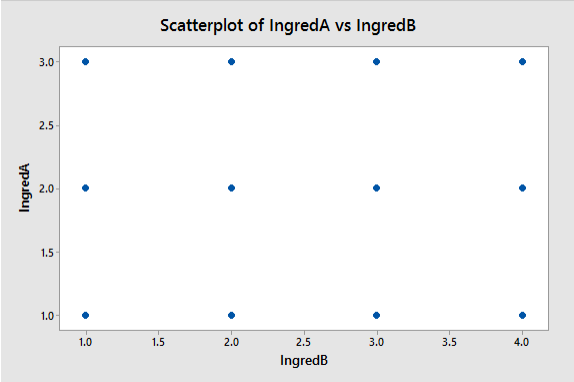
Homework 5

Roly Vicaría
STAT501 Fall 2015

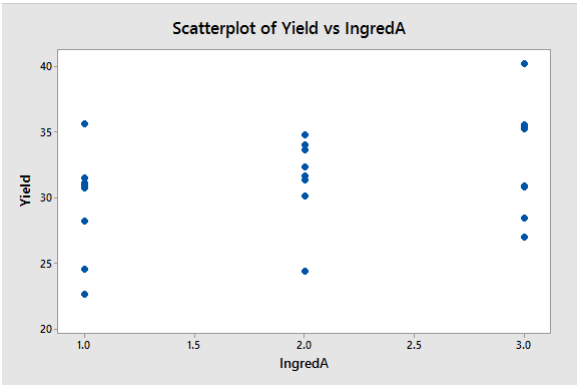
September 27, 2015

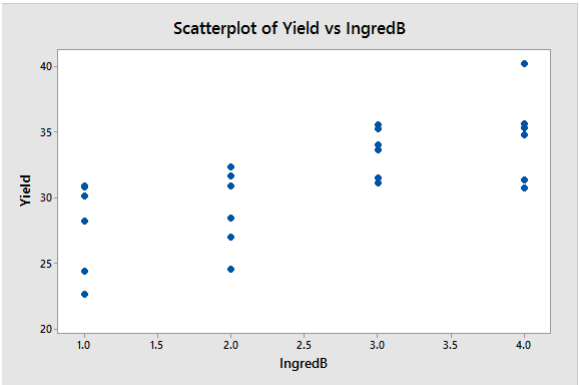
0.0.1 Question 1

a) Based on the plot, I would say that there is no correlation between ingredA and ingredB. You can see the fixed points corresponding to all the possible combinations of ingredA (1, 2, 3) and ingredB (1, 2, 3, 4).



b) The plot of ingredA vs Yield shows a slight indication of a positive linear relationship. It's not a strong one because the Yield has a lot of variability for each level of ingredA. The plot of ingredB vs Yield appears to have a stronger positive correlation. It looks to me that ingredB is a stronger predictor.





c) $Yield = 27.75 + \ 1.763\ IngredA$

i) The slope is 1.763. This can be interpreted as saying that increasing ingredA by 1 unit would increase yield by about 1.763 units.
ii) The R^2 value is 13.05%
iii) The p-value of the regression is .083 which indicates that it may not be statistically significant.

d) $Yield = 25.07 + \ 2.482\ IngredB$

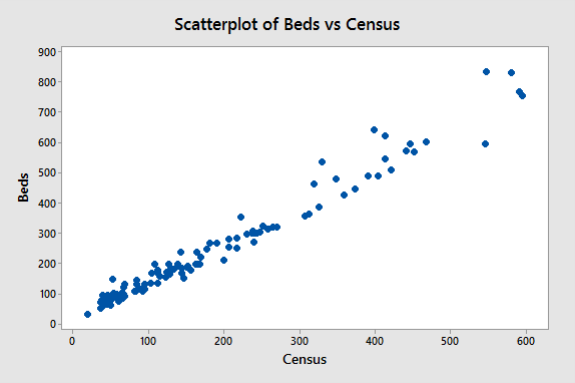
i) The slope is 2.482. This can be interpreted as saying that increasing ingredB by 1 unit would increase yield by about 2.482 units.
ii) The R^2 value is 48.50%
iii) The p-value of the regression is is 0.000 which indicates it is statistically significant.

e) $Yield = 21.54 + \ 1.763\ IngredA + \ 2.482\ IngredB$

i) The values of the coefficients are 1.763 and 2.482. This occurs because ingredA and ingredB are not correlated.
ii) The R^2 value is 61.55%. This is indeed equal to the sum of 13.05% and 48.5%. This occurs for the same reason as part (i), namely that the two predictor variables are not correlated.
iii) In the multiple regression, ingredA has a p-value of .014, which is statistically significant.

0.0.2 Question 2

a) Based on the scatterplot, there appears to be a strong positive correlation between number of beds and census. This makes sense since more beds means the hospital has more capacity.



b) $InfctRsk = 3.716 + \ 0.002457\ Beds$
The p-value of the regression is 0.000 which indicates a statistically significant linear relationship between the two variables.

c) $InfctRsk = 3.697 + \ 0.003374\ Census$
The p-value of the regression is 0.000 which indicates a statistically significant linear relationship between the two variables.

d) $InfctRsk = -0.692 + \ 0.3157\ Stay + \ 0.02050\ Xray - \ 0.00019\ Beds + \ 0.00216\ Census$

i) The p-value of the regression is 0.000, so we can conclude that there is a statistically significant relationship.
ii) The p-value of beds is 0.950 so we can conclude that it is not statistically significant in this model.
iii) The p-value of census is 0.587 so we can conclude that it too is not statistically significant in this model.

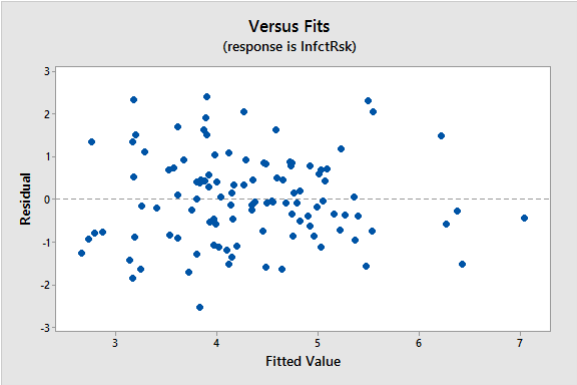
iv) The most likely reason that the significance results for beds and census changed from the simple model to the multiple model is that the other variables, xray and stay, were much better predictors of infection risk. Therefore, in the presence of these other predictors, beds and census did not have any significant effect on the residuals of the model.

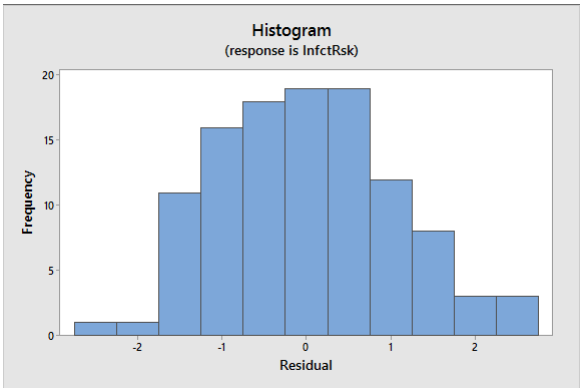
e) $InfctRsk = -0.701 + \ 0.3166\ Stay + \ 0.02049\ Xray + \ 0.001916\ Census$

i) In this model, every predictor has a p-value less than 0.05, so we can conclude that they are all statistically significant.

ii) The MSE for this model is 1.091. This is a bit lower than the MSE for the 4-variable model: 1.1017. This is evidence that the 3-variable model is preferable, because it indicates less variation of the data points around the regression line.

f) The residual vs fit plot shows a random bounce of values around the residual = 0 line which indicates good linear fit and consistent variance of the error terms. The histogram shows a near normal distribution of the residuals which shows the error distribtion is close to normal.





0.0.3 Question 3

a) $Y = 10.200 + \ 4.000\ X$

```
In [1]: # part (b)
        import numpy as np
        from numpy.linalg import inv

        X = np.matrix([[1,1], [1,0], [1,2], [1,0], [1,3], [1,1], [1,0], [1,1], [1,2], [1,0]])
        Y = np.matrix([[16], [9], [17], [12], [22], [13], [8], [15], [19], [11]])

In [2]: # part b-i)
        inv(X.transpose() * X)

Out[2]: matrix([[ 0.2, -0.1],
                [-0.1,  0.1]])

In [3]: # part b-ii)
        b = inv(X.transpose() * X) * X.transpose() * Y
        b

Out[3]: matrix([[ 10.2],
                [  4. ]])

In [4]: # part b-iii)
        e = Y - X * b
        e
```

```
Out[4]: matrix([[ 1.8],
                [-1.2],
                [-1.2],
                [ 1.8],
                [-0.2],
                [-1.2],
                [-2.2],
                [ 0.8],
                [ 0.8],
                [ 0.8]])

In [5]: # part b-iv)
        SSE = Y.transpose() * Y - b.transpose() * X.transpose() * Y
        SSE

Out[5]: matrix([[ 17.6]])

In [6]: # part b-v)
        MSE = (SSE / (10 - 2)).item(0,0)
        se_squared_b = MSE * inv(X.transpose() * X)
        se_squared_b

Out[6]: matrix([[ 0.44, -0.22],
                [-0.22,  0.22]])

In [7]: # part b-vi)
        X_h = np.matrix([[1], [2]])
        Y_hat = X_h.transpose() * b
        Y_hat

Out[7]: matrix([[ 18.2]])

In [8]: # part b-vii)
        se_squared_Y_hat = X_h.transpose() * se_squared_b * X_h
        se_squared_Y_hat

Out[8]: matrix([[ 0.44]])

In [9]: # part (c)
        cov_b0_b1 = se_squared_b.item(0,1) # -0.22
        s_b0 = np.sqrt(se_squared_b.item(0,0))
        s_b1 = np.sqrt(se_squared_b.item(1,1))
        corr_b0_b1 = cov_b0_b1 / (s_b0 * s_b1)
        corr_b0_b1

Out[9]: -0.70710678118654746
```

0.0.4 Question 4

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

b) The estimated standard deviation of the regression 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**.