COSC 515 Assignment #2

T.J. Liggett

2/13/2021

First, we must import relevant datasets

```
realestate <- read.csv("realestate.csv")
re_description <- read.csv("realestate description.csv")
```

1. Go through each variable predict the effect it may have on the price per square foot. Would it be associated with an increase, decrease, or should we not use it in the model?

```
re_description
```

```
##      Variable                                                  Description
## 1          No                        Identifier for the transaction number
## 2 transaction                             Date the transaction took place
## 3       price                The dollar price per square foot of the house
## 4         age                                  How old the house is in years
## 5         mrt Distance in yards from the nearest Mass Rapid Transit station
## 6         con           The number of convenience stores in close proximity
## 7         lat                          The latitude coordinates of the house
## 8        long                        The longitute coordinates of the house
```

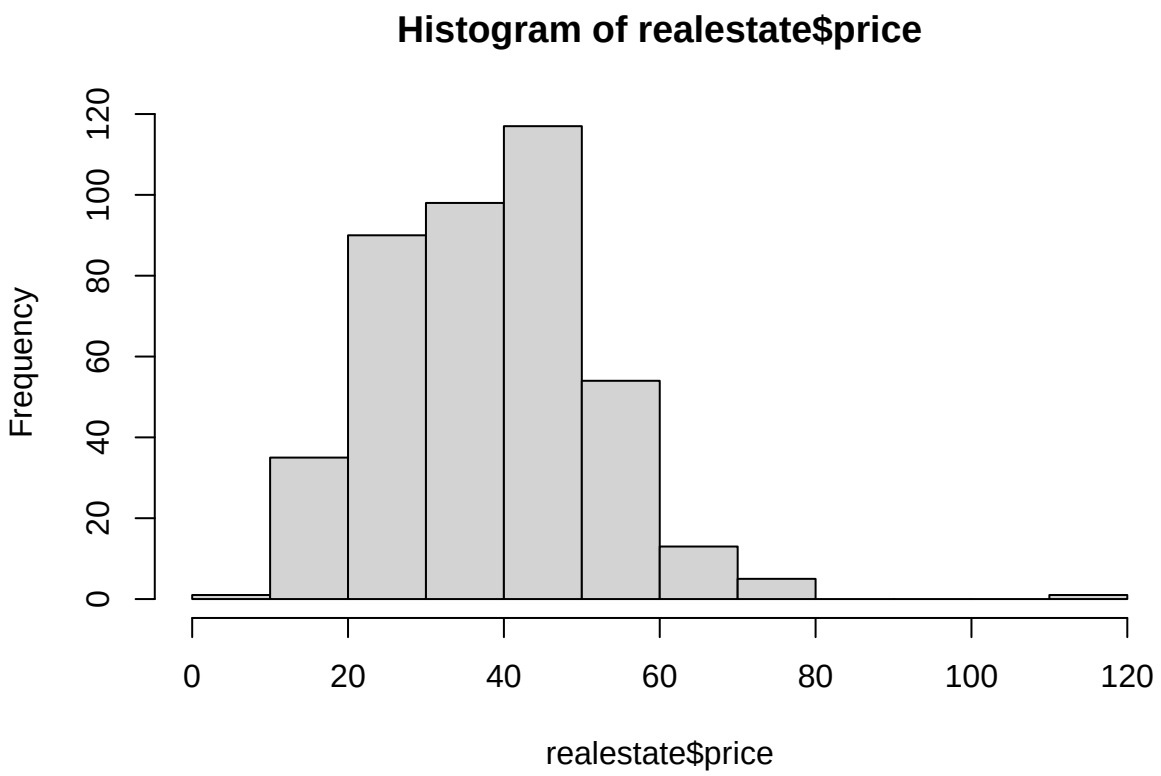
- **No:** this is the identifier for the transaction number. It is highly unlikely this number would impact the price of the house, could be correlated with the date of the transaction [transaction]
- **transaction:** The date the transaction took place. Further research shows this value to be between late 2012 and 2014. I am unfamiliar with the housing market during this time, but I would predict no correlation with price per square foot [price], assuming no market crash.
- **price:** this is the variable being modeled, the price per square foot
- **age:** this is the age of the house in years. I would predict that the age of the house would have a negative correlation (decrease) with the price per square foot [price] of the house.
- **mrt:** the distance in yards from the nearest Mass Rapid Transit station. This would likely add value to the property and I would predict a positive correlation
- **con:** the number of convenience stores in close proximity. This might has a negative correlation, as more convenience stores might mean a higher population and higher crime in the area.
- **lat:** the latitude coordinates of the house. I would say this likely does not correlate, unless this is coastal real estate where one direction is closer to the water.
- **long:** the longitude coordinates of the house. I would say this likely does not correlate, unless this is coastal real estate where one direction is closer to the water.

Prediction Summary

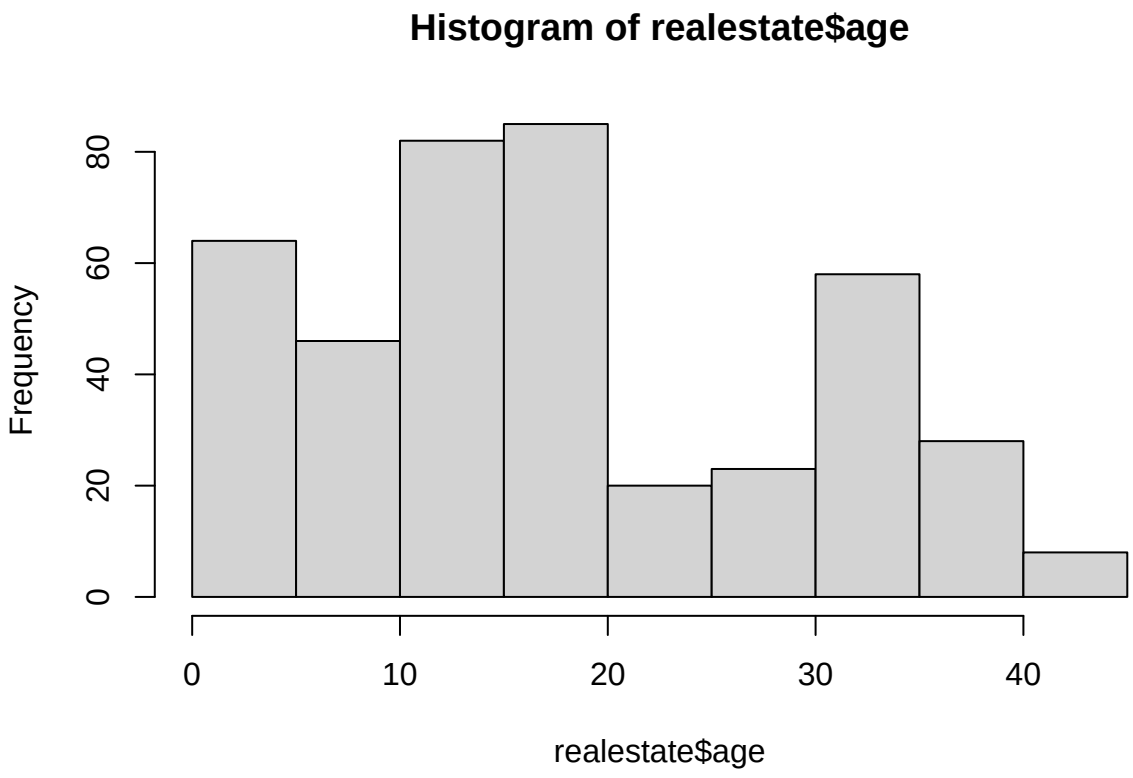
postive correlation: mrt **negative correlation:** age, con **no correlation:** no, lat, long, transaction

2. For all of the variables you are using in your model (both dependent and independent), create histograms to look at each separately and scatter plots to see how they interact with one another. (Hint: Try the pairs() function from the class example)

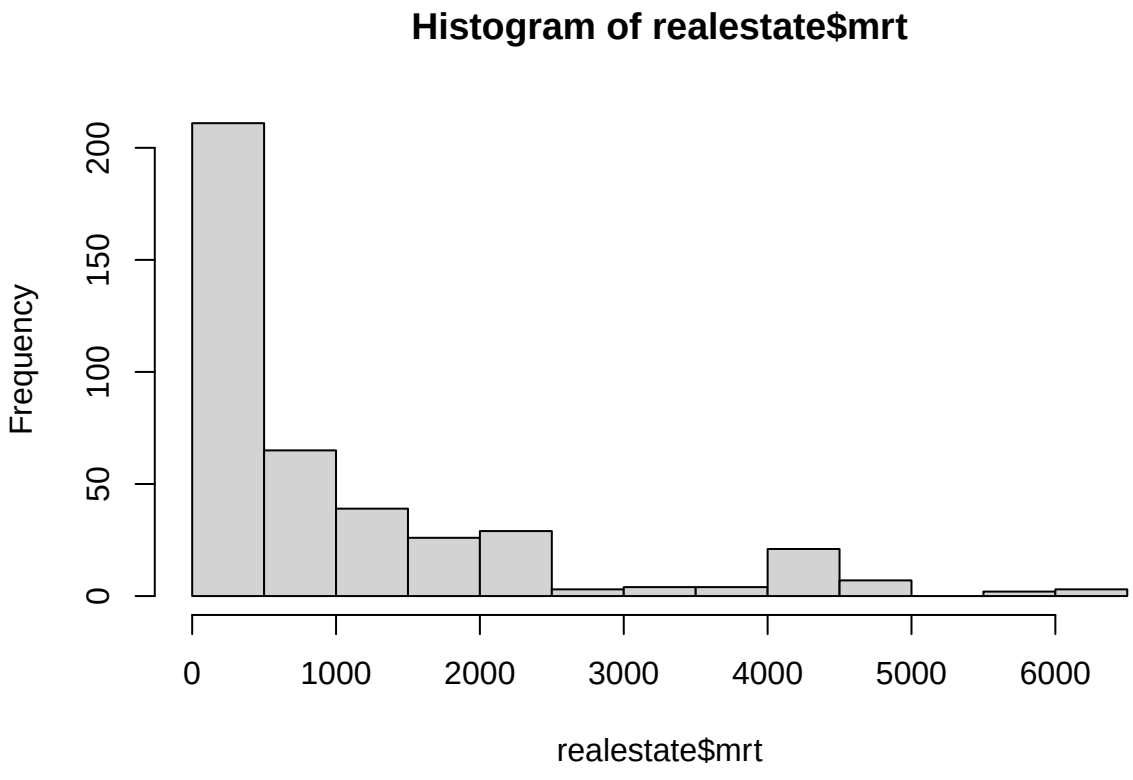
```
hist(realestate$price)
```



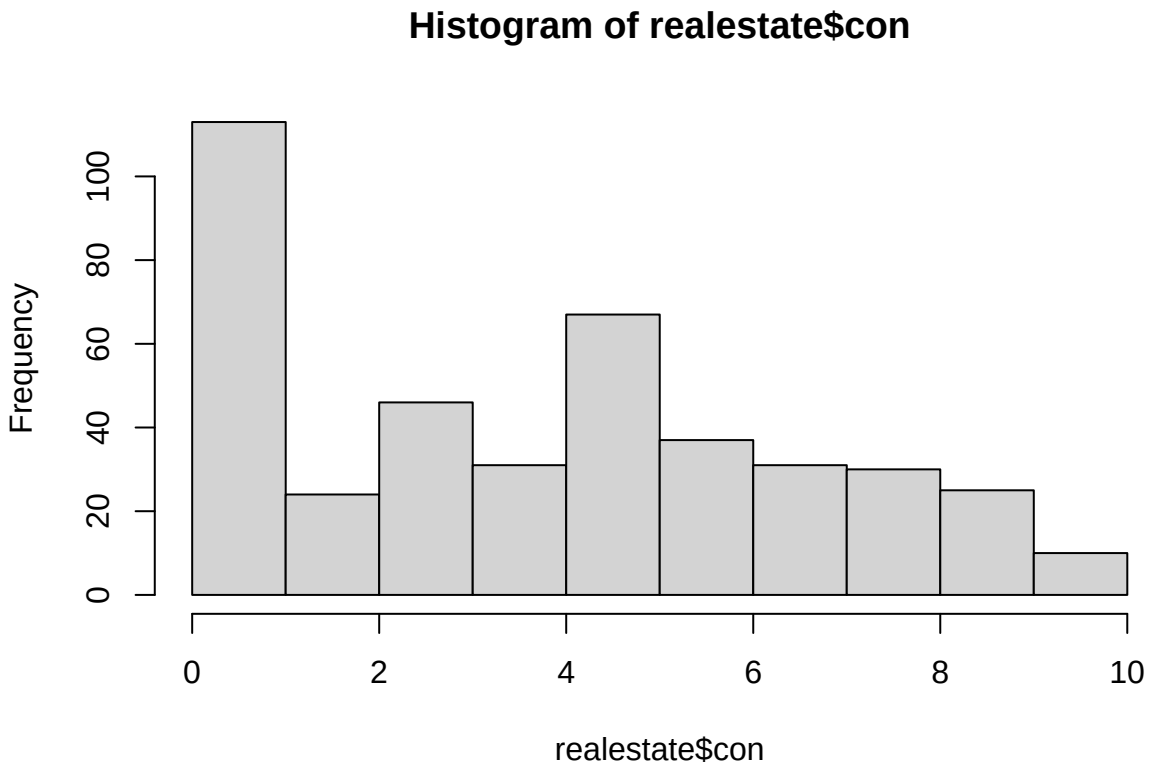

```
hist(realestate$age)
```



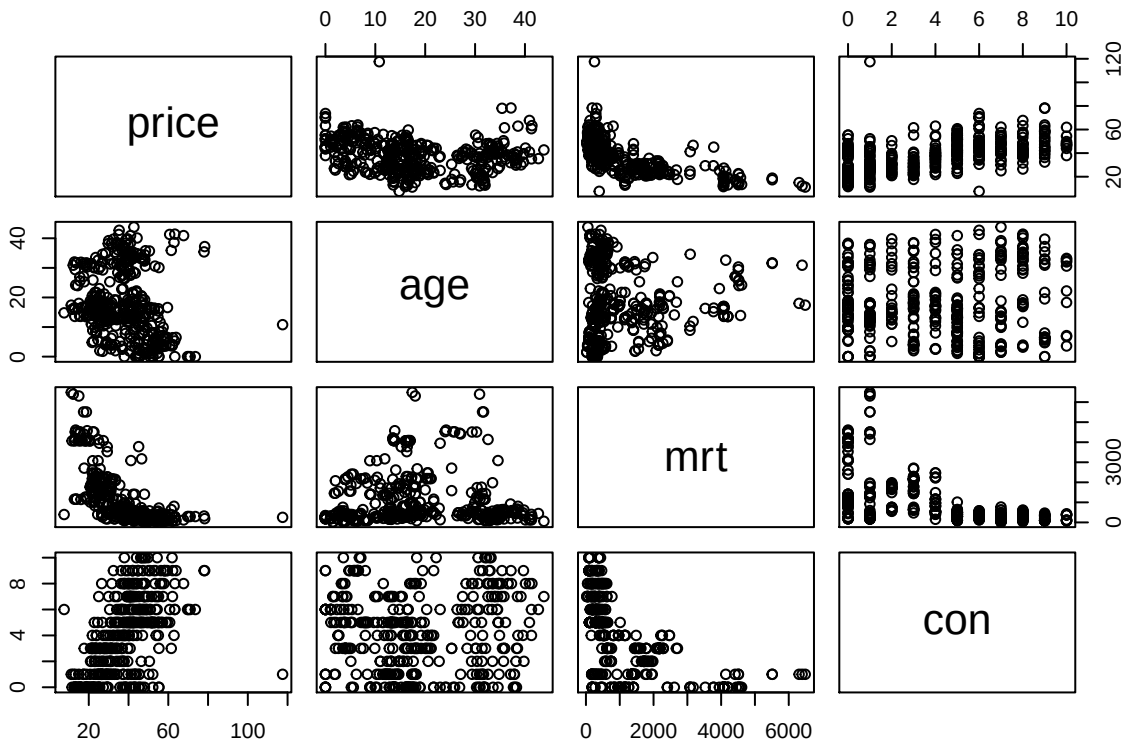

```
hist(realestate$mrt)
```




```
hist(realestate$con)
```




```
pairs(realestate[,c(3,4,5,6)])
```



3. Create the linear regression model with price as your dependent variable and all of your independent variables.

```
fit<-lm(price~age+mrt+con,data=realestate)
summary(fit)
```

```
##
## Call:
## lm(formula = price ~ age + mrt + con, data = realestate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.304  -5.430  -1.738   4.325  77.315
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.977286   1.384542  31.041  < 2e-16 ***
## age         -0.252856   0.040105  -6.305 7.47e-10 ***
## mrt         -0.005379   0.000453 -11.874  < 2e-16 ***
## con          1.297443   0.194290   6.678 7.91e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.251 on 410 degrees of freedom
## Multiple R-squared:  0.5411, Adjusted R-squared:  0.5377
## F-statistic: 161.1 on 3 and 410 DF,  p-value: < 2.2e-16
```

4. Interpret what each coefficient means for the price of your house. Do these make sense in terms of the impact on house price? Are they different than your predictions in part 1?

age -0.252856 mrt -0.005379
con 1.297443

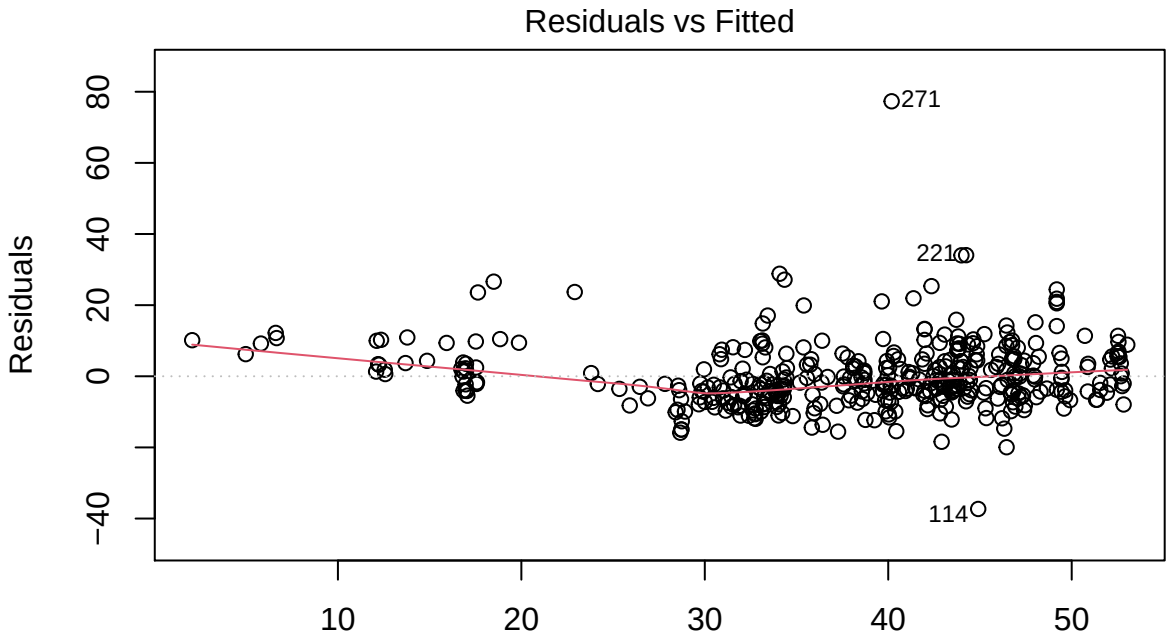
The age of the house had a negative coefficient, revealing a negative correlation with the price of the house, which was as predicted. The distance from Mass Rapid Transit had a very slight negative coefficient, revealing a very slight negative correlation with the price of the house. This was a contrast with my prediction in part 1. The number of convenience stores in the area had a positive coefficient and positive correlation with the price of the house. This also went against my prediction in part 1, perhaps more convenience stores mean a better location and a higher price.

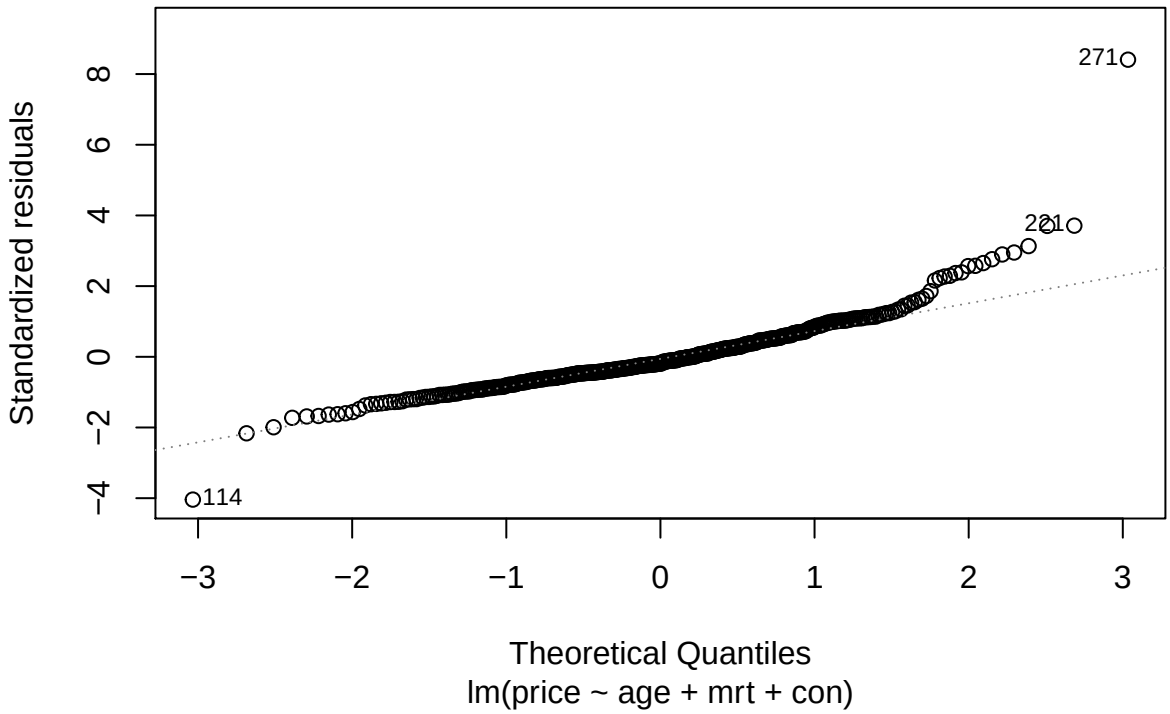
5. What is the fit of the model?

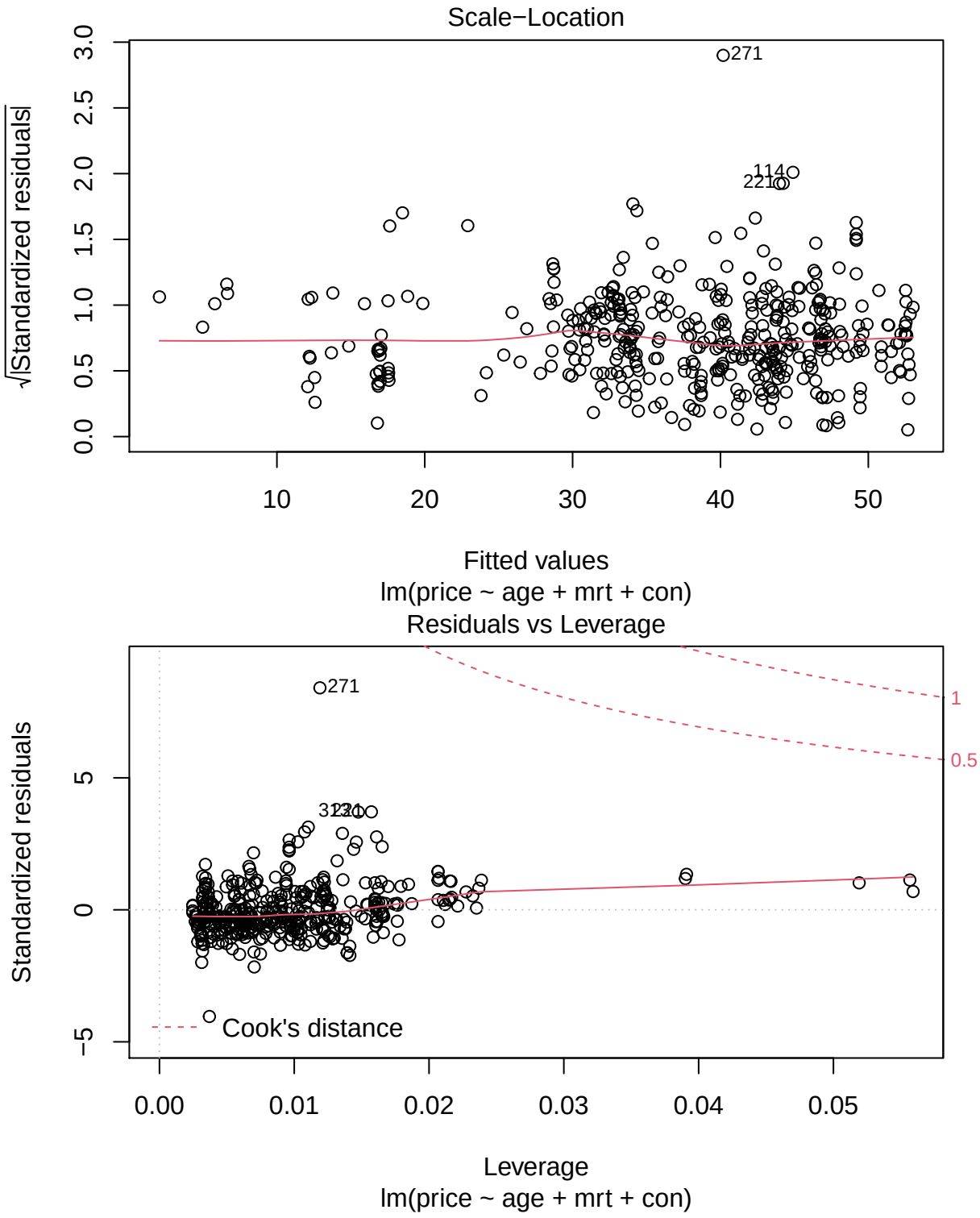
With an R^2 value of 0.54 we have an okay (not great) fit of the model. A little over 50% of the variance can be explained by the model.

6. Create the post-analysis graphs and determine which (if any) assumptions may be violated.

```
plot(fit)
```


Residuals vs Fitted
Residuals
Fitted values
lm(price ~ age + mrt + con)
271
221
114
−40
0
20
40
60
80
10
20
30
40
50


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(price ~ age + mrt + con)
271
221
114
−4
−2
0
2
4
6
8
−3
−2
−1
0
1
2
3



Assumption #1: Linearity

For our linearity assumption, there is a slight kink in the red line, but probably not enough for us to say that this assumption is violated.

Assumption #2: Normality

Our observations follow the dotted line very well, with a slight deviation on the right side. This suggests the normality assumption is intact. There is an exception of one observation (271) that might be an outlier.

Assumption #3: Homoskedasticity

The points appear to be dispersed pretty evenly across the red line (except for that pesky 271 again). This suggests the homoskedasticity assumption is intact.

Assumption #4: Outliers

While observation 271 appears to be an outlier, by the Cook's distance lines it doesn't appear to outside of the range of reasonable observations.

7. What are 3 variables you think should be included in this analysis that are not currently present?

1) **distance from schools**: a house that is closer to schools will likely have a higher value, as this is desirable for buyers with kids.
2) **lot size**: the lot size of the property wouldn't be included in the square footage, but would likely impact the price.
3) **number of baths/beds:** a house with more bathrooms might be valued higher than another house with a similar square footage.