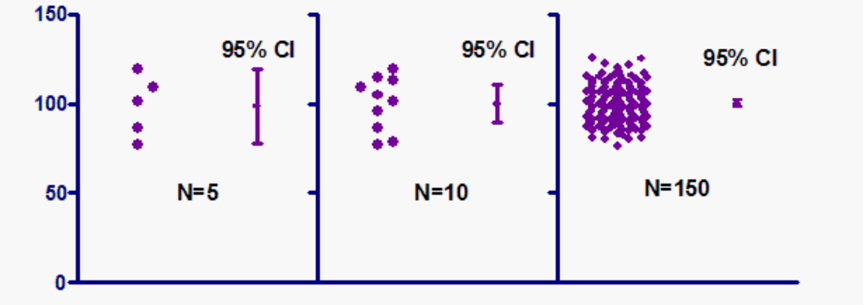
Kitu Komya (UID: 404-491-375)
Statistics 101A (Discussion 3A)
Homework 4 (due: 10/20/16)

Please note: I'm concurrently taking Stats 20, so I'm still familiarizing myself with R.

QA. RSS (residual sum of squares) is the spread of y about the least squares regression line. The equation SYY = SS_{reg} + RSS shows that the relationship between SS_{reg} and RSS is inverse since SYY, the total variation in y, is constant. Moreover, SS_{reg} is the explained variability while RSS is the unexplained variability by the regression line. Thus, if regression is useful, RSS will be small and SS_{reg} will be big.

As evident by the two models, Model 1 has a more useful least squares regression line for predicting its y-values than the least squares regression line in Model 2 since Model 1's spread of observed y-values about the line is smaller than Model 2's, indicating that its regression line explains a higher proportion of variation in the y-variable than Model 2's regression line. Thus, the RSS is smaller and SS_{reg} larger for Model 1 than for Model 2. This corresponds to answer **(d)**.

QB. A 95% confidence interval does not indicate that 95% of the observations fall within its interval, which is a very common misconception. Instead, a 95% confidence interval in this situation calculates an interval that has a 95% likelihood of containing the true population mean (not sample value, but population mean!) of the y-value at that particular x-value. With a large data set, the confidence interval will be narrow since the mean can be calculated more precisely, but most of the sample observations will likely be outside of the interval. A prediction interval would contain more observations than the confidence interval since its interval is larger as it takes into account individual variability. A useful image is shown below:



[source: www.tinyurl.com/stats101a-confidence]

As seen, with increasing sample sizes, the confidence interval becomes narrower and thus more precise in estimating the true population mean. However, the majority of the individual sample observations will lie outside of this confidence interval since the confidence interval is not a tool to measure the likelihood of individual sample observations within its interval (this is when a prediction interval comes handy), but rather a tool to measure the likelihood that the confidence interval contains the true population mean, so it's entirely possible for 95% of the observations to fall outside the 95% confidence interval.

QC. (a) Since RSE = $\sqrt{\frac{RSS}{df}}$ where

RSE = residual standard error = 2.418 (provided in R output)

df = degrees of freedom = 33 (provided in R output

then RSS = residual sum of squares = df * $(\text{RSE})^2$

= 33 * $(2.418)^2$ = 33 * 5.846724 = **<u>192.941892</u>**

(b) Since F = $\frac{\text{SSreg}/_1}{RSS/_{(n-2)}}$ where

F = F statistic = 87.17 (provided in R output)

RSS = 192.941892 (calculated from part (a))

n = sample size = degrees of freedom + 2 (since 2 variables) = 33 + 2 = 35 (provided in R output)

then SS_{reg} = F * $\frac{RSS}{n-2}$

= 87.17 * $\frac{192.941892}{35-2}$ = 87.17 * $\frac{192.941892}{33}$ = 87.17 * 5.846724 = **<u>509.6589</u>**

(c) Since mean SS_{reg} = SS_{reg}/k where

SS_{reg} = 509.6589 (calculated from part (b))

k = number of predictors = 1 (only one response variable as used in part (b))

then mean SS_{reg} = SS_{reg}/1 = 509.6589/1 = **<u>509.6589</u>**

(d) Since total SS = SYY = SS_{reg} + RSS where

SS_{reg} = 509.6589 (calculated from part (b))

RSS = 192.941892 (calculated from part (a))

then total SS = SS_{reg} + RSS = 509.6589 + 192.941892 = **<u>702.600792</u>**

(e) Since r = $\sqrt{r^2}$ where

r^2 = coefficient of determination (r-square) = 0.7254 (provided in R output)

then r = $\sqrt{r^2}$ = $\sqrt{0.7254}$ = **<u>0.8517</u>**

QD. (a) To use regression, the assumption that must be made is that the model is linear. Despite providing evidence that the model is or is not valid based on its scatterplot, residual plot, and QQ plot, the business analyst assumes that the model is valid. That is his first and largest mistake.

Moreover, his claim that the regression coefficient of the predictor value, Distance, is highly statistically significant is unsubstantiated since he is misinterpreting the data. He is referring to the small p-value of $< 2 * 10^{-16}$ which only claims that at the 95% significance level, the slope is not zero. It makes no claims on the significance of the predictor value, Distance.

His claim making, although seems innocent, is inherently wrong since it is based off assumptions that have been violated. This means that for larger values, the model will deviate largely from the observed values since it does not follow a linear trend, and thus the model cannot be used to effectively predict future values of Fare.

(b) At a first glance from the scatterplot, the model seems to correlate well with the observed values, further corroborated by a high r-square value. However, upon a closer look, one can see that the observed values actually wrap around the model parabolically. The residual plot confirms this; the residual plot has a clear parabolic pattern which demonstrates non-linearity, since it should be randomly scattered with no trend. Thus, the straight-line regression model does not fit the data well.

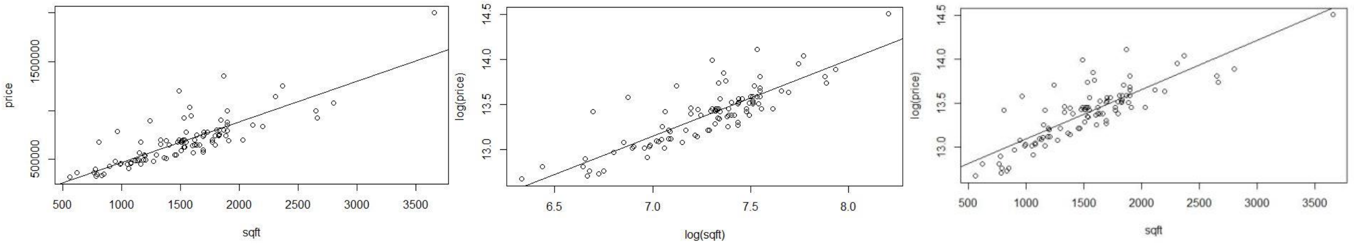
Since the data is most likely quadratic, or some other form such as a polynomial, the model can be improved by using a different model to fit the observed values, such as the quadratic or polynomial functions. To ensure that the model fit the data, a residual plot that has no trend as well as a QQ plot to show normality from all points lying in a straight line is further needed to demonstrate that another form of model fits better. Using the R output is simply not enough to make any claim since they are based on assumptions regarding the model (depending on whether it is a linear, quadratic, or any other polynomial regression).

QE. (a) As seen from the R output, $E(Y|x) = 55902.55 + 414.20x$. The y-intercept of the equation tells us that the starting house value for a house is \$55902.55 before taking into account the size or square footage of it. The slope of the equation says that for each additional square foot, the price of the house increases by \$414.20.

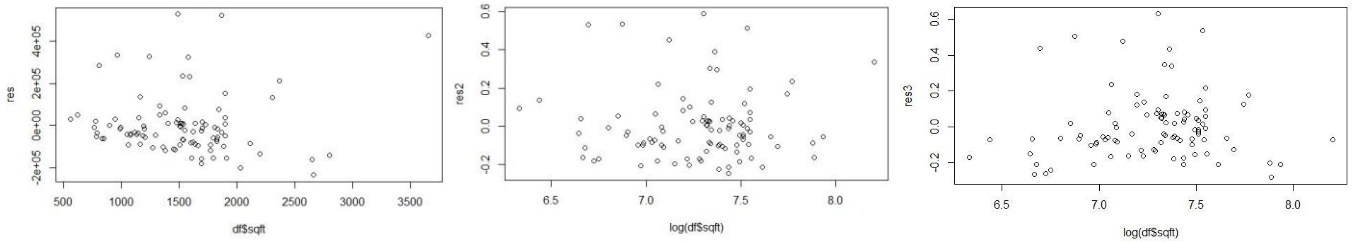
(b) As seen from the R output, $E(\log(Y)|\log(x)) = 7.20862 + 0.84860 * \log(x)$. The slope of the equation says that if we increase the square footage of a house by 1%, then we expect the price of the house to increase by 0.84860%.

(c) As seen from the R output, $E(\log(Y)|x) = 12.53 + 0.0005605x$.

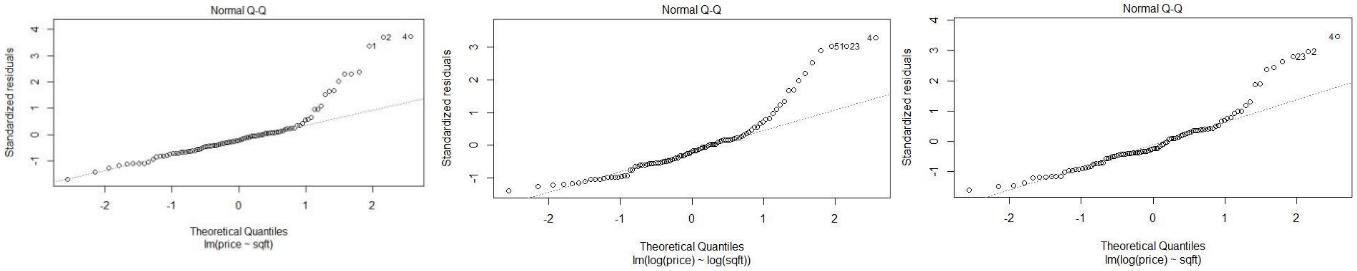
(d) Here are the scatterplots (with their regression lines) of models a, b, c, respectively:



Here are the residual plots of models a, b, c, respectively:



Here are the QQ plots of models a, b, c, respectively:



To analyze the scatterplots, we see that all three models have observed values near the regression line, which means that they all have approximately the same small RSS and large SS_{reg} since the model fits the data well. Thus, using the scatterplot, all three models seem valid since the regression line is appropriate for the observed values.

To analyze the residual plots, we see that models b and c are more randomly scattered with no trend, whereas model a is not as evenly distributed and has a clear negative trend. This means that model a is an invalid fit since the residual plot should not have any trends. Thus, using the residual plot, models b and c are valid.

To analyze the QQ plots, we see that at the tails of the line, all three models have observations that largely deviate from the regression line. Since we ruled out model a from the residual plot, we look between models b and c. Of the two, model c has smaller deviation of the observations at the higher tail from the regression line, indicating normality of the errors. Thus, model c is the most valid model.