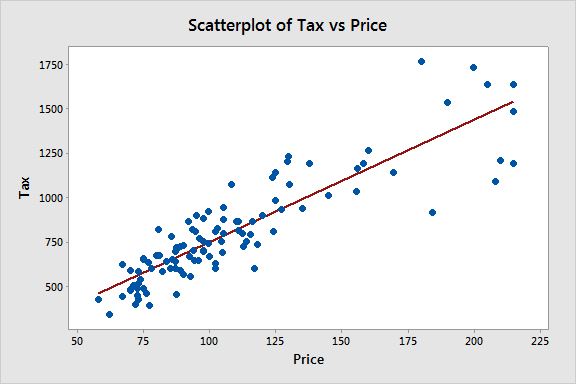
**STAT 501 – Homework Solutions 3 – Fall 2015**

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.



**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 because the sample values of *Tax* are less variable for low values of *Price* than for high values of *Price*.**

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 95% prediction intervals and widths are:**

**i. ($466, $1032), width=$566**

**ii. ($808, $1377), width=$569**

**iii. ($1146, $1727), width=$581**

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 exponentiated 95% prediction intervals and widths are:**

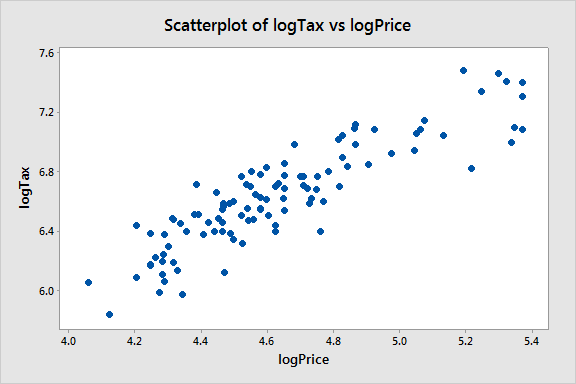
**i. ($534, $1018), width=$484**

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

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

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

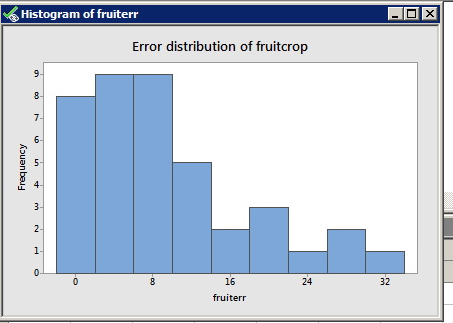
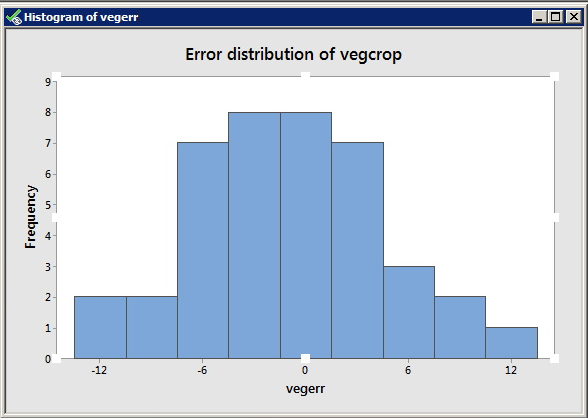
**Fertlevel 1 280300 280300 9103.98 0.000**

**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).

**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).

**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).

**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).

**This PI may not be okay since the errors associated with the fruitcrop values are not normal.**

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

**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. (7**x4 = 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.



Fitted Regression Line

Distance = 576.7 - 3.007 Age

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 (b1) = –3.007**

**This means that for every year, the distance that a driver can see decreases by 3.007 feet on average.**

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.

**Variable Setting Fit SE Fit 95% CI 95% PI**

**Age 75 351.169 13.6476 (323.213, 379.125)(245.473, 456.865)**

**The 95% CI for the mean distance when age=75 is (323.213, 379.125).**

**The 95% PI for the predicted distance when age=75 is (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.)

**Fit = = b0 + b1 \* age = 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.

**We can be 95% confident that the average distance 75 year-old drivers can see is between 323.213 feet and 379.125 feet.**

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.

**We can be 95% confident that the distance that a randomly chosen 75 year-old driver can see will be between 245.473 feet and 456.865 feet.**

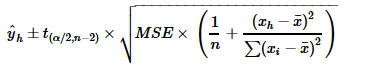
1. The intercept value of this fitted regression equation is not meaningful in any practical sense, simply because a person at birth cannot see a distance of 576.7 feet. Does this mean that the validity of the fitted straight line is questionable?

**No. 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. The CI formula is**

****

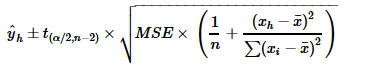
**The second term of this expression which determines the width will be smallest when xh = , since the second term within the parenthesis will equal zero for this value of xh and other quantities remain the same for all predictor values.**

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. In the second term of the PI formula that determines the width, there is an extra 1 inside the parenthesis compared to the same second term in the CI formula.**

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

**TRUE. Since the intercept is μY for xh = 0, substituting xh = 0 in the CI formula**

****

**gives the CI formula for β0, which is**

****

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

**TRUE. Observe in the PI formula, which is shown above in part (d), the last term depends on xh only through (xh – ) and this quantity is the same for predictor values equidistant from the mean.**

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.

***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.

**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?

**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.)**