STA302 Assignment 2

Manyi Luo - 1003799419

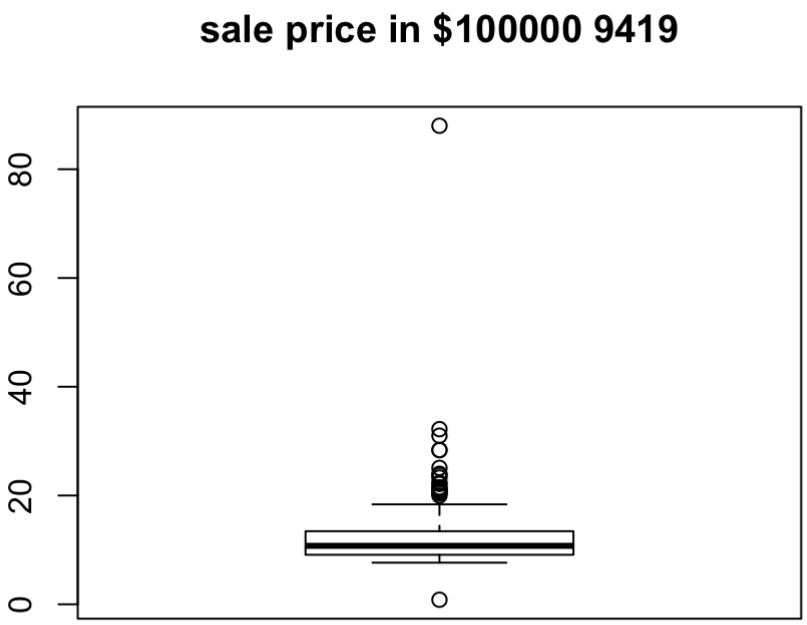
2019/10/12

Solution

1.

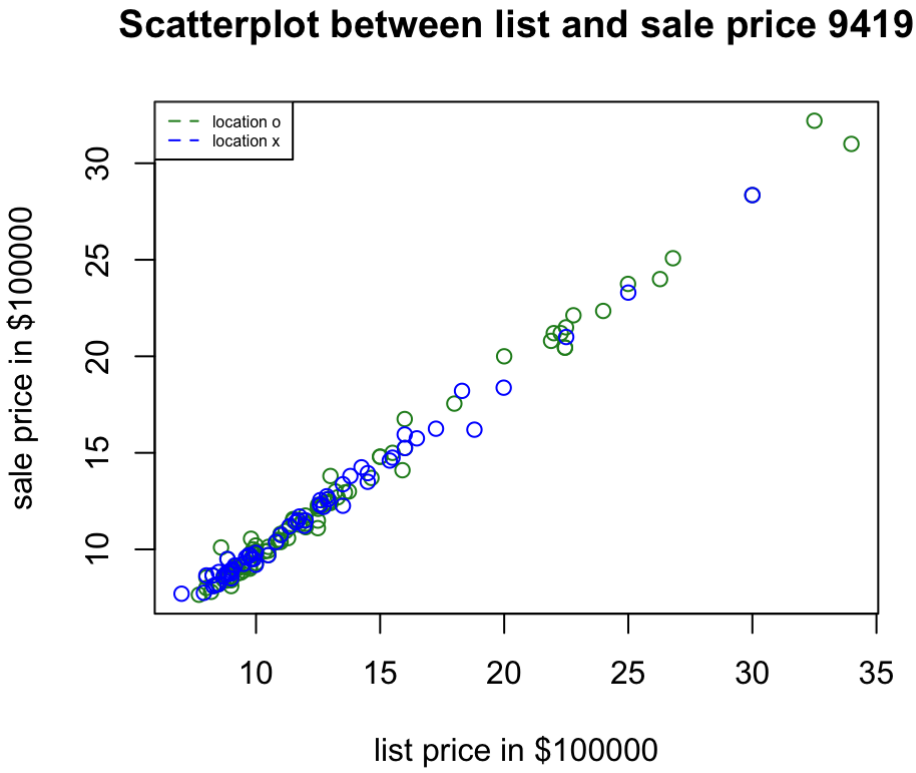
```
## 'data.frame':    163 obs. of  5 variables:
##  $ Case_ID              : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ sale.price.in..100000: num  16.8 9.3 10.4 11.3 11.5 ...
##  $ list.price.in..100000: num  16 10 10.8 12 11.7 ...
##  $ taxes                : int  6683 6119 6477 6500 6494 10631 5352 4607 4714 5706 ...
##  $ location             : Factor w/ 2 levels "O","X": 1 1 1 1 1 1 1 1 1 1 ...
```

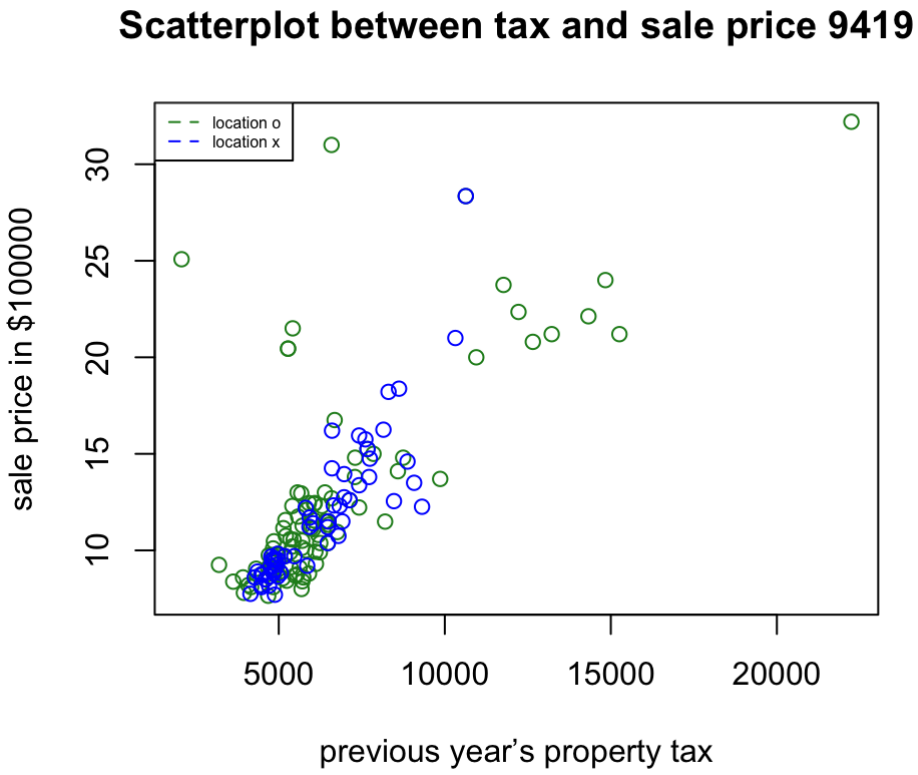
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.835   9.100  10.750  12.720  13.438  87.990
```



In order to exaimne the data, I first calculate its min, 1st quantile, median, mean, 3rd quantile and max values. The max value is 87.99 and the min value is 0.835. I also further look into its distribution by creating a boxplot. Outliers are extreme values located further than the 1st quartile and 3rd quartile. From the boxplot of sale price in

$100000, I can identify many outliers, and the largest and smallest points are my best choice to remove from the original data set, since the biggest and smallest points are the most extreme values among other outliers.

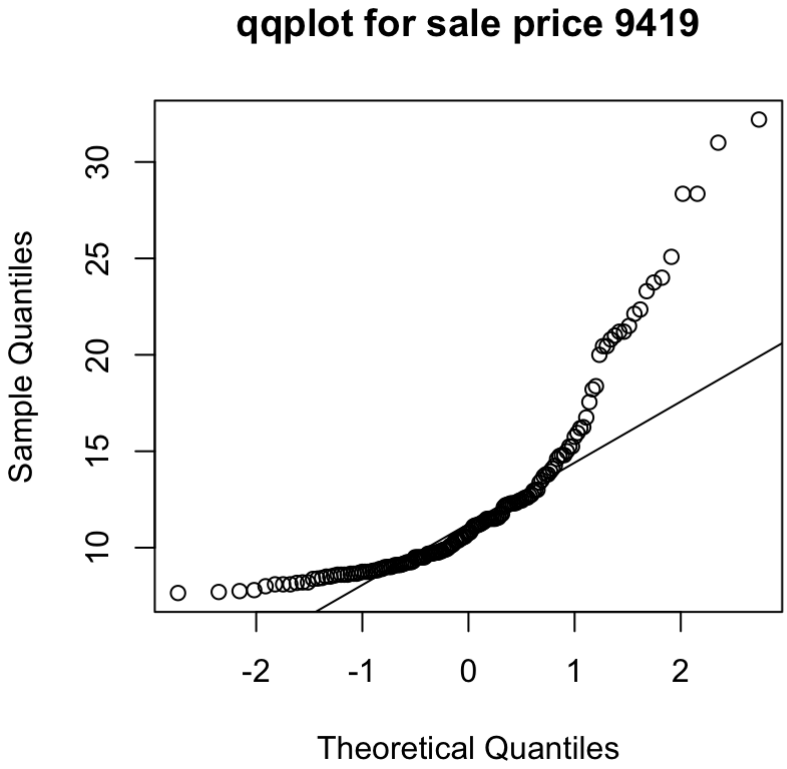




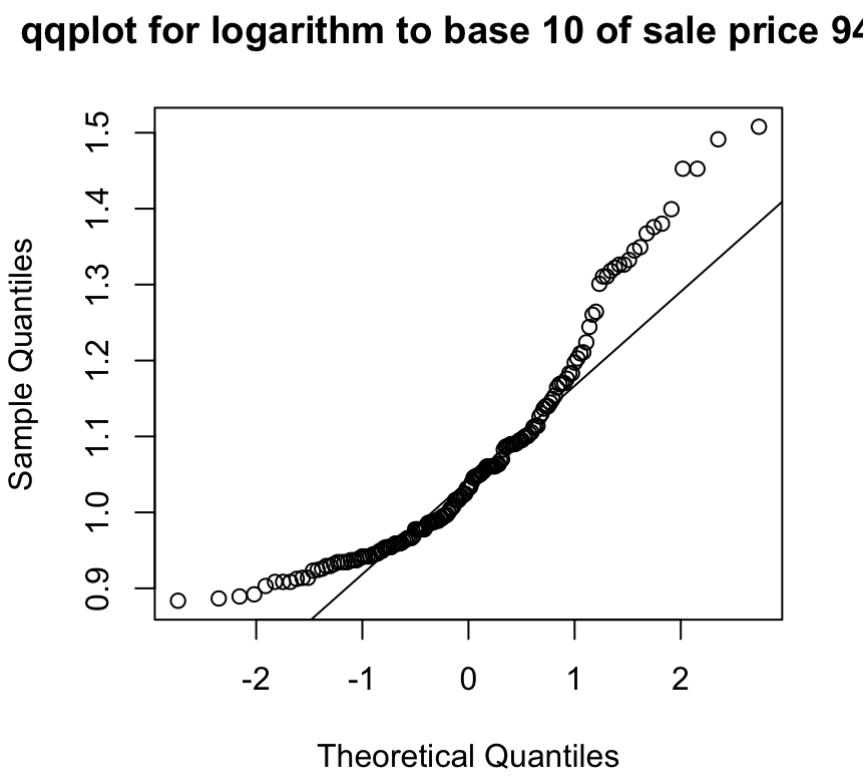
Comparing two models, the linear regression model for sale price and list price is more appropriate. This is because its standard deviations are nearly the same across all values of the independent variable. This implies homoscedasticity, which is prefered. In contrast, the linear regression model for sale price and tax has different variances among those points (heteroscedasticity), making it appropriate.

2.

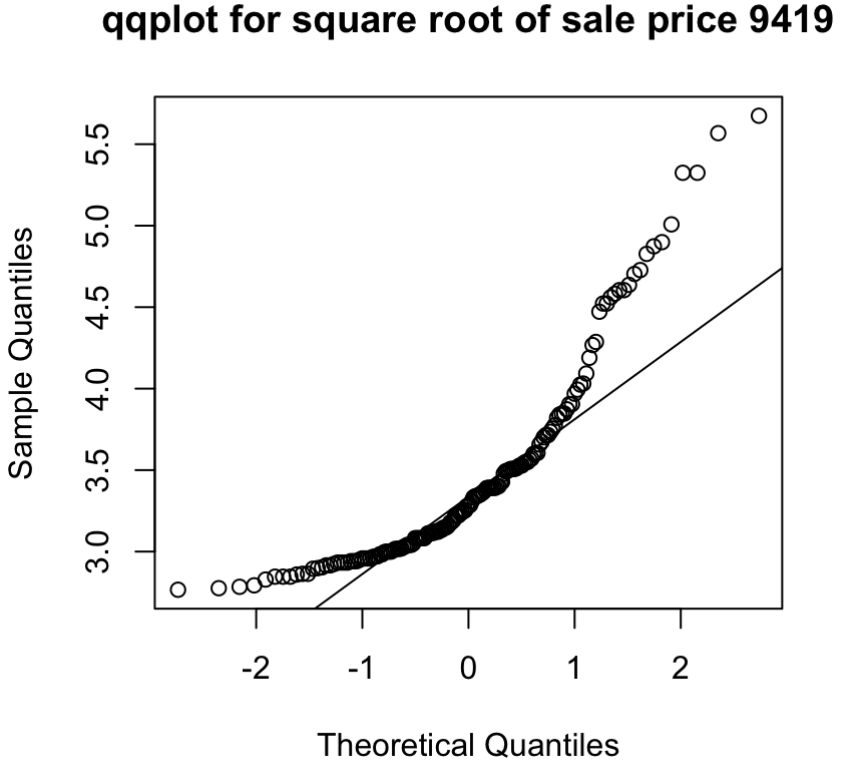
(a) sale price



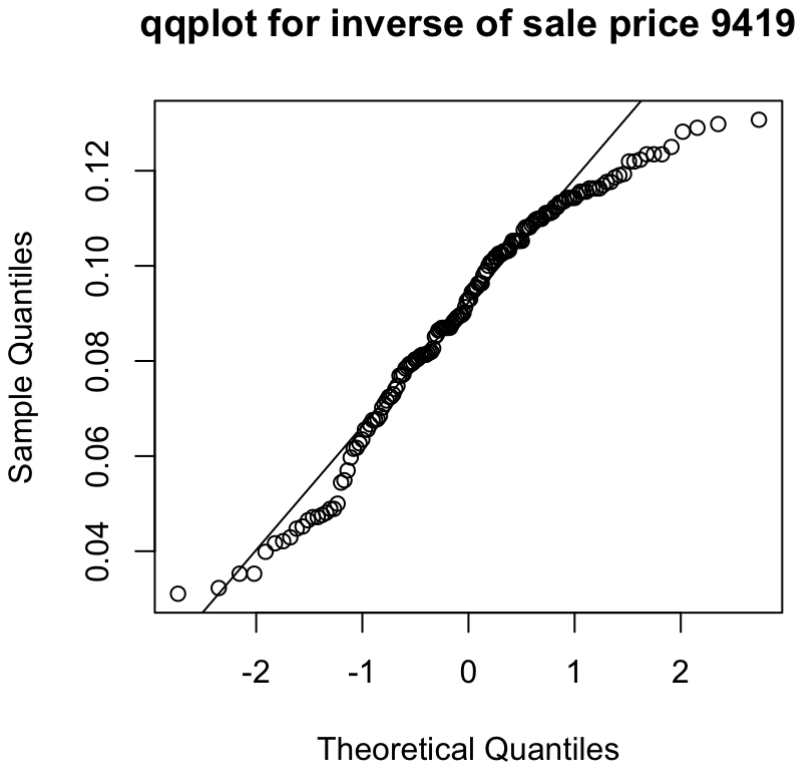
(b) logarithm to base 10 of sale price



(c) square root of sale price



(d) inverse of sale price



For a normal QQ plot, if most of the data points lie on the QQ fitted line, then the distribution presented by the normal QQ plot will approximate to normal distribution. From the four plots generated above, we can see that none of them show a perfect fit toward the QQ line. Among them, the QQ plot for the inverse of sale price performs best, which can be recognized as approximately normal, since a majority of its points are distributed on the qqline, for example: theoretical quantiles (-1, 1). In contrast, the distribution of sale price, the distribution of sale price with logarithm to base 10, and the distribution of sale price with square root do not approximate to normal, since most of their points deviate a lot from the qqline, especially on theoretical quantiles (-2, -1) and (1, 2).

3.

```
##
## Call:
## lm(formula = new_reale_manyi$sale.price.in..100000 ~ new_reale_manyi$list.price.in..1
00000)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.68330 -0.21387 -0.02146  0.16470  1.72149
##
## Coefficients:
##                                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)                          0.597466   0.095682   6.244 3.72e-09
## new_reale_manyi$list.price.in..100000 0.919459   0.006948 132.335  < 2e-16
##
## (Intercept)                          ***
## new_reale_manyi$list.price.in..100000 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4575 on 159 degrees of freedom
## Multiple R-squared:  0.991,  Adjusted R-squared:  0.9909
## F-statistic: 1.751e+04 on 1 and 159 DF,  p-value: < 2.2e-16
```

```
##                                         2.5 %    97.5 %
## (Intercept)                          0.4084937 0.7864386
## new_reale_manyi$list.price.in..100000 0.9057369 0.9331813
```

```
##
## Call:
## lm(formula = new_realeX_manyi$sale.price.in..100000 ~ new_realeX_manyi$list.price.i
n..100000)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.58064 -0.19451 -0.01204  0.14104  0.87977
##
## Coefficients:
##                                       Estimate Std. Error t value
## (Intercept)                            0.84505    0.15801   5.348
## new_realeX_manyi$list.price.in..100000  0.90083    0.01203  74.886
##                                       Pr(>|t|)
## (Intercept)                            1.7e-06 ***
## new_realeX_manyi$list.price.in..100000  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4042 on 56 degrees of freedom
## Multiple R-squared:  0.9901, Adjusted R-squared:  0.9899
## F-statistic:  5608 on 1 and 56 DF,  p-value: < 2.2e-16
```

```
##                                           2.5 %    97.5 %
## (Intercept)                            0.5285138 1.1615905
## new_realeX_manyi$list.price.in..100000 0.8767317 0.9249268
```

```
##
## Call:
## lm(formula = new_realeO_manyi$sale.price.in..100000 ~ new_realeO_manyi$list.price.i
n..100000)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.12708 -0.25692 -0.01229  0.14794  1.64368
##
## Coefficients:
##                                        Estimate Std. Error t value
## (Intercept)                            0.499979   0.120666   4.144
## new_realeO_manyi$list.price.in..100000 0.926232   0.008548 108.360
##                                        Pr(>|t|)
## (Intercept)                            7.12e-05 ***
## new_realeO_manyi$list.price.in..100000  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.483 on 101 degrees of freedom
## Multiple R-squared:  0.9915, Adjusted R-squared:  0.9914
## F-statistic: 1.174e+04 on 1 and 101 DF,  p-value: < 2.2e-16
```

```
##                                           2.5 %    97.5 %
## (Intercept)                            0.2606113 0.7393471
## new_realeO_manyi$list.price.in..100000 0.9092760 0.9431889
```

```
##          R_square est_intercept est_slope est_variance (error)    pvalue
## alldata 9.910e-01     5.975e-01 9.195e-01            4.575e-01 2.000e-16
## OnlyX   9.901e-01     8.451e-01 9.008e-01            4.042e-01 2.000e-16
## OnlyO   9.915e-01     5.000e-01 9.262e-01            4.830e-01 2.000e-16
##         upper_bond lower_bond
## alldata  9.057e-01  9.332e-01
## OnlyX    8.767e-01  9.249e-01
## OnlyO    9.093e-01  9.432e-01
```

4.

R^2 measures the percentage of response variable's variation that is explained by the linear regression model; in other words, R^2 shows how much the data is fitted to the regression model. In this case, the R^2 for sale price of all data is 0.991, the R^2 for sale price of neighbourhood X is 0.9901, and the R^2 for sale price of neighbourhood O is 0.9915. The three R^2 values appear to be similar, and the sale price of neighbourhood O performs best among the three, followed by sale price of all data and sale price of neighbourhood X. They appears to be similar (with small deviations) because they might come from the same population.

5.

Essentially, we are using t-test to compare the difference between the slopes of 2 regression models seperated by locations and we would like to determine if they really came from the same population. Our null hypothesis in this case is to assume the two slopes are equal, which is $\beta_1 locationX = \beta_1 locationO$, meaning that they come from the same population. The alternative hypothesis in contrast is $\beta_1 locationX \neq \beta_1 locationO$.

Assuming the null hypothesis is true, t test statistics will be $(\beta_1 locationX - \beta_1 location0)/s_p * \sqrt{(1/n_X) + (1/n_O)}$, where s_p is the sample standard deviation, which equals to the square root of $s_p^2 = \frac{(nX-2)*s_X^2+(nO-2)*s_O^2}{nX+nO-4}$. The t test statistics we got follows $T_(n_X + n_O - 4)$.

Therefore, we can compute the pooled estimate standard deviation by using the square of $\beta_1 locationX$'s standard error (0.01203), and the square of $\beta_1 locationO$'s standard error (0.008548), and then take a square root of the entire value. In this case, n_X is 58, n_O is 103. So the pooled standard deviation equals to 0.00993.

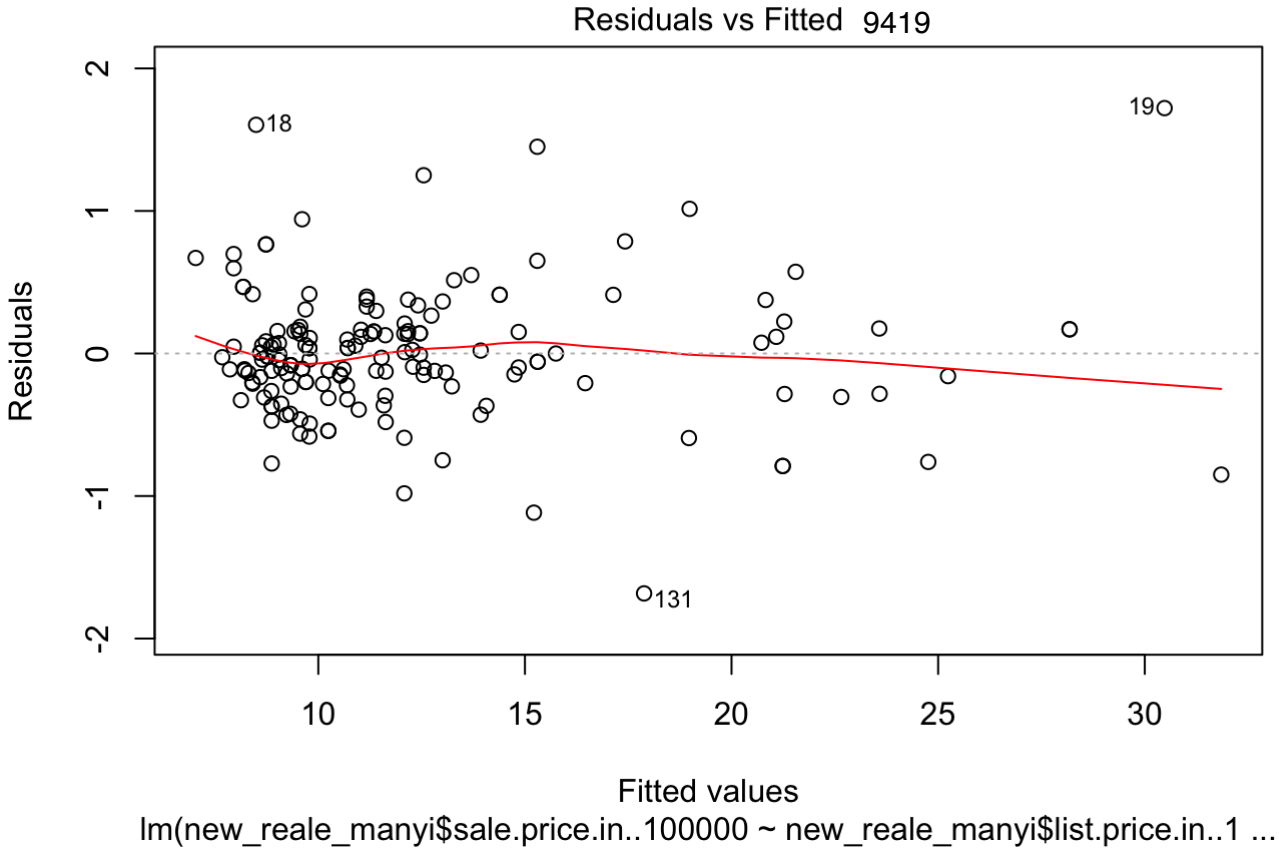
Using the pooled standard deviation and the formula for t test statistics provided above, we can further calculate the t test statistic. In this case, $\hat{\beta_X} = 0.90083$ and $\hat{\beta_O} = 0.92632$, so t = −15.58. This follows a degree of freedom of 157. The value of t test statistics represents the size of deviation comparing to the null hypothesis over the standard error of sample data.

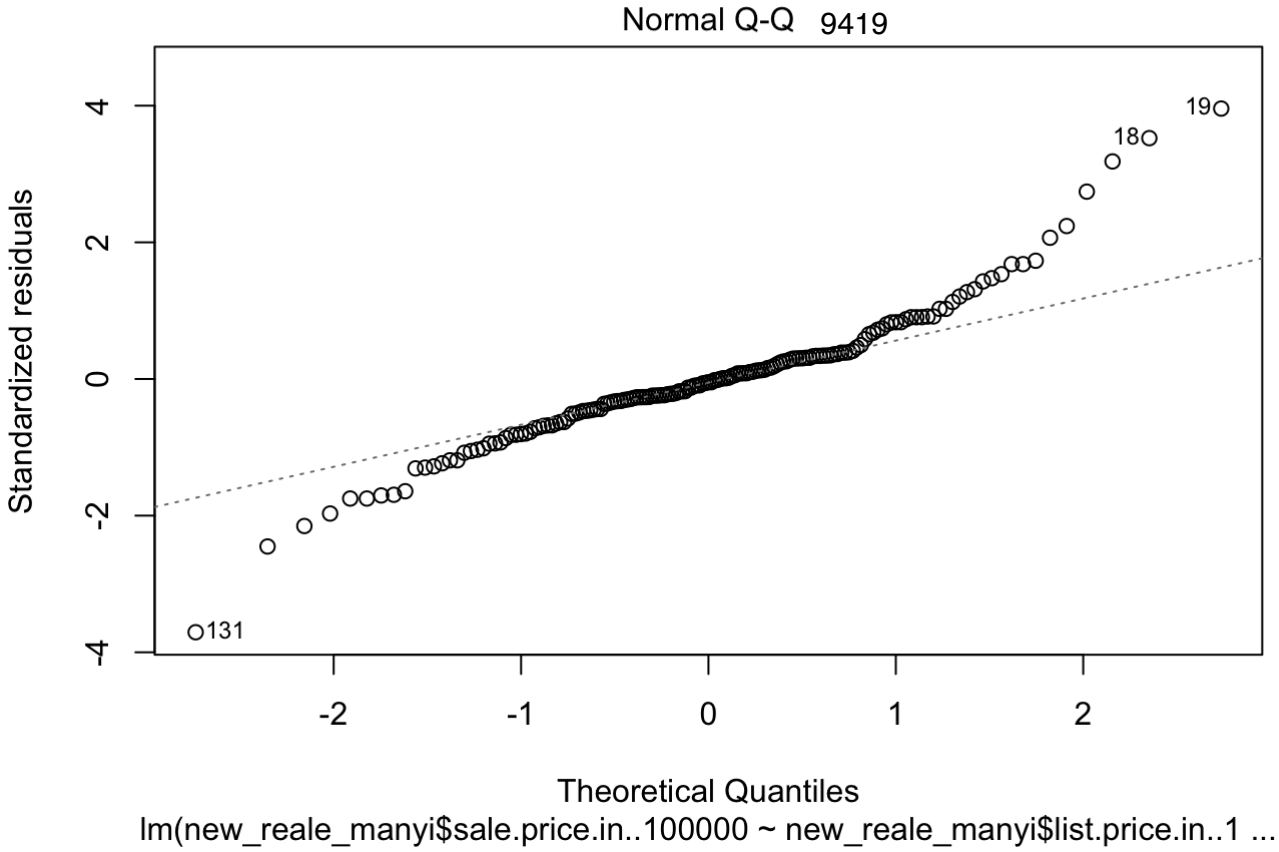
```
## [1] 1.11881e-33
```

Now we can find the p value, using the result of t test statistics and degree of freedom, which is approximately zero. Since our p value is smaller than 0.05, we can conclude that there is a strong evidence to reject our null hypothesis that there's no difference between the slopes of two location's linear regression models and a significant difference does exist.

Also, its crucial to note that since we have a sufficiently large amount of data, we can assume the population parameters and sampling estimators are normally distributed, and they are independent of each other. The variance of estimated slope of location X and O are approximately zero as well. Therefore, we are able to conduct two sample t-test by pooling their variance and comparing the estimated slopes.

6.


Residuals vs Fitted 9419
Residuals
Fitted values
lm(new_reale_manyi$sale.price.in..100000 ~ new_reale_manyi$list.price.in..1 ...
18
19
131
2
1
0
-1
-2
10
15
20
25
30


Normal Q-Q 9419
Standardized residuals
Theoretical Quantiles
lm(new_reale_manyi$sale.price.in..100000 ~ new_reale_manyi$list.price.in..1 ...
19
18
131
4
2
0
-2
-4
-2
-1
0
1
2

The model I selected is the simple linear regression for sale price toward list price for all data. This is because the other two regression models are grouped by dummy variable (location, which only produce two catagorical results: X and O) from all data, making them insignificant.

In order to find out any violations of the normal error SLR assumptions, I create a scatterplot of “residuals versus fits” and a normal probability plot of the residuals. We can see from the “residuals versus fits plot” that most of the points are aggregated on the left, rather than scattered with out any pattern. This means that the variance of the residuals is not the same across all values of the x-axis (fitted values). So this might violate the assumption of equal variance.

In the second plot (Normal Q-Q) for residuals, the observations around therotical quantile (-1.5, -2) and (1.5, 2) deviate a lot from the fitted line. Approximately, this means that the residuals are not perfectly normal distributed, which violate the assumption of normality of errors.

7.

Other two potential numeric predictors that could be used to fit a multiple linear regression for sale price might be housing size and housing age from construction.

Appendix

```
Q1
reale_manyi <- read.csv("/Users/meow/Desktop/STA302/A2/reale.csv")
str(reale_manyi)
sale_price = reale_manyi$sale.price.in..100000
summary(sale_price)
boxplot(sale_price, main = "sale price in $100000 9419")
new_reale_manyi = subset(reale_manyi, reale_manyi$sale.price.in..100000 != max(reale_man
yi$sale.price.in..100000) & reale_manyi$sale.price.in..100000 != min(reale_manyi$sale.pr
ice.in..100000))

new_realeO_manyi = subset(new_reale_manyi, new_reale_manyi$location == "O" )
new_realeX_manyi = subset(new_reale_manyi, new_reale_manyi$location == "X" )
plot(new_reale_manyi$list.price.in..100000, new_reale_manyi$sale.price.in..100000, col =
ifelse(new_reale_manyi$location == 'O',"forestgreen","blue"), xlab = "list price in $100
000", ylab = "sale price in $100000", main = "Scatterplot between list and sale price 94
19")
legend("topleft", legend = c("location o ", "location x"), col = c("forestgreen", "blu
e"), lty = 2, cex = 1)
plot(new_reale_manyi$taxes, new_reale_manyi$sale.price.in..100000, col = ifelse(new_real
e_manyi$location == 'O',"forestgreen","blue"), xlab = "previous year's property tax", yl
ab = "sale price in $100000", main = "Scatterplot between tax and sale price 9419")
legend("topleft", legend = c("location o ", "location x"), col = c("forestgreen", "blu
e"), lty = 2, cex = 1)

Q2
#(a).
qqnorm(new_reale_manyi$sale.price.in..100000, main = "qqplot for sale price 9419")
qqline(new_reale_manyi$sale.price.in..100000)

#(b).
log_sale = log10(new_reale_manyi$sale.price.in..100000)
qqnorm(log_sale, main = "qqplot for logarithm to base 10 of sale price 9419")
qqline(log_sale)

#(c).
sqrt_sale = sqrt(new_reale_manyi$sale.price.in..100000)
qqnorm(sqrt_sale, main = "qqplot for square root of sale price 9419")
qqline(sqrt_sale)

#(d).
inverse_sale = 1/(new_reale_manyi$sale.price.in..100000)
qqnorm(inverse_sale, main = "qqplot for inverse of sale price 9419")
qqline(inverse_sale)

Q3
alldata = lm(new_reale_manyi$sale.price.in..100000 ~ new_reale_manyi$list.price.in..1000
00)
summary(alldata)
confint(alldata)

onlyX = lm(new_realeX_manyi$sale.price.in..100000 ~ new_realeX_manyi$list.price.in..1000
00)
summary(onlyX)
```

```
confint(onlyX)

onlyO = lm(new_realeO_manyi$sale.price.in..100000 ~ new_realeO_manyi$list.price.in..1000
00)
summary(onlyO)
confint(onlyO)

manyitab<-matrix(c(0.991,0.5975,0.9195,0.4575,2e-16,0.9057,0.9332,0.9901,0.8451,0.9008,
0.4042,2e-16,0.8767,0.9249,0.9915,0.5000,0.9262,0.4830,2e-16,0.9093,0.9432),ncol=7,byrow
=TRUE)
colnames(manyitab)<-c("R_square", "est_intercept","est_slope","est_variance (error)","pv
alue","upper_bond","lower_bond")
rownames(manyitab)<-c("alldata","OnlyX", "OnlyO")
manyitab<-as.table(manyitab)
manyitab

Q5
p_stat <- 2*pt(-abs(15.58), df = 157)
p_stat

Q6
plot(alldata, 1)
plot(alldata, 2)
```