Code

R Notebook

2/12/23

This is an R Markdown (http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code.

Try executing this chunk by clicking the *Run* button within the chunk or by placing your cursor inside it and pressing *Ctrl+Shift+Enter*.

Linear regression is a statistical method that is used to model the relationship between a dependent variable and one or more independent variables.The goal is to find a line of best fit that shows the relationship between the independent and dependent varibles. The pros of linear regression are that it can easily handle multiple independent variables. It is also fast to train. The cons are that it assumes a linear relationship between the independent and dependent variables and is sensitive to outliers.

Hide

```
data(Paris)
str(Paris)
plot(Paris$squareMeters~Paris$price, xlab="square-meters", ylab="price")
abline(lm(Paris$squareMeters~Paris$price), col="red")
dim(Paris)
head(Paris)
```

The plot shows a distinctly linear relationship between square feet and price.

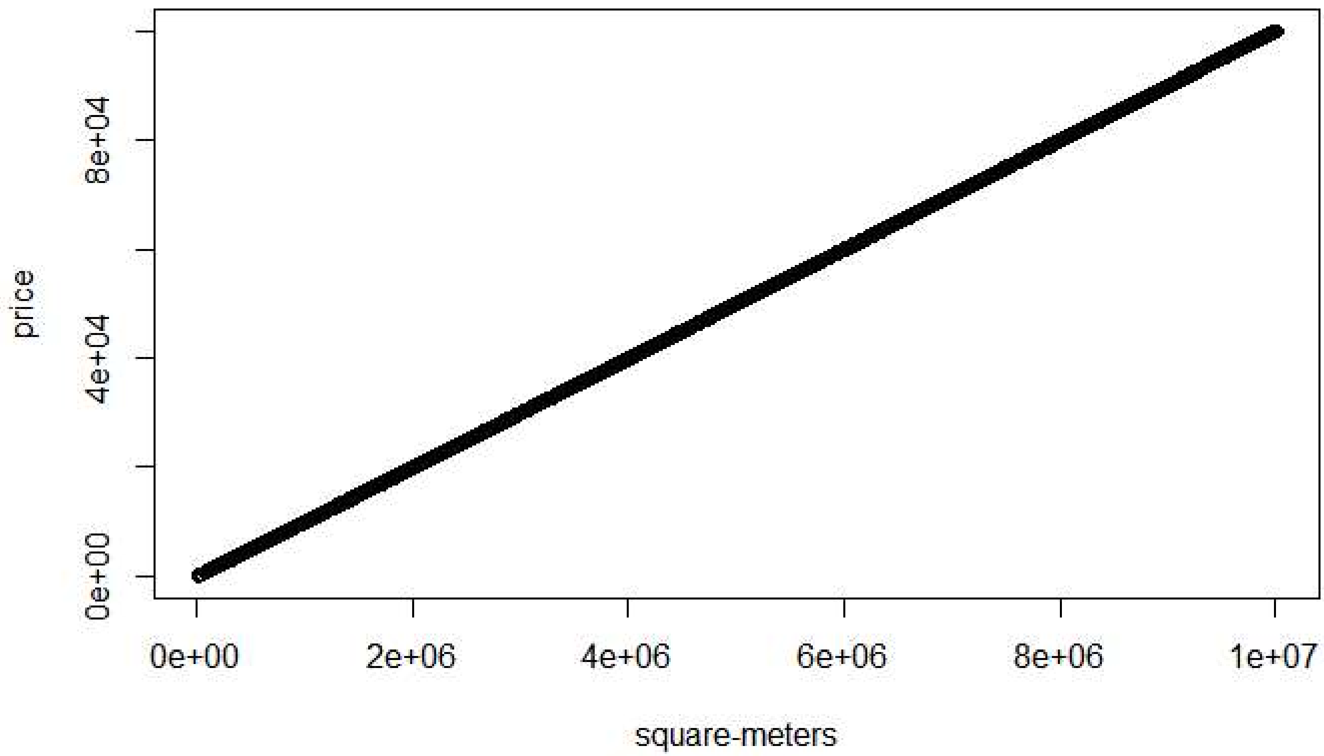
Hide

```


set.seed(1234)
i <- sample(1:nrow(Paris), nrow(Paris)*0.80, replace=FALSE)
train <- Paris[i,]
test <- Paris[-i,]
```

Dividing train and test data; 80% train data, 20% test data

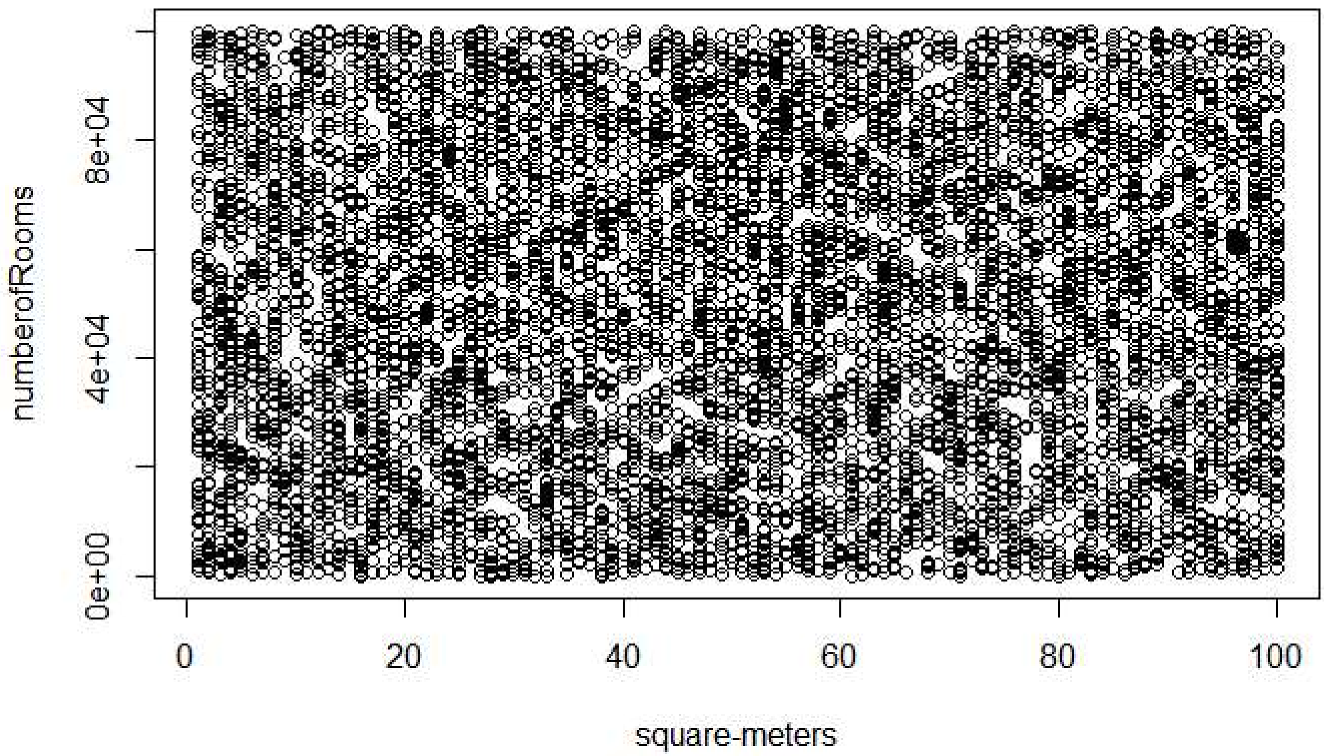
Hide

```
plot(train$squareMeters~train$price, xlab="square-meters", ylab="price")
```



Hide

```
plot(train$squareMeters~train$numberOfRooms, xlab="square-meters", ylab="numberofRooms")
```



Hide

```
lm1 <- lm(price~squareMeters, data=train)
lm1
```

```
Call:
lm(formula = price ~ squareMeters, data = train)

Coefficients:
 (Intercept)  squareMeters
        6420           100
```

Hide

```
pred <- lm1$fitted.values
cov(pred, train$squareMeters) / (sd(pred) * sd(train$price))
```

```
[1] 0.00999997
```

BUilding 1-variable linear model of the data. Data results show high correlation

Hide

```
summary(lm1)
```

```
Call:
lm(formula = squareMeters ~ price, data = train)

Residuals:
     Min       1Q   Median       3Q      Max 
-123.766  -21.970    2.918   24.127   61.270 

Coefficients:
              Estimate Std. Error  t value Pr(>|t|)    
(Intercept) -6.413e+01  7.290e-01   -87.97   <2e-16 ***
price        1.000e-02  1.266e-07 79012.21   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 32.62 on 7998 degrees of freedom
Multiple R-squared:      1, Adjusted R-squared:      1 
F-statistic: 6.243e+09 on 1 and 7998 DF,  p-value: < 2.2e-16
```

The summary first shows the residuals, which are errors that quantify how far off from the regression line the actual values are. Next, there are coefficients that represent the model. The estimate is low for the intercept meaning that there is a very negative correlation. THe Std error for the price is much lower than for the intercept meaning that the coeffient for the slope has less variance. The R squared is 1 meaning that there is a erfect linear coreelatioon between the variables, and the p value is low indicating a high statistical significance.

Hide

```
pred <- predict(lm1, newdata=test)
correlation <- cor(pred, test$price)
print(paste("correlation: ", correlation))
```

```
[1] "correlation:  0.999999347987235"
```

Hide

```
mse <- mean((pred - test$squareMeters)^2)
print(paste("mse: ", mse))
```

```
[1] "mse:  32697904792390.3"
```

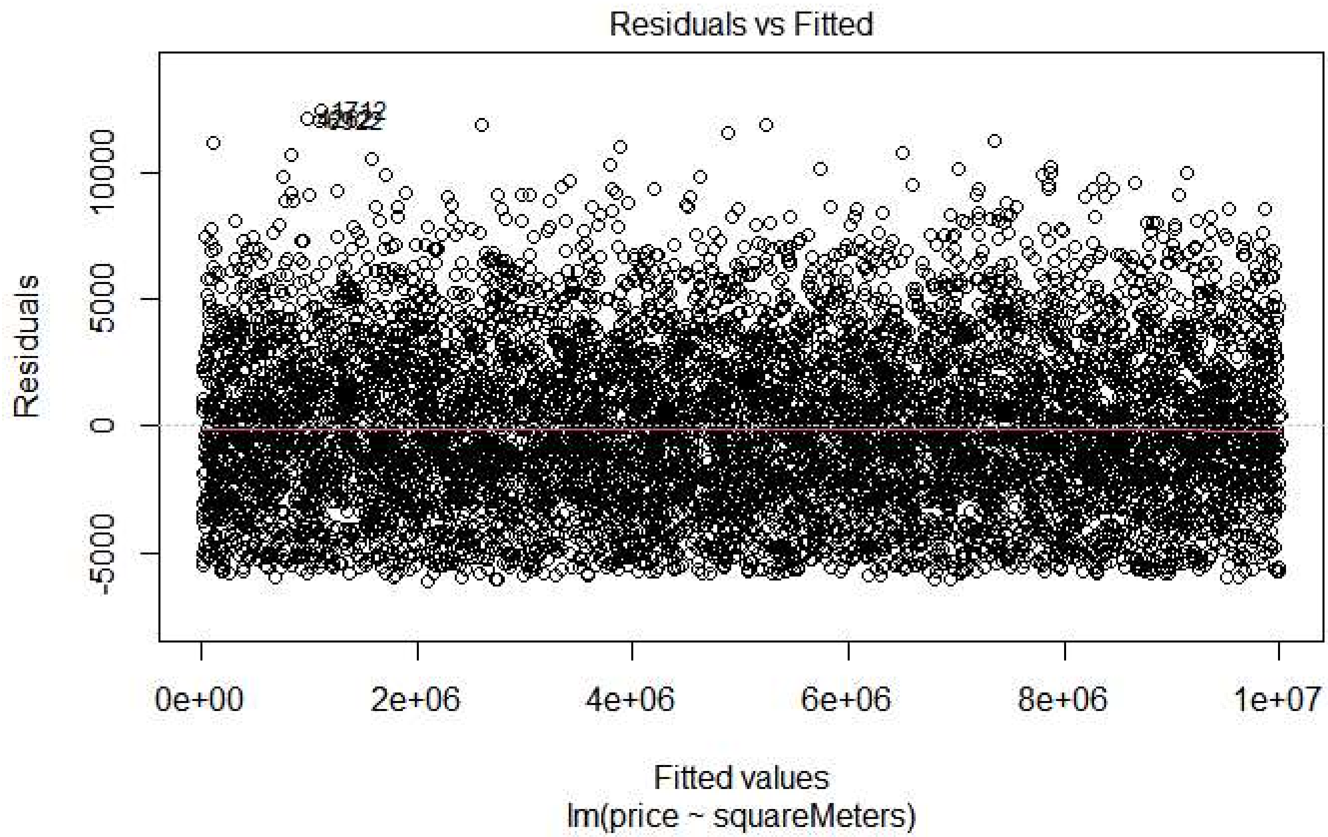
Hide

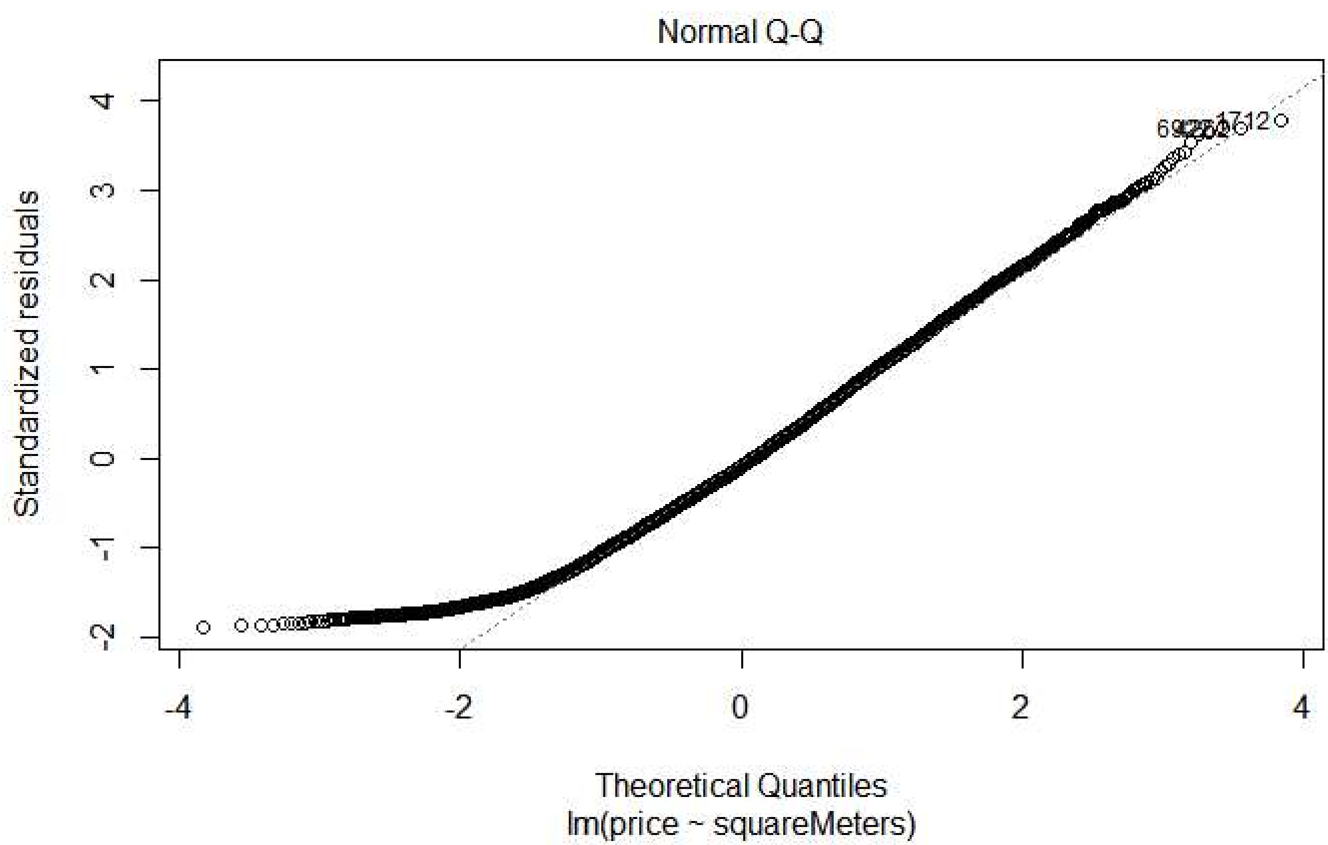
```
rmse <- sqrt(mse)
print(paste("rmse: ", rmse))
```

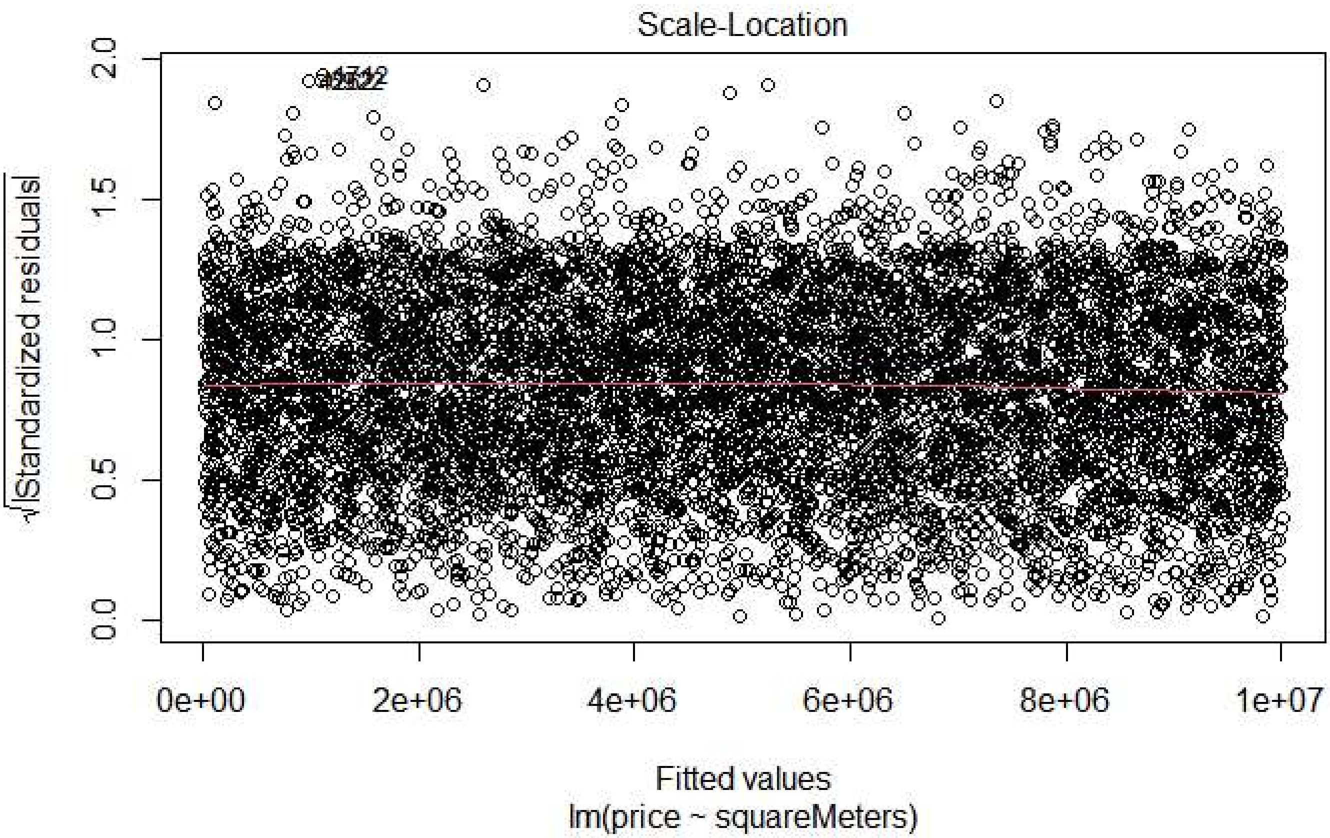
```
[1] "rmse:  5718208.18022485"
```

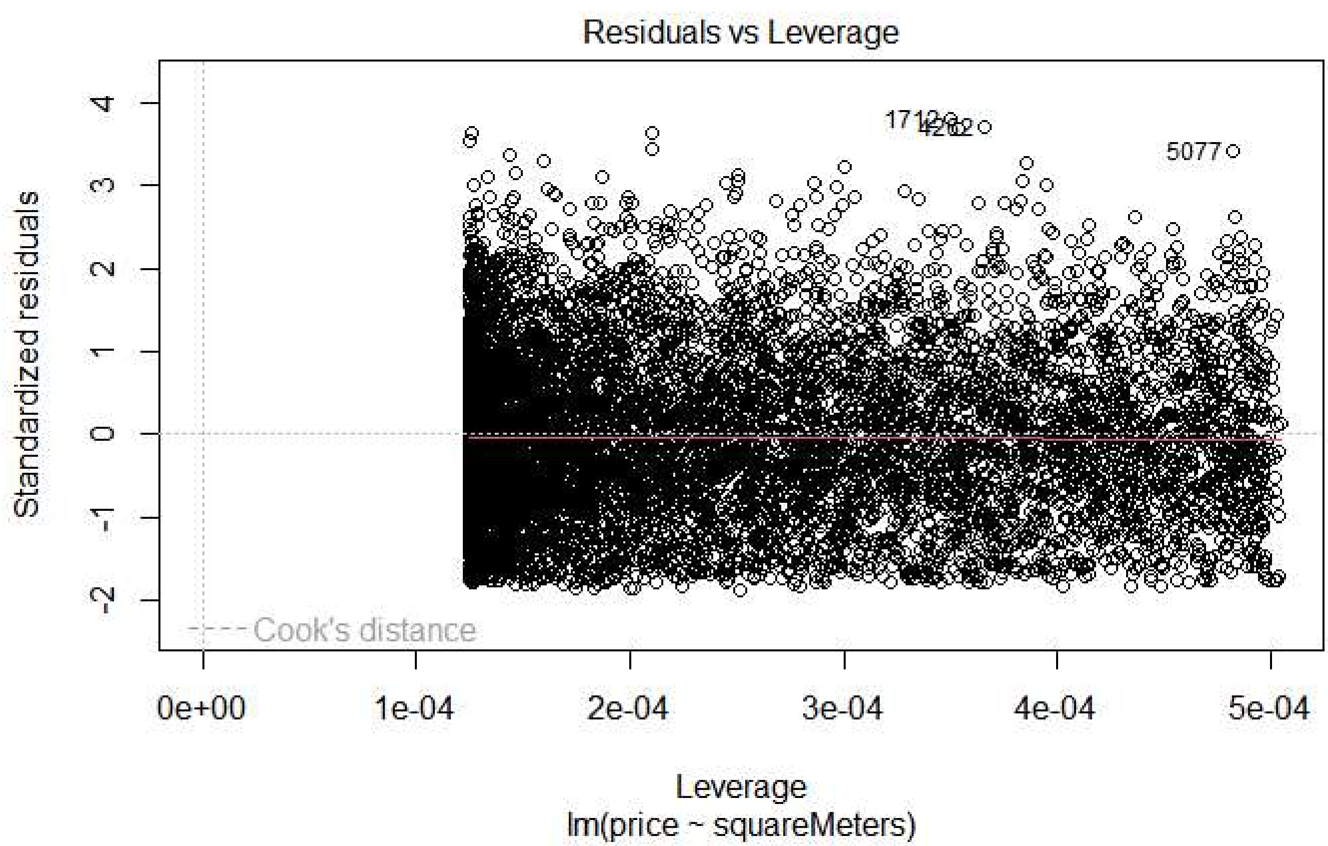
Hide

```
plot(lm1)
```


Residuals vs Fitted
Residuals
-5000
0
5000
10000
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
Fitted values
lm(price ~ squareMeters)
1712


Normal Q-Q
Standardized residuals
-2
-1
0
1
2
3
4
-4
-2
0
2
4
Theoretical Quantiles
lm(price ~ squareMeters)
1712


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
Fitted values
lm(price ~ squareMeters)

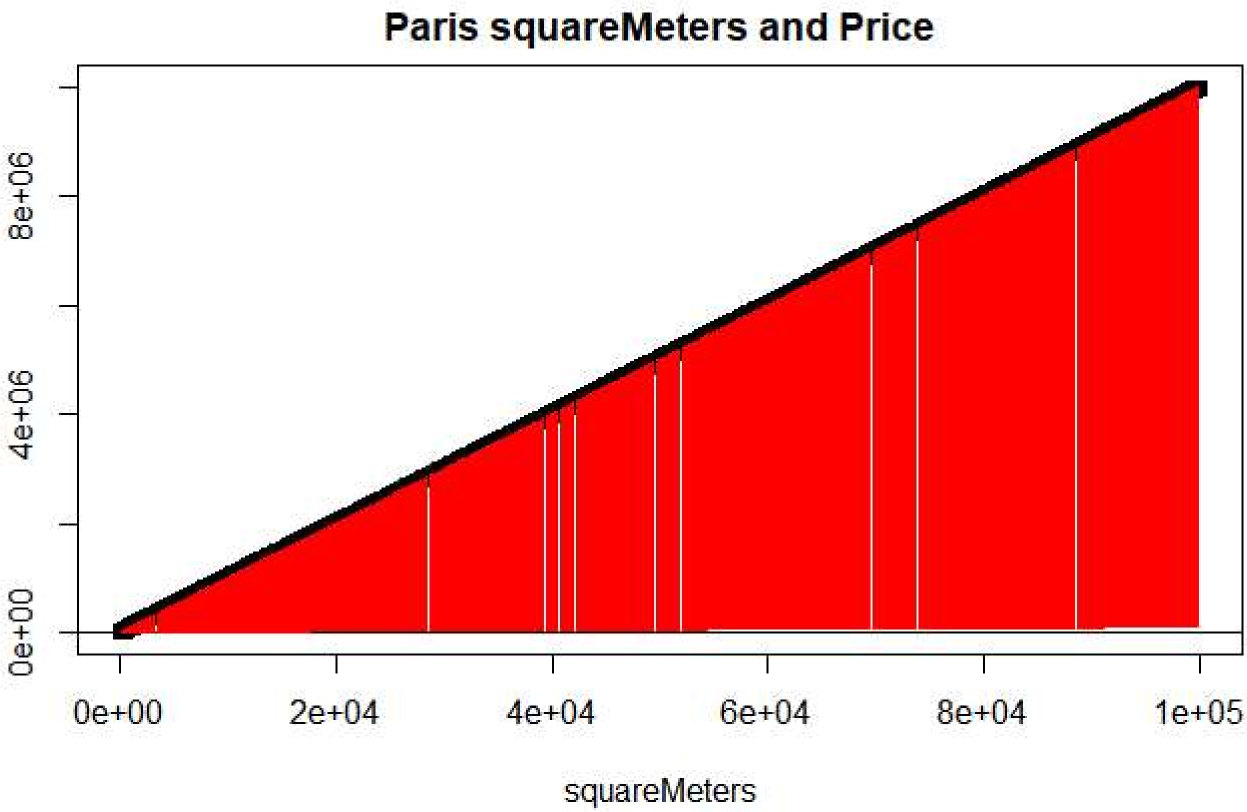

Residuals vs Leverage
Standardized residuals
-2
-1
0
1
2
3
4
1712
4262
5077
Cook's distance
0e+00
1e-04
2e-04
3e-04
4e-04
5e-04
Leverage
lm(price ~ squareMeters)

Hide

```
plot(Paris$squareMeters, Paris$price, main="Paris squareMeters and Price",
     xlab="squareMeters", ylab="")
abline(lm1)
```

Hide

```
points(test$squareMeters, test$price, pch=0)
segments(test$squareMeters, test$price, test$squareMeters, pred, col="red")
```



The residual plot shows a distinct linear relationship between price and squareMeters. The residuals are below the line of best fit, which means that they are negative. They are also very close to the line of best fit, which means that the line is a good approximator for the data.There is a perfect linear coreelation between the variables price and squareMeters which is shown in the graph.

Hide

```
lm2 <- lm(price~squareMeters+numberOfRooms+cityCode+floors, data=train)
summary(lm2)
```

```
Call:
lm(formula = price ~ squareMeters + numberOfRooms + cityCode + 
    floors, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-7186.5 -2179.5  -125.8  1917.2  9752.4 

Coefficients:
               Estimate Std. Error   t value Pr(>|t|)    
(Intercept)   3.628e+03  1.145e+02    31.690   <2e-16 ***
squareMeters  1.000e+02  1.105e-03 90493.465   <2e-16 ***
numberOfRooms 6.966e-01  1.106e+00     0.630    0.529    
cityCode      1.330e-05  1.093e-03     0.012    0.990    
floors        5.513e+01  1.103e+00    49.969   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2848 on 7995 degrees of freedom
Multiple R-squared:      1,	Adjusted R-squared:      1 
F-statistic: 2.048e+09 on 4 and 7995 DF,  p-value: < 2.2e-16
```

Adjusted R squared is the same as model 1.

Hide

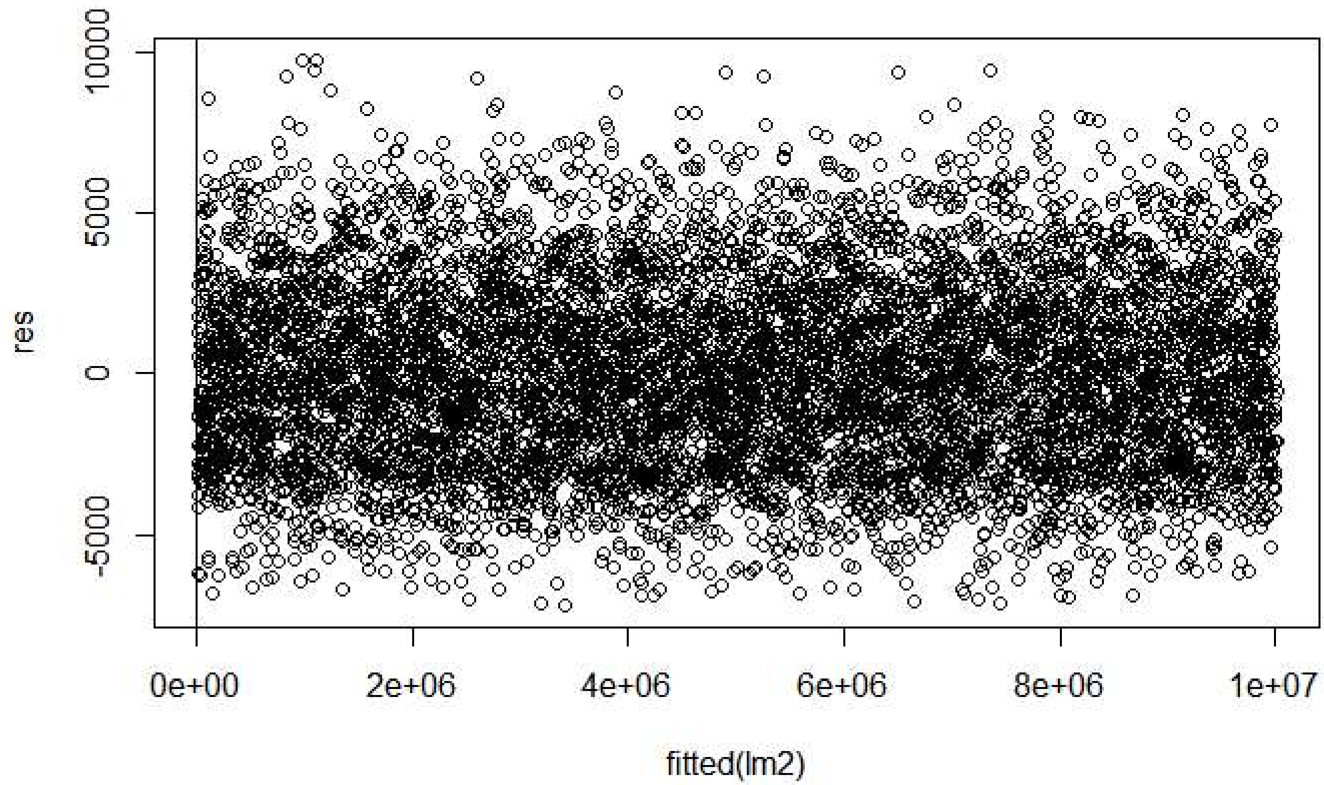
```
anova(lm1, lm2)
```

```
Analysis of Variance Table

Model 1: price ~ squareMeters
Model 2: price ~ squareMeters + numberOfRooms + cityCode + floors
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)    
1   7998 8.5106e+10                                   
2   7995 6.4833e+10  3 2.0273e+10 833.33 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hide

```
res<- resid(lm2)
plot(fitted(lm2), res)
abline(lm2)
```

```
Warning: only using the first two of 5 regression coefficients
```



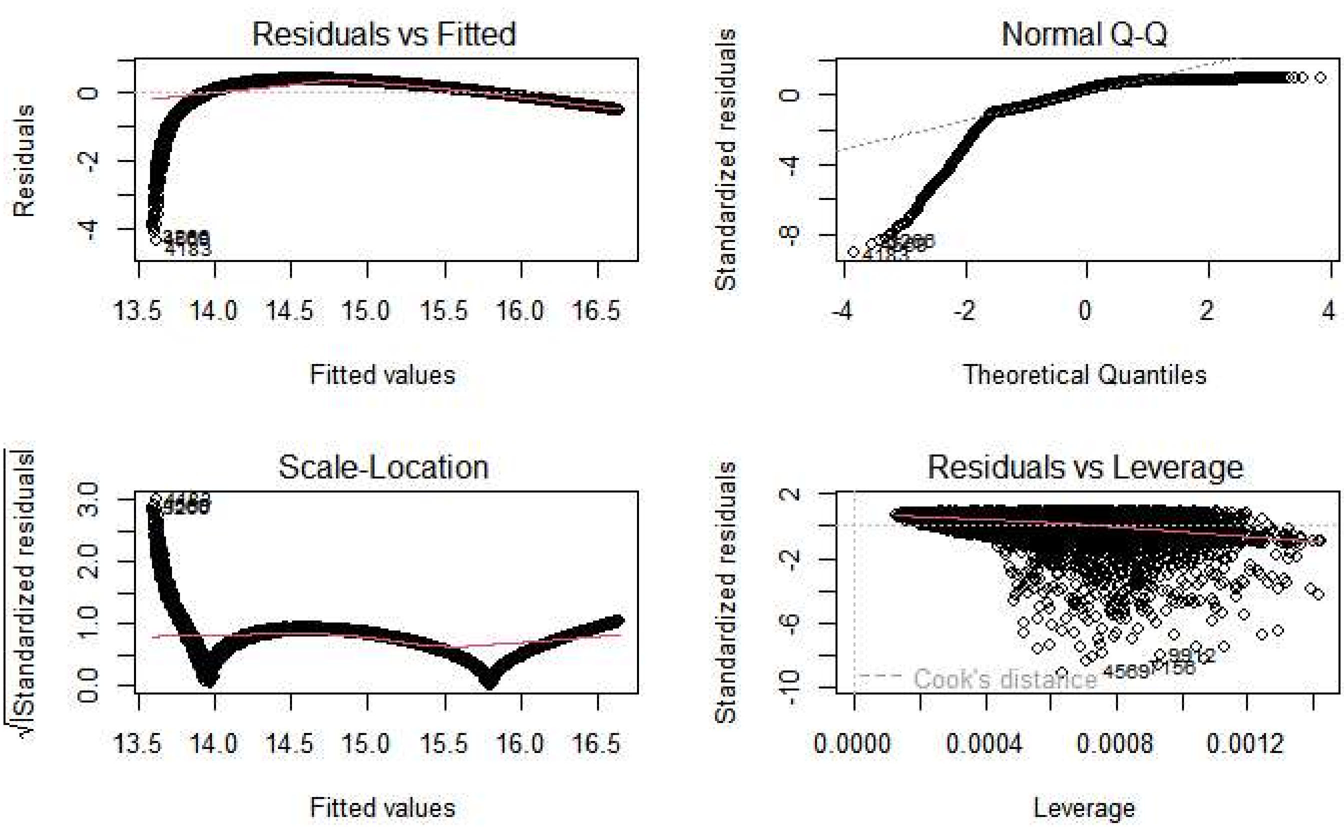
Hide

```
lm3 <- lm(log(price)~squareMeters+numberOfRooms+cityCode+floors, data=train)
summary(lm3)
```

```
Call:
lm(formula = log(price) ~ squareMeters + numberOfRooms + cityCode +
    floors, data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-4.3807 -0.1931  0.1324  0.3310  0.4291

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.358e+01  1.937e-02 701.405   <2e-16 ***
squareMeters  3.011e-05  1.869e-07 161.079   <2e-16 ***
numberOfRooms 2.051e-04  1.871e-04   1.096    0.273
cityCode      1.332e-07  1.849e-07   0.721    0.471
floors        2.923e-04  1.866e-04   1.566    0.117
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4818 on 7995 degrees of freedom
Multiple R-squared:  0.7646,    Adjusted R-squared:  0.7644
F-statistic:  6491 on 4 and 7995 DF,  p-value: < 2.2e-16
```

Hide

```
par(mfrow=c(2,2))
plot(lm3)
```



Model 2 was the best, then Model 1, and then Model 3. Model 1 and 2 both had an adjusted R squared of 1 which means that they were both very accurate models. Also the plot showed a clear linear relationship in the data so the log model in model 3 is not as accurate as the two other models. Model 2 had a slightly lower degrees of freedom which would make it very slightly more accurate. However, the models are neck and neck since squarefeet and price have a clear and distinct linear relationship.

Hide

```
pred <- predict(lm1, newdata=test)
correlation <- cor(pred, test$price)
print(paste("correlation: ", correlation))
```

```
[1] "correlation:  0.999999347987235"
```

Hide

```
mse <- mean((pred - test$price)^2)
print(paste("mse: ", mse))
```

```
[1] "mse:  10676692.7171376"
```

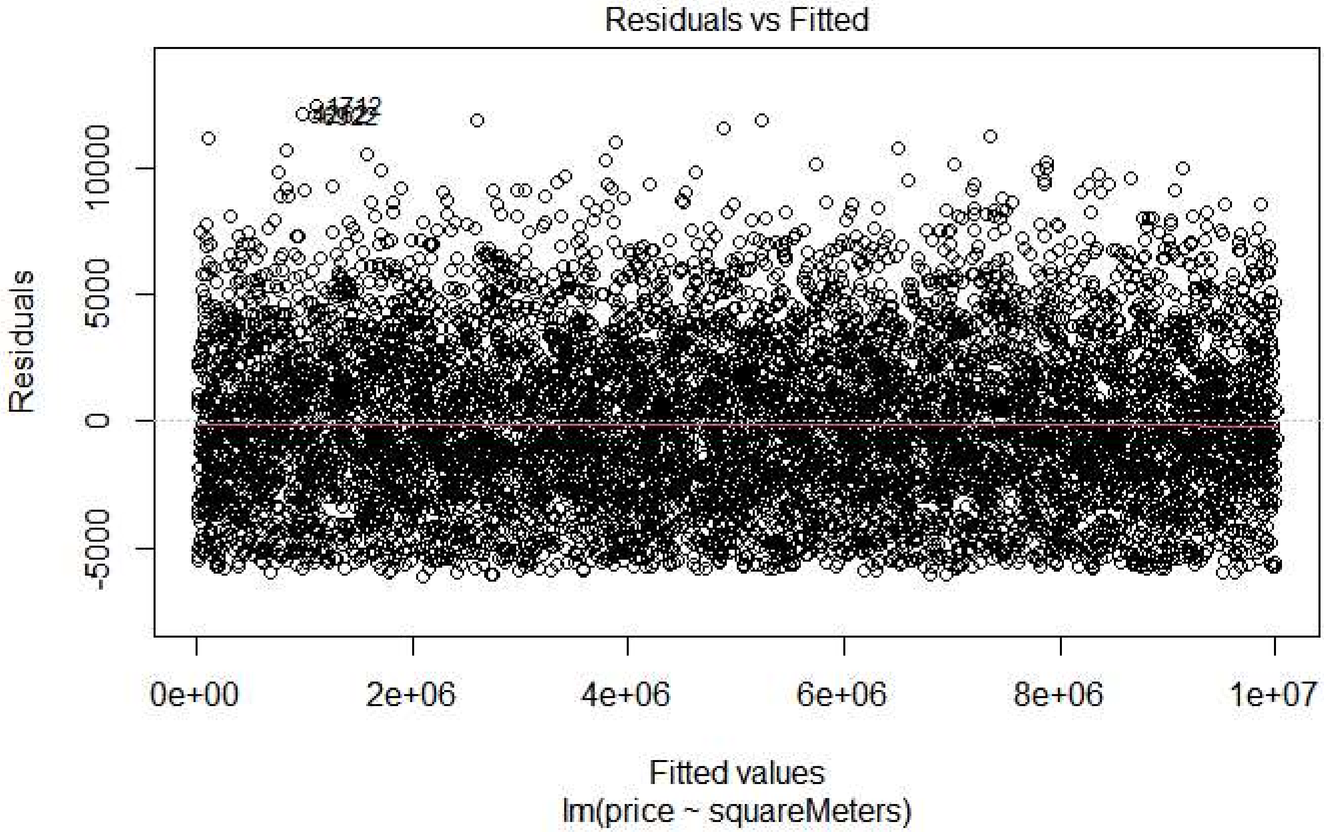
Hide

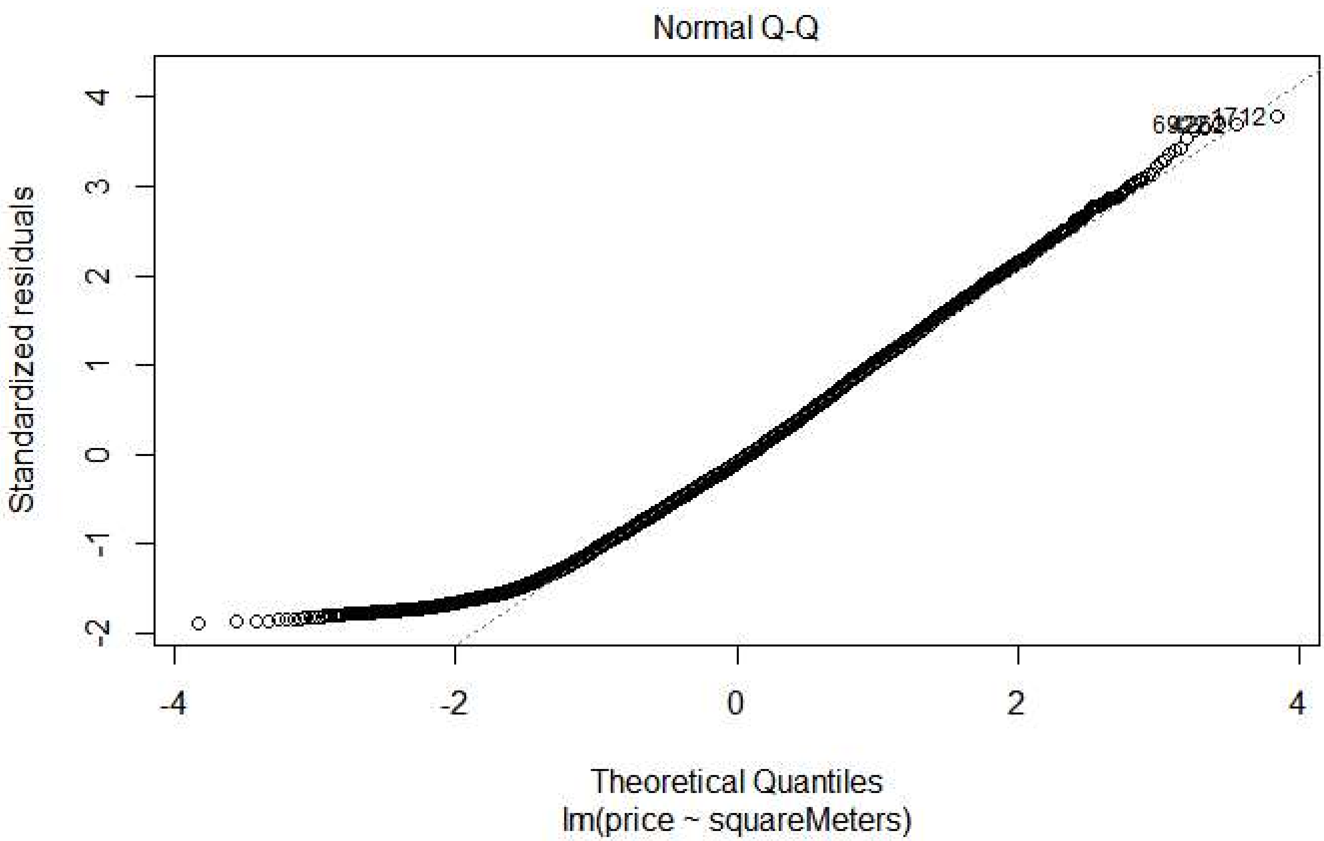
```
rmse <- sqrt(mse)
print(paste("rmse: ", rmse))
```

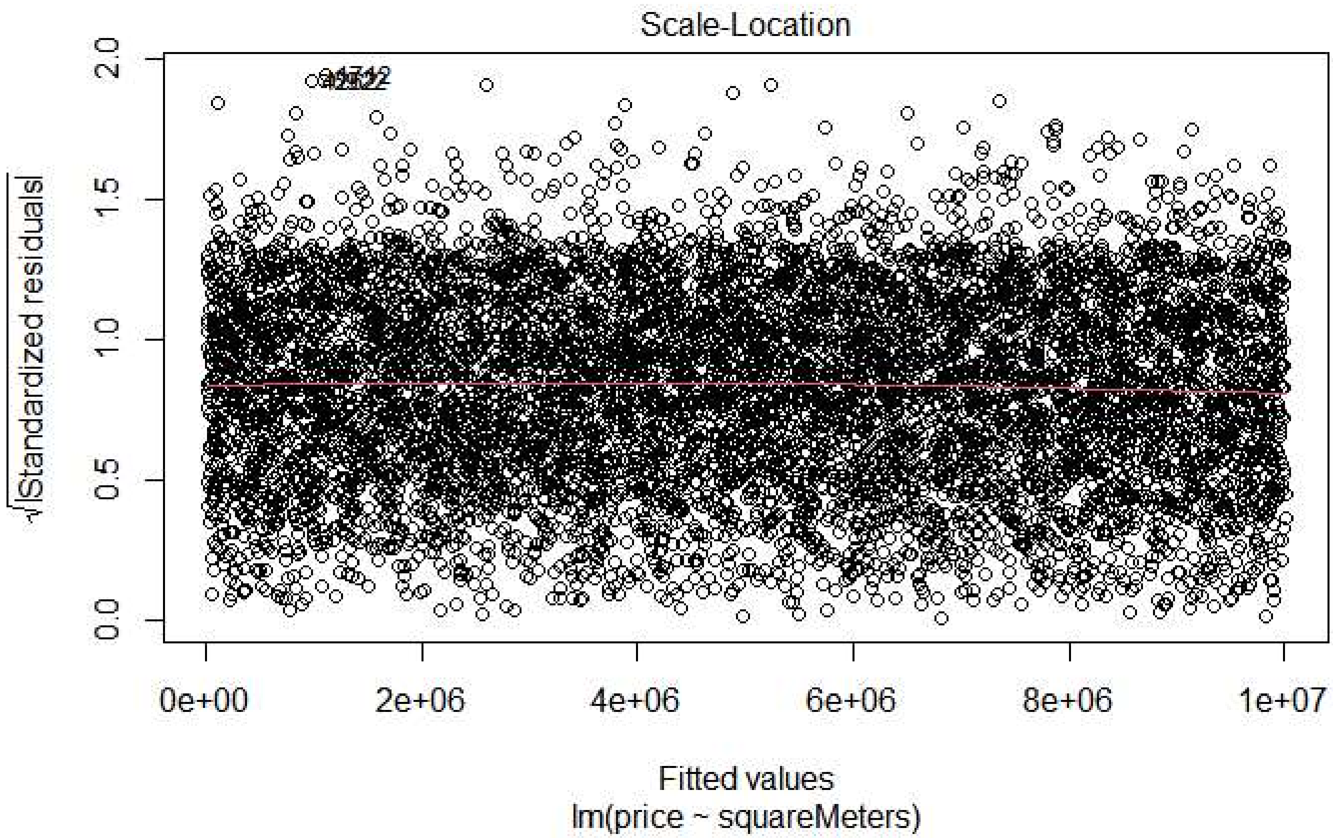
```
[1] "rmse:  3267.52088243329"
```

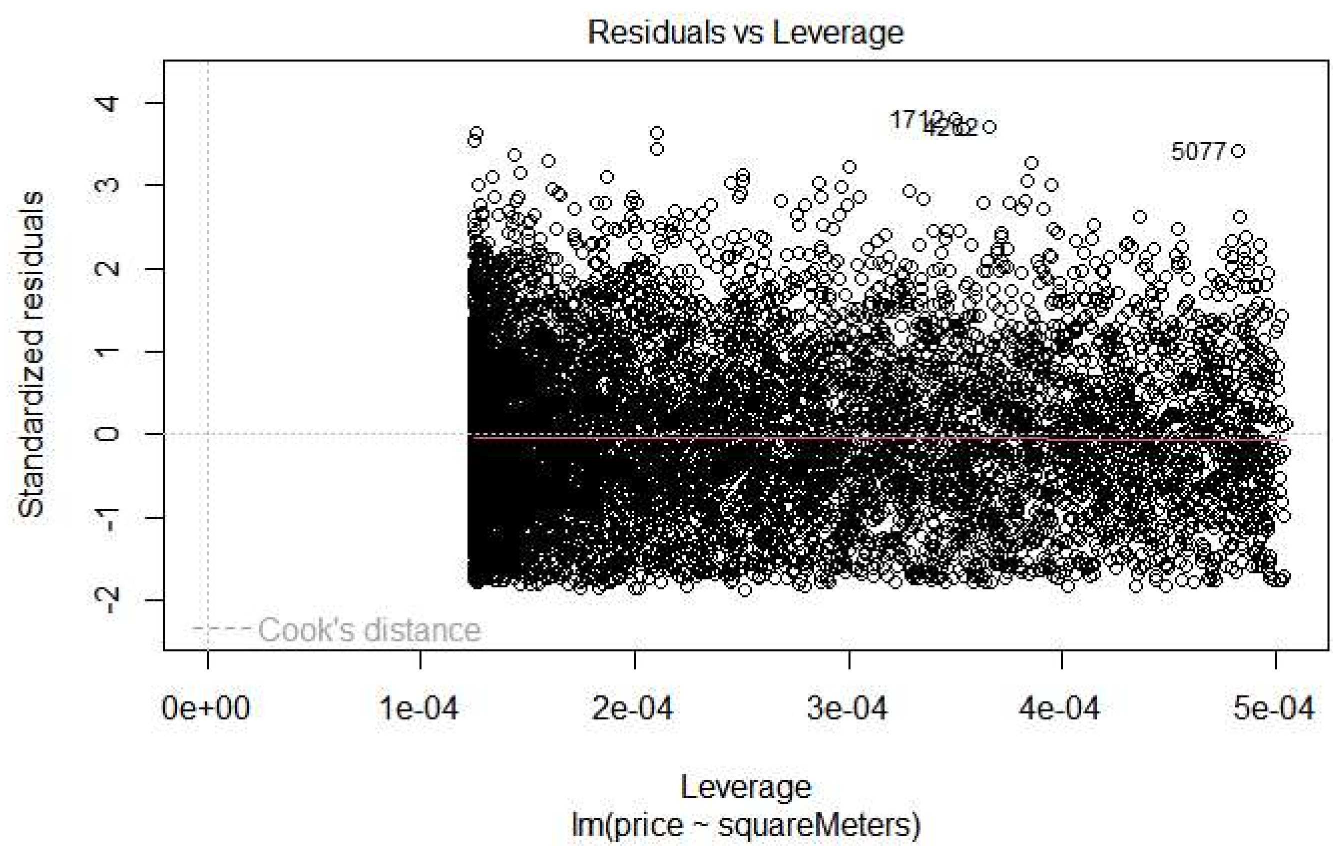
Hide

```
plot(lm1)
```


Residuals vs Fitted
Residuals
10000
5000
0
-5000
1712
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
Fitted values
lm(price ~ squareMeters)


Normal Q-Q
Standardized residuals
4
3
2
1
0
-1
-2
1712
-4
-2
0
2
4
Theoretical Quantiles
lm(price ~ squareMeters)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
Fitted values
lm(price ~ squareMeters)


Residuals vs Leverage
Standardized residuals
-2
-1
0
1
2
3
4
1712
5077
Cook's distance
0e+00
1e-04
2e-04
3e-04
4e-04
5e-04
Leverage
lm(price ~ squareMeters)

Hide

```
pred <- predict(lm2, newdata=test)
correlation <- cor(pred, test$price)
print(paste("correlation: ", correlation))
```

```
[1] "correlation:  0.999999479365376"
```

Hide

```
mse <- mean((pred - test$price)^2)
print(paste("mse: ", mse))
```

```
[1] "mse:  8520992.94560919"
```

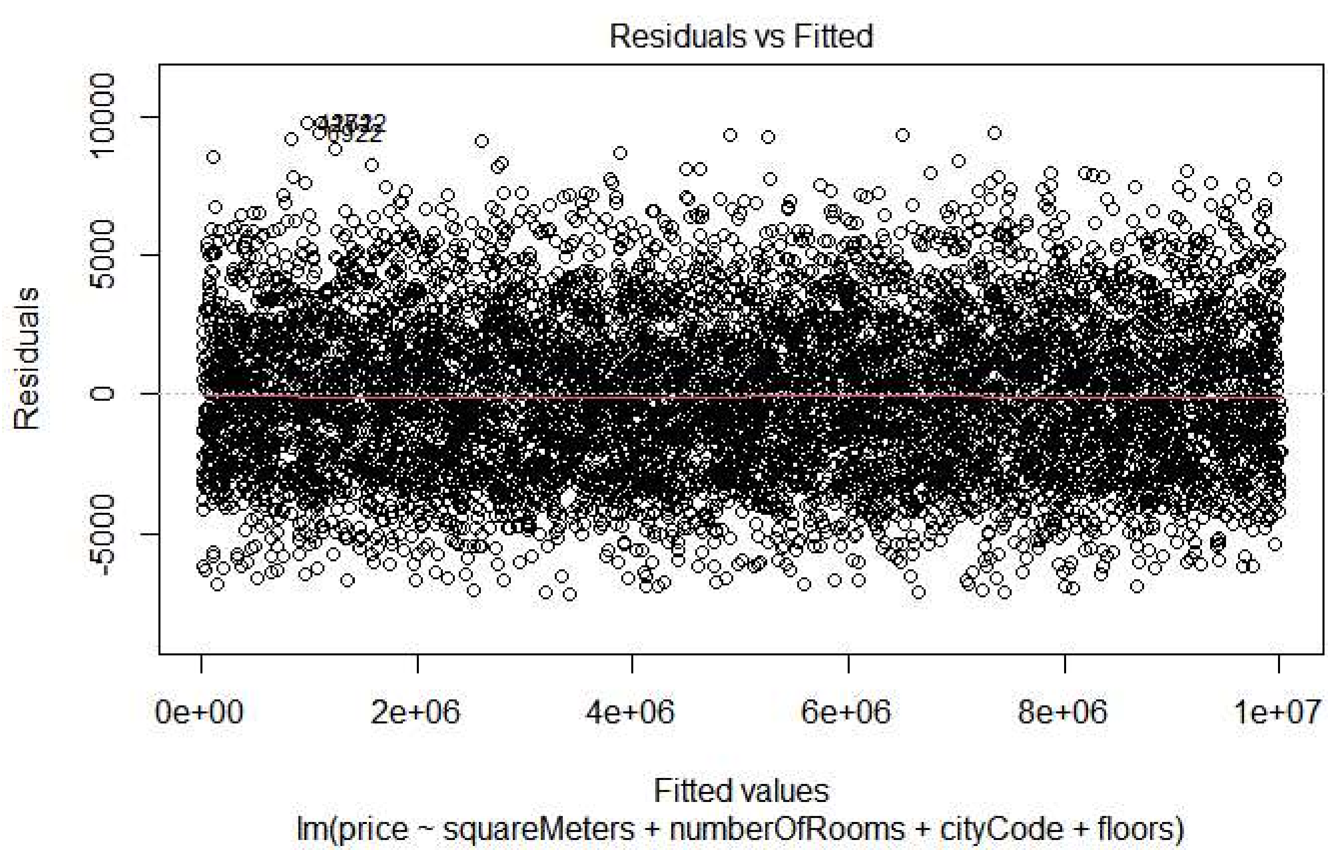
Hide

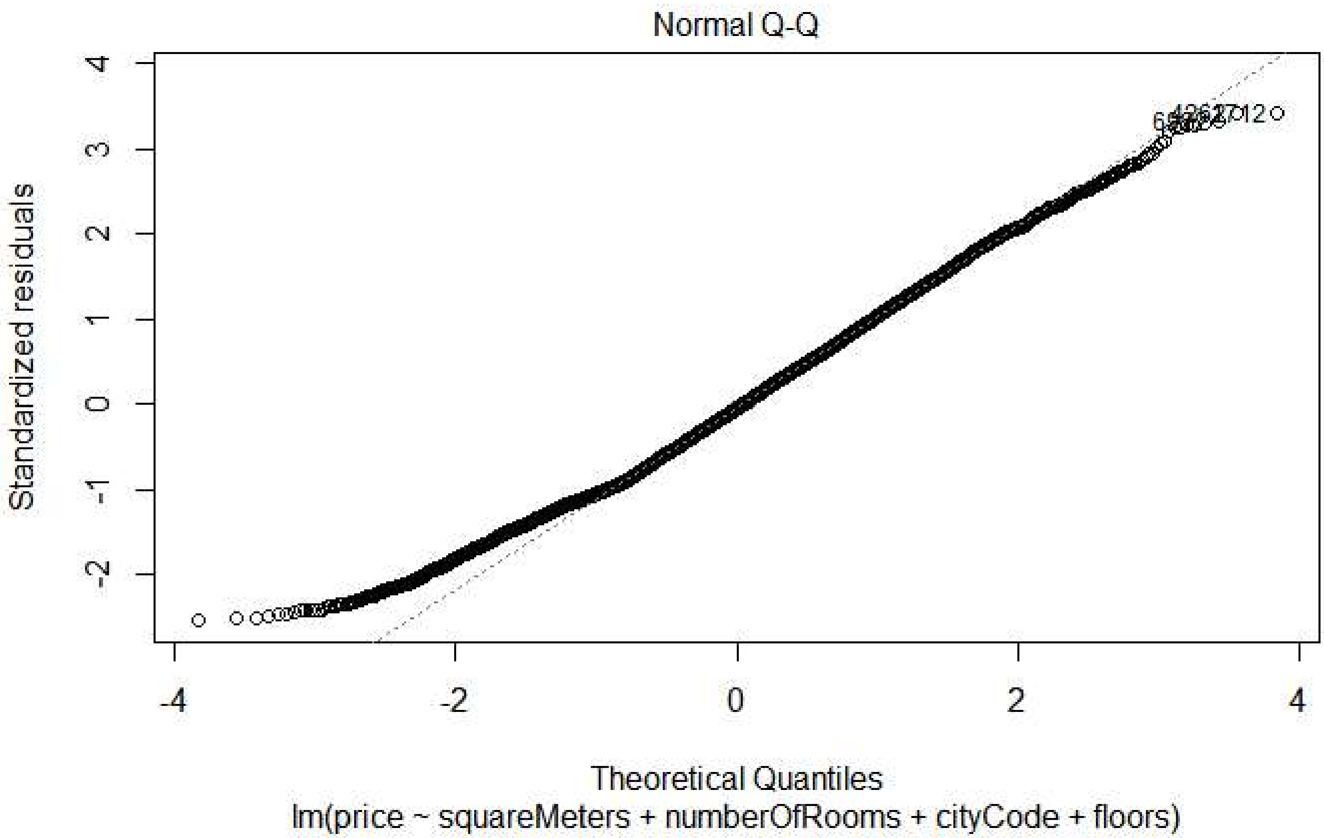
```
rmse <- sqrt(mse)
print(paste("rmse: ", rmse))
```

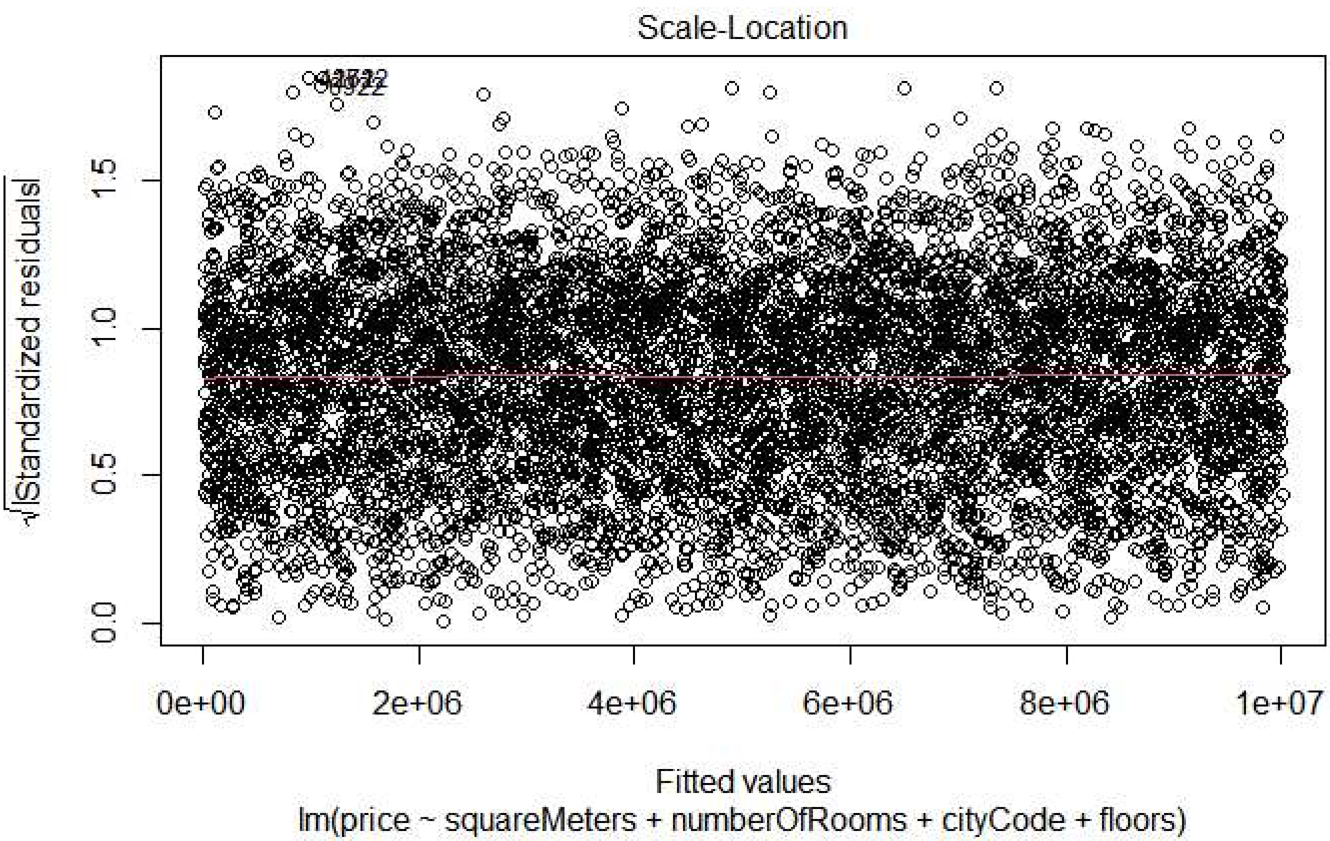
```
[1] "rmse:  2919.07398769014"
```

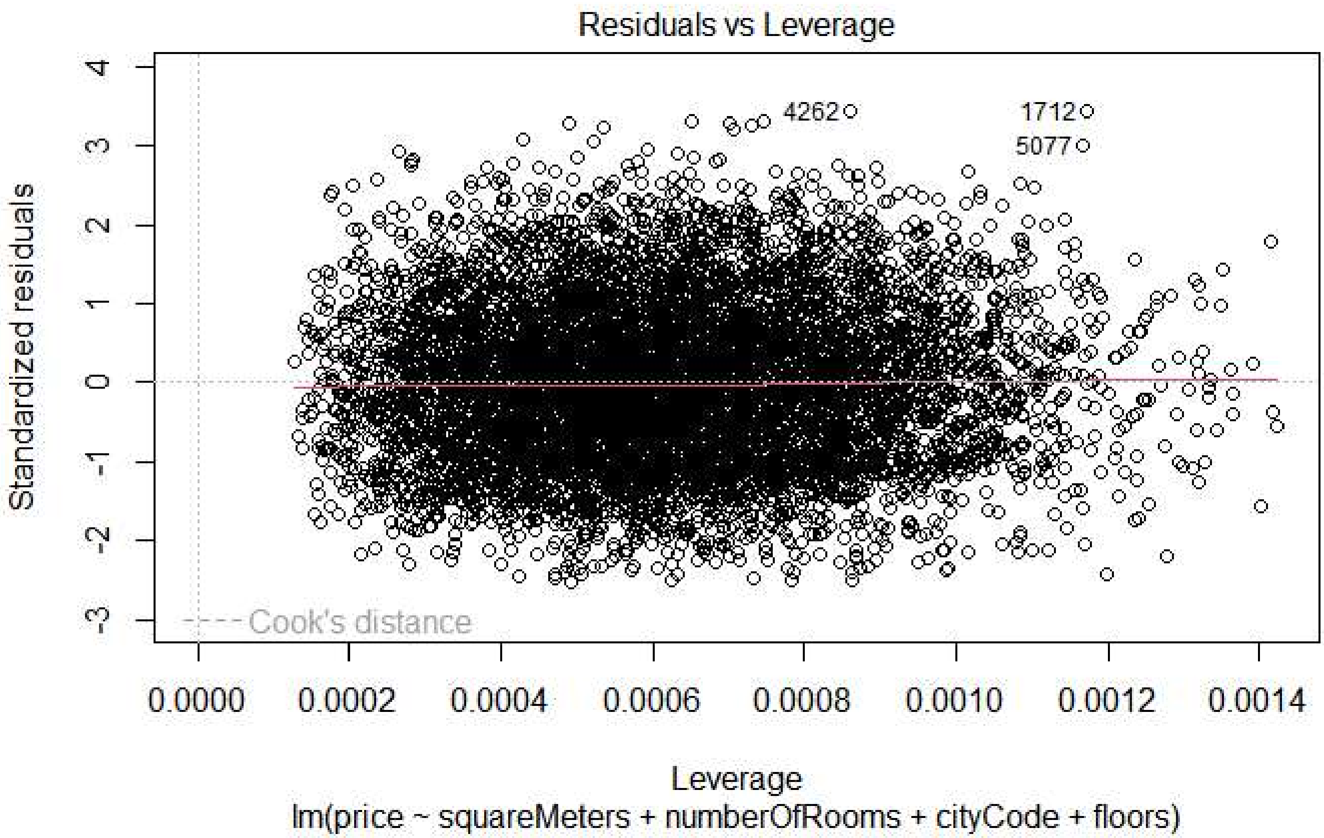
Hide

```
plot(lm2)
```




Normal Q-Q
Standardized residuals
-2
-1
0
1
2
3
4
-4
-2
0
2
4
Theoretical Quantiles
lm(price ~ squareMeters + numberOfRooms + cityCode + floors)


Scale-Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
Fitted values
lm(price ~ squareMeters + numberOfRooms + cityCode + floors)



Hide

```
pred <- predict(lm3, newdata=test)
correlation <- cor(pred, test$price)
print(paste("correlation: ", correlation))
```

```
[1] "correlation:  0.999919274893544"
```

Hide

```
mse <- mean((pred - test$price)^2)
print(paste("mse: ", mse))
```

```
[1] "mse:  33360519544773.2"
```

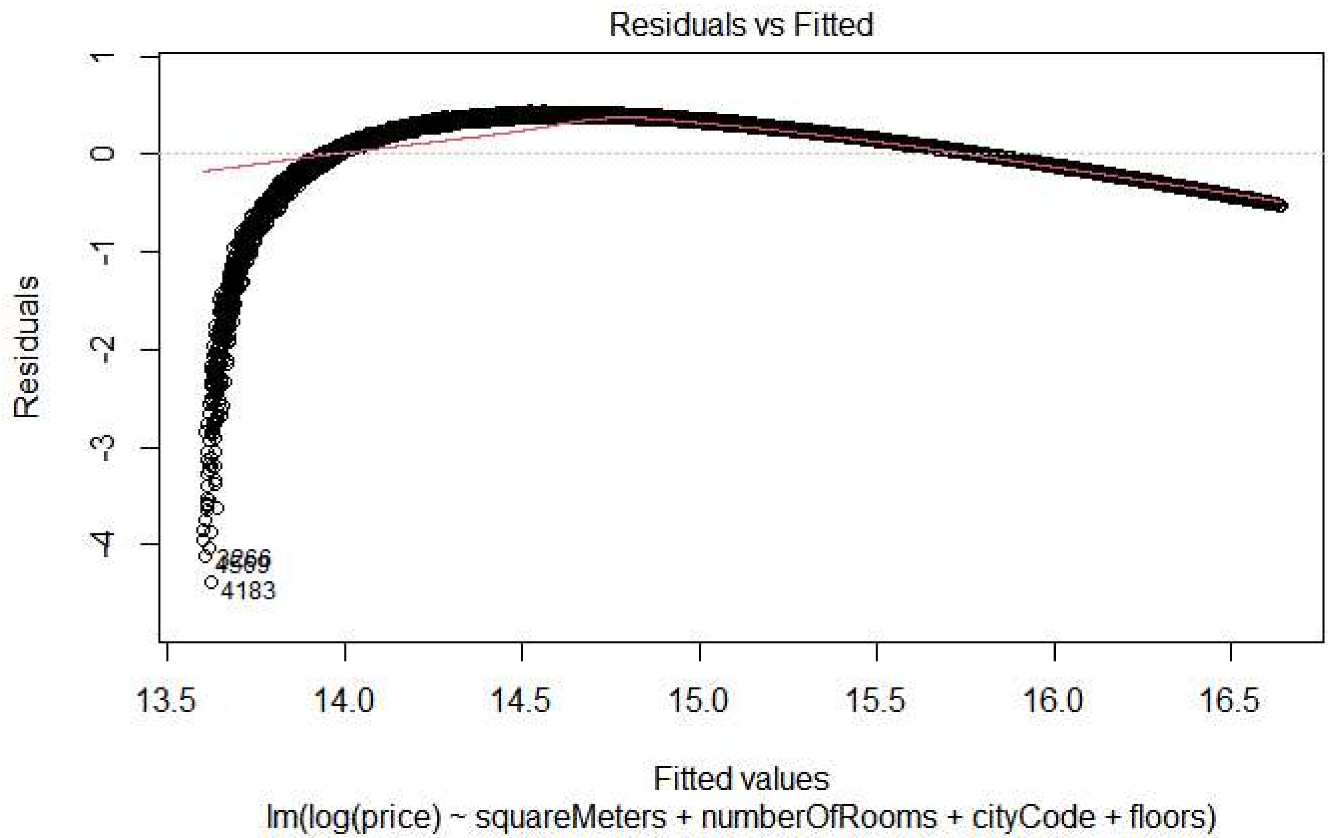
Hide

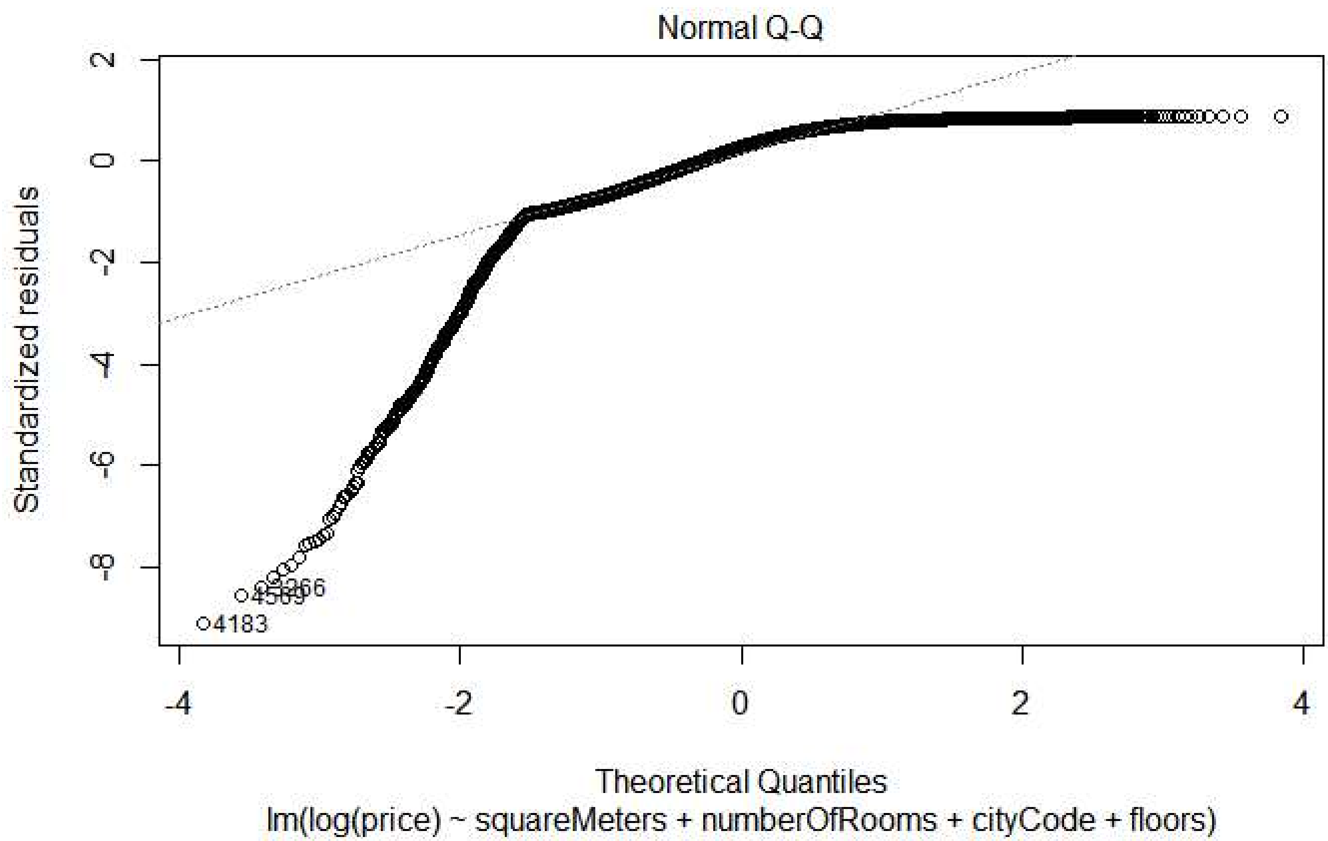
```
rmse <- sqrt(mse)
print(paste("rmse: ", rmse))
```

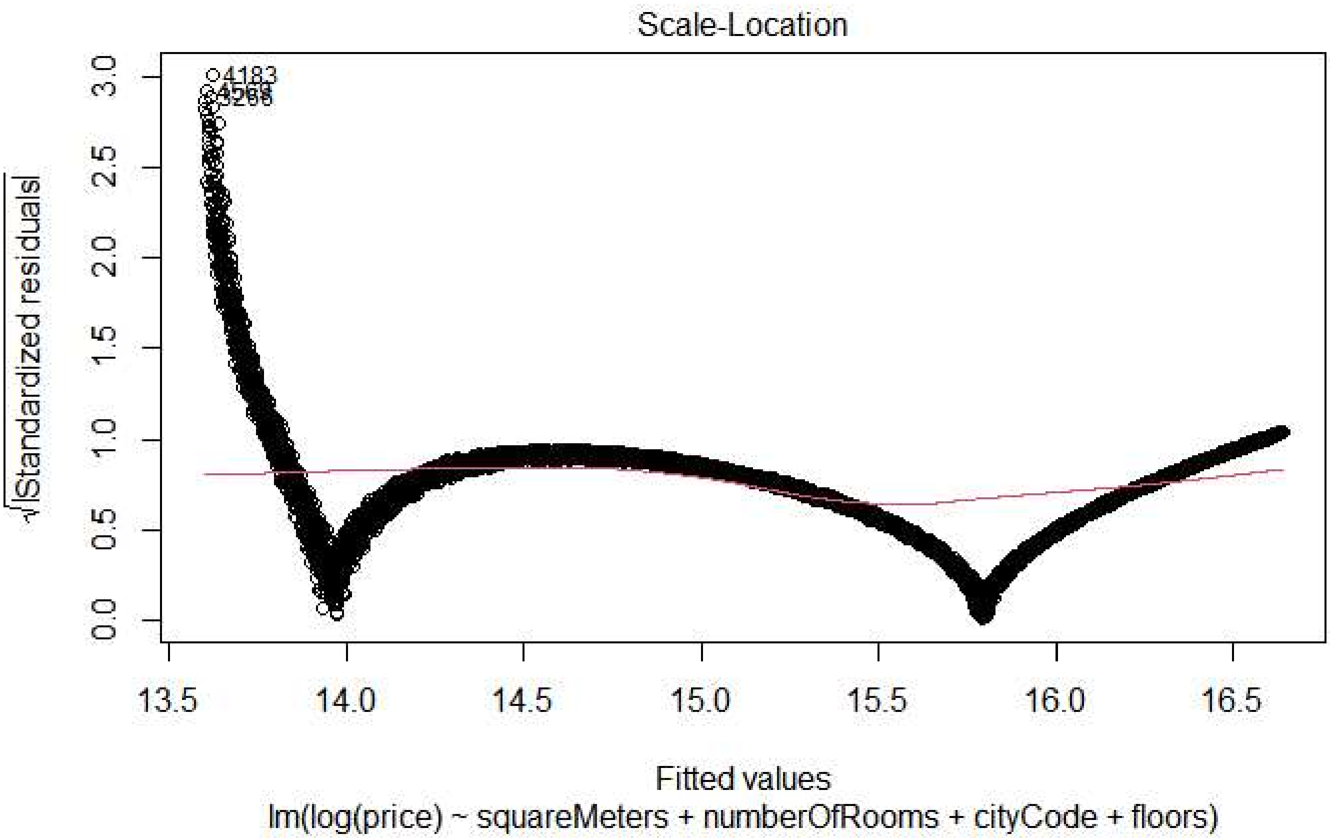
```
[1] "rmse:  5775856.60701278"
```

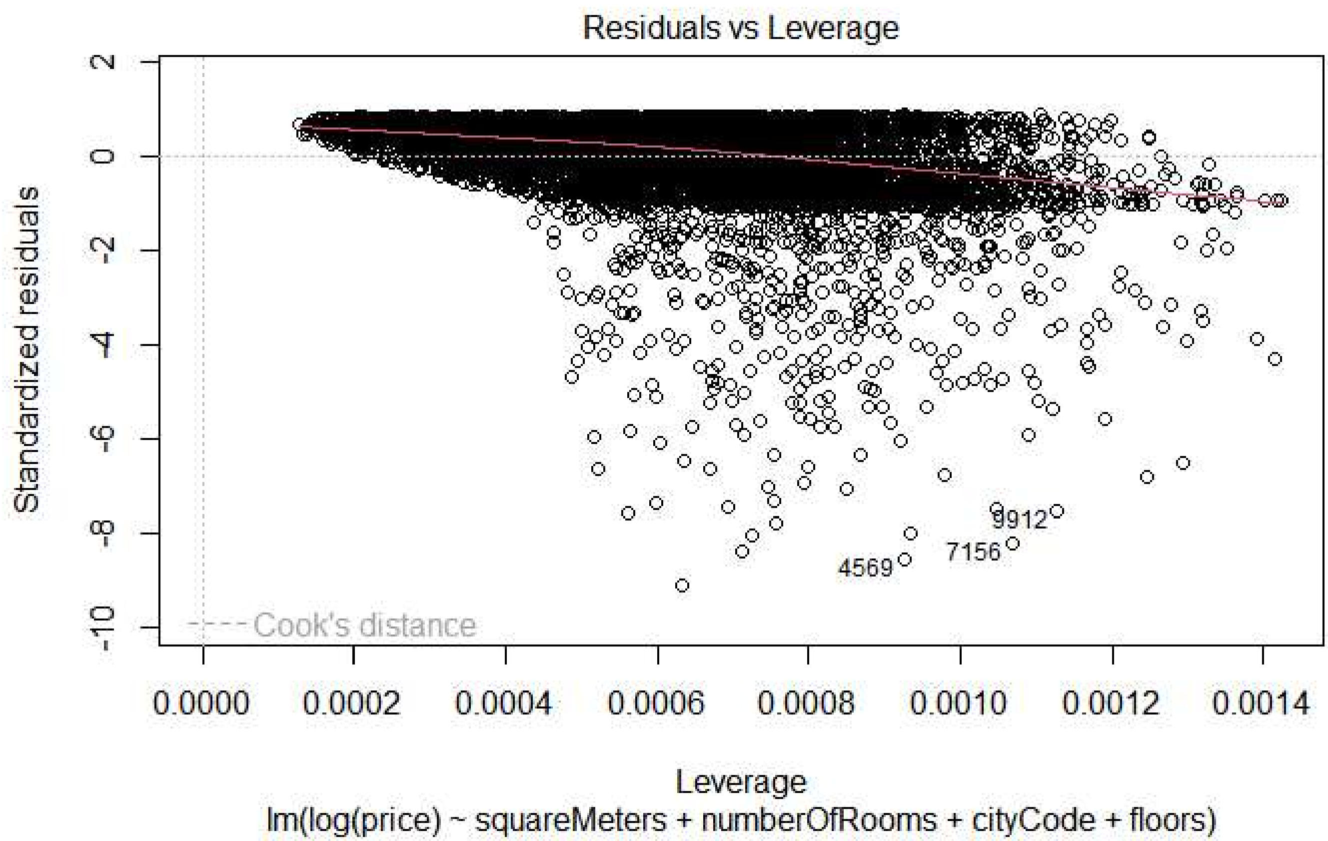
Hide

```
plot(lm3)
```


Residuals vs Fitted
Residuals
1
0
-1
-2
-3
-4
3266
4569
4183
13.5
14.0
14.5
15.0
15.5
16.0
16.5
Fitted values
lm(log(price) ~ squareMeters + numberOfRooms + cityCode + floors)


Normal Q-Q
Standardized residuals
2
0
-2
-4
-6
-8
3266
4569
4183
-4
-2
0
2
4
Theoretical Quantiles
lm(log(price) ~ squareMeters + numberOfRooms + cityCode + floors)





As expected, the model 2 shows the highest correlation, with Model 1 coming next up and then Model 3. The rmse is greatest for model 3 since it is the worst fir model, and Model 2 has the lowest rsme. This happenend because a log model is not as accurate because the variables price and square meters have a perfect linear correlation.

Model 2 is slightly more accurate since it takes into account other features that have an effect on the price so the accuracy in predicitng the price goes up in Model 2.