**STAT 501 – Homework 5 – Fall 2015 – Due Sep 27**

**Instructions**: Use Word to type your answers within this document. Then, submit your answers in the appropriate dropbox in ANGEL by the due date. The point distribution is located next to each question.

1. (**3+4+6+6+6 = 25 points**) Use the “Crop Yield” dataset for this problem. The *y*-variable is Yield = the yield of a crop during a growing season. There are two *x*-variables. The variable IngredA = the amount of ingredient A put into a soil treatment used to help grow the crop and the variable IngredB = the amount of ingredient B put into a soil treatment. Twelve combinations of IngredA (which had 3 different levels) and IngredB (which had 4 different levels) were considered. Two fields were treated with each combinations so the sample size of the dataset is *n* = 24.

1. Using statistical software, plot IngredA versus IngredB. This is a plot of the two *x*-variables. On the basis of this plot, describe in words the amount of correlation between the two *x*-variables?
2. Now, graph Yield versus IngredA and separately graph Yield versus IngredB. Describe the important features of each plot. For instance, are the relationships linear, are there any outliers, which *x*-variable is the stronger predictor, and so on?
3. Fit a simple linear regression model with *y* = Yield and *x* = IngredA.
   1. What is the value of the slope? Write a sentence that interprets this slope.
   2. What is the value of *R*2 for this regression?
   3. On the basis of this regression, can we say that there is a statistically significant linear relationship between Yield and IngredA? Explain why or why not.
4. Fit a simple linear regression model with *y* = Yield and *x* = IngredB.
   1. What is the value of the slope? Write a sentence that interprets this slope.
   2. What is the value of *R*2 for this regression?
   3. On the basis of this regression, can we say that there is a statistically significant linear relationship between Yield and IngredB? Explain why or why not.
5. Fit a multiple linear regression model with *y* = Yield using predictors *x*1 = IngredA and *x*2 = IngredB.
   1. What are the values of the coefficients that multiply the two *x*-variables? Explain why these are the same values found in the previous two parts of this question.
   2. What is the value of *R*2 for this regression? Verify that this *R*2 is the sum of the *R*2 values for the simple regressions done in the previous two parts of this question, and explain why this relationship holds here.
   3. On the basis of this regression, can we say that there is a statistically significant relationship between Yield and IngredA? Explain why or why not.

2. (**4+4+4+8+6+4 = 30 points**) Use the “Hospital Infections” dataset. This is data from *n* = 113 hospitals in the United States. The *y*-variable is InfctRsk = percentage of patients who get an infection while in the hospital. The four *x*-variables are Stay = average length of patent stay in each hospital, Xrays = a measure of how often X-rays are given in the hospital, Beds = number of beds in the hospital, and Census = average daily number of patients in the hospital. There are two suspect data observations with very high values of Stay (ID 47 and 112). For the purpose of this exercise we’ll first remove these two points. In Minitab select Data > Delete Rows, then type “47, 112” in “Rows to delete” and “ID-Facilities” in “Columns from which to delete these rows.” **IT IS IMPORTANT THAT YOU DO THIS FIRST.**

1. Draw a scatterplot showing the relationship between the two *x*-variables Beds and Census. Briefly describe the correlation.
2. Fit a simple linear regression model with *y* = InfctRsk and *x* = Beds. Is there a statistically significant linear relationship between the two variables? Explain.
3. Fit a simple linear regression model with *y* = InfctRsk and *x* = Census. Is there a statistically significant linear relationship between the two variables? Explain.
4. Fit a multiple linear regression model with *y* = InfctRsk using the four *x*-variables Stay, Xrays, Beds and Census as the predictors.
   1. In the Analysis of Variance table, what is the *p*-value? On the basis of this *p*-value, what can we conclude?
   2. On the basis of the *p*-value for testing the statistical significance of Beds, what can we conclude?
   3. On the basis of the *p*-value for testing the statistical significance of Census, what can we conclude?
   4. What is the most likely reason that the significance results for Beds and Census in this regression differ from what we found in the simple linear regressions of the previous two parts.
5. Of the two variables Beds and Census, Beds may be a slightly weaker predictor so let us drop that variable. Fit a multiple linear regression model with *y* = InfctRsk using the three predictors Stay, Xrays and Census.
   1. Explain whether each variable is a significant predictor within this model.
   2. What is the value of MSE for this model? Compare this value to the MSE for the 4-variable model examined in the previous part and explain why this is evidence that the 3-variable model is preferable.
6. For the 3-variable model of the previous part, create a plot of residuals versus fits, a histogram of the residuals, and write an interpretation of both of these plots.

3. (**4 + 7x3 + 4 = 29 points**) The table below gives ten observations on X = Number of times cartons were transferred from one aircraft to another over the shipment route and Y = the number of ampules found to be broken upon arrival.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 1 | 0 | 2 | 0 | 3 | 1 | 0 | 1 | 2 | 0 |
| Y | 16 | 9 | 17 | 12 | 22 | 13 | 8 | 15 | 19 | 11 |

For the above data,

1. Find the fitted regression equation.
2. Use matrix methods to obtain
3. (**X**T**X**)-1 (a 2 by 2 matrix),
4. **b = (XTX)–1XT*Y*** (a 2 by 1 column-vector),
5. **e** (a 10 by 1 column-vector),
6. SSE (a scalar),
7. se2(**b**) (a 2 by 2 matrix),
8. when Xh = 2 (a scalar),
9. se2() when Xh = 2 (a scalar).
10. Use se2(**b**) to calculate Correlation(b0, b1) (a scalar).

You may find the document, “Matrix Calculations in Minitab,” helpful.

4. (**8x2 = 16 points**) In a small-scale experimental study of the relation between degree of brand liking (Y) and moisture content (X1) and sweetness (X2) of the product, the following results were obtained from the experiment based on a completely randomized design (data are coded):

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X1 | 4 | 4 | 4 | 4 | 6 | 6 | 6 | 6 | 8 | 8 | 8 | 8 | 10 | 10 | 10 | 10 |
| X2 | 2 | 4 | 2 | 4 | 2 | 4 | 2 | 4 | 2 | 4 | 2 | 4 | 2 | 4 | 2 | 4 |
| Y | 64 | 73 | 61 | 76 | 72 | 80 | 71 | 83 | 83 | 89 | 86 | 93 | 88 | 95 | 94 | 100 |

A multiple linear regression model was fit with the following results:

Analysis of Variance

Source DF Adj SS Adj MS F-Value P-Value

Regression 2 1872.70 936.35 129.08 0.000

X1 1 1566.45 1566.45 215.95 0.000

X2 1 306.25 306.25 42.22 0.000

Error 13 94.30 7.25

Total 15 1967.00

Model Summary

S R-sq R-sq(adj) R-sq(pred)

2.69330 95.21% 94.47% 92.46%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 37.65 3.00 12.57 0.000

X1 4.425 0.301 14.70 0.000 1.00

X2 4.375 0.673 6.50 0.000 1.00

Regression Equation

Y = 37.65 + 4.425 X1 + 4.375 X2

Use the results to complete the following sentences:

1. \_\_\_\_\_\_\_\_\_\_\_\_\_ % of the variation in degree of brand liking (Y) is accounted for by moisture content (X1) and sweetness (X2).
2. The estimated standard deviation of the regression errors is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.
3. The F-statistic of \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ with a p-value of \_\_\_\_\_\_\_\_\_\_\_\_\_ indicates that the model containing X1 and X2 is more useful in predicting Y than not taking into account the two predictors.
4. The t-statistic of \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ with a p-value of \_\_\_\_\_\_\_\_\_\_\_\_\_ indicates that the slope parameter for X1 is significantly different from 0 in this model.
5. The t-statistic of \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ with a p-value of \_\_\_\_\_\_\_\_\_\_\_\_\_ indicates that the slope parameter for X2 is significantly different from 0 in this model.
6. We estimate that E(Y) increases by \_\_\_\_\_\_\_\_\_\_\_\_\_ unit(s) when X1 increases by \_\_\_\_\_\_\_\_\_\_\_\_\_\_ unit(s) and X2 is held constant.
7. We estimate that E(Y) increases by \_\_\_\_\_\_\_\_\_\_\_\_\_ unit(s) when X2 increases by \_\_\_\_\_\_\_\_\_\_\_\_\_\_ unit(s) and X1 is held constant.
8. We predict that the degree of brand liking when moisture content is 7 and sweetness is 3 is \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.