0, x_6 = 0$
$\hat{y} = 3.09 + 0.08x_1 + 0.49x_2 + 0.24x_3 + 2.48x_4 + 1.27x_5 + 0.76x_6$

$= 3.09 + 0.08 \times 55 + 0.49 \times 200 = 105.49$

The expected average length of stay in a hospital for given values x is 105.49 days.

d. [10] Perform an F-test to determine if there is a useful linear relationship between y and at least one of $x_1, \dots, x_6$

(1) State the null and alternative hypotheses
H_0: $\beta_1 = \beta_2 = \cdots = \beta_6 = 0$
H_1: At least one $\beta_j \neq 0$ for $j = 1, \cdots, 6$

(2) Specify the significance level, α = 0.05
(2.5) ANOVA assumptions are met.

(3) Compute the test statistic

ANOVA table for the multiple linear regression model

	SS	df	MS	F	p-value
Model	185.67494	6	30.95	14.67	< 0.001
Error	223.53544	106	2.11		
Total	409.21	112			

$MSM = \frac{SSM}{df_1} = \frac{185.67}{6}$ =30.95
$MSE = \frac{SSE}{df_2} = \frac{223.54}{106} = 2.11$
$F = \frac{MSM}{MSE} = \frac{30.95}{2.11} = 14.67; F \sim F(6{,}106)$

(4) Generate the decision rule
Given α = 0.05,
Reject H_0 if $F \geq F_{1-\alpha}(df_1, df_2) = F_{0.95}(6{,}106) = 2.19$ or if $p \leq 0.05$

(5) Draw a statistical conclusion, and state the conclusion in words in the context of the problem.
$F = 14.67 \geq 2.19 \;\; \rightarrow Reject\ H_0$
$or\ p = P(F \geq 14.67) < 0.001 \rightarrow Reject\ H_0$
Conclusion: There is evidence to reject H_0 and conclude there is a significant linear relationship between the average length of stay in a hospital and at least one explanatory variable (age, daily census, affiliation, geography) ($p < 0.001$).

e. [15] Complete the “t” and “p-value” columns in Table 2. Interpret the result of the test of β_1 (the slope associated with AGE), β_3 (the slope associated with AFFILIATION), and β_4 (the slope associated with the dummy variable for NE geographic region) in the context of the research question.

Table 2

	Estimate	Standard Error	t	p-value
Intercept	3.08827	1.73997	1.775	0.08
AGE x_1	0.08187	0.03177	2.577	0.01
DAILYCENSUS x_2	0.49310	0.11561	4.265	0.00
AFFILIATION x_3	0.23772	0.49671	0.479	0.63
NE x_4	2.47809	0.46339	5.348	0.00
NC x_5	1.27053	0.45343	2.802	0.01
S x_6	0.75886	0.44322	1.712	0.09

β_1 (the slope associated with AGE): We reject H_0 and conclude that there is a significant linear association between age and average length of stay in a hospital ($b_1 = 0.08, p = 0.01$) after daily census, affiliation and geography are controlled.

β_3 (the slope associated with AFFILIATION): We fail to reject H_0 and conclude that there is not a significant linear association between affiliation and average length of stay in a hospital ($b_3 = 0.24, p = 0.63$) after age, daily census and geography are controlled.

β_4 (the slope associated with the dummy variable for NE geographic region): We reject H_0 and conclude that there is a significant difference in average length of stay in a hospital ($b_4 = 2.48, p < 0.01$) in NE geographic region vs. W geographic region after age, daily census and affiliation are controlled.

f. [5] Report the 95% confidence interval for β_1. Does this CI support your conclusion from part (e)?

95% CI for β_1: $b_1 \pm t_{106,0.975} \times s_{b_1} = 0.08 \pm 1.98 \times 0.03 = (0.02, 0.14)$
Since the CI does not include null, it supports my conclusion from part (e) that we reject H_0.

g. [10] Calculate R^2 and the adjusted R^2. How does this compare to R^2? Will the adjusted R^2 always increase if more independent variables are added to the model? Explain.

Table 2. ANOVA table for the multiple linear regression model

	SS	df	MS	F	p-value
Model	185.67494	6	30.95	14.67	< 0.001
Error	223.53544	106	2.11		
Total	409.21	112			

$$R^2 = \frac{SSM}{SST} = \frac{185.67}{409.21} = 0.45$$
$$R_a{}^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)} = \frac{223.54/106}{409.21/112} = 0.58 > R^2$$

The adjusted R^2 $is\ larger\ than\ R^2$. The adjusted R^2 will not always increase if more independent variables are added to the model because it compensates for added complexity of a larger model by accounting for the number of predictors in the model.

A smaller model was fit that did not include geographic region (and thus did not include x_4, x_5, x_6 in the regression model). The ANOVA table from this model is shown in Table 3.

Table 3. ANOVA Table [x_1, x_2, x_3 included in model of average length of stay]

	SS	df	MS	F	p-value
Model	111.43521	3	37.14507	13.60	
Error	297.77517	109	2.73188		

h. [15] Perform an F-test to determine if geographic region is an important predictor in this model.

(1) State the null and alternative hypotheses
H_0: $\beta_4 = \beta_5 = \beta_6 = 0$
H_1: $\beta_4, \beta_5, \beta_6$ not all 0

(2) Specify the significance level, α = 0.05

(2.5) ANOVA assumptions are met.

(3) Compute the test statistic

ANOVA table for the full model

	SS	df	MS	F	p-value
Model	185.67494	6	30.95	14.67	< 0.001
Error	223.53544	106	2.11		
Total	409.21	112			

ANOVA table for the reduced model

	SS	df	MS	F	p-value
Model	111.43521	3	37.14507	13.60	
Error	297.77517	109	2.73188		

$$F_0 = \frac{\frac{SSM(F)-SSM(R)}{\#\,params\,tested\,under\,H_0}}{\frac{SSE(F)}{df_2(F)}} = \frac{\frac{185.67-111.44}{3}}{\frac{223.54}{106}} = \frac{24.74}{2.11} = 11.73;\ F_0 \sim F(3{,}106)$$

(4) Generate the decision rule

Given α = 0.05,

Reject H_0 if $F_0 \geq F_{1-\alpha}(df_1, df_2) = F_{0.95}(3{,}106) = 2.69$ or if $p \leq 0.05$

(5) Draw a statistical conclusion, and state the conclusion in words in the context of the problem.

$F = 11.73 \geq 2.69 \ \rightarrow Reject\ H_0$
$or\ p = P(F \geq 11.73) < 0.001 \rightarrow Reject\ H_0$

Conclusion: There is evidence to reject H_0 and conclude there is at least one significant difference in the effect of age, daily census and affiliation on the average length of stay in a hospital in the NE vs. W or NC vs. W or NE vs. W or S vs. W ($p < 0.001$), that is, geographic region is an important predictor in this model.