**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?



The plot between IngredA versus IngredB is the classical appearance of a scatter plot for the experimental conditions. The plot suggests that there is no correlation at all between the two variables.

1. 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?





Main features:

* Yield and IngredA seemingly have a curvilinear relationship
* Yield and IngredB seemingly have a linear relationship
* IngredB appears to be a stronger predictor than IngredA

**For both plots, the pattern looks to be linear with a positive association and no major outliers. Ingredient B is the stronger predictor.**

1. 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.

The regression output from minitab:

**Regression Analysis: Yield versus IngredA**

Analysis of Variance

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

Regression 1 49.703 49.7025 3.30 0.083

IngredA 1 49.703 49.7025 3.30 0.083

Error 22 331.247 15.0567

Lack-of-Fit 1 0.701 0.7008 0.04 0.835

Pure Error 21 330.546 15.7403

Total 23 380.950

Model Summary

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

3.88029 13.05% 9.09% 0.00%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 27.75 2.10 13.24 0.000

IngredA 1.763 0.970 1.82 0.083 1.00

Regression Equation

Yield = 27.75 + 1.763 IngredA

1. The value of the slope is 1.763. It is interpreted as: the mean Yield will increase by 1.763 for every additional unit of IngredA.
2. Value of R2 is 13.05%. We can say that 13.05% of the variation in the Yield is reduced by taking into account IngredA. Or, we can say that 13.05% of the variation in Yield is 'explained by' IngredA.
3. On the basis of this regression, we can say that there is **not** a statistically significant linear relationship between Yield and IngredA. We can see from the regression that the slope coefficient has the p-value = 0.08 > alpha (assume 0.05) and therefore we can’t reject the null hypothesis. We don’t have sufficient evidence to reject the null hypothesis that slope = 0.
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.

The regression output from minitab:

**Regression Analysis: Yield versus IngredB**

Analysis of Variance

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

Regression 1 184.76 184.760 20.72 0.000

IngredB 1 184.76 184.760 20.72 0.000

Error 22 196.19 8.918

Lack-of-Fit 2 11.86 5.932 0.64 0.536

Pure Error 20 184.33 9.216

Total 23 380.95

Model Summary

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

2.98625 48.50% 46.16% 37.34%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 25.07 1.49 16.79 0.000

IngredB 2.482 0.545 4.55 0.000 1.00

Regression Equation

Yield = 25.07 + 2.482 IngredB

1. The value of the slope is 2.482. It is interpreted as: the mean Yield will increase by 2.482 for every additional unit of IngredB
2. Value of R2 is 48.5%. We can say that 48.5% of the variation in the Yield is reduced by taking into account IngredB. Or, we can say that 48.5% of the variation in Yield is 'explained by' IngredB
3. On the basis of this regression, we can say that there is a statistically significant linear relationship between Yield and IngredB. We can see from the regression that the slope coefficient has the p-value < 0.001 < alpha (assume 0.05) and therefore we reject the null hypothesis. We have sufficient evidence to reject the null hypothesis that slope = 0 and conclude the alternative.
4. 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.

The regression output from minitab:

**Regression Analysis: Yield versus IngredA, IngredB**

Analysis of Variance

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

Regression 2 234.46 117.231 16.81 0.000

IngredA 1 49.70 49.703 7.13 0.014

IngredB 1 184.76 184.760 26.49 0.000

Error 21 146.49 6.976

Lack-of-Fit 9 62.51 6.946 0.99 0.493

Pure Error 12 83.97 6.998

Total 23 380.95

Model Summary

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

2.64113 61.55% 57.88% 48.92%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 21.54 1.87 11.53 0.000

IngredA 1.763 0.660 2.67 0.014 1.00

IngredB 2.482 0.482 5.15 0.000 1.00

Regression Equation

Yield = 21.54 + 1.763 IngredA + 2.482 IngredB

1. The value of the coefficients is 1.763 (for IngredA) and 2.482 (for IngredB). The values are the same because there is no correlation (interaction) between the two predictors i.e. IngredA and IngredB.
2. Value of R2 is 61.55%. The R2 is sum of the two R2 computed above:

48.5 + 13.05 = 61.55

We can say that 61.55% of the variation in the Yield is reduced by taking into account IngredA and IngredB. Or, we can say that 61.55% of the variation in Yield is 'explained by' IngredA and IngredB.

The relationship holds here because there is no correlation (interaction) between the two predictors i.e. IngredA and IngredB.

1. On the basis of this regression, we can say that there is a statistically significant linear relationship between Yield and IngredA. We can see from the regression that the slope coefficient has the p-value < 0.014 < alpha (assume 0.05) and therefore we reject the null hypothesis. We have sufficient evidence to reject the null hypothesis that slope = 0 and conclude the alternative.

This is in contrast to the result above from the simple linear regression between Yield and IngredA. The variable IngredA is not statistically significant in the simple regression, but it is in the multiple regression. This is a benefit of doing a multiple regression. By putting both variables into the equation, we have reduced the standard deviation of the residuals. This in turn reduces the standard errors of the coefficients, leading to greater t-values and smaller p-values.

**Yes, ingredient A is significant in this model. The p-­‐value is 0.014 for testing that the associated beta is 0. (Note: In the multiple linear regression model, the inclusion of both variables reduces MSE and that affects the standard errors of all coefficients. Thus, ingredient A was able to achieve significance in the multiple linear regression when it could not in the simple linear regression.)**

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.



The correlation is very high and +ve

1. Fit a simple linear regression model with *y* = InfctRsk and *x* = Beds. Is there a statistically significant linear relationship between the two variables? Explain.

The minitab output is:

**Regression Analysis: InfctRsk versus Beds**

Analysis of Variance

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

Regression 1 23.05 23.051 14.67 0.000

Beds 1 23.05 23.051 14.67 0.000

Error 109 171.22 1.571

Lack-of-Fit 94 150.70 1.603 1.17 0.383

Pure Error 15 20.51 1.368

Total 110 194.27

Model Summary

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

1.25332 11.87% 11.06% 8.62%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 3.716 0.198 18.79 0.000

Beds 0.002457 0.000641 3.83 0.000 1.00

Regression Equation

InfctRsk = 3.716 + 0.002457 Beds

There is statistically significant linear relationship since the p-value for coefficient of slope is < 0.001 < alpha (assume 0.05) and therefore we can reject the null hypothesis that the population slope equals 0.

Alternatively, we also see that the p-value associated with the F test is also less than alpha (assume 0.05) and therefore we can state that there is statistically significant linear relationship

1. Fit a simple linear regression model with *y* = InfctRsk and *x* = Census. Is there a statistically significant linear relationship between the two variables? Explain.

The minitab output is:

**Regression Analysis: InfctRsk versus Census**

Analysis of Variance

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

Regression 1 25.92 25.920 16.78 0.000

Census 1 25.92 25.920 16.78 0.000

Error 109 168.35 1.544

Lack-of-Fit 92 140.48 1.527 0.93 0.609

Pure Error 17 27.87 1.639

Total 110 194.27

Model Summary

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

1.24277 13.34% 12.55% 10.38%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 3.697 0.193 19.17 0.000

Census 0.003374 0.000823 4.10 0.000 1.00

Regression Equation

InfctRsk = 3.697 + 0.003374 Census

There is statistically significant linear relationship, since the p-value for coefficient of slope is < 0.001 < alpha (assume 0.05) and therefore we can reject the null hypothesis that the population slope equals 0.

Alternatively, we also see that the p-value associated with the F test is also less than alpha (assume 0.05) and therefore we can state that there is statistically significant linear relationship

1. 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.

The minitab output is:

**Regression Analysis: InfctRsk versus Stay, Xray, Beds, Census**

Analysis of Variance

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

Regression 4 77.492 19.3731 17.59 0.000

Stay 1 16.088 16.0876 14.60 0.000

Xray 1 14.445 14.4453 13.11 0.000

Beds 1 0.004 0.0044 0.00 0.950

Census 1 0.327 0.3266 0.30 0.587

Error 106 116.776 1.1017

Total 110 194.268

Model Summary

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

1.04960 39.89% 37.62% 34.01%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant -0.692 0.708 -0.98 0.331

Stay 0.3157 0.0826 3.82 0.000 1.48

Xray 0.02050 0.00566 3.62 0.000 1.19

Beds -0.00019 0.00301 -0.06 0.950 31.41

Census 0.00216 0.00397 0.54 0.587 32.61

Regression Equation

InfctRsk = -0.692 + 0.3157 Stay + 0.02050 Xray - 0.00019 Beds + 0.00216 Census

1. The P-value for the analysis of variance F-test (P < 0.001) suggests that the model containing Stay, Xray, Beds and Census is more useful in predicting InfctRsk than not taking into account the predictors. **It suggests that at least one of the four predictors is significant (correction)**
2. On the basis of this regression, we can say that there is **not** a statistically significant linear relationship between InfctRsk and Beds. We can see from the regression that the slope coefficient has the p-value = 0.95 > alpha (assume 0.05) and therefore we can’t reject the null hypothesis. The variable Beds is not a useful predictor within this model that includes Stay, Xray, Census

**The p-value = 0.950 for testing the statistical significance of Beds, so we conclude that in this model, Beds is not a significant predictor.**

1. On the basis of this regression, we can say that there is **not** a statistically significant linear relationship between InfctRsk and Census. We can see from the regression that the slope coefficient has the p-value = 0.587 > alpha (assume 0.05) and therefore we can’t reject the null hypothesis. The variable Census is not a useful predictor within this model that includes Stay, Xray, Beds
2. One of the reasons to explain this difference is the high correlation between beds and census. Additionally the two scatter plots below indicate moderate +ve correlation between beds and stay & census and stay.



**The variables Beds and Census are strongly correlated so both might essentially provide about the same information for predicting InfctRsk (so both aren’t necessary in the model together).**



1. 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.

The minitab output is:

**Regression Analysis: InfctRsk versus Stay, Xray, Census**

Analysis of Variance

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

Regression 3 77.488 25.829 23.67 0.000

Stay 1 16.724 16.724 15.32 0.000

Xray 1 14.441 14.441 13.23 0.000

Census 1 6.782 6.782 6.21 0.014

Error 107 116.780 1.091

Total 110 194.268

Model Summary

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

1.04470 39.89% 38.20% 35.24%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant -0.701 0.688 -1.02 0.310

Stay 0.3166 0.0809 3.91 0.000 1.43

Xray 0.02049 0.00563 3.64 0.000 1.19

Census 0.001916 0.000769 2.49 0.014 1.23

Regression Equation

InfctRsk = -0.701 + 0.3166 Stay + 0.02049 Xray + 0.001916 Census

1. On the basis of this regression, we can say that all the variables are statistically significant at an alpha = 0.05 since all the p-values are < alpha = 0.05

However the variable census is not significant at an alpha = 0.01 since p-value = 0.014 > alpha = 0.01

1. MSE for this model (SXC) = 1.091. As compared to MSE of the previous model (SXBC) = 1.1017

MSE\_SXC = 1.091 < MSE\_SXBC = 1.1017 therby indicating that the 3-variable model is better than the 4-variable model.

**MSE = 1.091 is smaller than MSE = 1.102 for the 4-­‐variable model for this model. A low MSE is good so the 3-­‐variable model looks better. Note: In the calculus of least squares, the SSE cannot decrease when we delete a variable. However, its increase may be small relative to the change in the degrees of freedom, so the MSE can either increase or decrease.**

1. 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.



This plot looks good in that the variance is roughly the same all the way across and there are no worrisome patterns. There seems to be no difficulties with the model or data.



The histogram is roughly bell-shaped so it is an indication that it is reasonable to assume that the errors have a normal distribution.

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.

The fitted regression equation is: Yhat = 10.200 + 4.000 X

**Regression Analysis: Y versus X**

Analysis of Variance

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

Regression 1 160.000 160.000 72.73 0.000

X 1 160.000 160.000 72.73 0.000

Error 8 17.600 2.200

Lack-of-Fit 2 0.933 0.467 0.17 0.849

Pure Error 6 16.667 2.778

Total 9 177.600

Model Summary

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

1.48324 90.09% 88.85% 85.44%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 10.200 0.663 15.38 0.000

X 4.000 0.469 8.53 0.000 1.00

Regression Equation

Y = 10.200 + 4.000 X

1. Use matrix methods to obtain
2. (**X**T**X**)-1 (a 2 by 2 matrix),

Matrix M5

0.2 -0.1

-0.1 0.1

1. **b = (XTX)–1XT*Y*** (a 2 by 1 column-vector),

Matrix M7

10.2

4.0

1. **e** (a 10 by 1 column-vector),

resid

1.8

-1.2

-1.2

1.8

-0.2

-1.2

-2.2

0.8

0.8

0.8

1. SSE (a scalar),

17.6

1. se2(**b**) (a 2 by 2 matrix),

se2(b0) = Var(b0) = 2.200 \* 0.2 = 0.44

se2(b1) = Var(b1) = 2.200 \* 0.1 = 0.22

1. when Xh = 2 (a scalar),

= 10.200 + 4.000 \* 2 = 18.2

1. se2() when Xh = 2 (a scalar).

0.440

1. Use se2(**b**) to calculate Correlation(b0, b1) (a scalar).

Cov(b0, b1) = 2.200 \* -0.1 = -0.22

Corr(b0, b1) = covariance divided by product of standard errors

= -0.22 / sqrt(se2(b0)) \* sqrt(se2(b1)) = -0.7071

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. \_\_95.21\_\_\_\_\_\_\_\_\_ % 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 \_2.69330\_\_\_\_\_\_\_.
3. The F-statistic of \_\_\_\_129.08\_\_\_\_\_\_\_\_\_\_\_ with a p-value of \_\_0.000\_\_\_\_\_\_\_\_\_ 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 \_\_\_14.70\_\_\_\_\_\_\_\_\_\_\_\_\_\_ with a p-value of \_\_\_0.000\_\_\_\_\_\_\_\_ indicates that the slope parameter for X1 is significantly different from 0 in this model.
5. The t-statistic of \_\_\_\_\_6.50\_\_\_\_\_\_\_\_\_\_\_ with a p-value of \_\_\_\_0.000\_\_\_\_\_\_\_\_\_ indicates that the slope parameter for X2 is significantly different from 0 in this model.
6. We estimate that E(Y) increases by \_\_\_4.425\_\_\_\_\_\_\_\_\_\_ unit(s) when X1 increases by \_\_1\_\_\_\_\_\_\_\_\_\_\_\_ unit(s) and X2 is held constant.
7. We estimate that E(Y) increases by \_\_4.375\_\_\_\_\_\_\_\_\_\_\_ unit(s) when X2 increases by \_\_1\_\_\_\_\_\_\_\_\_\_\_\_ 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 \_\_\_\_81.75\_\_\_\_\_\_\_\_\_\_\_\_.