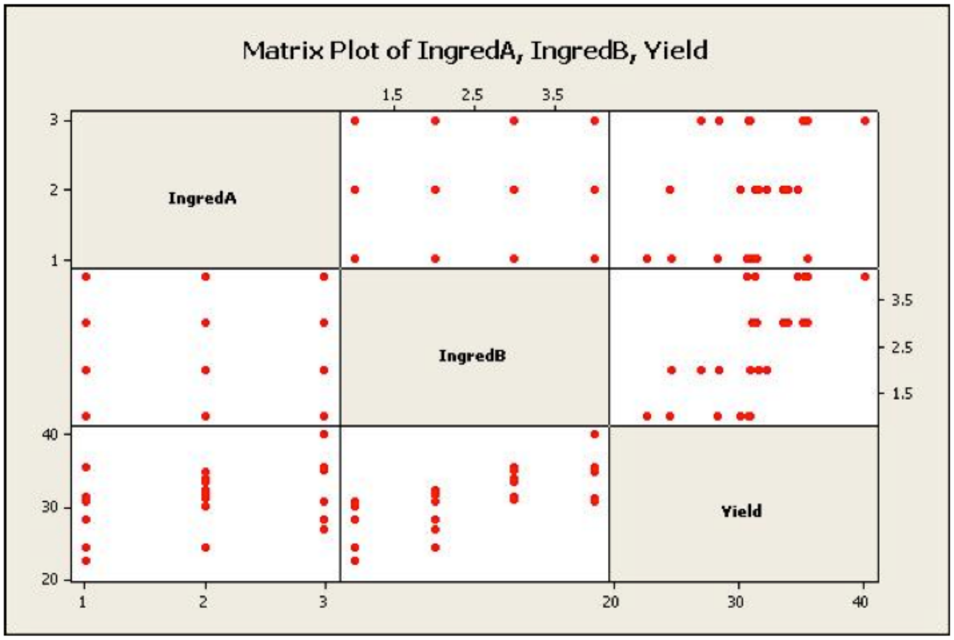
STAT 501 – Homework 5 Solutions – Fall 2015

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.

This question had parts that asked you to graph each variable versus the other variables. As a compact way to present them, I used a matrix plot in Minitab to produce the following:



This can be done by following Graph > Matrix Plot > Simple and then selecting all of the variables you wish to plot against one another.

(a) 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 observations for the two *x*-variables are on a rectangular grid. The correlation between the variables is 0.

(b) 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?

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

(c) Fit a simple linear regression model with y = `Yield` and x = `IngredA`.

i. What is the value of the slope? Write a sentence that interprets this slope.

The slope is 1.76. The average (or, predicted) yield increases 1.76 units for each one-unit increase in the amount of ingredient A.

ii. What is the value of R^2 for this regression?

R^2 = 13.0%.

iii. 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 relationship is not quite significant at the 0.05 level. The p-value is 0.083 for testing the null hypothesis that the slope is 0.

(d) Fit a simple linear regression model with y = `Yield` and x = `IngredB`.

i. What is the value of the slope? Write a sentence that interprets this slope.

The slope is 2.48. The average (or, predicted) yield increases 2.48 units for each one-unit increase in the amount of ingredient B.

ii. What is the value of R^2 for this regression?

R^2 = 48.5%.

iii. 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 relationship is significant at the 0.05 level. The *p*-value is 0.000 for testing the null hypothesis that the slope is 0.

(e) Fit a multiple linear regression model with y = `Yield` using predictors x_1 = `IngredA` and x_2 = `IngredB`.

i. 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.

The values of the slopes are 1.76 and 2.48 for ingredients A and B, respectively. These are the same as in the simple linear regressions because the *x*-variables have correlation equal to 0.

ii. 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.

R^2 = 61.5% = 13.0% + 48.5%. This relationship holds because the *x*-variables have 0 correlation.

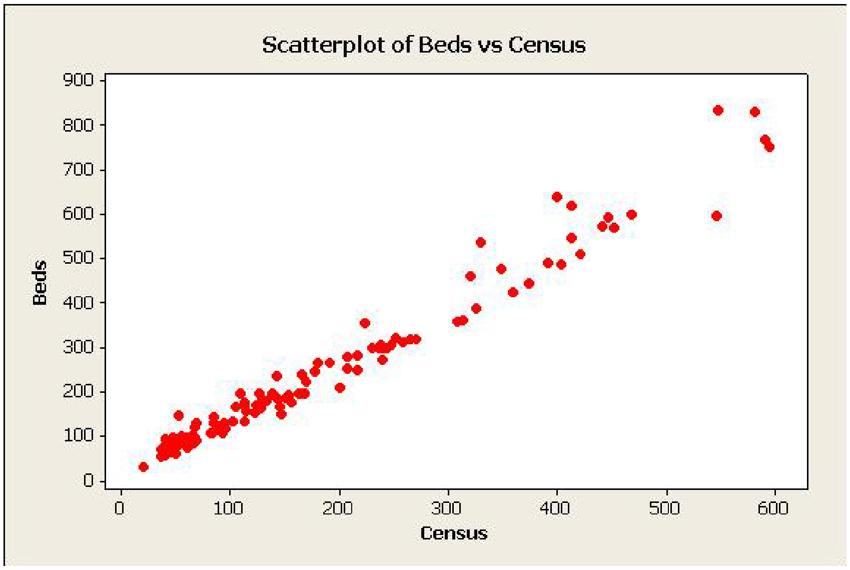
iii. On the basis of this regression, can we say that there is a statistically significant relationship between `Yield` and `IngredA`? Explain why or why not.

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.**

(a) Draw a scatterplot showing the relationship between the two *x*-variables `Beds` and `Census`. Briefly describe the correlation.

There is a strong positive correlation between `Beds` and `Census` when looking at the scatterplot for the hospital data.



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

There is a statistically significant linear relationship between the two variables (*p*-value = 0.000).

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

There is a statistically significant linear relationship between the two variables (*p*-value = 0.000).

(d) Fit a multiple linear regression model with *y* = `InfctRsk` using the four *x*-variables `Stay`, `Xrays`, `Beds` and `Census` as the predictors.

i. In the Analysis of Variance table, what is the *p*-value? On the basis of this *p*-value, what can we conclude?

The *p*-value = 0.000 so we conclude that at least one of the *x*-variables is a significant predictor of `InfctRsk`.

ii. On the basis of the *p*-value for testing the statistical significance of `Beds`, what can we conclude?

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.

iii. On the basis of the *p*-value for testing the statistical significance of `Census`, what can we conclude?

The *p*-value = 0.587 for testing the statistical significance of `Census`, so we conclude that in this model, `Census` is not a significant predictor.

iv. 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 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).

(e) 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`.

i. Explain whether each variable is a significant predictor within this model.

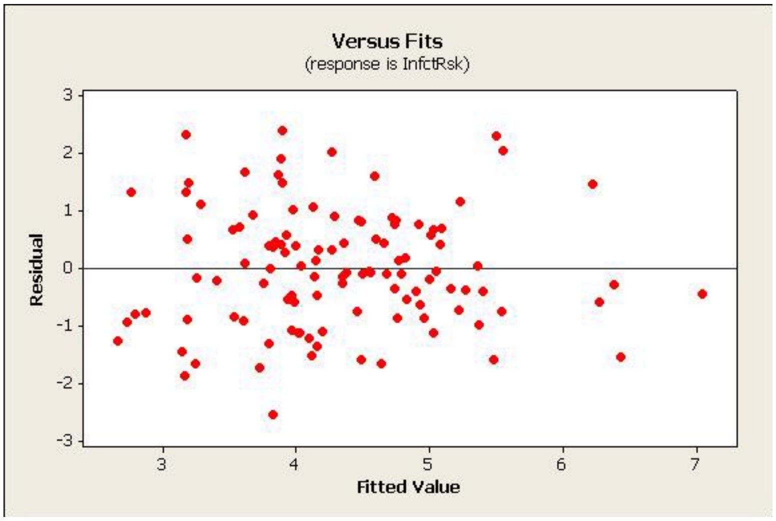
All variables are significant (*p*-values are 0.000 for `Stay` and `Xrays` and 0.014 for `Census`.)

ii. 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.

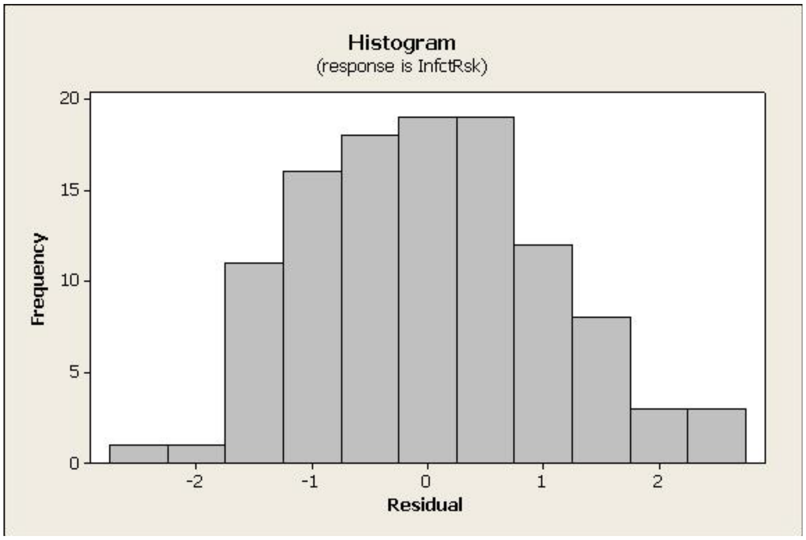
MSE = 1.091 is smaller than MSE = 1.102 for the 4-variable model for this model. A low MSE is good so the 3-variable model looks better. Note: In the calculus of least squares, the SSE cannot decrease when we delete a variable. However, its increase may be small relative to the change in the degrees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

(f) 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.

The residual plot below looks about as it should - a "random" horizontal band of points.



The histogram is roughly bell-shaped:



The assumption of normally distributed errors seems to be reasonable. No difficulties with the data or the model are indicated.

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,

(a) Find the fitted regression equation.

$\boldsymbol{\hat{y} = 10.2 + 4.0\, x}$.

(b) Use matrix methods to obtain

i. $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ (a 2 by 2 matrix),

$$(\boldsymbol{X^T X})^{-1} = \begin{pmatrix} \mathbf{0.2} & \mathbf{-0.1} \\ \mathbf{-0.1} & \mathbf{0.1} \end{pmatrix}$$

ii. $\mathbf{b = (X^TX)^{-1}X^T}\boldsymbol{Y}$ (a 2 by 1 column-vector),

$$\boldsymbol{b} = \begin{pmatrix} \mathbf{10.2} \\ \mathbf{4.0} \end{pmatrix}$$

iii. $\mathbf{e}$ (a 10 by 1 column-vector),

$$\boldsymbol{e^T} = (\mathbf{1.8, -1.2, -1.2, 1.8, -0.2, -1.2, -2.2, 0.8, 0.8, 0.8, 0.8})$$

iv. SSE (a scalar),

SSE = 17.60

v. $\mathrm{se}^2(\mathbf{b})$ (a 2 by 2 matrix),

$$\mathbf{se^2(b)} = \begin{pmatrix} \mathbf{0.44} & \mathbf{-0.22} \\ \mathbf{-0.22} & \mathbf{0.22} \end{pmatrix}$$

vi. $\widehat{Y_h}$ when $X_h = 2$ (a scalar),

$\boldsymbol{\widehat{Y_h}}$ **at** $\mathbf{X_h = 2}$ **is 18.2**

vii. $\mathrm{se}^2(\widehat{Y_h})$ when $X_h = 2$ (a scalar).

$$\boldsymbol{se^2\left(\widehat{Y_h}\right) = 0.44}$$

(c) Use $\mathrm{se}^2(\mathbf{b})$ to calculate Correlation(b_0, b_1) (a scalar).

$$\mathbf{Correlation(b_0, b_1)} = \frac{\mathbf{-0.22}}{\sqrt{\mathbf{(0.44)(0.22)}}} = \mathbf{-0.707}$$

4. (**8x2 = 16 points**) In a small-scale experimental study of the relation between degree of brand liking (Y) and moisture content (X_1) and sweetness (X_2) of the product, the following results were obtained from the experiment based on a completely randomized design (data are coded):

X_1	4	4	4	4	6	6	6	6	8	8	8	8	10	10	10	10
X_2	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:

a) 95.21 % of the variation in degree of brand liking (Y) is accounted for by moisture content (X_1) and sweetness (X_2).

b) The estimated standard deviation of the regression errors is 2.69330.

c) The F-statistic of 129.08 with a p-value of 0.000 indicates that the model containing X_1 and X_2 is more useful in predicting Y than not taking into account the two predictors.

d) The t-statistic of 14.70 with a p-value of 0.000 indicates that the slope parameter for X_1 is significantly different from 0 in this model.

e) The t-statistic of 6.50 with a p-value of 0.000 indicates that the slope parameter for X_2 is significantly different from 0 in this model.

f) We estimate that E(Y) increases by 4.425 units when X_1 increases by 1 unit and X_2 is held constant.

g) We estimate that E(Y) increases by 4.375 units when X_2 increases by 1 unit and X_1 is held constant.

h) We predict that the degree of brand liking when moisture content is 7 and sweetness is 3 is 81.75.