Regression

2/8/2023

Regression Notebook

Dinesh Angadipeta(DXA190032), Alejo Vinluan(ABV210001)

2/8/2023

What is Linear Regression?

Linear regression is a measure that takes multiple independent and dependent variables and creates a model of their relationship in the form of a linear relationship. This is represented by a equation that is calculated, and can be used to make further predictions and estimates based off the relationship.

The strengths of linear regression is that it is a very easy way to interpret the relationship between multiple variables. It is very easy to understand in graph form as well, and can be used to make reasonable predictions.

The weaknesses of linear regression lies in the existence of outliers. Not every data set is has a clean set of data, and with outliers comes skewed data at times. Linear regression is very sensitive to outliers at times, and this could lead to future predictions being slightly off at times.

Data set

```r
data <- read.csv("melb_data.csv")
```

For this notebook we will be using the data set of 2017 Melbourne housing prices/sales. The source for the original Kaggle page of the data set is here. The code segment above reads the data set in.

```r
set.seed(1)
sample <- sample(c(TRUE, FALSE), nrow(data), replace=TRUE, prob=c(0.8,0.2))
train  <- data[sample, ]
test   <- data[!sample, ]
dim(train)
```

```
## [1] 10775    21
```

```r
dim(test)
```

```
## [1] 2805   21
```

Here we are dividing into a 80/20 train/test.

Data Exploration

```
names(train)
```

names function

```
##  [1] "Suburb"        "Address"       "Rooms"         "Type"
##  [5] "Price"         "Method"        "SellerG"       "Date"
##  [9] "Distance"      "Postcode"      "Bedroom2"      "Bathroom"
## [13] "Car"           "Landsize"      "BuildingArea"  "YearBuilt"
## [17] "CouncilArea"   "Lattitude"     "Longtitude"    "Regionname"
## [21] "Propertycount"
```

Here we are listing the names of the variables in the data set. This helps plan out what variables will be useful for data exploration.

```
str(train)
```

str function

```
## 'data.frame':    10775 obs. of  21 variables:
##  $ Suburb       : chr  "Abbotsford" "Abbotsford" "Abbotsford" "Abbotsford" ...
##  $ Address      : chr  "85 Turner St" "25 Bloomburg St" "5 Charles St" "55a Park St" ...
##  $ Rooms        : int  2 2 3 4 2 1 2 2 3 2 ...
##  $ Type         : chr  "h" "h" "h" "h" ...
##  $ Price        : num  1480000 1035000 1465000 1600000 1636000 ...
##  $ Method       : chr  "S" "S" "SP" "VB" ...
##  $ SellerG      : chr  "Biggin" "Biggin" "Biggin" "Nelson" ...
##  $ Date         : chr  "3/12/2016" "4/02/2016" "4/03/2017" "4/06/2016" ...
##  $ Distance     : num  2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 2.5 ...
##  $ Postcode     : num  3067 3067 3067 3067 3067 ...
##  $ Bedroom2     : num  2 2 3 3 2 1 3 2 3 2 ...
##  $ Bathroom     : num  1 1 2 1 1 1 1 2 2 2 ...
##  $ Car          : num  1 0 0 2 2 1 2 1 2 1 ...
##  $ Landsize     : num  202 156 134 120 256 0 220 0 214 0 ...
##  $ BuildingArea : num  NA 79 150 142 107 NA 75 NA 190 94 ...
##  $ YearBuilt    : num  NA 1900 1900 2014 1890 ...
##  $ CouncilArea  : chr  "Yarra" "Yarra" "Yarra" "Yarra" ...
##  $ Lattitude    : num  -37.8 -37.8 -37.8 -37.8 -37.8 ...
##  $ Longtitude   : num  145 145 145 145 145 ...
##  $ Regionname   : chr  "Northern Metropolitan" "Northern Metropolitan" "Northern Metropolitan" "North
##  $ Propertycount: num  4019 4019 4019 4019 4019 ...
```

Here we are using the “str” function to see how the data set is structured.

```
colSums(is.na(train))
```

colSums function using is.na

```
##        Suburb       Address         Rooms          Type         Price 
##             0             0             0             0             0 
##        Method       SellerG          Date      Distance      Postcode 
##             0             0             0             0             0 
##      Bedroom2      Bathroom           Car      Landsize  BuildingArea 
##             0             0            52             0          5137 
##     YearBuilt   CouncilArea     Lattitude    Longtitude    Regionname 
##          4288             0             0             0             0 
## Propertycount 
##             0
```

Here we are looking at the number of missing values in each of the variables of the data set. This can cause problems with the missing data values, so we can replace all the missing values with the mean values of the columns to make the data calculations a little more accurate.

```
test$Car[is.na(test$Car)]<-mean(test$Car,na.rm=TRUE)
test$YearBuilt[is.na(test$YearBuilt)]<-mean(test$YearBuilt,na.rm=TRUE)
train$Car[is.na(train$Car)]<-mean(train$Car,na.rm=TRUE)
train$YearBuilt[is.na(train$YearBuilt)]<-mean(train$YearBuilt,na.rm=TRUE)
```

```
dim(train)
```

dim function

```
## [1] 10775    21
```

As used before when creating the test and training data, the dim function helps how the number of rows and columns.

```
head(train)
```

head function

```
##       Suburb            Address Rooms Type   Price Method SellerG      Date
## 1 Abbotsford       85 Turner St     2    h 1480000      S  Biggin 3/12/2016
## 2 Abbotsford    25 Bloomburg St     2    h 1035000      S  Biggin 4/02/2016
## 3 Abbotsford       5 Charles St     3    h 1465000     SP  Biggin 4/03/2017
## 5 Abbotsford        55a Park St     4    h 1600000     VB  Nelson 4/06/2016
## 8 Abbotsford      98 Charles St     2    h 1636000      S  Nelson 8/10/2016
## 9 Abbotsford 6/241 Nicholson St     1    u  300000      S  Biggin 8/10/2016
##   Distance Postcode Bedroom2 Bathroom Car Landsize BuildingArea YearBuilt
```

```
## 1      2.5     3067        2        1   1      202           NA  1964.564
## 2      2.5     3067        2        1   0      156           79  1900.000
## 3      2.5     3067        3        2   0      134          150  1900.000
## 5      2.5     3067        3        1   2      120          142  2014.000
## 8      2.5     3067        2        1   2      256          107  1890.000
## 9      2.5     3067        1        1   1        0           NA  1964.564
##   CouncilArea Lattitude Longtitude           Regionname Propertycount
## 1       Yarra  -37.7996   144.9984 Northern Metropolitan          4019
## 2       Yarra  -37.8079   144.9934 Northern Metropolitan          4019
## 3       Yarra  -37.8093   144.9944 Northern Metropolitan          4019
## 5       Yarra  -37.8072   144.9941 Northern Metropolitan          4019
## 8       Yarra  -37.8060   144.9954 Northern Metropolitan          4019
## 9       Yarra  -37.8008   144.9973 Northern Metropolitan          4019
```

The head function helps look at the first 6 rows.

```
summary(train)
```

summary function

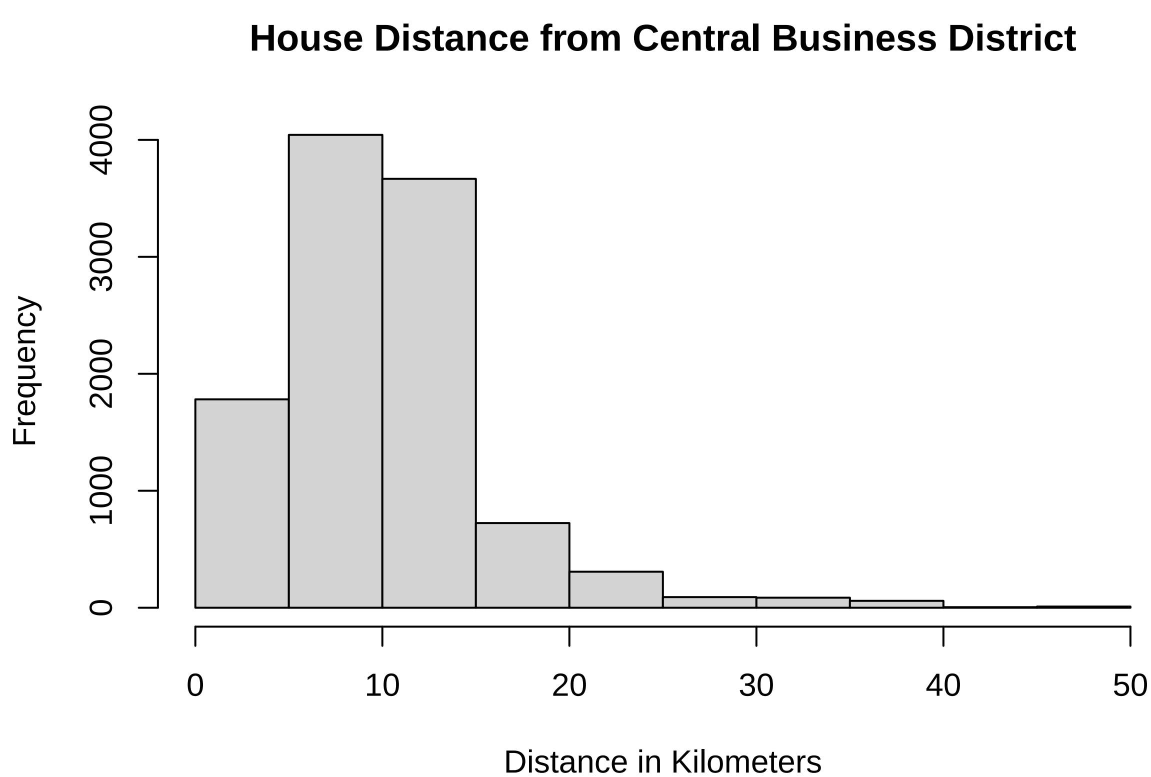
```
##     Suburb            Address              Rooms            Type          
##  Length:10775       Length:10775       Min.   : 1.000   Length:10775      
##  Class :character   Class :character   1st Qu.: 2.000   Class :character  
##  Mode  :character   Mode  :character   Median : 3.000   Mode  :character  
##                                        Mean   : 2.931                     
##                                        3rd Qu.: 3.000                     
##                                        Max.   :10.000                     
##                                                                           
##      Price            Method             SellerG              Date          
##  Min.   : 131000   Length:10775       Length:10775       Length:10775      
##  1st Qu.: 650000   Class :character   Class :character   Class :character  
##  Median : 901000   Mode  :character   Mode  :character   Mode  :character  
##  Mean   :1073697                                                           
##  3rd Qu.:1329000                                                           
##  Max.   :9000000                                                           
##                                                                            
##     Distance       Postcode       Bedroom2         Bathroom    
##  Min.   : 0.0   Min.   :3000   Min.   : 0.000   Min.   :0.000  
##  1st Qu.: 6.1   1st Qu.:3046   1st Qu.: 2.000   1st Qu.:1.000  
##  Median : 9.2   Median :3084   Median : 3.000   Median :1.000  
##  Mean   :10.1   Mean   :3105   Mean   : 2.909   Mean   :1.533  
##  3rd Qu.:13.0   3rd Qu.:3149   3rd Qu.: 3.000   3rd Qu.:2.000  
##  Max.   :47.3   Max.   :3977   Max.   :20.000   Max.   :8.000  
##                                                                
##       Car            Landsize         BuildingArea      YearBuilt   
##  Min.   : 0.000   Min.   :     0.0   Min.   :   0.0   Min.   :1196  
##  1st Qu.: 1.000   1st Qu.:   173.0   1st Qu.:  92.0   1st Qu.:1960  
##  Median : 2.000   Median :   431.0   Median : 125.0   Median :1965  
##  Mean   : 1.599   Mean   :   563.8   Mean   : 144.2   Mean   :1965  
##  3rd Qu.: 2.000   3rd Qu.:   650.0   3rd Qu.: 173.3   3rd Qu.:1973  
##  Max.   :10.000   Max.   :433014.0   Max.   :3558.0   Max.   :2018  
```

```
##                                    NA's   :5137
##   CouncilArea          Lattitude        Longtitude     Regionname
##   Length:10775       Min.   :-38.18   Min.   :144.4   Length:10775
##   Class :character   1st Qu.:-37.86   1st Qu.:144.9   Class :character
##   Mode  :character   Median :-37.80   Median :145.0   Mode  :character
##                      Mean   :-37.81   Mean   :145.0
##                      3rd Qu.:-37.76   3rd Qu.:145.1
##                      Max.   :-37.41   Max.   :145.5
##
##   Propertycount
##   Min.   :  249
##   1st Qu.: 4380
##   Median : 6543
##   Mean   : 7467
##   3rd Qu.:10331
##   Max.   :21650
##
```

This function gives an overview of the statistics of each variable.

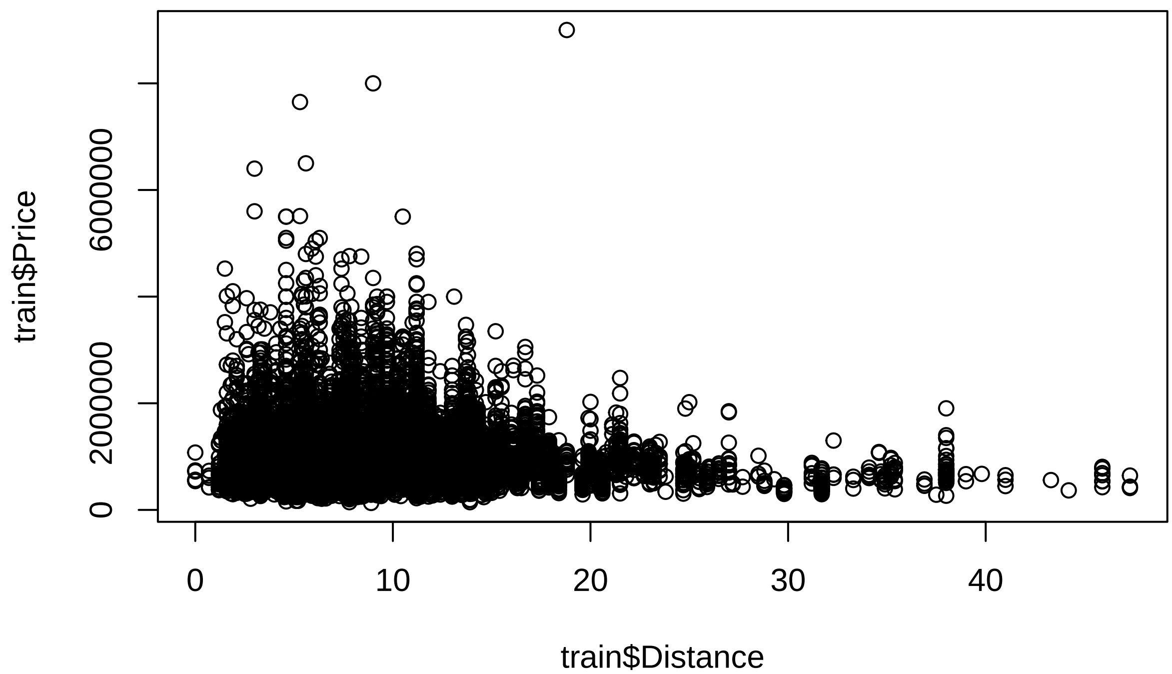
Informative Graphs

```
options(scipen=5)
hist(train$Distance, main = "House Distance from Central Business District", xlab = "Distance in Kilome
```



This graph shows how many houses are located at certain distances from the Central Business District in Melbourne.

```
options(scipen=5)
plot(train$Distance, train$Price)
```



This graph shows how the house price relates to the distance from the Central Business District.

Linear Regression Model

```
lm1 <- lm(Price~Distance, data = train)
summary(lm1)
```

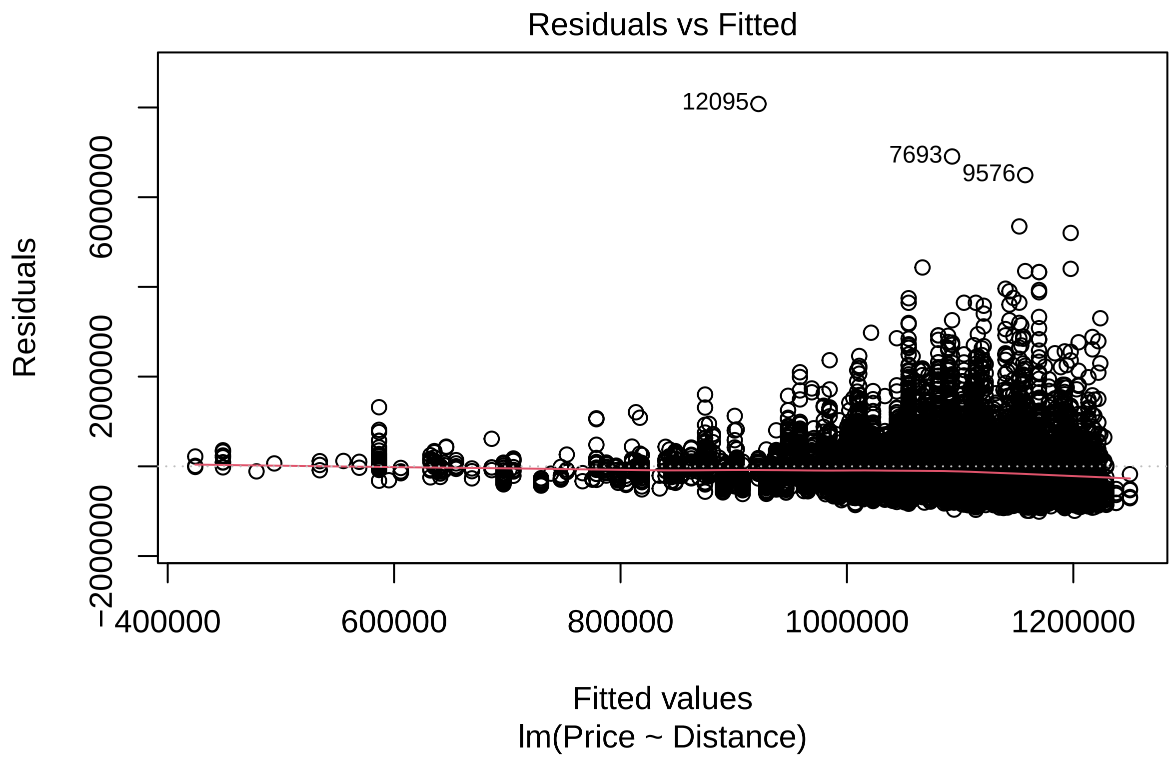
```
##
## Call:
## lm(formula = Price ~ Distance, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1009733  -401394  -153461   249918  8078148
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

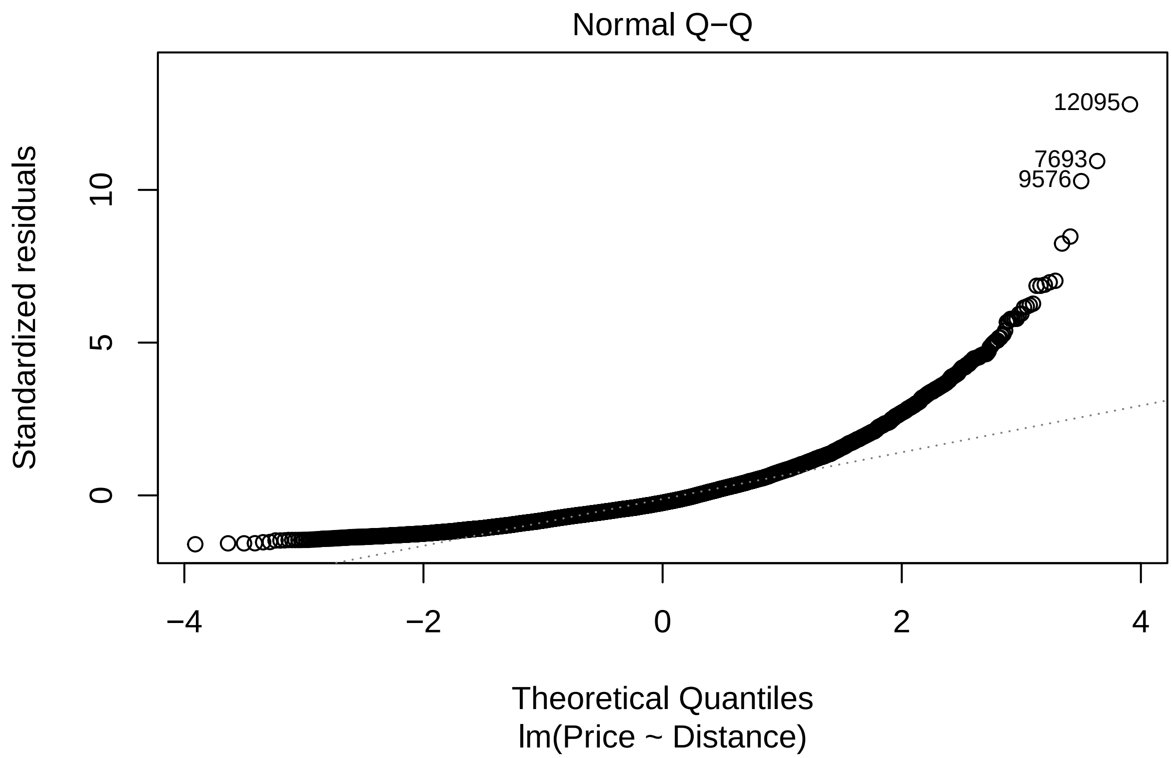
```
## (Intercept)  1250032      12216  102.33   <2e-16 ***
## Distance      -17456       1049  -16.64   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 631200 on 10773 degrees of freedom
## Multiple R-squared:  0.02507,    Adjusted R-squared:  0.02498
## F-statistic:   277 on 1 and 10773 DF,  p-value: < 2.2e-16
```

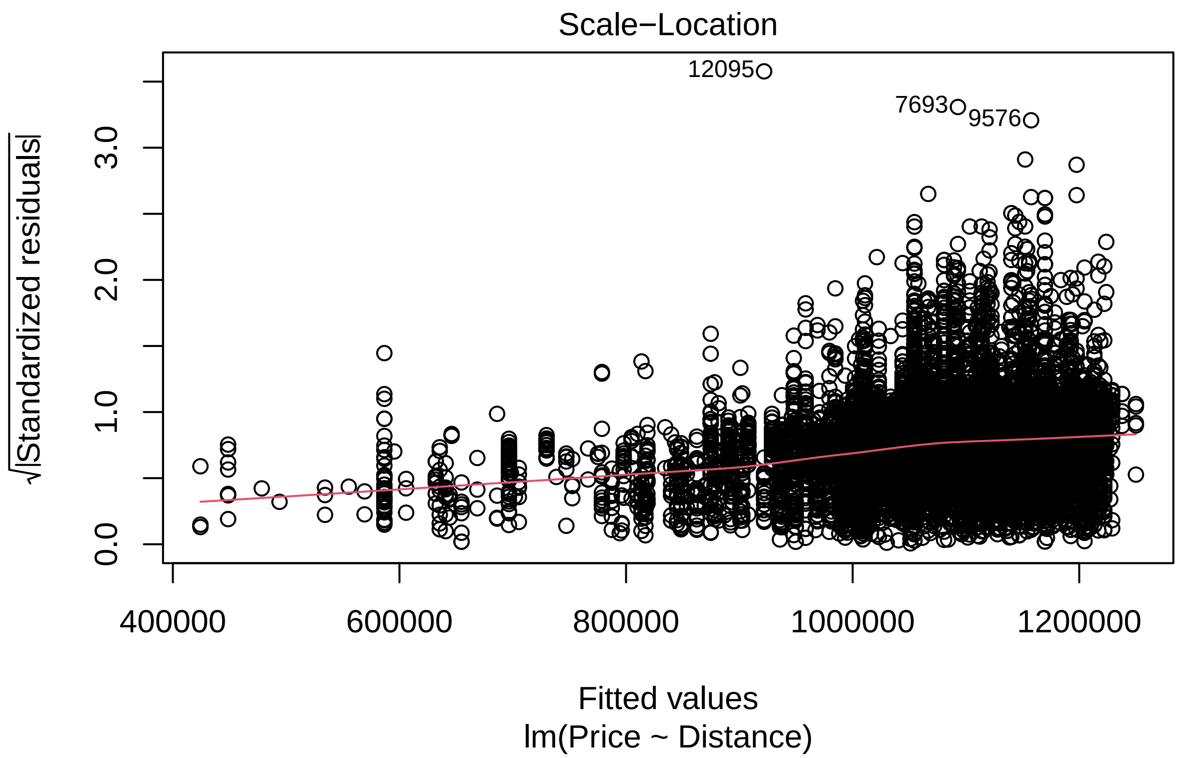
This code segment builds a simple linear regression model. For linear regression, the parameters can be defined by w and b, where w stands for the slope of the line and b stands for the intercept. Here w = -17456 and b = 1250032. So that means that for every kilometer increase in away from the Central Business District, a house in Melbourne drops around $17,456 in value on average. The intercept helps show that the average price of a house located at the Central Business District is around $1,250,032. Looking at the linear regression values, we can actually see some problems. For example, the R-squared statistic is quite far from the value 1, showing that this may not have that strong of a correlation. The RSE shows that the model is about 631200 y units off, which is still pretty big despite the relatively large scale of the numbers used in the data. The p-value is low, which does show some signs of it being a decent model. All in all, this model may need some more variables to help find a better correlation.

Residual Plot

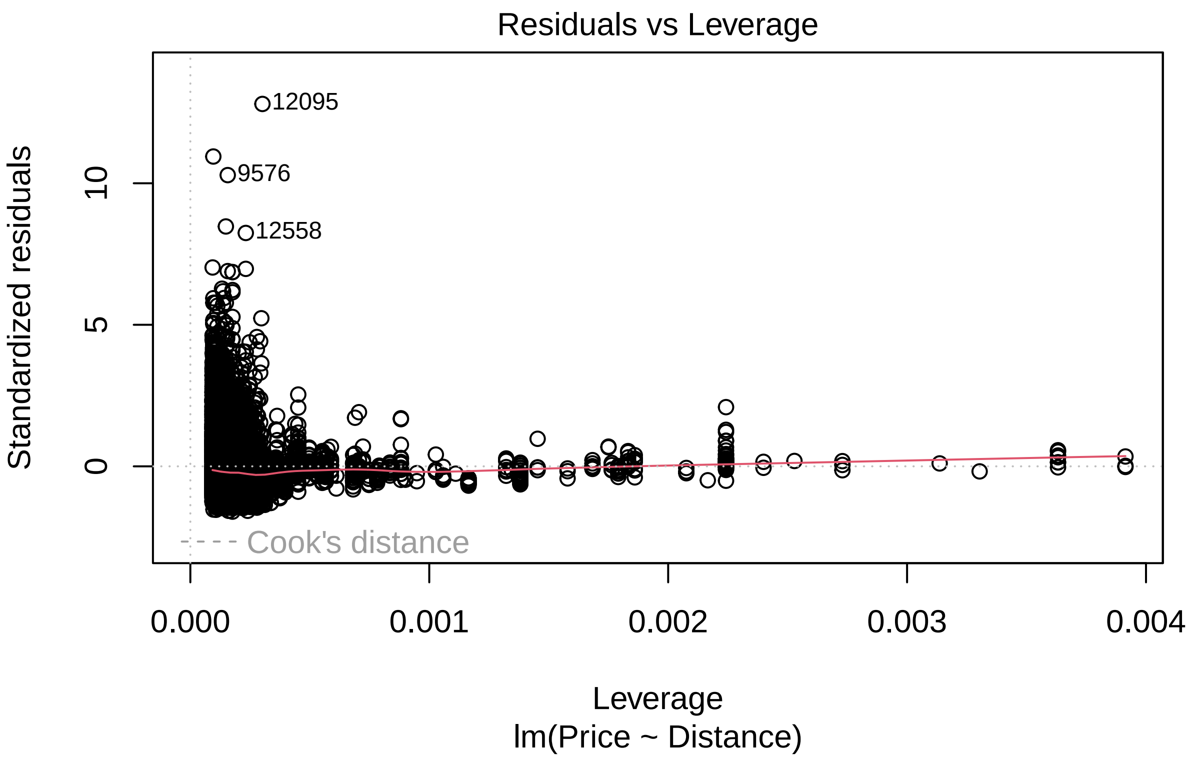
```r
lm1 <- lm(Price~Distance, data = train)

plot(lm1)
```


Residuals vs Fitted
12095
7693
9576
Residuals
−2000000
2000000
6000000
400000
600000
800000
1000000
1200000
Fitted values
lm(Price ~ Distance)


Normal Q−Q
12095
7693
9576
Standardized residuals
0
5
10
−4
−2
0
2
4
Theoretical Quantiles
lm(Price ~ Distance)



Scale−Location
12095
7693
9576
3.0
2.0
1.0
0.0
√|Standardized residuals|
400000
600000
800000
1000000
1200000
Fitted values
lm(Price ~ Distance)



Residuals vs Fitted

The Residual vs Fitted plot shows no distinct patterns, so it is a good indication that there aren't non-linear relationships.

Normal Q-Q

The Normal Q-Q plot isn't an exact perfect line and it skews lightly upward near the end. The residuals are generally normally distributed here, but problems could arise around #9576, #7693, and #12095.

Scale-Location

The Scale-Location plot shows a about horizontal line with around equal residuals on either side. This shows that the residuals are spread about equally along the ranges of the predictors.

Residuals vs Leverage

Since the Residuals vs Leverage plot barely shows the Cook's distance, there are not many influential outliers that will truly skew and affect the data.

Multiple Linear Regression Models

Using Distance and YearBuilt as predictors.

```
lm2 <- lm(Price~Distance+YearBuilt, data = train)
summary(lm2)
```

```
##
```

```
## Call:
## lm(formula = Price ~ Distance + YearBuilt, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3762287  -385234  -146842   235223  8012157
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11271962.7   403242.1   27.95   <2e-16 ***
## Distance      -12559.5     1038.9  -12.09   <2e-16 ***
## YearBuilt      -5126.5      206.2  -24.86   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 613800 on 10772 degrees of freedom
## Multiple R-squared:  0.07798,    Adjusted R-squared:  0.07781
## F-statistic: 455.6 on 2 and 10772 DF,  p-value: < 2.2e-16
```

Using Rooms, Bathroom, Distance, Car, YearBuilt, Landsize, Propertycount, and Bedroom2 as predictors

```
lm3 <- lm(Price~Bedroom2+Rooms+Bathroom+Distance+Car+YearBuilt+Propertycount+Landsize, data = train)
summary(lm3)
```

```
##
## Call:
## lm(formula = Price ~ Bedroom2 + Rooms + Bathroom + Distance +
##     Car + YearBuilt + Propertycount + Landsize, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3604093  -271488   -80423   186595  8327920
##
## Coefficients:
##                   Estimate   Std. Error t value Pr(>|t|)
## (Intercept)   10008747.939   325287.518  30.769  < 2e-16 ***
## Bedroom2         26375.803    14308.243   1.843 0.065298 .
## Rooms           219106.442    14680.102  14.925  < 2e-16 ***
## Bathroom        256248.403     8567.721  29.909  < 2e-16 ***
## Distance        -31115.634      870.377 -35.750  < 2e-16 ***
## Car              60286.890     5396.119  11.172  < 2e-16 ***
## YearBuilt        -4997.721      165.766 -30.149  < 2e-16 ***
## Propertycount       -1.675        1.058  -1.584 0.113271
## Landsize             3.724        1.049   3.551 0.000386 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 479800 on 10766 degrees of freedom
## Multiple R-squared:  0.4369, Adjusted R-squared:  0.4365
## F-statistic:  1044 on 8 and 10766 DF,  p-value: < 2.2e-16
```

Findings: As we can see, all 3 of the models bring very different summaries to the table. The linear model using only Distance as a predictor showed poor correlation statistics, showcasing that the distance from the

Central Business District in Melbourne was not the best predictor for house prices in the area, and that it needed more. In the multiple linear regression model using only Distance and the YearBuilt variables, it is seen that even just these two variables yield a poor correlation result, even though it showed a better correlation than the linear model using a single predictor, showcasing that the housing market of Melbourne is affected by many more variables added together. Finally, the third linear regression model showcasing the use of Rooms, Bathroom, Distance, Car, YearBuilt, Landsize, Propertycount, and Bedroom2 yielded a drastically better result, having an R-squared value of 0.4369 which is the best of all the models. This ultimately shows that the housing market of Melbourne is not dominated by certain factors, and only shows a correlation when factoring all the important statistics together. All in all the third linear regression model is the best on to use, with a lower RSE, and higher R-squared value.

Predictions

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

Predictions for 1st linear model

```
## [1] "correlation: -1"
```

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

```
## [1] "mse: 1157689720417.46"
```

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

```
## [1] "rmse: 1075959.90651021"
```

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

Predictions for 2nd linear model

```
## [1] "correlation: -0.971584181265245"
```

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

```
## [1] "mse: 1168740221587.22"
```

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

```
## [1] "rmse: 1081082.89302311"
```

```r
pred <- predict(lm3, newdata = test)
correlation <- cor(pred, test$Bedroom2+test$Rooms+test$Bathroom+test$Distance+test$Car+test$YearBuilt+te
print(paste("correlation:", correlation))
```

Predictions for 3rd linear model

```
## [1] "correlation: -0.0561826205776105"
```

```r
mse <- mean((pred-(test$Bedroom2+test$Rooms+test$Bathroom+test$Distance+test$Car+test$YearBuilt+test$Pr
print(paste("mse:",mse))
```

```
## [1] "mse: 1322176978672.71"
```

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

```
## [1] "rmse: 1149859.54736773"
```

Conclusion From the shown predictions, it can be seen that the first 2 linear models show a large negative correlation. Despite that, the third linear model shows a very weak negative correlation. The reason why the third linear model could show a weak negative correlation could be due to the fact that some of the factors used had opposing affects on other factors. Conflicting correlations caused the correlation to ultimately level our. The high rmse could be a byproduct of the very high price values used in the data.