**STAT 501 – Homework 3 – Fall 2015 - Due September 13th**

**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. (**4x5 =** **20 points**) Use the “HomeTax” dataset, which contains the annual taxes in dollars (*Tax*) and sale price in thousands of dollars (*Price*) for a sample of 104 homes sold in Albuquerque, New Mexico in 1993. The dataset also includes *logTax* = natural logarithm of *Tax* and *logPrice* = natural logarithm of *Price*.

1. Draw a scatterplot with *Tax* on the vertical axis and *Price* on the horizontal axis. Consider predicting annual taxes for homes based on their sale prices. *Based only on what you see in the plot*, explain why there should be less uncertainty (i.e., narrower prediction intervals) for homes with lower sale prices and more uncertainty (i.e., wider prediction intervals) for homes with higher sale prices.



We can explain why there should be less uncertainty for homes with lower sale prices based on the following reasons:

* The variance of the prediction error σ2{pred} has two components:
  1. The variance of the distribution of Y at X = Xh, namely σ2
  2. The variance of the sampling distribution of YhHat namely σ2{YhHat}

It is clear from the scatter plot that the variance of the distribution of Y at particular values of X is lower for the homes with lower sale prices as compared to homes with higher sale prices. From the breakdown above the first item will have lesser contribution to the uncertainty of the response variable at lower home prices.

* The mean XBar which the density of the plot indicates will be towards the left. We know that the variability of the sampling distribution of YhHat is affected by how far Xh is from XBar (mean). So as we go rightwards the variability of the sampling distribution of YhHat increases. This is the 2nd component elaborated above.

1. Fit a simple linear regression model with response variable, *Tax*, and predictor variable, *Price*. Use Minitab to find 95% prediction intervals for *Price*=100, *Price*=150, and *Price*=200. Calculate the width of each interval.

The model is:

**Regression Analysis: Tax versus Price**

Analysis of Variance

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

Regression 1 7167176 7167176 355.92 0.000

Price 1 7167176 7167176 355.92 0.000

Error 102 2053964 20137

Lack-of-Fit 86 1827203 21247 1.50 0.182

Pure Error 16 226761 14173

Total 103 9221140

Model Summary

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

141.905 77.73% 77.51% 76.03%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 61.3 42.0 1.46 0.147

Price 6.876 0.364 18.87 0.000 1.00

Regression Equation

Tax = 61.3 + 6.876 Price

The minitab output with the highlighted prediction intervals is:

Variable Setting

Price 100

Fit SE Fit 95% CI 95% PI

748.905 14.2684 (720.604, 777.207) (466.019, 1031.79)

Variable Setting

Price 150

Fit SE Fit 95% CI 95% PI

1092.72 20.5102 (1052.04, 1133.40) (808.331, 1377.11)

Variable Setting

Price 200

Fit SE Fit 95% CI 95% PI

1436.54 36.0832 (1364.97, 1508.11) (1146.12, 1726.96) X

The width in the different cases is:

Prediction interval width when Price=100 is: 565.771

Prediction interval width when Price=150 is: 568.779

Prediction interval width when Price=200 is: 580.84

1. Fit a simple linear regression model with response variable, *logTax*, and predictor variable, *logPrice*. Use Minitab to find 95% prediction intervals for *logPrice*=ln(100)=4.605, *logPrice*=ln(150)=5.011, and *logPrice*=ln(200)=5.298. Exponentiate the end-points of each interval to express the intervals in dollars (since exp(ln(y)) = y). Calculate the width of each exponentiated interval.

The model is

**Regression Analysis: logTax versus logPrice**

Analysis of Variance

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

Regression 1 9.7103 9.71035 371.53 0.000

logPrice 1 9.7103 9.71035 371.53 0.000

Error 102 2.6659 0.02614

Lack-of-Fit 86 2.3458 0.02728 1.36 0.247

Pure Error 16 0.3201 0.02001

Total 103 12.3763

Model Summary

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

0.161667 78.46% 78.25% 77.43%

Coefficients

Term Coef SE Coef T-Value P-Value VIF

Constant 2.076 0.237 8.76 0.000

logPrice 0.9830 0.0510 19.28 0.000 1.00

Regression Equation

logTax = 2.076 + 0.9830 logPrice

The minitab output with the highlighted prediction intervals is:

Variable Setting

logPrice 4.605

Fit SE Fit 95% CI 95% PI

6.60323 0.0159329 (6.57163, 6.63484) (6.28101, 6.92545)

Variable Setting

logPrice 5.011

Fit SE Fit 95% CI 95% PI

7.00234 0.0248297 (6.95309, 7.05159) (6.67791, 7.32677)

Variable Setting

logPrice 5.298

Fit SE Fit 95% CI 95% PI

7.28447 0.0372852 (7.21051, 7.35842) (6.95538, 7.61355)

The width in the different cases is:

Prediction interval width when Price=100 is: 0.64444

Prediction interval width when Price=150 is: 0.64886

Prediction interval width when Price=200 is: 0.65817

1. Describe if and how results from parts (b) and (c) confirm your answer in part (a).

We can see from the results above that the width increases by 2.5% as Price goes from 100 to 200.

In the case where we take ln (c), we see it increasing by 2% as Price goes from 100 to 200.

2. (**5x4=20 points**) Consider three different datasets where the response variable, y, is Corn, Vegcrop, and Fruitcrop, respectively. In each dataset the predictor variable, x, is fertilizer level. Note the observed predictor range for each dataset goes up to x = 80 units. Based on the following ANOVA outputs and descriptive graphics, comment on the validity of the confidence and prediction intervals given below. In other words, which confidence intervals/prediction intervals are okay to be used? Please justify your answers.

**Analysis of Variance: Corn**

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

**Regression 1 41349.8 41349.8 19.80 0.000**

**Fertlevel 1 41349.8 41349.8 19.80 0.000**

**Error 20 41759.3 2088.0**

**Lack-of-Fit 15 41509.3 2767.3 55.35 0.000**

**Pure Error 5 250.0 50.0**

**Total 21 83109.1**

**Analysis of Variance: Vegcrop**

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

**Regression 1 285910 285910 4201.91 0.000**

**Fertlevel 1 285910 285910 4201.91 0.000**

**Error 38 2586 68**

**Lack-of-Fit 15 1252 83 1.44 0.210**

**Pure Error 23 1334 58**

**Total 39 288496**

**Analysis of Variance: Fruitcrop**

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

**Regression 1 280300 280300 9103.98 0.000**

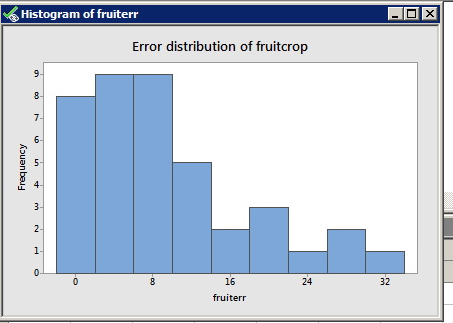
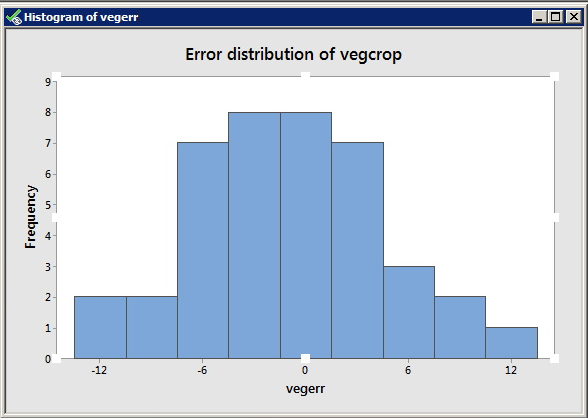
**Fertlevel 1 280300 280300 9103.98 0.000**

**Error 38 1170 31**

**Lack-of-Fit 15 567 38 1.44 0.209**

**Pure Error 23 603 26**

**Total 39 281470**



1. A 95% confidence interval (CI) for the mean corn yield at fertilizer level = 40 is (398.538, 464.193).

Valid.

1. A 95% confidence interval for the mean corn yield at fertilizer level = 85 is (265.074, 343.804).

Invalid. The Xh = 85 is out of scope of the model.

1. A 95% prediction interval (PI) for vegcrop value at fertilizer level = 70 is (411.185, 434.023).

Valid.

1. A 95% prediction interval for fruitcrop value at fertilizer level = 40 is (232.032, 266.877).

Invalid. Prediction intervals are sensitive to Normality assumptions.

1. A 95% confidence interval for the mean fruitcrop at fertilizer level = 40 is (244.486, 254.423).

Valid. Confidence intervals are not very sensitive to Normality assumptions.

3. (**7x4 = 28 points**) Use the “AgeDist” dataset that contains data from *n* = 30 individuals, including the age of a driver in years (ages) and the distance the driver can see in feet (distance).

1. Use Minitab to obtain a scatterplot with Y=distance on the vertical axis and X=age on the horizontal axis. Add the estimated simple linear regression line to the plot.
2. What is the slope estimate of the regression line? Write a sentence that interprets this value in the context of this situation.
3. Obtain a 95% confidence interval for the mean distance when age = 75 together with a 95% prediction interval for the predicted distance at age = 75.
4. Show how the value under “Fit” (351.169) in your Minitab output was calculated. (There may be some round-off error in your calculation.)
5. Write a sentence that interprets the interval represented by “95% CI”. In your interpretation, include the numerical values of the interval and use the fact that this applies to age=75 years.
6. Write a sentence that interprets the interval represented by “95% PI”. In your interpretation, include the numerical values of the interval and use the fact that this applies to age=75 years.
7. The intercept value of this fitted regression equation is not meaningful in any practical sense, simply because a person at birth cannot see a distance of 576.7 feet. Does this mean that the validity of the fitted straight line is questionable?

4. (**4x3=12 points**) Say whether the following statements about the simple linear regression model are true or false? Explain your answers in terms of the confidence and prediction interval formulas.

1. The confidence interval width for the mean value of Y (μY) is smallest at the predictor sample mean.
2. For any given Xh within the scope of the model, the confidence interval width for the mean value of Y (μY) is smaller than the prediction interval width of a new response (Ynew).
3. The confidence interval formula for the population intercept (β0) can be derived from the confidence interval formula of μY.
4. The prediction interval widths of Ynew corresponding to predictor values equidistant from the sample mean are equal.

5. (**4x5 = 20 points**) The “HospitalInfectionL3” dataset gives characteristics of *n* = 58 hospitals in the eastern and north central areas of the United States. The overall purpose for the data set is to analyze factors that predict *InfctRsk*, the infection risk for patients staying in the hospital. The infection risk value is the percentage of patients who get an infection while they are hospitalized. The variable *Stay* is the average length of stay (days) for patients at the hospital. Assume that a first-order regression model is appropriate.

1. Calculate a 95% interval estimate of the mean infection risk for patients staying in hospitals with an average length of stay of 10 days. Interpret your confidence interval.
2. Mercy Hospital has an average length of stay of 10 days. Predict the infection risk for patients staying in Mercy Hospital using a 95% prediction interval. Interpret your prediction interval.
3. Determine the boundary values of the 95% confidence band for the regression line when *Stayh* = 10. (You’ll have to calculate this by hand using the formula from Section 2.6 in the textbook.) Is your-confidence band wider at this point than the confidence interval in part (a)? Should it be?
4. Test, at 0.05 significance level, whether the mean infection risk E(Y\_h), when Stayh = 8, is less than 4 days. Write down the test statistic, decision rule, and your conclusion.