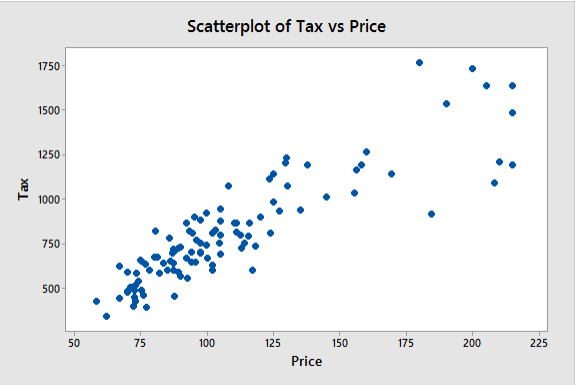
Homework 3

Roly Vicaría
STAT501 Fall 2015

September 13, 2015

0.0.1 Question 1

a. Based on the scatterplot, I would expect less uncertainty for homes with lower sale prices because I see a lot more data points for lower sale prices than there are for homes with higher sale prices. And not only are there more data points for lower sale prices, but the points are a lot more dense (i.e., closer in y value), whereas the data points for higher sale price homes are more scattered.



b. $Tax = 61.3 + \ 6.876\ Price$

Price	95% Prediction Interval	Interval Width
100	(466.019, 1031.79)	565.771
150	(808.331, 1377.11)	568.779
200	(1146.12, 1726.96)	580.84

c. $logTax = 2.076 + \ 0.9830\ logPrice$

logPrice	95% Prediction Interval	Interval Width
4.605	(534.323, 1017.852)	483.529
5.011	(794.657, 1520.463)	725.806
5.298	(1048.777, 2025.456)	976.679

d. The results from parts (b) and (c) confirm my answer from part (a) because as the prices (and logPrices) increased, the width of the prediction intervals increased also.

0.0.2 Question 2

a. The Lack-of-Fit test for the corn yield indicates that there is a lack of linear fit for the model. Without knowing more about the model and/or data, I would be hesitant to use this confidence interval.

b. In addition to the above, the usefulness of this confidence interval is questionable based solely on the fact that the fertizer level is outside of the range of observed levels in the dataset.

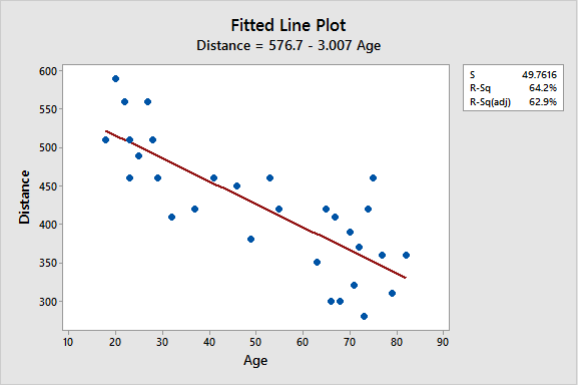
c. This prediction interval is OK to use since the error distribution appears close to normal and it seems to meet all the other LINE criteria + the fertilizer level is within the scope of the model.

d. This prediction interval is NOT OK to use since the error distributino does not appear normal and the prediction interval is strongly dependent on that.

e. This confidence interval is OK to use since the sample size is large enough so that the error distribution isn't an issue. And it satisfies the other LINE criteria.

0.0.3 Question 3

a. $Distance = 576.7 - \ 3.007\ Age$



b. The slope of the regression line is -3.007. This can be interpreted as saying that for every additional year a driver gets older, they see about 3 ft less.

c. Confidence interval and prediction interval for 75 yrs old

Age	95% Confidence Interval	95% Prediction Interval
75	(323.213, 379.125)	(245.473, 456.865)

d. The "Fit" value is calculated by inserting 75 into the regression equation: $576.7 - 3.007(75) \approx 351.175$

e. The confidence interval can be interpreted as saying that we can be 95% confident that the mean distance seen by all people age 75 is between 323.213 and 379.125 ft.

f. The prediction interval can be interpreted as saying that for a random 75 yr old person, we can be 95% confident that the distance they can see will be between 245.473 and 456.865 ft.

g. The fact that the intercept value of this fitted regression is not meaningful, does not invalidate the fitted straight line. It just means that Age = 0 does not fall within the scope of this model. The data used for this model was ages between 18 and 82. We can't really trust the regression equation for values outside those ages.

0.0.4 Question 4

Confidence interval formula: $\hat{y}_h \pm t_{(\alpha/2,n-2)} \times \sqrt{MSE \times \left(\frac{1}{n} + \frac{(x_h-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$

Prediction interval formula: $\hat{y}_h \pm t_{(\alpha/2,n-2)} \times \sqrt{MSE \times \left(1 + \frac{1}{n} + \frac{(x_h-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}$

a. True. We can see in the formula above, that the second term inside the square root becomes zero when $x_h = \bar{x}$

b. True. The prediction interval formula has an extra MSE term that the confidence interval formula does not, so it will always necessarily be wider.

c. True. If you compute the confidence interval for $E(Y_h)$ with $X_h = 0$, it would be equal to the confidence interval formula for β_0, since $\hat{Y}_h$ would be equal to b_0 and $s(b_0) = s(\hat{Y}_h)$ when $X_h = 0$.

d. True. Since the last term in the prediction interval formula squares the difference between x_h and $\bar{x}$, it would come out to the same value for predictor values equidistant from the sample mean.

0.0.5 Question 5

a. The 95% confidence interval when Stay = 10 days is: (4.25921, 4.79849). This interval can be interpreted as saying that we can be 95% confident that the infection risk for patients staying in hospitals with an average length of stay of 10 days is between 4.259 and 4.798 percent.

b. The 95% prediction interval when Stay = 10 days is: (2.45891, 6.59878). This interval can be interpreted as saying that we can be 95% confident that the infection risk for patients at a random hospital with average stay of 10 days (Mercy Hospital), is between 2.459 and 6.599 percent.

c. 95 percent confidence band when Stay = 10 days

$\hat{Y}_h = -1.160 + \ 0.5689(10) = 4.529$
$s\{\hat{Y}_h\} = \sqrt{1.0496\left[\frac{1}{58} + \frac{(10-10.049)^2}{118.371}\right]} \approx 0.1346$
$W^2 = 2F(.95, 2, 56) \approx 2(3.162) \approx 6.324$
$W \approx 2.515$

Boundary values for confidence band at $X_h = 10$ are $4.529 \pm 2.515(0.1346)$ or (4.1905, 4.8675). This confidence band is wider at this point than the confidence interval from part (a). It is expected to be wider since the W multiple in this computation is larger than the t multiple in the confidence interval computation.

d. Test if $E(Y_h)$ when $Stay_h = 8$ is less than 4%

$H_0 : E(Y_h) \geq 4$
$H_a : E(Y_h) < 4$
$\hat{Y}_h = -1.160 + \ 0.5689(8) = 3.3912$
$s\{\hat{Y}_h\} = \sqrt{1.0496\left[\frac{1}{58} + \frac{(8-10.049)^2}{118.371}\right]} \approx 0.2352$
Test statistic: $t = \frac{\hat{Y}_h - E(Y_h)}{s(Y_h)} = \frac{3.3912-4}{0.2352} = -2.588$
p-value: P(t < -2.588) = 0.00614

Since p-value is less than .05, we can conclude at the .05 significance level that the mean infection risk $E(Y_h)$ when Stay = 8 is less than 4%.