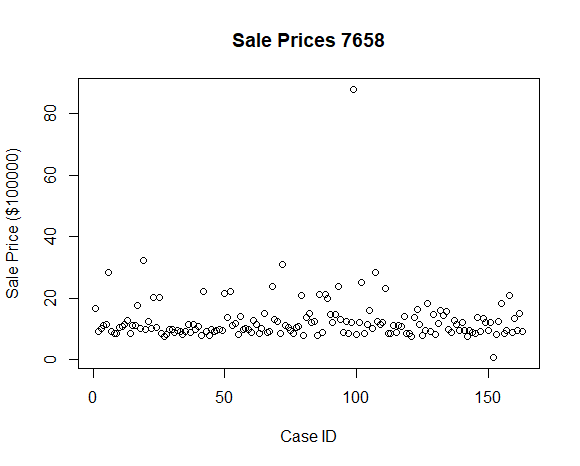
**STA302 Assignment 2**

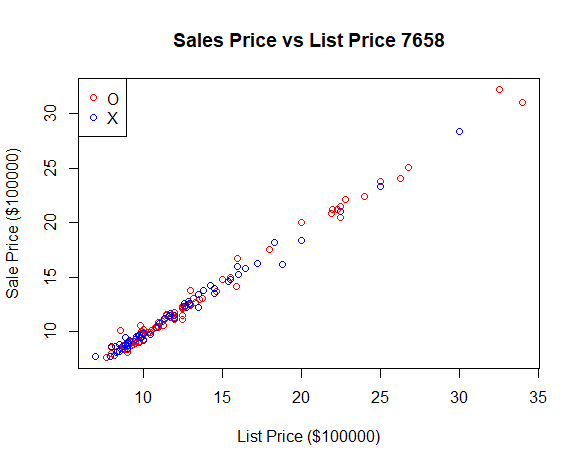
***Question 1***

First, we create a scatterplot involving all data points to see if any data points are skewing the general correlation trend.

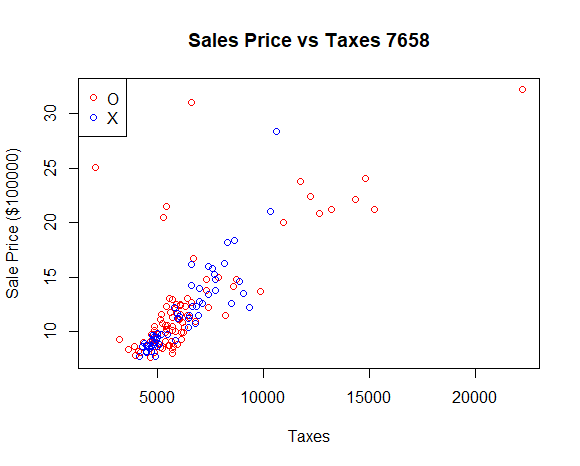


Clearly, there are two points around case 100 and around case 150 that have sale prices that are skewing the model. Thus, we remove those.

Using the new data we plot a scatterplot assessing list price against sales price.



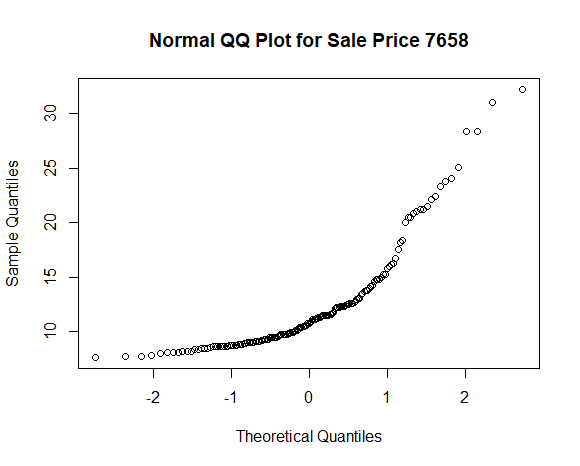
Next, we assess taxes against sale prices, using the same modified data.



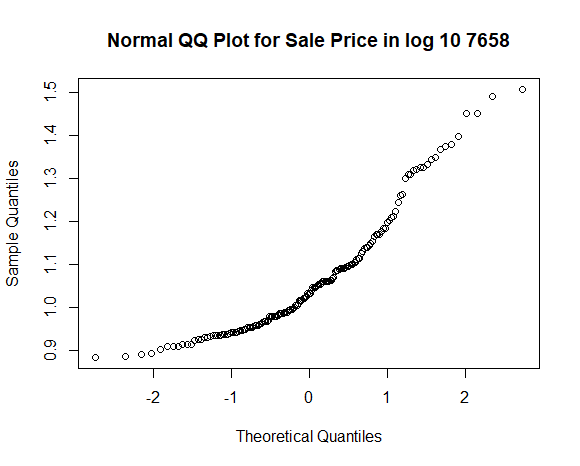
It’s pretty evident that there is a visible correlation between taxes and sale price, as well as list price and sale price. Thus, we can conclude that a linear regression would be appropriate.

***Question 2***

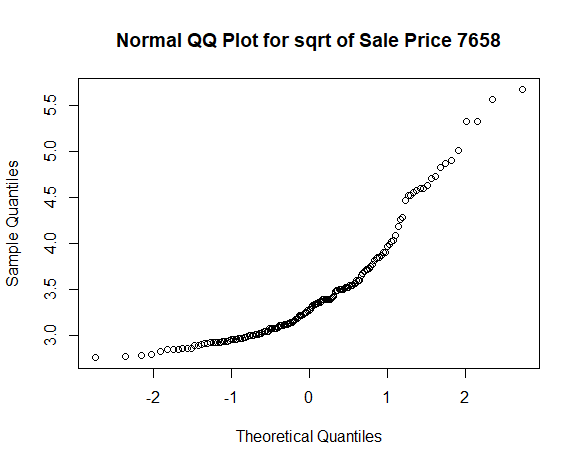
1. We can see that this model does not quite fit a normal distribution as the data seems very curved, indicating a non linear relationship between the theoretical data points and the sample data points.



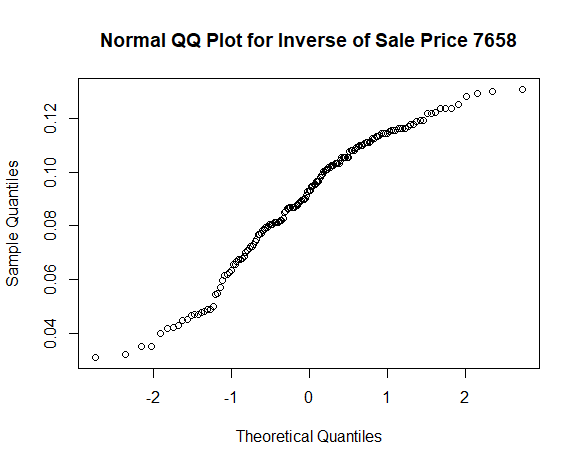
1. The sale price in log 10 seems to be getting closer to a normal distribution as the center of the data points show more evidence of being a linear relationship between theoretical and sample data points.



1. The square root of the sale prices doesn’t seem to be indicative of a normal distribution as well, since the relationship between theoretical and sample data points is even further than that of the relationship in log 10. However, it is closer than the original sale price relationship so there is slight evidence of a normal relationship.



1. Of the four relationships, the one involving the inverse of sale price seems to be the one that’s most indicative of a normal relationship as the points line up fairly linearly, relative to the other plots given above.



***Question 3***

(See appendix question 3 for R Code.)

names XandO Xonly Oonly

1 R^2 0.991 0.9901 0.9915

2 B0 estimate 0.5974 0.845 0.4999

3 B1 estimate 0.9194 0.9008 0.9262

4 Estimated error variance 0.2 0.16 0.23

5 p-value < 2.2e-16 < 2.2e-16 < 2.2e-16

6 slope 95% CI (0.9057,0.9331) (0.8767, 0.9249) (0.8767, 0.9249)

***Question 4***

The R squared values of all of the different regression models are slightly similar. However, it makes sense that for the model that represents both neighbourhoods X and O to have an R Squared value that lies in between neighbourhood X only and neighbourhood O only.

The R squared values appear slightly different as the models don’t share the exact same data so it’s trivial that they wouldn’t have the same coefficient of determination. However, they do appear similar since each of the models have a strong linear relationship and therefore the R squared value is higher, as it is able to reliably predict values within the data points.

***Question 5***

*H0* : β1 for X = β1 for O

*HA* : β1 for X ≠ β1 for O

Paired t-test

data: simon\_lmX$coefficients and simon\_lmO$coefficients

t = 0.86286, df = 1, p-value = 0.5468

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-2.193838 2.513508

sample estimates:

mean of the differences

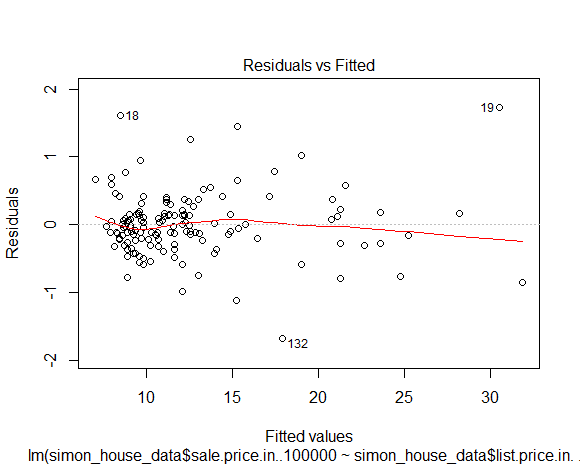
0.1598349

Since the p value is greater than 0.05, we fail to reject the null hypothesis.

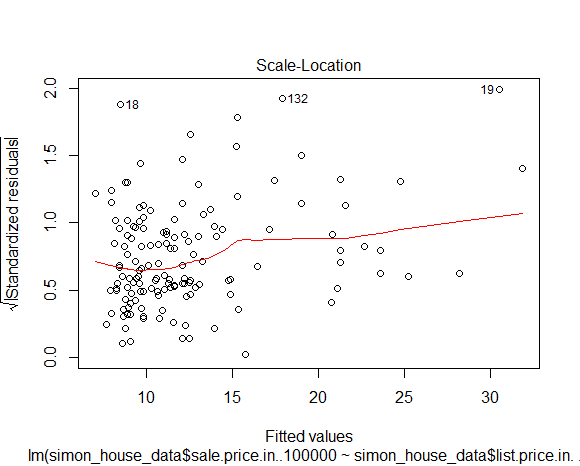
The purpose of a two sample T test is to compare whether the average difference between two slopes in simple linear regression models is really significant or if it is due to random chance. It is relevant here as that is the exact statistic we are testing. However, it would require us to do the final calculation separate from the test itself whereas the paired T test gives it immediately.

***Question 6***

We will test the model that assesses both neighbourhoods X and O in one model. We will specifically be testing for linearity of data, and whether we had outliers or high leverage points affecting our data.



The first model, testing linearity, we can see that there is no pattern in the residual plot, thus we can assume there is a linear relationship between the predictor and response variables.



The purpose of this second plot is to assess whether there were any outliers or high leverage points in our data. The plot outlines 3 points that were relatively extreme compared to the rest, however as we can see there are no points that lie outside 3 standard deviations of the mean, so it’s safe to assume the data was no affected.

Taking these plots into consideration we can conclude that there were no major violations of the normal SLR assumptions.

***Question 7***

A variable that could be previous sale price. For example, if the house was sold before, we could assess the current sale price based on the last sale price and see how prices have (likely) grown.

Another variable that may be relevant for a multiple regression test could be square footage or size of the house. Seeing whether there is a correlation between size of house/plot of land and the sale price could be very relevant to home buyers and also be present in a linear regression model.

*Appendix*

*Question 1*

> house\_data = read.csv("C:\\Users\\Simon Shen\\Documents\\sta302\_a2\\reale.csv")

> plot(house\_data$sale.price.in..100000, xlab = "Case ID", ylab = "Sale Price ($100000)", main = "Sale Prices 7658")

> simon\_house\_data <- house\_data[-c(99, 152),]  
> attach(simon\_house\_data);plot(simon\_house\_data$list.price.in..100000, simon\_house\_data$sale.price.in..100000, main = "Sales Price vs List Price 7658", xlab = "List Price ($100000)", ylab = "Sale Price ($100000)", col=c("red","blue")[location]); detach(simon\_house\_data)

> legend(x="topleft", legend = levels(simon\_house\_data$location), col=c("red","blue"), pch=1)

> attach(simon\_house\_data);plot(simon\_house\_data$taxes, simon\_house\_data$sale.price.in..100000, main = "Sales Price vs Taxes 7658", xlab = "Taxes", ylab = "Sale Price ($100000)", col=c("red","blue")[location]); detach(simon\_house\_data)

> legend(x="topleft", legend = levels(simon\_house\_data$location), col=c("red","blue"), pch=1)

*Question 2*

|  |
| --- |
| > qqnorm(simon\_house\_data$sale.price.in..100000, main="Normal QQ Plot for Sale  Price 7658")  > qqnorm(log(simon\_house\_data$sale.price.in..100000, 10), main="Normal QQ Plot  for Sale Price in log 10 7658")  > qqnorm(sqrt(simon\_house\_data$sale.price.in..100000), main="Normal QQ Plot for  sqrt of Sale Price 7658")  > qqnorm(1/(simon\_house\_data$sale.price.in..100000), main="Normal QQ Plot for  Inverse of Sale Price 7658")  *Question 3*  > plot(simon\_house\_data$list.price.in..100000, simon\_house\_data$sale.price.in..  100000, main="Sales Price vs List Prince in X and O 7658", xlab = "Sale Price  ($100000)", ylab = "List Price ($100000)")  > simon\_lm <- lm(simon\_house\_data$sale.price.in..100000 ~ simon\_house\_data$list  .price.in..100000, data= simon\_house\_data)  > abline(simon\_lm, col="red")  > plot(house\_dataX$list.price.in..100000, house\_dataX$sale.price.in..100000, main="Sales Price vs List Prince in X 7658", xlab = "Sale Price ($100000)", ylab = "List Price ($100000)")  > simon\_lmX <- lm(house\_dataX$sale.price.in..100000 ~ house\_dataX$list.price.in..100000, data= house\_dataX)  > abline(simon\_lmX, col="red")  > plot(house\_dataO$list.price.in..100000, house\_dataO$sale.price.in..100000, main="Sales Price vs List Prince in O 7658", xlab = "Sale Price ($100000)", ylab = "List Price ($100000)")  >  > simon\_lmO <- lm(house\_dataO$sale.price.in..100000 ~ house\_dataO$list.price.in..100000, data= house\_dataO)  > abline(simon\_lmO, col="red")  > summary(simon\_lm)  > confint(simon\_lm, level = 0.95)  > summary(simon\_lmX)  > confint(simon\_lmX, level = 0.95)  > summary(simon\_lmO)  > confint(simon\_lmO, level = 0.95)  > XandO <-c(0.991,0.5974,0.9194,"0.2","< 2.2e-16", "(0.9057,0.9331)")  > Xonly <- c(0.9901, 0.8450, 0.9008, "0.16", "< 2.2e-16", "(0.8767, 0.9249)")  > Oonly <- c(0.9901, 0.8450, 0.9008, "0.23", "< 2.2e-16", "(0.8767, 0.9249)")  > names <- c( 'R^2', 'B0 estimate', "B1 estimate","Estimated error variance", "p-value", "slope 95% CI")  > simon\_q3<- data.frame(names, XandO, Xonly, Oonly)  > print(simon\_q3)  *Question 4*  None.  *Question 5*  > t.test(simon\_lmX$coefficients, simon\_lmO$coefficients, var.equal = T, paired = T)  *Question 6*  > plot(simon\_lm, 1)  > plot(simon\_lm, 3)  *Question 7*  None. |
|  |
| |  | | --- | |  | |