STA 108 Project I

Junyao Lu, Fengshuo Song
STA108 SectionB
Instructor:Dr. JoAnna C. Whitener

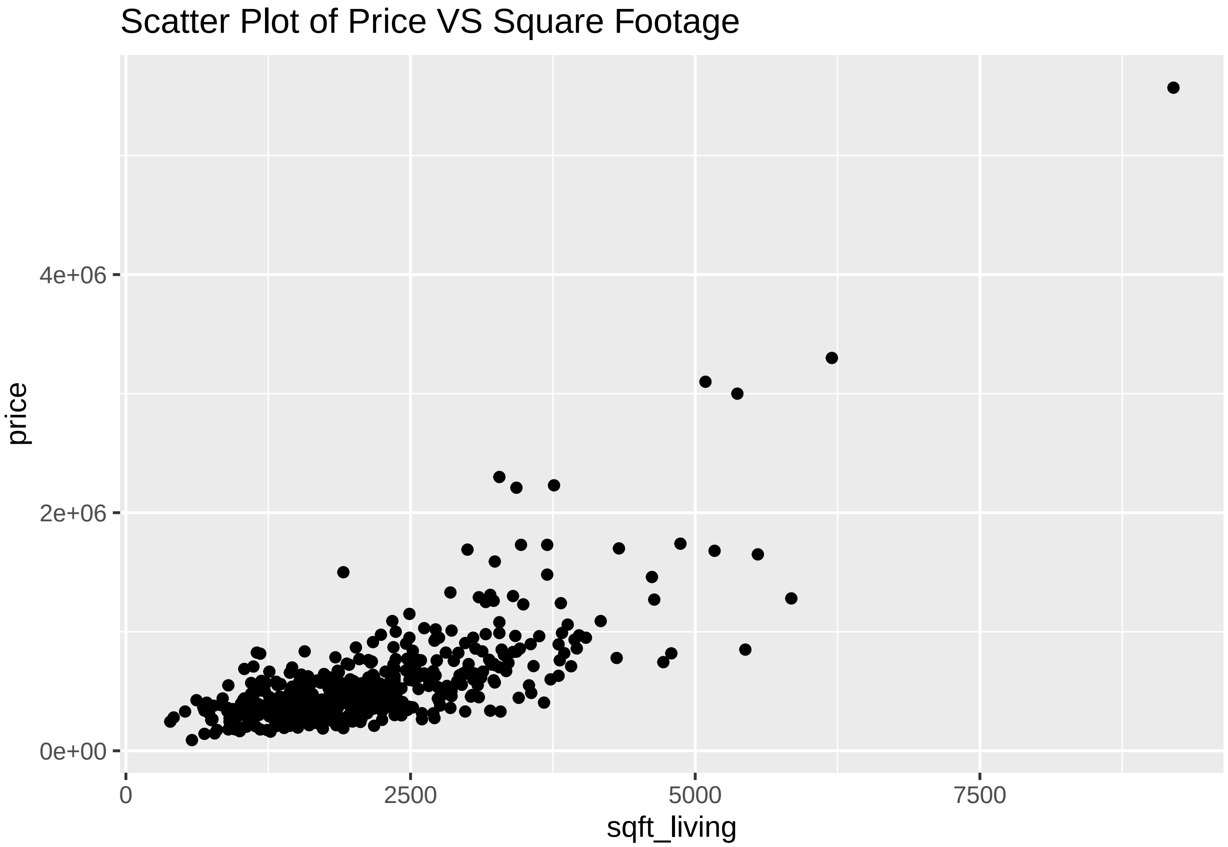
10/25/2019

I. Introduction

Our question of interest is to find out whether there is a significant positive linear relationship between the price and the living square footage for houses sold in King County, Washington in a particular year. We are going to build a regression model to help predict sale prices based on the living square footage of the houses. Based on the 500 sets of randomly selected data, we are going to use a linear regression model to answer our question. We identify price as the independent variable X and living square footage as the dependent variable Y. We are interested in this question because we want to figure out the predicted prices for living square footage of 2800, 3200, 8000. In the following report, we are going to utilize multiple approaches with the use of Rstudio, including scatterplots, histograms, normal Q-Q plots, confidence intervals, hypothesis tests and so on.

II. Summary of The Data

The scatter plot of prices and living square footages is shown as following:



The figure above is a scatter plot of the price (y) against living square footage (x). It plots 500 sets of data through which we can see if there is a linear relationship between price and living square footage. From the scatter plot, we can clearly observe that there is an uphill pattern as x increases. Also, the pattern generally looks like a straight line. These observations roughly indicate that there is a positive linear relationship between living square footage and price, which means as the living square footage increases, the price of the house tends to increase.

The summary values of the living square footage (X) is shown as following:

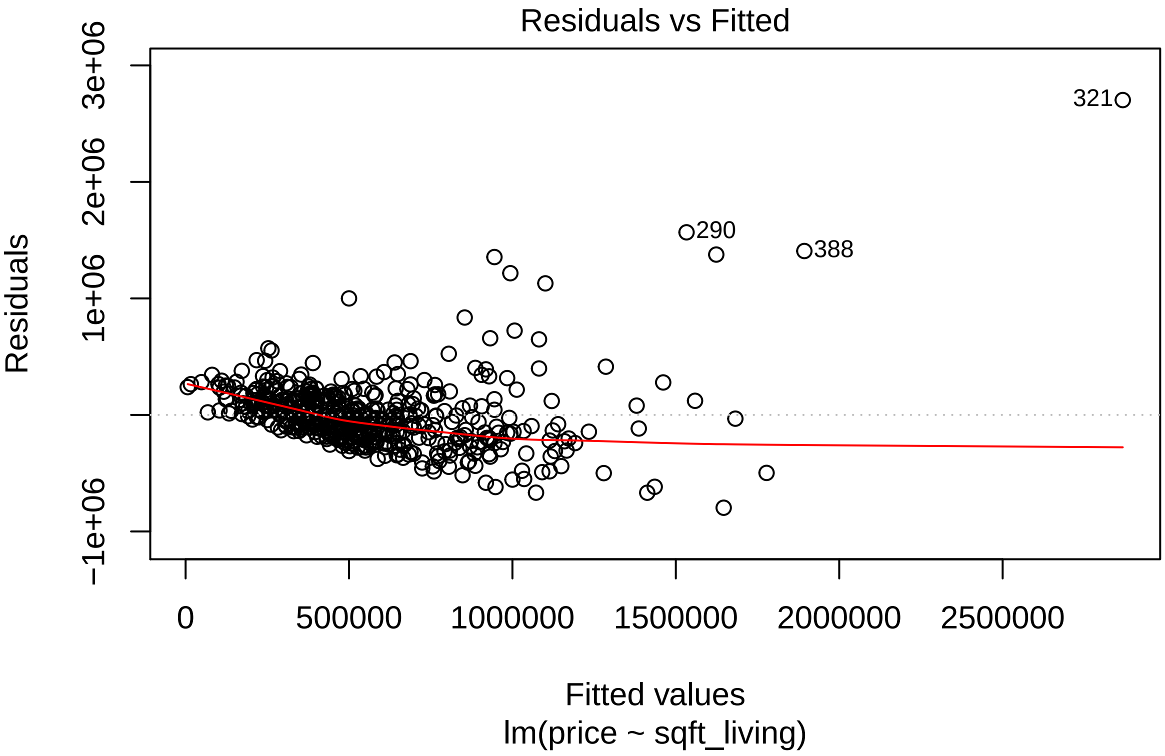
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     390    1460    1940    2115    2562    9200
```

The summary values of the price (Y) is shown as following:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   90000  333722  460000  566594  646250 5570000
```

III. Diagnostics

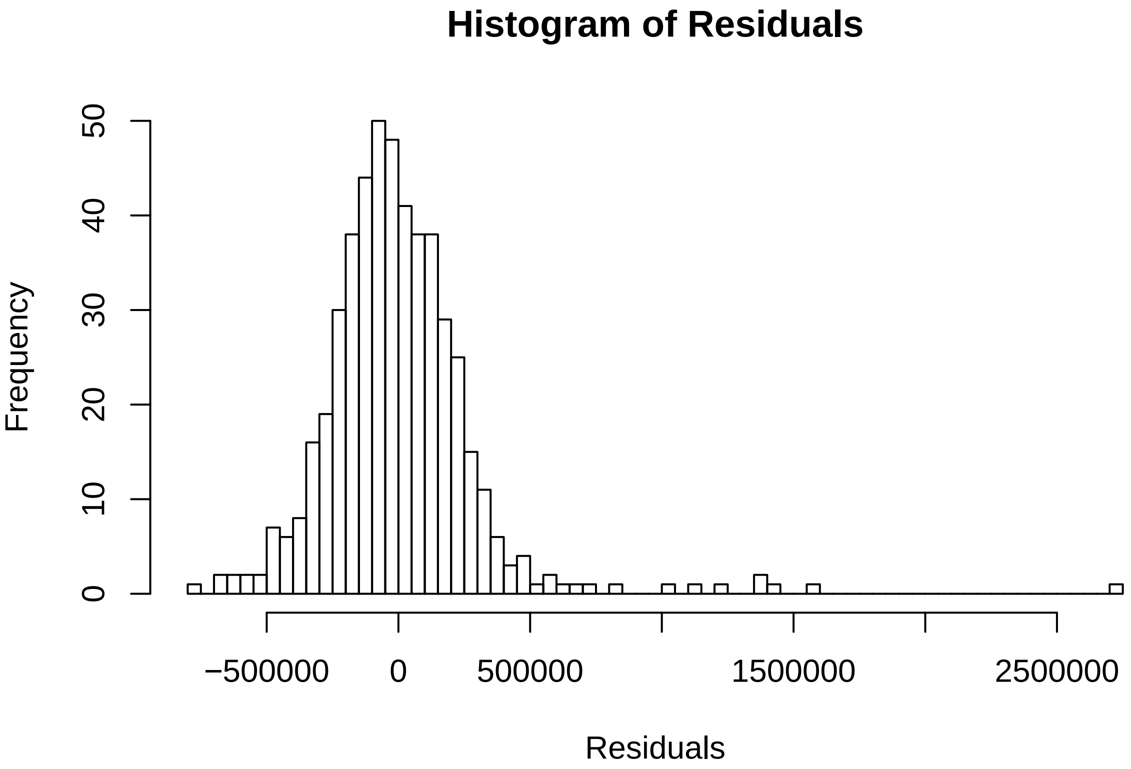
To do our regression model diagnostics, we first plot the residuals VS fitted plot. The residuals vs fitted plot is shown as following:



For the plot above, the points are not evenly distributed vertically. Also, a pattern could be observed in the plot. There is a slight positive to negative trend as moving along the fitted line. The variances are not constant as well, as the plot points show an increasing spread from left to right. Most importantly, we could observe 3 outliers from the plot, which could be the reason for non-constant variances and non-linearity.

Then, we plot the histogram of the residuals.

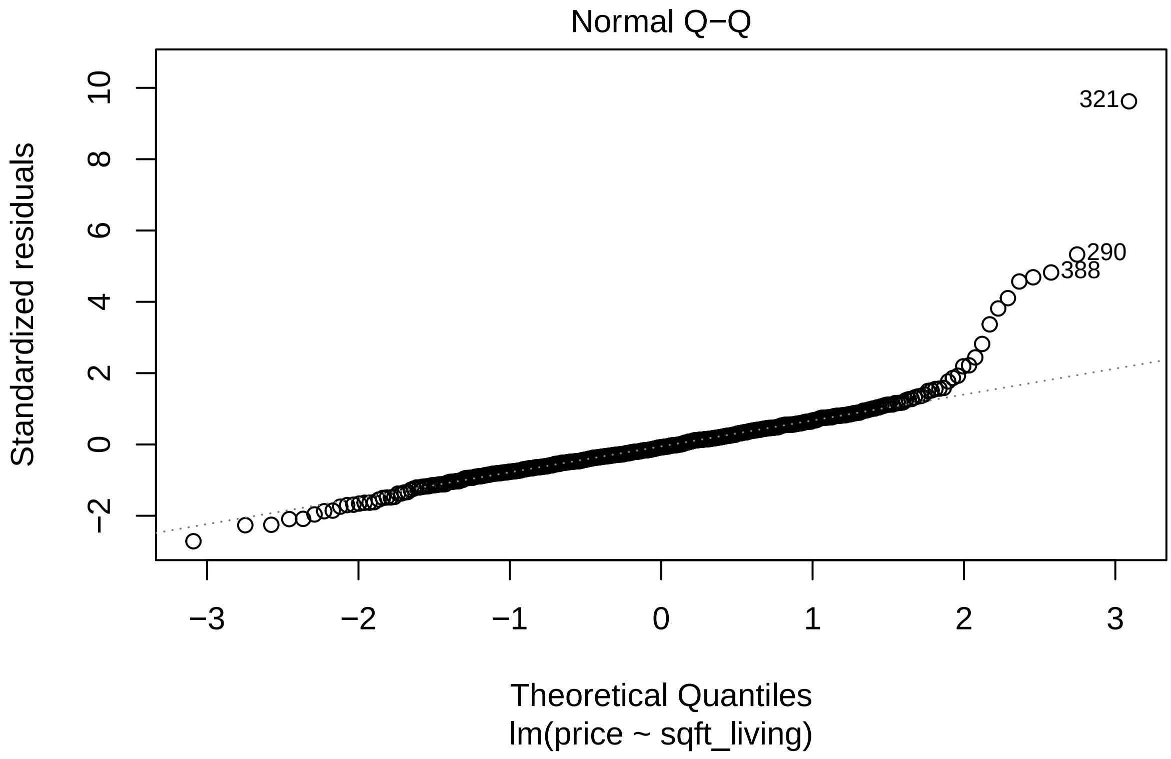
The histogram is shown as following:



From the histogram, we can observe a slight positive skewness, which indicates that the residuals may not be normally distributed. Also, we could observe outliers from the histogram.

Last but not least, we plot the normal Q-Q plot.

The normal Q-Q plot is shown as following:



In the Q-Q plot, we can see that most of the points lie on the diagonal line. However, on both tails, points are deviating significantly from the line. Also, there are 3 outliers observed. Therefore, non-normality is observed from the Normal Q-Q plot.

In conclusion, all of the three plots show that there are assumptions violated. The assumptions violated includes:

1. The residuals do not have a constant variance;
2. The residuals are not normally distributed;
3. There are outliers in the data.

IV. Analysis

The summary of the model fit is shown as following:

```
## 
## Call:
## lm(formula = price ~ sqft_living, data = house.data)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -796455 -161412  -23841  129386 2702388 
## 
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)    
## (Intercept) -120325.52   31724.19  -3.793 0.000167 ***
## sqft_living     324.78      13.62  23.847  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 297200 on 498 degrees of freedom
## Multiple R-squared:  0.5331, Adjusted R-squared:  0.5322 
## F-statistic: 568.7 on 1 and 498 DF,  p-value: < 2.2e-16
```

β 0 and β 1 are:

```
##  (Intercept)  sqft_living 
## -120325.5225     324.7759
```

The estimated linear model is Y = -120325.5225 + 324.7759X.

In order to prove whether there is a significant positive linear relationship between living square footage and price, we decide to conduct a hypothesis test.

The null hypothesis and alternative hypothesis are stated as following:

H0: β 1 ≤ 0

Ha: β 1 > 0

From the summary above, we can find the test statistic is -3.793 and the p-value is 0.000167. If there is no significant linear relationship between living square footage and price, we would observe our data or more extreme less than 0.0167% of the time.

Assuming $\alpha = 0.05$, we reject H0 because p-value < 0.05.

Therefore, at a 5% significant level, there is sufficient evidence to support our claim that there is a significant positive linear relationship between living square footage and price.

Then, we want to find the confidence interval for the slope of the regression line.

```
##                    2.5 %      97.5 %
## (Intercept) -182655.2793 -57995.7656
## sqft_living     298.0178    351.5339
```

We are 95% confident that when the living square footage increases by 1 unit, the estimated change in price is between 298.0178 and 351.5339 on average.

For the information gained from the confident interval, both lower bound and upper bound are positive numbers, which indicates a significant positive linear relationship. We can also see that the β 1 we get from our regression line lies between the confidence interval, therefore, it supports our claim that there is a significant positive linear correlation between living square footage and price.

V. Interpretation

From the hypothesis test, we conclude that at 5% significant level, there is sufficient evidence to support the claim that there is a significant positive linear relationship between living square footage and price.

From the confidence interval, we conclude that we are 95% confident that when the living square footage increases by 1 unit, the estimated change in price is between 298.0178 and 351.5339 on average.

To conclude, we successfully prove that there is a significant positive linear relationship between sqaure footage and price. The estimated regression line is Y = -120325.5225 + 324.7759X.

To interpret the slope, we can say that when the sqaure footage increases by 1 unit, we expect the price would increase by 324.7759 on average. The interpret of the interception is not applicable, because when living square footage is 0, the price would be -120325.5225, which is not possible in reality. So, there is no actual meaning of the interception for our regression line.

VI. Prediction Results

The prediction results are following:

1.

```
##        fit      lwr     upr
## 1 789046.9 204211.9 1373882
```

We predict that the average house price for houses with living square footage 2800 is between 204211.9 and 1373882.

2.

```
##        1
## 918957.3
```

We predict that the price of a particular house with living square footage 3200 is 918957.3.

3.

```
##       fit     lwr     upr
## 1 2477881 1872495 3083268
```

We predict that the average houses price for houses with living square footage 8000 is between 1872495 and 3083268.

VII. Conclusion

All in all, with all the work done above, we are able to conclude that there is a significant positive linear relationship between living square footage and the price of the houses sold in King County, Washington in a particular year.

First of all, we roughly observe a positive linear relationship between the living square footage and the price from the scatter plot. Then, to further investigate the relationship between them, we conduct a diagnostics on our regression model. Then, with the help of the residuals vs fitted values plot, the histogram of residuals and the normal Q-Q plot, we conclude that several assumptions were violated, including non-normality, non-constant variances, as well as outliers. Outliers could affect the accuracy of our analysis. To further analyze the model, a hypothesis test is done and the result supports our claim that at a 5% significance level, there is a significant positive linear relationship. Also, we calculate the confidence interval for β 1 to double-check the linear relationship. Both lower bound and upper bound are positive numbers, which represent a significant positive linear relationship. The result also supports our claim as β 1 lies between the confidence interval.

After all the approaches have done on our regression line, we are then confident to use our estimated regression line to predict the price information for houses with living square footage of 2800, 3200 and 8000. The results show that as the living square footage increases, the price of the house also tends to increases, which fits our conclusion.

Appendix Code

```r
library(ggplot2)
house.data = read.csv("~/Desktop/KingCounty.csv")
# II
ggplot(data = house.data, mapping = aes(x=sqft_living, y=price)) +
  geom_point() + ggtitle("Scatter Plot of Price VS Square Footage")
summary(house.data$sqft_living)
summary(house.data$price)
# III
house.model=lm(price~sqft_living, data=house.data)
plot(house.model, which=1)
hist(house.model$residuals, breaks = 50, main="Histogram of Residuals", xlab="Residuals")
plot(house.model, which=2)
# IV
lm.fit=lm(price~sqft_living, data=house.data)
summary(lm.fit)
lm.fit$coefficients
confint(house.model, level=0.95)
# VI
data1 = data.frame(sqft_living = 2800)
predict(house.model, data1, interval="prediction")
data2 = data.frame(sqft_living = 3200)
predict(house.model, newdata=data2)
data3 = data.frame(sqft_living = 8000)
predict(house.model, data3, interval="prediction")
```