**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~~

**The exponentiated 95% prediction intervals and widths are:**

**i. ($534, $1018), width=$484**

**ii. ($795, $1520), width=$726**

**iii. ($1049, $2025), width=$977**

exp(6.92545) - exp(6.28101) = 484

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

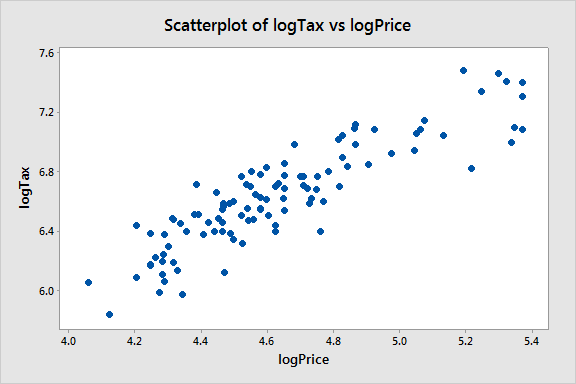
~~We can see from the results above that the width increases by 2.5% as Price goes from 100 to 200.~~

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

~~We don’t have the tools as yet to state whether this difference is significant or not. However intuitively the changes appear to be small when compared to 100% increase in the price.~~

~~The results do show an increase as stipulated in part a.~~

**The results confirm the answer for part (a) since the intervals in part (b) remain approximately the same width for each value of *Price*, but the intervals in part (c) increase in width as *Price* increases. The model in part (b) does not support the constant variance assumption and so is not an appropriate model for this dataset. By looking at the scatterplot of *logTax* vs *logPrice* below we can see that the model in part (c) does support the LINE conditions (including the constant variance assumption) and so is a more appropriate model for this dataset.**



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**

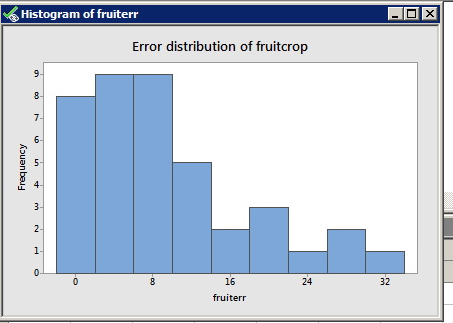
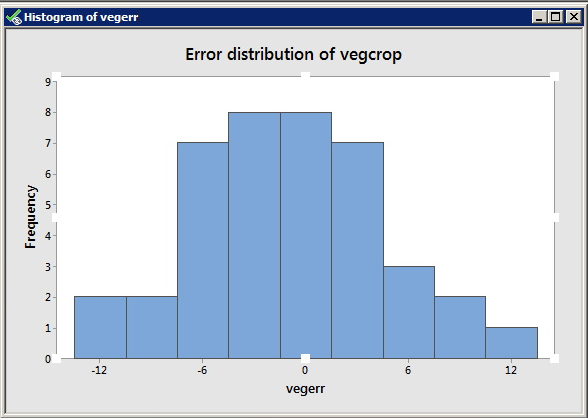
**Fertlevel 1 280300 280300 9103.98 0.000**

**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. The Xh (fertilizer level) = 40 is within the scope of the model.~~

**The statistically significant lack of linear fit indicates that the linearity condition is violated. Hence this CI may not be okay.**

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

Invalid. The Xh (fertilizer level) = 85 is out of scope of the model. We can’t use the model in this case.

**This CI also may not be okay for two reasons: (i) The violation of the linearity condition as explained above, and (ii) X=85 is beyond the scope of this model.**

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

Valid. The Xh (fertilizer level) = 70 is within the scope of the model. The error distribution of vegcrop appears to be normal.

**This PI may be okay provided it is reasonable to assume that the vegcrop values are independent and their error variances are equal. Note that that there is no significant lack of linear fit and also the vegcrop errors are normally distributed.**

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 and we see that the error distribution of fruitcrop doesn’t appear to be normal. The Xh (fertilizer level) = 40 is within the scope of the model.

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

Valid. Even though the error distribution of fruitcrop doesn’t appear to be normal, confidence intervals are not very sensitive to Normality assumptions. The Xh (fertilizer level) = 40 is within the scope of the model.

**This CI should be okay. Once again it should be assumed that y = fruitcrop values are independent and their errors have equal variances. Here n=40 is large enough to ensure that is normal even though the errors are not. Also note that there is no significant lack of linear fit indicating that linear model is adequate.**

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.



1. What is the slope estimate of the regression line? Write a sentence that interprets this value in the context of this situation.

Slope estimate = -3.007

Here, it tells us that we predict that the mean distance the driver can see (in feets) will **decrease** by 3.007 feets for every additional year of age.

1. 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.

When age = 75, we have:

95% confidence interval: 323.213, 379.125

95% prediction interval: 245.473, 456.865

1. Show how the value under “Fit” (351.169) in your Minitab output was calculated. (There may be some round-off error in your calculation.)

The value comes from the regression equation given by the model: Distance = 576.7 - 3.007 \* Age

So, Fit = yHat = 576.7 - 3.007 \* 75 = 351.175

1. 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.

The interval represents the average distance the driver can see for individuals with age 75 years.

1. 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.

The interval represents the distance the driver can see for a randomly selected individual of age 75 years.

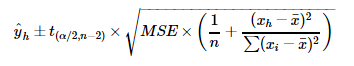
1. The intercept value of this fitted regression equation is not meaningful in any practical sense, simply because a person at birth cannot see a distance of 576.7 feet. Does this mean that the validity of the fitted straight line is questionable?

No, here we have to differentiate between validity and meaningfulness. While the regression equation is correctly stated to be not meaningful because a person at birth cannot see a distance of 576.7 feet, it doesn’t deem the straight line as invalid. It is also important to stress that it is not reasonable to use the regression equation to infer the response far outside the range of the predictor variable. Therefore we shouldn’t use the equation to predict the response at or close to age = 0.

**No. The linear model was fitted to accommodate the observed age data, which are all far from the origin. In other words, linearity is assumed only within the scope of the model and there is no intention to extrapolate this assumed linear relationship over an expanded range to include age equal to 0.**

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.

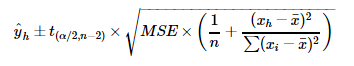
1. The confidence interval width for the mean value of Y (μY) is smallest at the predictor sample mean. TRUE

With the formula  we can see that the width depends on (xh – xBar)^2. This term is the smallest when xh = xBar. It will increase as xh moves away from xBar in either direction.

1. 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). TRUE

Since standard error of the prediction has an extra MSE term added to the standard error of the fit, this makes the prediction intervals wider than the confidence intervals.

1. The confidence interval formula for the population intercept (β0) can be derived from the confidence interval formula of μY. TRUE

The formula for latter is: 

This when evaluated at Xh = 0, we will get the formula for β0 i.e.



yhHat

1. The prediction interval widths of Ynew corresponding to predictor values equidistant from the sample mean are equal. TRUE

This is true as becomes evident from the prediction interval width formula:



The value (xh – xBar)^2 will evaluate to be the same when xh is equidistant from the sample mean. Therefore the width will be the same.

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.

The minitab output is:

**Prediction for InfctRsk**

Regression Equation

InfctRsk = -1.160 + 0.5689 Stay

Variable Setting

Stay 10

Fit SE Fit 95% CI 95% PI

4.52885 0.134602 (4.25921, 4.79849) (2.45891, 6.59878)

The 95% interval estimate of the mean infection risk is given under 95% CI

Interpretation: with 95% confidence we can estimate that in hospitals in which the average length of stay is 10 days, the mean infection risk is between 4.25921 and 4.79849.

Width: 0.53928

***ŷh* = 4.52885, se(*ŷh*) = 0.134602, t(0.975; 56) = 2.00324.**

**So, the 95% confidence interval for E(Yh) is given by  
4.52885 ± 2.00324 (0.134602) = (4.2592, 4.7985).**

**We can be 95% confident that the mean infection risk for patients staying in hospitals with an average length of stay of 10 days is between 4.2592 and 4.7985 percent.**

1. 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.

The 95% prediction interval for the infection risk for patients staying at Mercy Hospital is given under 95% PI, (2.45891, 6.59878)

Interpretation: we say with 95% confidence that for any future hospital where the average length of stay is 10 days, the infection risk is between 2.45891 and 6.59878.

**se(pred) = √(MSE + se(*ŷh*)2) = √(1.0496 + 0.1346022) = 1.033304, so the 95% prediction interval for Yh(new) is given by  
4.52885 ± 2.00324 (1.033304) = (2.4589, 6.5988).**

**We can be 95% confident that the infection risk for patients staying in Mercy Hospital, which has an average stay length of 10 days, will be between 2.4589 and 6.5988 percent.**

1. 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?

We know from the previous minitab output that YhHat = 4.52885 and s{ YhHat } = 0.1346 when Stayh = xh = 10

W2 = 2F(.95; 2, 56) = 2\* 3.16186 = 6.32372

Therefore W = 2.5147

Hence, the boundary values of the confidence band for the regression line at Xh = 10 are 4.52885 ± 2.5147 (0.1346), and the confidence band there is:

4.19037138 <= β0 + β1 Xh <= 4.86732862 for Xh = 10

Width = 0.67695724

We can see that the confidence band width = 0.67695724 is wider than the confidence interval of width 0.53928. It should be wider since the W multiple is larger than the t multiple because the confidence band must encompass the entire regression line, whereas the confidence limits for E{Yh } at Xh apply only at the single level Xh.

**To construct confidence band, we need W2 = 2 F(0.95; 2; 56) = 2 (3.16186) = 6.32372.**

**So, W = = 2.5147 and the 95% confidence band for the regression line is given by 4.52885 ± 2.5147 (0.134602) = (4.1904, 4.8673).**

**Yes, it is wider. It is wider because the confidence band has to encompass the entire regression line, whereas the confidence limits for E(Yh) at *Stayh* = 10 apply only at the *Stay*-value of 10.**

1. 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.

**H0: E(Yh|X = 8) = 4 vs. Ha: E(Yh|X = 8) < 4.**

**Test statistic: t = (**

**Decision rule: Reject H0 if t ≤ −t(0.95; 96) = −1.66088.**

**Conclusion: Since we do reject H0.**

**(Or, *p*-value = *P*(*T* ≤ − 2.588) = 0.0055758 << 0.05 and make same conclusion.)**

When Stayh = 8, we get yHat = InfctRskHat = -1.160 + 0.5689 \* 8 = 3.3912

The minitab output is:

**Prediction for InfctRsk**

Regression Equation

InfctRsk = -1.160 + 0.5689 Stay

Variable Setting

Stay 8

Fit SE Fit 95% CI 95% PI

3.39111 0.235232 (2.91989, 3.86234) (1.28541, 5.49681)

The 95% CI for the mean infection risk E(Y\_h) is (2.91989, 3.86234), when Stayh = 8

The interval doesn’t contain 4 and moreover is less than 4. Therefore the conclusion is that at the 0.05 significance level, the mean infection risk E(Y\_h), when Stayh = 8, is less than 4 days

**Alternatively using the hypothesis testing approach**

We have the hypothesis setup as:

H0: E(Y\_h) >= 4 and Ha : E(Y\_h) < 4

This is a one-sided test and α = 0.05. So tcritical = t(.95, 56) = - 1.67252

Test Statistic: E(Y\_h) – 4 / se(E(Y\_h)) = (3.39111 – 4) / 0.235232 = -2.5885

Since test statistic < t\_critical, we reject the null hypothesis and conclude that the mean infection risk E(Y\_h), when Stayh = 8, is less than 4 days.

We can also use the p-value: calculating the probability that a t-random variable with n-2 = 56 degrees of freedom would be smaller than -2.5885. Using minitab:

**Cumulative Distribution Function**

Student’s t distribution with 56 DF

x P( X ≤ x )

-2.5885 0.0061291

p-value: 0.0061291 < alpha = 0.05: we can reject the null hypothesis and conclude the alternative.

**H0: E(Yh|X = 8) = 4 vs. Ha: E(Yh|X = 8) < 4.**

**Test statistic: t = (**

**Decision rule: Reject H0 if t ≤ −t(0.95; 96) = −1.66088.**

**Conclusion: Since we do reject H0.**

**(Or, *p*-value = *P*(*T* ≤ − 2.588) = 0.0055758 << 0.05 and make same conclusion.)**