Building a Linear Regression model for housing dataset

Nithya S

2023-11-10

Introduction

Linear regression is a machine learning algorithm which estimates how a model is following a linear relationship between one response variable (denoted by y) and one or more explanatory variables (denoted by X_1, X_2, X_3…., X_n). The response variable will dependent on how the explanatory variables changes and not the other way round. Response variable is also known as target or dependent variable while the explanatory variable is known as independent or predictor variables.

There are two types of linear regression:

1. Simple Linear Regression

2. Multiple Linear Regression

Simple Linear Regression: It is a type of linear regression model where there is only independent or explanatory variable.

Objective

To choose any dataset for simple linear regression and examine the following

1. To comment about the different steps involved in building a simple linear regression model

2. To plot the scatter diagram for the data and find coefficient of correlation. What inference can be drawn from the scatter plot.

3. To estimate the parameters of a simple linear regression model and fit a regression line. To interpret the results.

4. To test the significance of the regression coefficient and interpret the results.

5. Different ways in which we can assess the quality of the fit.

Different steps involved in building a simple linear regression model

1. Reading and understanding the data We need to understand the data properly in order to proceed to further analysis.

2. Identifying the dependent and independent variables. Understanding or identifying the independent and dependent variables becomes crucial when it comes to building a linear regression model. 3.Visualizing the data We need to plot the data in order to depict the relationship the variables.
3. Building a linear model We then proceed to build a model in order to estimate the parameters of the model
4. Residual analysis Here we check if the fitted values and the predicted values collide in order to validate our model.
5. Testing of hypotheses Based on our objective and interest we test the hypotheses for intercept and slope parameters and draw inference for the dataset from the model.
6. Goodness of fit. From the results obtained we check for the significance of the test and evaluate the model.

To plot the scatter diagram for the data and find coefficient of correlation. Inference to be drawn from the scatter plot.

The housing dataset is taken from Kaggle

URL for the dataset:

https://www.kaggle.com/datasets/ashydv/housing-dataset/data?select=Housing.csv

The dataset has information about the price, area, different rooms, furniture style. Here the interested variables are price and area. Since the price increases or decreases with the decrease in area, we conclude that the independent variable is area (X) and the dependent variable is price(Y). We now proceed to check the kind of relationship between the variables of interest. First, we import the data and then proceed further

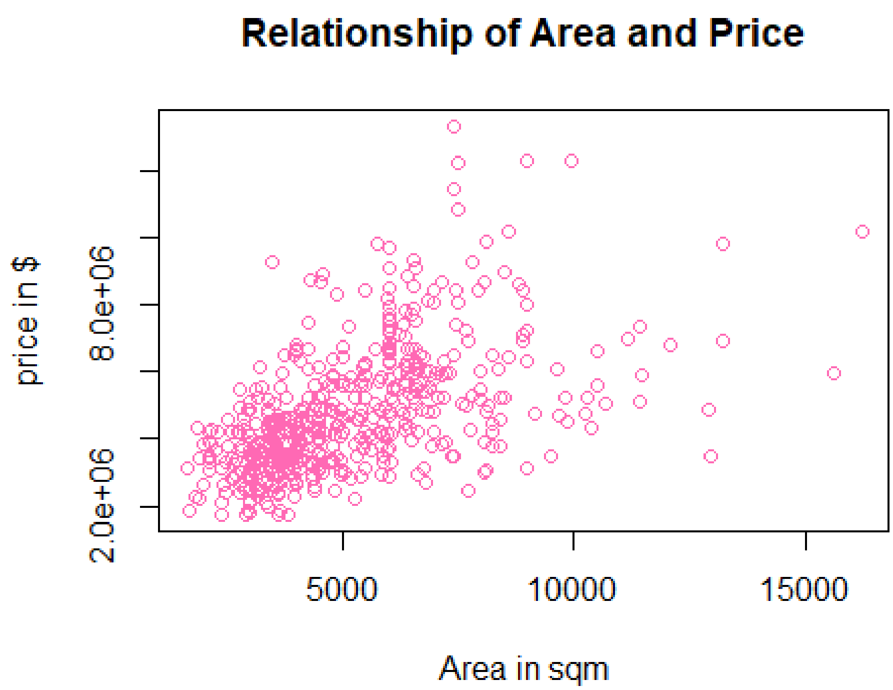
```
library(readr)
Housing_dataset <- read_csv("G:/My Drive/Linear
Regression/Datasets/Housing_dataset.csv")
```

```
## Rows: 545 Columns: 13
## ── Column specification
────────────────────────────────────────────────────────
## Delimiter: ","
## chr (7): mainroad, guestroom, basement, hotwaterheating, airconditioning,
pr...
## dbl (6): price, area, bedrooms, bathrooms, stories, parking
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this
message.
```

```
View(Housing_dataset)
attach(Housing_dataset)
```

We now plot the graph as follows:

```
plot(Housing_dataset$area,Housing_dataset$price,col="hotpink",main="Relations
hip of Area and Price",xlab="Area in sqm",ylab="price in $")
```



From the scatter plot we interpret as: the independent variable (X) is area in square meters and the dependent variable (Y) is price in dollars. The plot shows that there is a moderately linear positive relationship between the variables under study.

The relationship can be understood better numerically by calculating the Karl Pearson's co-efficient. We find the correlation of coefficient also known as Karl Pearson's correlation coefficient as follows:

```
cor(Housing_dataset$area,Housing_dataset$price)

## [1] 0.5359973
```

The correlation co-efficient between the variables is 0.5359973 So we say that there is a moderately positive linear relationship between the variables under study.

To estimate the parameters of a simple linear regression model and fit a regression line. To interpret the results.

In order to estimate the parameters, we fit the model. We build the model as below:

```
reg_model=lm(Housing_dataset$price~Housing_dataset$area)
reg_model
```

```
##
## Call:
## lm(formula = Housing_dataset$price ~ Housing_dataset$area)
##
## Coefficients:
##          (Intercept)  Housing_dataset$area
##              2387308                   462
```

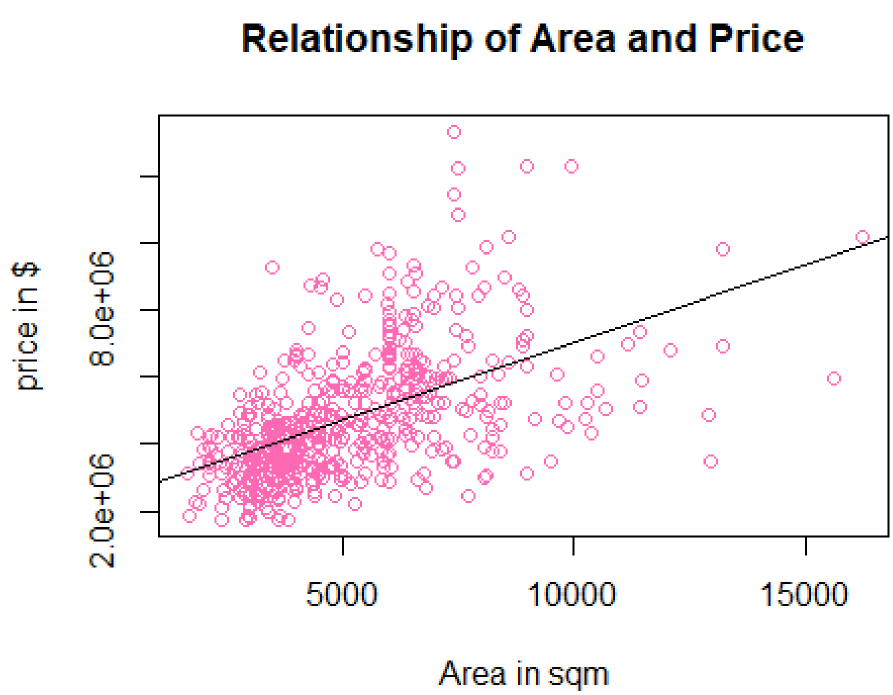
The model here is

$$Y = B_0 + B_1 X + U$$

From the model table, we see that the intercept parameter(B_0) is 2387308 and the slope parameter(B_1) is 462. The intercept term (B_0) depicts that if X sometimes becomes zero [Hypothetical situation], the intercept is simply the expected value of Y at that value. So, if B_1 becomes zero the average value of the price will be $2387308. The slope term (B_1) depicts that for every 1 unit increase in X, the value of Y increases by $462.

To fit a regression line

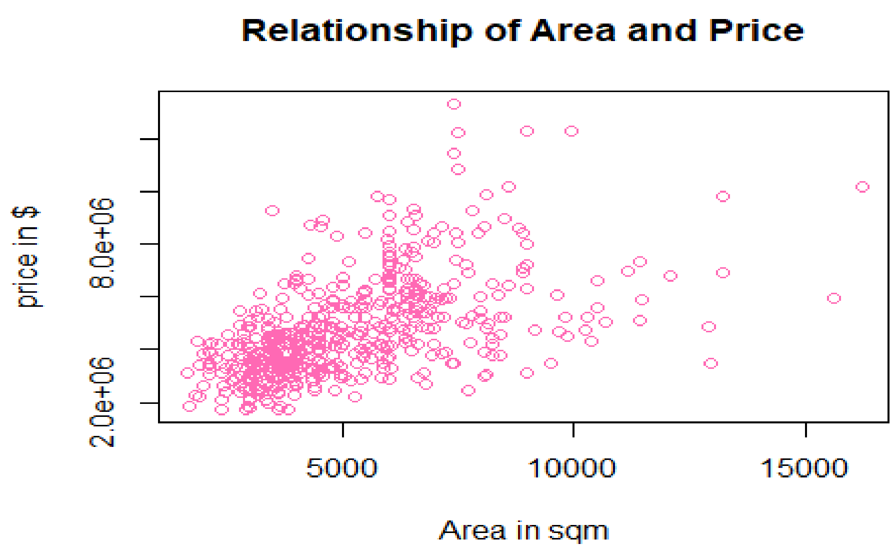
```
plot(Housing_dataset$area,Housing_dataset$price,col="hotpink",main="Relations
hip of Area and Price",xlab="Area in sqm",ylab="price in $")

abline(reg_model)
```



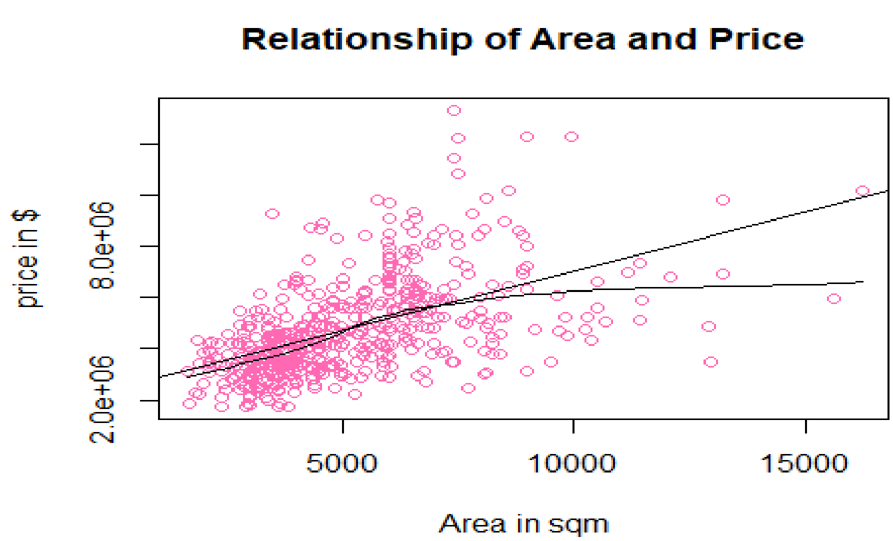
The interpretation that can be drawn from the above graph is that, the above graph is the best possible fit for the dataset. We also see that there are outliers in the dataset from the above graph.

We can also have a smooth line connecting the points which gives us a better visual in order to understand the data better.

```
plot(Housing_dataset$area,Housing_dataset$price,col="hotpink",main="Relations
hip of Area and Price",xlab="Area in sqm",ylab="price in $")
```




```
scatter.smooth(Housing_dataset$area,Housing_dataset$price,col="hotpink",main=
"Relationship of Area and Price",xlab="Area in sqm",ylab="price in $")

abline(reg_model)
```



To test the significance of the regression coefficient and interpret the results.

We give the hypotheses as follows:

H_0: B_1=B_{10} [zero]

v/s

H_1: B_1=!B_{10} [zero]

We quickly check the summary of the model as foll0ws:

```
summary(reg_model)

##
## Call:
## lm(formula = Housing_dataset$price ~ Housing_dataset$area)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4867112 -1022228  -200135   683027  7484838
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.387e+06  1.745e+05   13.68   <2e-16 ***
## Housing_dataset$area 4.620e+02  3.123e+01   14.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1581000 on 543 degrees of freedom
## Multiple R-squared:  0.2873, Adjusted R-squared:  0.286
## F-statistic: 218.9 on 1 and 543 DF,  p-value: < 2.2e-16
```

From the above table the p-value in the line of units is <2e-16 which is less than 0.05 level of significance. Hence, we reject null hypothesis and hence B_1=!0. Thus, these exist a significant linear relationship between the variables. In other words, there is a significant relationship between the area and price of the houses.

Goodness of fit-Analysis

Here t calculated value 14.79 is significantly large than t (table value) = 1.9643 [two tailed test with 344 degrees of freedom and 5% significance level], therefore there exist a very strong linear relation. Since p-value is very small compared to 0.05 there exist a very strong linear relation.

Different ways in which we can assess the quality of the fit.

We fit the observed and the predicted values as follows

```
Fitted_values=fitted.values(reg_model)
Fitted_values
```

```
##       1       2       3       4       5       6       7       8       9
10
## 5815162 6526604 6988578 5852120 5815162 5852120 6351053 9871302 6129305
5043664
##      11      12      13      14      15      16      17      18      19
20
## 8485377 5159158 5413244 4004221 5990713 5159158 5436343 6314095 4512393
5353187
##      21      22      23      24      25      26      27      28      29
30
## 4383040 5692739 6106206 4493914 6452688 5408624 5159158 6487336 6060009
4928170
##      31      32      33      34      35      36      37      38      39
40
## 5840571 5621133 4641746 5140679 5547217 5621133 5843805 6545083 5159158
5159158
##      41      42      43      44      45      46      47      48      49
50
## 5413244 5325469 5380906 5159158 5159158 5159158 5159158 5436343 4373801
5824402
##      51      52      53      54      55      56      57      58      59
60
## 5824402 5309300 5159158 4766479 5159158 5159158 7672301 6545083 5935276
5159158
##      61      62      63      64      65      66      67      68      69
70
## 5159158 6489646 5270032 5325469 7549878 6489646 8485377 5944515 5159158
7972585
##      71      72      73      74      75      76      77      78      79
80
## 4235208 5159158 4706422 5436343 4253687 4355322 5353187 5390145 5020565
5159158
##      81      82      83      84      85      86      87      88      89
90
## 5159158 4235208 7238045 5159158 4124334 6198601 5468681 4216729 5810542
6351053
##      91      92      93      94      95      96      97      98      99
100
## 4697183 5505639 4604788 5713528 5159158 4281406 6545083 5343948 5436343
5159158
##     101     102     103     104     105     106     107     108     109
110
## 5436343 4928170 4928170 5320849 4928170 4466196 4905072 5353187 3884107
5443272
##     111     112     113     114     115     116     117     118     119
120
## 5436343 6254962 4373801 6831507 5528738 6083108 5574935 4096616 5353187
5630372
##     121     122     123     124     125     126     127     128     129
130
```

```
## 5408624 5727849 5276499 5768965 5401695 9594117 5695049 5390145 4928170
7681541
##     131     132     133     134     135     136     137     138     139
140
## 4604788 5079698 4789578 4604788 5621133 5159158 4881973 4530872 4697183
5325469
##     141     142     143     144     145     146     147     148     149
150
## 5066763 5464061 7238045 4604788 4558590 4697183 7238045 4928170 5325469
5436343
##     151     152     153     154     155     156     157     158     159
160
## 4760012 4419998 4881973 3911826 4073517 5205355 5574935 3688692 6073868
3842529
##     161     162     163     164     165     166     167     168     169
170
## 5256173 5205355 5436343 5540287 5487160 5367047 5990713 4512393 4355322
5408624
##     171     172     173     174     175     176     177     178     179
180
## 4928170 7131329 6267898 4835775 4142813 6914662 6323335 5182257 5660401
3856389
##     181     182     183     184     185     186     187     188     189
190
## 4466196 5713528 3962643 6073868 3773233 3773233 7658442 5205355 5029805
4022700
##     191     192     193     194     195     196     197     198     199
200
## 5898318 7330440 5436343 4604788 6152404 4424618 5938048 3680838 5135135
4327603
##     201     202     203     204     205     206     207     208     209
210
## 4475435 4279096 4290645 4881973 4590929 5297750 5066763 3773233 3759374
5491780
##     211     212     213     214     215     216     217     218     219
220
## 4533644 8346785 3967263 4694873 4396899 4309124 5177637 5557380 4611718
5621133
##     221     222     223     224     225     226     227     228     229
230
## 6129305 3967263 6621770 5307452 7117931 5362427 4775719 5159158 4064277
6853220
##     231     232     233     234     235     236     237     238     239
240
## 4881973 4383040 4117404 4309124 4179771 5011326 3713176 4701803 4470815
4235208
##     241     242     243     244     245     246     247     248     249
250
## 4161292 4124334 4068897 3565344 4845015 4863494 4013460 6267898 4281406
4692563
```

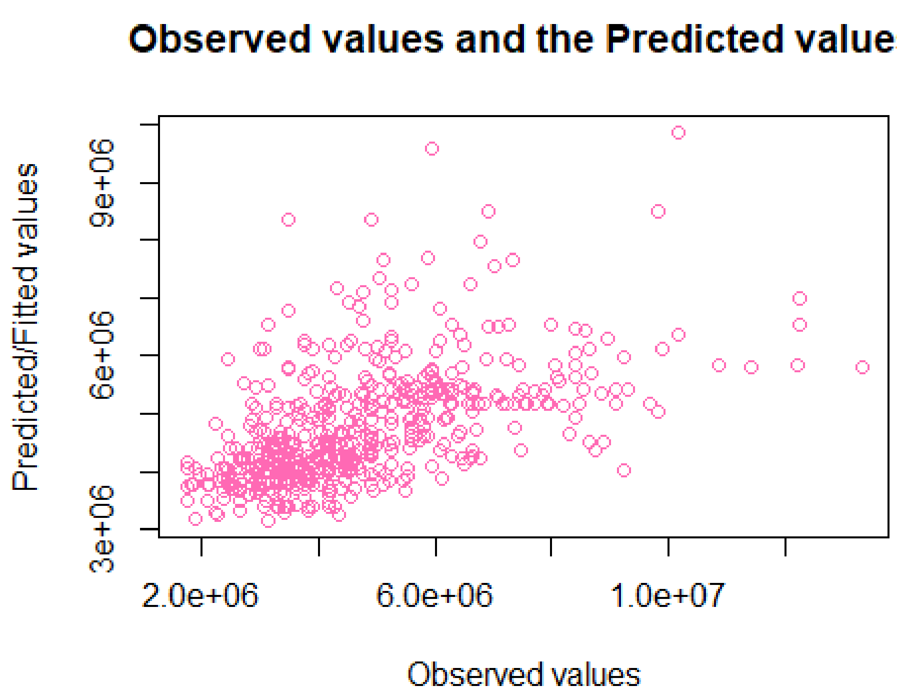
```
##     251     252     253     254     255     256     257     258     259 
260 
## 4008840 3981122 6942381 4013460 4470815 5106031 4235208 6198601 4253687 
5325469 
##     261     262     263     264     265     266     267     268     269 
270 
## 3848073 4008840 4119714 4220425 4650985 3717796 4641746 4660225 4674084 
4189011 
##     271     272     273     274     275     276     277     278     279 
280 
## 4466196 3267371 4269856 4004221 5367047 4249991 4419998 7173368 3958023 
5325469 
##     281     282     283     284     285     286     287     288     289 
290 
## 5325469 4466196 3392104 4401519 5976853 5459442 3674833 4928170 4715662 
5089862 
##     291     292     293     294     295     296     297     298     299 
300 
## 3593063 3751520 3656354 4424618 4235208 3461400 4512393 4068897 5066763 
5621133 
##     301     302     303     304     305     306     307     308     309 
310 
## 4271704 4013460 3378245 4466196 6198601 3981122 4623267 4272166 4256459 
4527176 
##     311     312     313     314     315     316     317     318     319 
320 
## 5152228 5186876 4050418 4087376 4253687 4974368 5112960 4693487 4392280 
3773233 
##     321     322     323     324     325     326     327     328     329 
330 
## 4383040 4064277 3985742 4881973 4466196 3985742 4281406 5380906 4466196 
4216729 
##     331     332     333     334     335     336     337     338     339 
340 
## 4258307 5741246 4928170 3773233 3907206 4150205 6120066 3378245 4133574 
3856389 
##     341     342     343     344     345     346     347     348     349 
350 
## 4835775 3856389 5691353 4272166 4165912 3318188 3392566 3934924 3842529 
4614027 
##     351     352     353     354     355     356     357     358     359 
360 
## 3967263 4050418 5080622 3706709 6267898 6198601 3551485 5588794 3994981 
4050418 
##     361     362     363     364     365     366     367     368     369 
370 
## 4253687 5168397 4258307 4043027 3828670 4905072 4064277 4064277 4992847 
4050418 
##     371     372     373     374     375     376     377     378     379 
380 
```

```
## 4364561 4036559 3856389 3773233 4013460 5140679 4295265 3703937 3438301
4013460
##     381     382     383     384     385     386     387     388     389
390
## 4466196 4235208 3842529 4466196 4466196 4068897 4165912 4346082 4073517
4512393
##     391     392     393     394     395     396     397     398     399
400
## 3373625 3789864 4230588 5817010 3994981 4050418 4068897 5112960 3828670
5782824
##     401     402     403     404     405     406     407     408     409
410
## 4009764 6776070 5103721 8367112 4650985 3800952 4845015 3378245 4235208
3858699
##     411     412     413     414     415     416     417     418     419
420
## 4165912 3378245 3593063 3288160 4253687 4597858 3981122 4068897 4004221
4678704
##     421     422     423     424     425     426     427     428     429
430
## 4290645 4581689 4105855 4119714 3819431 3858699 3634641 3378245 4253687
4593239
##     431     432     433     434     435     436     437     438     439
440
## 3542246 3856389 5186876 3994981 4139117 4253687 3378245 5103721 4466196
4202870
##     441     442     443     444     445     446     447     448     449
450
## 4068897 4406139 3627249 4383040 3828670 3981122 4228740 4004221 4279096
3149567
##     451     452     453     454     455     456     457     458     459
460
## 3981122 5505639 6545083 3805109 4466196 4925861 3495124 3773233 4165912
4004221
##     461     462     463     464     465     466     467     468     469
470
## 6129305 4678704 3385174 3814811 4466196 4142813 3814811 3884107 3697007
4512393
##     471     472     473     474     475     476     477     478     479
480
## 4732293 4119714 4064277 6106206 4397823 3773233 5089862 4678704 4050418
4078137
##     481     482     483     484     485     486     487     488     489
490
## 3994981 3634641 3842529 5443272 3791712 4064277 5159158 4881973 4789578
3911826
##     491     492     493     494     495     496     497     498     499
500
## 4396899 3606922 3611542 4216729 5528738 4235208 4235208 4204718 3311258
4064277
```

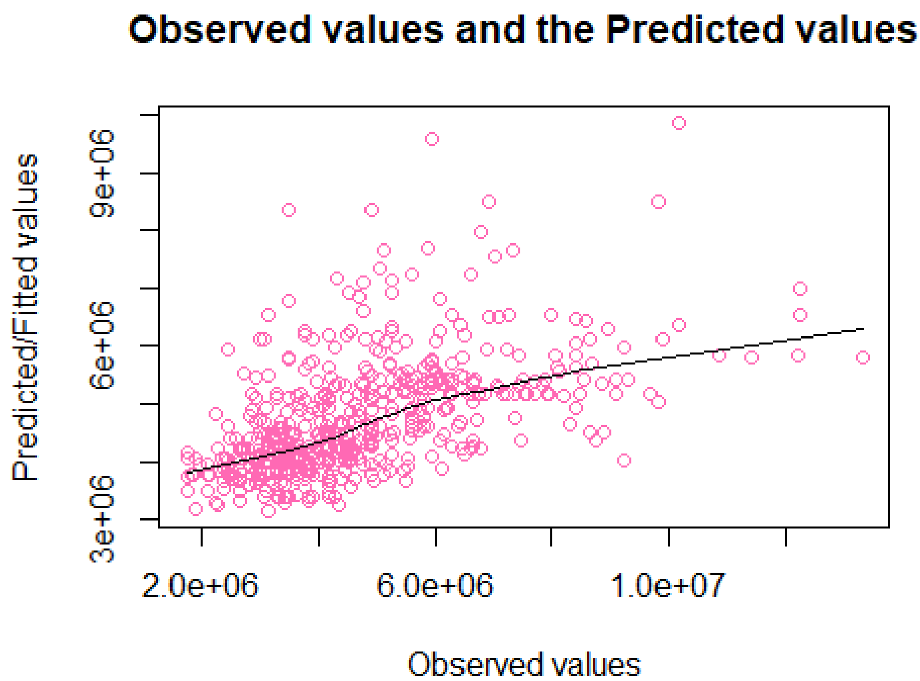
```
##      501      502      503      504      505      506      507      508      509
510
## 3680838 3509907 3994981 4235208 3858699 4235208 3731655 4050418 4419998
4050418
##      511      512      513      514      515      516      517      518      519
520
## 3717796 3856389 3773233 4419998 3773233 3870248 3884107 3773233 4004221
4623267
##      521      522      523      524      525      526      527      528      529
530
## 5944515 4066587 3530696 3674833 3895195 4068897 3856389 3235494 4221349
4221349
##      531      532      533      534      535      536      537      538      539
540
## 3288160 4835775 3773233 3496048 3773233 3939544 3967263 3172666 4073055
3768613
##      541      542      543      544      545
## 3773233 3496048 4059658 3731655 4165912
```

After calculating the fitted values with the help of the estimated coefficients B_0 and B_1 along with the independent variable area(X), we plot the observed and the predicted values. We also use “scatter.smooth” in order to get better understanding from the graph

```
plot(Housing_dataset$price,Fitted_values,col="hotpink",main="Observed values and the Predicted values",xlab="Observed values",ylab="Predicted/Fitted values")
```



```r
scatter.smooth(Housing_dataset$price,Fitted_values,col="hotpink",main="Observed values and the Predicted values",xlab="Observed values",ylab="Predicted/Fitted values")
```



When observed from the graph, the error is very small (y and y^{hat} are very close to each other.)

Coefficient of determination

Determining the coefficient is important in order to assess the quality of the fit We calculate the coefficient of determination (r) as follows:

```r
r=cor(Housing_dataset$price,Fitted_values)
r

## [1] 0.5359973

r^2

## [1] 0.2872932
```

r^2 = 0.2872932, we draw the inference as 29% of the data's total variability is explained by the independent variable (area) of the dependent variable(price). Although 50% of the data should be explained for the consideration of prediction for it to be model that can be considered as a good fit. According to the r^2, the model is a poor fit for the considered dataset.

Conclusion

We constructed the linear regression model and the estimated model obtained is

$$Y = 2387308 + 462X + U$$

The model has been tested for the hypotheses

H_0: B_1=B_{10} [zero]

v/s

H_1: B_1=!B_{10} [zero]

and the null hypothesis has been rejected and concluded that B_1 was not zero and thus there is a significant relationship between the variables, area and price.

The correlation of determination has been calculated and observed that 29% of the data's total variability is explained by area of the price which is a poor fit for the taken dataset.