Stat 363 HomeWork 1

BERFİN AKDEMİR, ZEYNEP KANAR

2024-08-11

Stat363 Linear Models I Homework 1

The data was obtained from Kaggle. Source link: https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques/data

Reading and observing the data

```
house <- read.csv("/Users/berfinakdemir/Desktop/STAT363/Homework/HW1/HousePriceData.csv")
head(house)
```

```
##           id            date   price bedrooms bathrooms sqft_living sqft_lot
## 1 7129300520 20141013T000000  221900        3      1.00        1180     5650
## 2 6414100192 20141209T000000  538000        3      2.25        2570     7242
## 3 5631500400 20150225T000000  180000        2      1.00         770    10000
## 4 2487200875 20141209T000000  604000        4      3.00        1960     5000
## 5 1954400510 20150218T000000  510000        3      2.00        1680     8080
## 6 7237550310 20140512T000000 1225000        4      4.50        5420   101930
##   floors waterfront view condition grade sqft_above sqft_basement yr_built
## 1      1          0    0         3     7       1180             0     1955
## 2      2          0    0         3     7       2170           400     1951
## 3      1          0    0         3     6        770             0     1933
## 4      1          0    0         5     7       1050           910     1965
## 5      1          0    0         3     8       1680             0     1987
## 6      1          0    0         3    11       3890          1530     2001
##   yr_renovated zipcode     lat     long sqft_living15 sqft_lot15
## 1            0   98178 47.5112 -122.257          1340       5650
## 2         1991   98125 47.7210 -122.319          1690       7639
## 3            0   98028 47.7379 -122.233          2720       8062
## 4            0   98136 47.5208 -122.393          1360       5000
## 5            0   98074 47.6168 -122.045          1800       7503
## 6            0   98053 47.6561 -122.005          4760     101930
```

```
dim(house)
```

```
## [1] 21613    21
```

We have observed that the data has over 21.000 rows, for practicality through this analysis we used a random sample, then proceeded to observe the sample

```
set.seed(1012)
samp <- sample(nrow(house), 250)
hsamp <- house[samp, ]

dim(hsamp)
```

```
## [1] 250  21
```

```
str(hsamp)
```

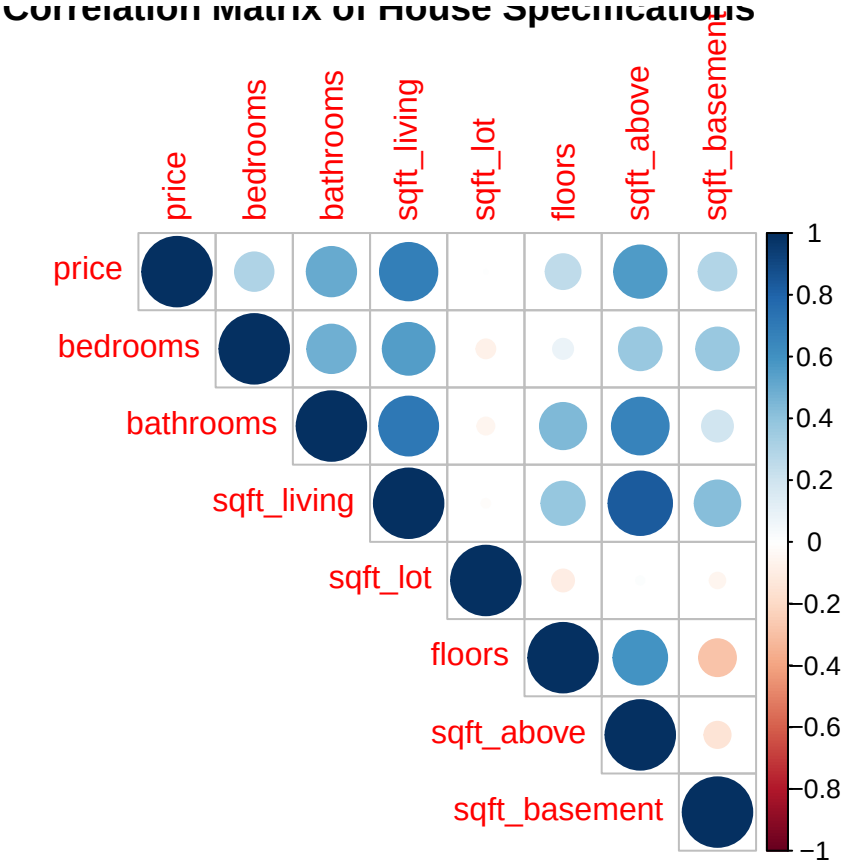
```
## 'data.frame':    250 obs. of  21 variables:
##  $ id           : num  2.20e+08 9.25e+09 5.46e+09 3.03e+08 3.96e+09 ...
##  $ date         : chr  "20150401T000000" "20140819T000000" "20150407T000000" "20140528T000000" ...
##  $ price        : num  599950 331000 1000000 245100 467000 ...
##  $ bedrooms     : int  3 2 3 3 3 4 5 3 2 3 ...
##  $ bathrooms    : num  2.5 1 1.75 1.75 2.5 2.5 2.5 1.75 1 3 ...
##  $ sqft_living  : int  1970 1480 2610 1300 3460 2170 2820 1840 1030 1850 ...
##  $ sqft_lot     : int  106722 6210 6360 7958 6590 7533 67518 11440 5072 19966 ...
##  $ floors       : num  1 1 2 1 2 2 2 1 1 1 ...
##  $ waterfront   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ view         : int  4 0 2 0 0 0 0 0 0 0 ...
##  $ condition    : int  3 3 3 3 3 3 3 4 3 4 ...
##  $ grade        : int  9 7 8 7 7 8 8 8 6 7 ...
##  $ sqft_above   : int  1970 1080 2130 1300 3460 2170 2820 1340 1030 1090 ...
##  $ sqft_basement: int  0 400 480 0 0 0 0 500 0 760 ...
##  $ yr_built     : int  1985 1950 1924 1996 2001 1991 1979 1977 1924 1992 ...
##  $ yr_renovated : int  0 0 0 0 0 0 0 0 1958 0 ...
##  $ zipcode      : int  98022 98133 98109 98092 98056 98059 98029 98059 98115 98038 ...
##  $ lat          : num  47.2 47.8 47.6 47.3 47.5 ...
##  $ long         : num  -122 -122 -122 -122 -122 ...
##  $ sqft_living15: int  2910 1290 3010 1640 2490 2170 2820 1940 1220 1410 ...
##  $ sqft_lot15   : int  101494 7509 6000 8698 6312 8728 48351 11440 6781 6715 ...
```

We obtained a correlation matrix with some of the numeric variables we have observed above to decide which variable is suitable to the respond variable (which is price as seen on the data) for regression analysis then review with a plot

```
numeric_data <- hsamp[, c("price", "bedrooms", "bathrooms", "sqft_living", "sqft_lot", "floors", "sqft_a
cor_matrix <- cor(numeric_data)
print(cor_matrix)
```

```
##                    price    bedrooms   bathrooms sqft_living     sqft_lot
## price         1.000000000  0.30621488  0.50371091  0.68059935  0.002562162
## bedrooms      0.306214883  1.00000000  0.48441080  0.55411830 -0.073263660
## bathrooms     0.503710908  0.48441080  1.00000000  0.71817167 -0.059519024
## sqft_living   0.680599354  0.55411830  0.71817167  1.00000000 -0.015400570
## sqft_lot      0.002562162 -0.07326366 -0.05951902 -0.01540057  1.000000000
## floors        0.250772305  0.08342732  0.44155262  0.38744078 -0.099481461
## sqft_above    0.566074536  0.37743268  0.66614483  0.83330701  0.013765367
## sqft_basement 0.294034548  0.37570893  0.19783916  0.42944582 -0.050066493
##                    floors  sqft_above sqft_basement
## price          0.25077230  0.56607454    0.29403455
## bedrooms       0.08342732  0.37743268    0.37570893
## bathrooms      0.44155262  0.66614483    0.19783916
## sqft_living    0.38744078  0.83330701    0.42944582
## sqft_lot      -0.09948146  0.01376537   -0.05006649
## floors         1.00000000  0.59690472   -0.28130996
## sqft_above     0.59690472  1.00000000   -0.14137885
## sqft_basement -0.28130996 -0.14137885    1.00000000
```

```
corrplot(cor_matrix, , type = "upper",
         title = "Correlation Matrix of House Specifications")
```

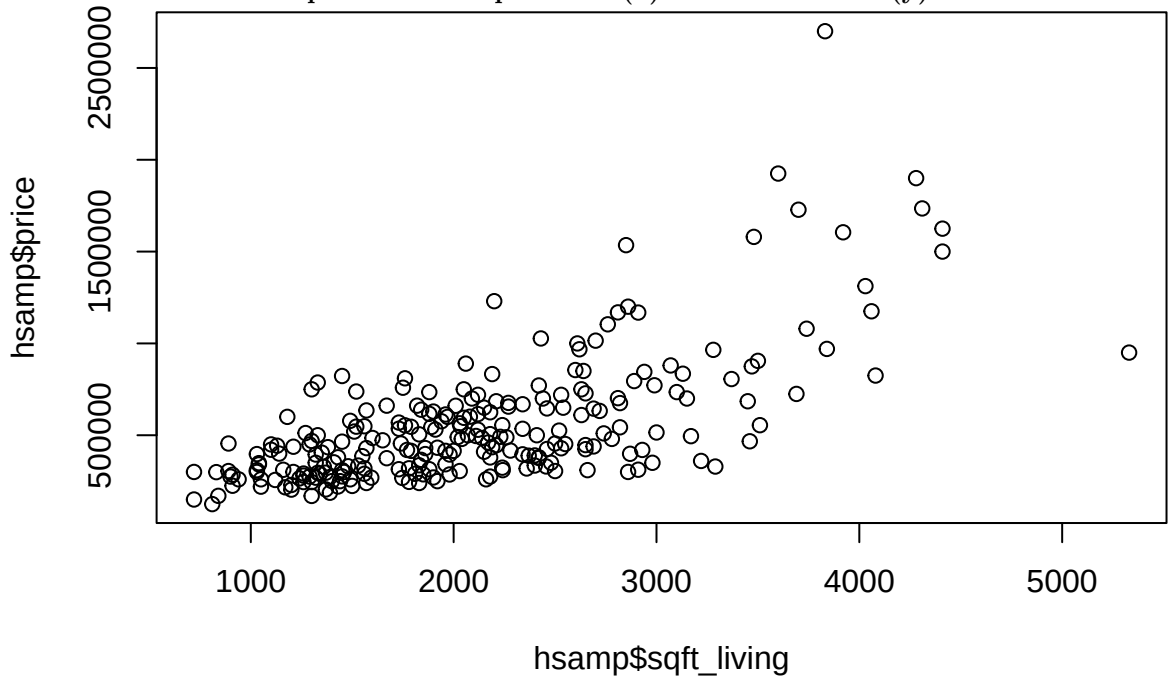


Checking assumptions

We proceeded to examine the sqft_living variable for the assumptions of the simple linear regression

```
plot(hsamp$sqft_living, hsamp$price)
abline(model5)
```

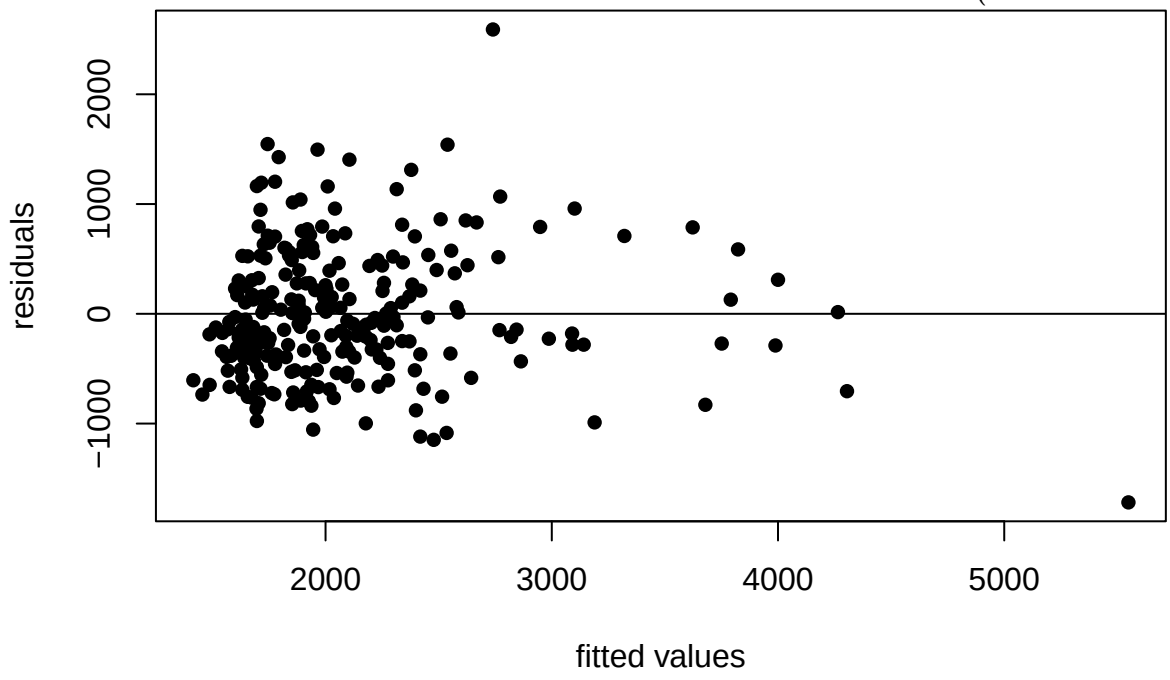
Linearity of the data: The relationship between the predictor (x) and the outcome (y) is assumed to be linear.



There is an acceptable linear relationship observed on the plot. We proceed to check the other assumptions by residulas since solely this plot is not sufficient.

```
yhat=predict(model5)  ### obtaining yhat by predict()
resid=model5$residuals
plot(yhat,resid,pch=16,xlab="fitted values",ylab="residuals")
abline(h=0)
```

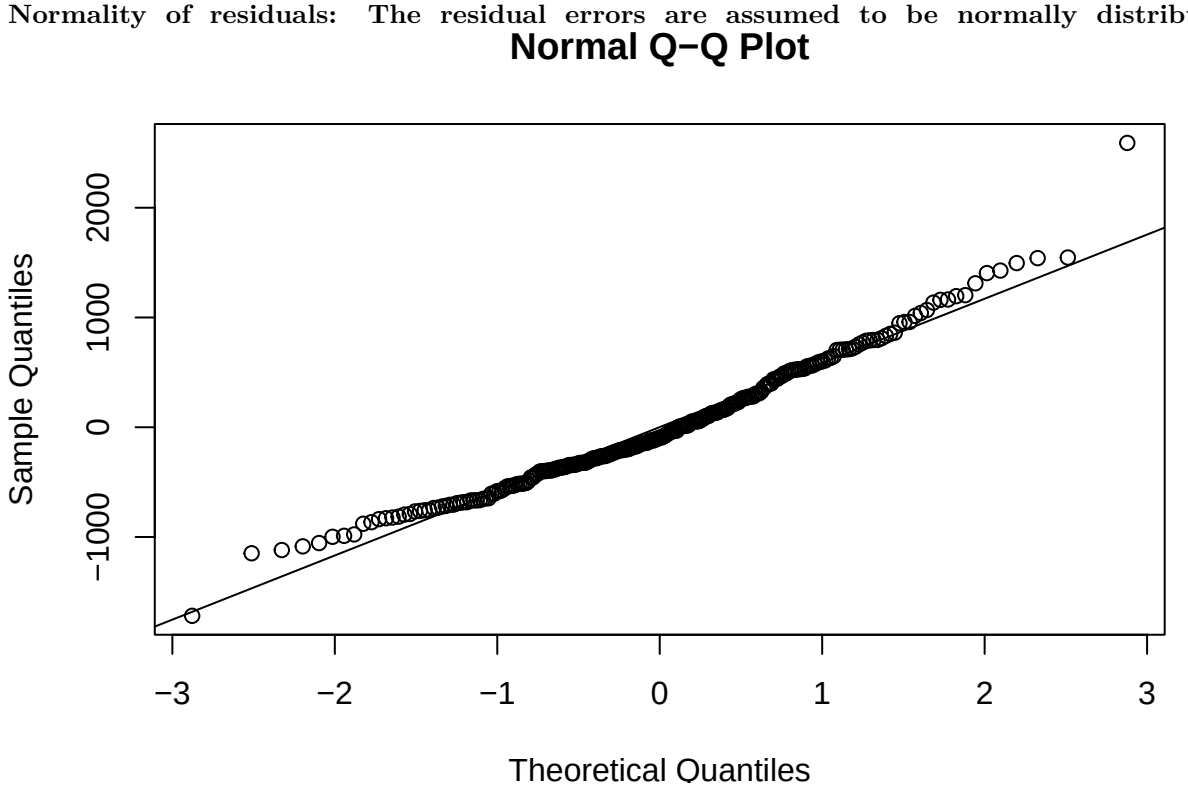
Homogeneity of residuals variance: The residuals are assumed to have a constant variance (homoscedasticity)



The residuals bounce randomly around the 0 line which suggests that the variances are constant.

```
qqnorm(resid)
qqline(resid)
```

Normality of residuals: The residual errors are assumed to be normally distributed.



The quantiles mostly lie on the straight line, which indicates the residuals are normally distributed.

Since we only observe one dependent and one independent variable there is no need to test for any hidden relationships between variable. We can say the independence assumption is also satisfied.

After the assumptions are seen to be satisfied, we observe the model

```
model5 <- lm(sqft_living  ~ price, hsamp)
model5
```

```
##
## Call:
## lm(formula = sqft_living ~ price, data = hsamp)
##
## Coefficients:
## (Intercept)        price
##   1.215e+03    1.605e-03
```

observing the analysis of variance table to test the significance of the model

H0 = model is not significant Ha = model is significant

```
anova(model5)
```

```
## Analysis of Variance Table
```

```
##
## Response: sqft_living
##            Df   Sum Sq  Mean Sq F value    Pr(>F)
## price       1 77315935 77315935  214.01 < 2.2e-16 ***
## Residuals 248 89595445   361272
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

F value is 214.01, p-value is clearly smaller than 0.05, so we reject the null hypothesis. It is possible to conclude that the model is significant.

we conduct a t-test to test the significance of the model

There is a hypothesis test computed here also;

H0 = model is not significant Ha = model is significant

```
summary(model5)
```

```
##
## Call:
## lm(formula = sqft_living ~ price, data = hsamp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1718.03  -393.12   -95.69   395.58  2590.18
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.215e+03  7.165e+01   16.95   <2e-16 ***
## price       1.605e-03  1.097e-04   14.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 601.1 on 248 degrees of freedom
## Multiple R-squared:  0.4632, Adjusted R-squared:  0.4611
## F-statistic:   214 on 1 and 248 DF,  p-value: < 2.2e-16
```

If we were to continue examining the outcome of the t-test the p-values for the significance of the coefficients ß0-hat and ß1-hat are less than 0.05 so we reject the null hypothesis and we can say ß0-hat and ß1-hat are significant therefore the model is significant. We also see that the coefficient of determination (adjusted R-squared) has a value of 0.4611. The value of the coefficient of determination here implies that our model explains ~46% of the total variance.