Property values

Anaid Villamizar
Lancaster University

a.villamizaralbornoz@lancaster.ac.uk

Introduction

The dataset includes information about the house sales in the city of Northampton in Massachusetts situated in the United States. A train route used to be in this city, but this suspended all its services in 1960. In 1984, part of the railway was converted to a multi-use track (for cycling, walking and running). There is a study about the association between rail trails and property values, which was conducted by Hartenian and Horton (2015). The dataset used offer house information about square footage, acre, number of bedrooms and other. In this report, I will analyze the factors which are associated with house prices in Northampton in 2014, concentrating on any effect of the distance of the house to the bike trail.

The dataset used contained information about 104 houses located Northampton in Massachusetts and house prices are estimated from 2014 by Zillow. The house prices for this study is the response variable and the unit is in thousand dollars.

The variables are given in the Table 1.

Name of variable	Description
housenum	unique house number
acre	number of acres that the property encompasses
bedgroup	1-2, 3, 4+ bedrooms
bikescore	bike friendliness (0-100 score, higher scores are better)
distance	distance (in feet) to the nearest entry point to the rail trail network
garage_spaces	number of garage spaces (0-4)
no_full_baths	number of full baths (includes shower or bathtub)
no_half_baths	number of half paths (no shower or bath – In UK English, a WC)
no_rooms	number of room
squarefeet	square footage of interior finished space (in thousands of square feet)
walkscore	walk friendliness (0-100 score, higher scores are better)
zip	location (1060 = Northampton: 1062 = Florence)

Table 1. Explanatory variables for the house prices in Northampton

The variable zip was converted to a categorical factor with two levels (1060 and 1062). The variable bedgroup is also treated as categorical. The other variables were treated as continuous. Missing values were not detected in the dataset.

Exploratory analysis

An exploratory analysis was carried out using scatter plots to look simple relationships between house prices and the following variables: distance, squarefeet, number of rooms and walkscore.

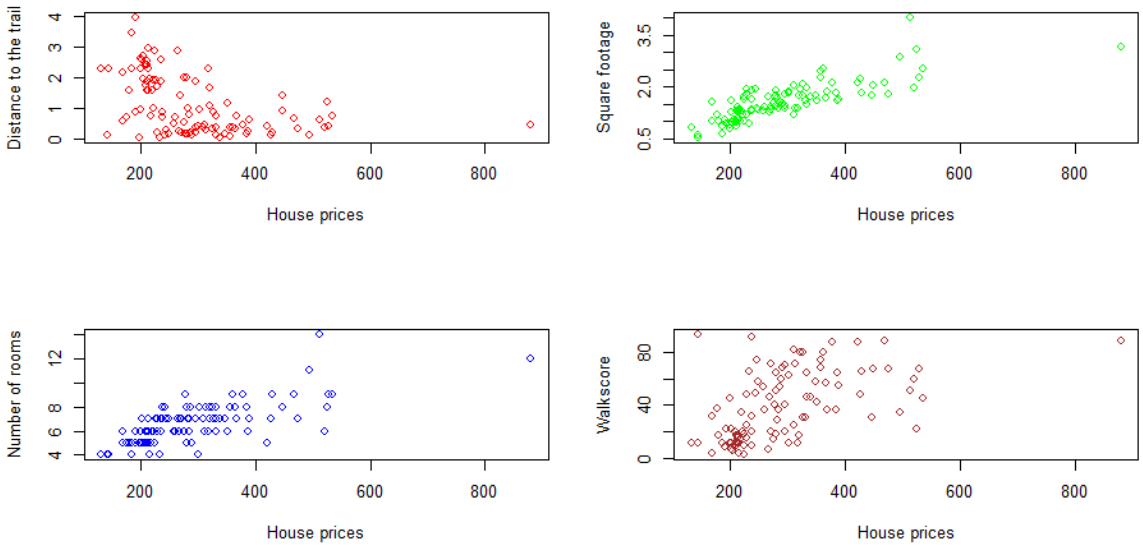

Figure 1: Scatter plots of the house prices and four covariates (distance to the trail, number of rooms, square footage and Walkscore)

In Figure 1, it can be observed that the house prices increase when variables such as number of rooms and square footage also increase, so there is a positive relationship between these variables and house prices. It can be seen that the relationship in square footage is stronger than the other variables such as distance and walkscore in house prices. Figure 1 also shows that the most expensive houses are closer to the trail and there is one instance where the price of the house is greater than 600 thousand when the prices of other houses range are between 132.13 and 534.

It can be seen in Figure 2 that houses which have more number of rooms, number of full baths, number of garage space and they are bigger are close to the trail.

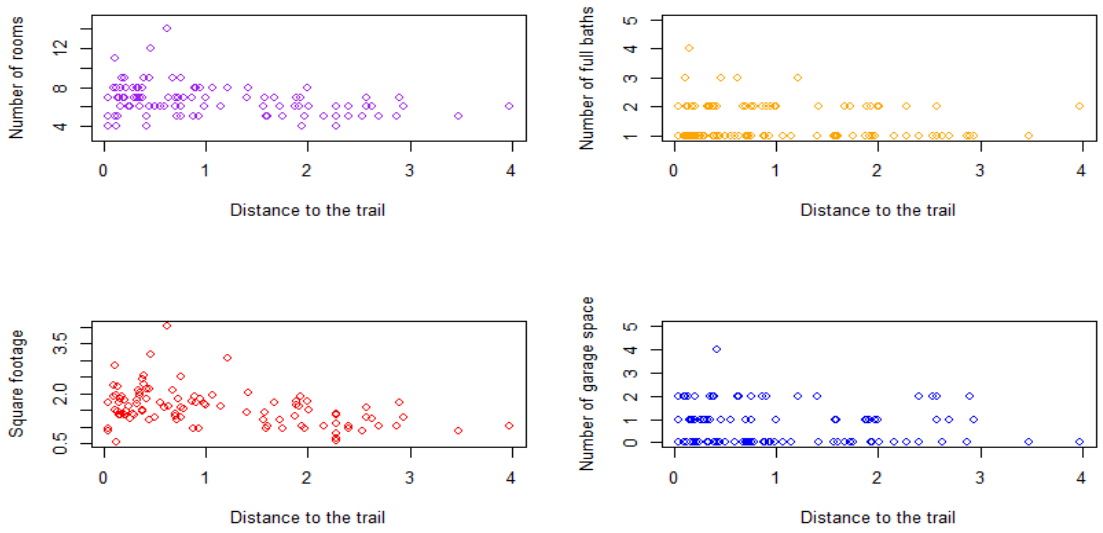

Figure 2: Scatter plots of the distance to the trail and four covariates (number of rooms, number of baths, square footage and number of garage spaces)

A summary statistic of the variables is given in the Table 2.

Name of variable	Mean	Range
price2014	293.09	132.135 - 879.328
acre	0.26	0.05 – 0.56
bikescore	57.27	18 -97
distance	1.11	0.03882576 - 3.97678
garage_spaces	0.76	0 - 4
no_full_baths	1.45	1 – 4
no_half_baths	0.22	0 - 1
no_rooms	6.62	4 - 14
squarefeet	1.57	0.524 -4.03
walkscore	38.88	2 - 94

Table 2. Summary statistic of the variables

Model fitting

Simple linear regression

A simple linear regression was used to evaluate the relationship between the distance to the trail and the house prices in 2014 using a critical value of p=0.05. The house prices were treated as a response variable and the distance as an explanatory variable. The form of the statistical model used is:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

In the statistical model, y is house prices (response variable), x is distance to the trail(explanatory variable). Normal distribution was assumed.

With $\epsilon \sim Normal\ (0, \sigma^2)$

```
model1=lm(price2014 ~ distance, data=hoPric)
summary(model1)

##
## Call:
## lm(formula = price2014 ~ distance, data = hoPric)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -202.74  -56.27  -15.91   31.53  551.45
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   352.16      15.18  23.192  < 2e-16 ***
## distance      -53.01      10.44  -5.079 1.72e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 99.55 on 102 degrees of freedom
## Multiple R-squared:  0.2018, Adjusted R-squared:  0.194
## F-statistic: 25.79 on 1 and 102 DF,  p-value: 1.72e-06
```

Figure 3 (a)

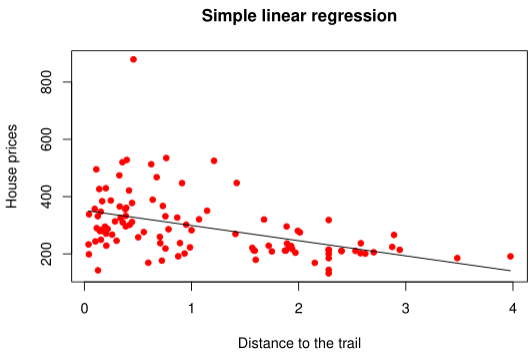

Figure 3 (b)

Figure 3. Summary of the model using simple linear regression and plotting the data superimposing the regression line

It can be seen in Figure 3, that the variable **distance** is highly significant (p<0.001). And it can be interpreted, therefore, that for every increase of a mile in distance to the trail, there is a corresponding decrease of 53.01 thousand dollars in the houses prices.

Multiple linear regression

Additionally, a multiple linear regression was carried out for this dataset including all explanatory variables (acre, bedgroup, bikescore, distance, garage_spaces, no_full_baths, no_half_baths, no_rooms, squarefeet, walkscore, and, zipF). The form of the statistical model used is:

$$y_i \sim \beta_0 + \sum_j \beta_j X_{ij} + \epsilon$$

In the statistical model, y is house prices (response variable), X_{ij} are the values explanatory variables. Normal distribution was assumed.

$$\text{With } \epsilon \sim Normal\ (0, \sigma^2)$$

```
#Multiple linear regression

reg1 <- lm(price2014 ~ acre + bedgroup + bikescore + distance + garage_spaces + no_full_baths + no_hal
          + factor(zip) , data = hoPric)
summary(reg1)

##
## Call:
## lm(formula = price2014 ~ acre + bedgroup + bikescore + distance +
##     garage_spaces + no_full_baths + no_half_baths + no_rooms +
##     squarefeet + walkscore + factor(zip), data = hoPric)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -163.646  -32.505   -0.563   26.937  283.909
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       85.9656    56.0872   1.533  0.12882
## acre             -37.6408    62.3184  -0.604  0.54734
## bedgroup3 beds    11.6855    20.0657   0.582  0.56176
## bedgroup4+ beds -10.9302    25.2252  -0.433  0.66582
## bikescore         -1.6153     0.9549  -1.692  0.09416 .
## distance          -7.9235    12.3622  -0.641  0.52317
## garage_spaces     15.7704     8.2838   1.904  0.06010 .
## no_full_baths     41.5999    13.0814   3.180  0.00201 **
## no_half_baths     16.7968    17.0457   0.985  0.32704
## no_rooms           9.0681     8.3055   1.092  0.27780
## squarefeet        86.9879    26.6944   3.259  0.00157 **
## walkscore          1.7077     0.6679   2.557  0.01222 *
## factor(zip)1062 -38.5635    17.6273  -2.188  0.03125 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 60.91 on 91 degrees of freedom
## Multiple R-squared:  0.7334, Adjusted R-squared:  0.6982
```

Figure 4. Summary of the model using multiple linear regression

It can be seen in the Figure 4 that the variables **no_full_baths** and **squarefeet** are significant ($p<0.01$) and they are followed by **walkscore** and **zip** ($p<0.05$). The other variables are not significant in this model. The R-squared of the variance indicates 73.34% of variance, so it could be assumed that each variable would explain around 6,11% of variance.

The estimated regression coefficient of **squarefeet** is 86.98 and it indicates that if all other independent variables are kept constant, then increase in one square feet of interior space in the house is associated with an increase of 86.98 thousand of dollars in the houses prices.

Additionally, the estimated regression coefficient of **no_full_bath** is 41.59 and it indicates that if all other independent variables are held constant, then increase in one full bathroom is associated with an increase of 41.59 thousand of dollars in the houses prices.

And the estimated regression coefficient of **walkscore** is 1.70 and it shows that if all other independent variables are kept constant, then increase in one score of walk friendliness is associated with an increase of 1.70 thousand of dollars in the houses prices.

However, the estimated regression coefficient of the **zip (1062)** is -38,56 and it indicates that houses ubicated in **zip 1062** is associated with an decrease of 38,56 thousand of dollars prices compared to houses located in **zip 1060**.

Backwards elimination

A backwards elimination was carried out using Anova() to obtain the least significant variable at each stage. In each stage the least significant variable was removed, using a critical value of p=0.05. This process was carried out 9 times (steps) until all terms were significant at the 5% level. In Table 3, it can be observed that a variable was removed in each stage.

Stage	Variable removed	*p*-value		Stage	Variable removed	*p*-value
1	- (model with all variables)			**6**	no_half_baths	0.233889
2	acre	0.547341		**7**	bikescore	0.136846
3	distance	0.517682		**8**	garage_spaces	0.144488
4	bedgroup	0.4227862		**9**	zipF	0.081973
5	no_rooms	0.313119				

Table 3. Variables removed in each stage using backwards elimination

The **distance** variable was removed in the third stage. This can be observed in Figure 5. The effect of the distance on the houses prices is non-significant, its p-value is 0.517682, so there is evidence of little effect of the distance to the trail on the house prices.

```
Anova Table (Type II tests)

Response: price2014
               Sum Sq Df F value   Pr(>F)
bedgroup         6928  2  0.9401 0.394319
bikescore       10505  1  2.8506 0.094725 .
distance         1554  1  0.4218 0.517682
garage_spaces   13023  1  3.5341 0.063284 .
no_full_baths   37618  1 10.2083 0.001916 **
no_half_baths    3286  1  0.8917 0.347480
no_rooms         3745  1  1.0162 0.316071
squarefeet      39070  1 10.6024 0.001582 **
walkscore       24788  1  6.7267 0.011051 *
zipF            28018  1  7.6031 0.007027 **
Residuals      339020 92
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 5. Anova table of the second model using backwards elimination

In the **final stage (stage 9)** that was carried out, there were three remaining variables: **no_full_baths**, **squarefeet** and **walkscore**. All of them were significant (p<0.05). So, there is

evidence that the effect of the **number of full baths, square footage of interior space** and **score of walk friendliness** varies the houses prices. The Anova table of the final model can be observed in Figure 6.

```
Anova Table (Type II tests)

Response: price2014
              Sum Sq  Df F value    Pr(>F)
no_full_baths  23125   1  6.0245 0.0158365 *
squarefeet    270045   1 70.3528 3.308e-13 ***
walkscore      51676   1 13.4628 0.0003924 ***
Residuals     383844 100
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 6. Anova table of the final model using backwards elimination

```
> summary(rega9)

Call:
lm(formula = price2014 ~ no_full_baths + squarefeet + walkscore,
    data = hoPric)

Residuals:
     Min       1Q   Median       3Q      Max
-136.635  -32.497   -1.176   28.674  297.549

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    23.0400    19.1630   1.202 0.232079
no_full_baths  29.3358    11.9519   2.454 0.015836 *
squarefeet    121.6575    14.5043   8.388 3.31e-13 ***
walkscore       0.9491     0.2587   3.669 0.000392 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 61.96 on 100 degrees of freedom
Multiple R-squared:  0.6969,    Adjusted R-squared:  0.6878
F-statistic: 76.65 on 3 and 100 DF,  p-value: < 2.2e-16
```

Figure 7. Summary of the final model using backwards elimination

It can be seen in the Figure 7 that the variables **squarefeet** and **walkscore** are highly significant ($p<0.001$) and they are followed by **no_full_baths** ($p<0.05$). The R-squared of the variance indicates 69.70% of variance, so it could be assumed that each variable would explain around 23,23% of variance.

The estimated regression coefficient of **squarefeet** is 121.6575 and it indicates that if all other independent variables are kept constant, then increase in one square feet of interior space in the house is associated with an increase of 121.66 thousand of dollars in the houses prices.

And the estimated regression coefficient of **walkscore** is 0.9491 and it shows that if all other independent variables are kept constant, then increase in one score of walk friendliness is associated with an increase of 0.95 thousand of dollars in the houses prices.

Additionally, the estimated regression coefficient of **no_full_bath** is 29.3358 and it indicates that if all other independent variables are held constant, then increase in one full bathroom is associated with an increase of is 29.34 thousand of dollars in the houses prices.

Model testing and diagnostics

Standardized residuals were used in the **final model (no_full_baths, squarefeet** and **walkscore)** to check whether the residuals follow a normal distribution. It can be observed in the Figure 8 that standardized residuals look normal. There is just one outlier far from the line for which the standardized residual value is greater than 4.

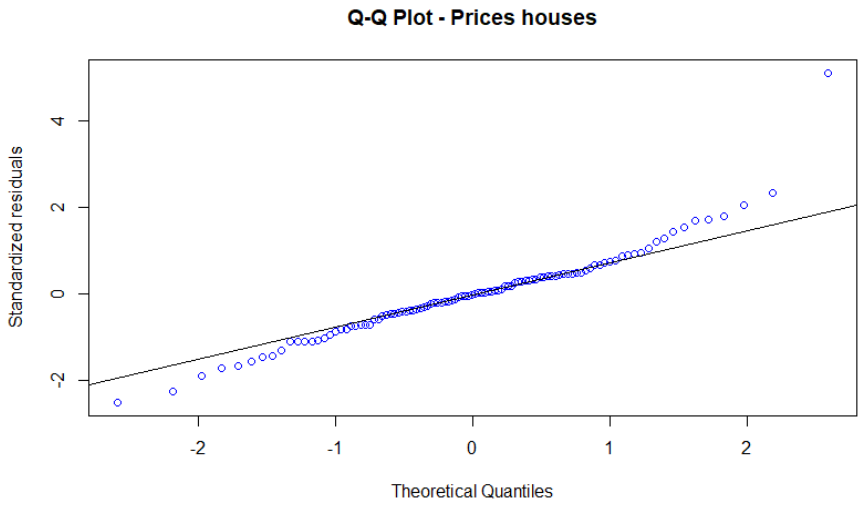

Figure 8: Q-Q plots of the standardized residuals using the final model

Another diagnostic test was carried out to identify aberrant and influential points, because it can be seen in Figure 8 that there is a point with large standardised residuals. In Figure 9, a plot of the standardized residuals can be observed against an index vector, so it can be identified in this plot that the house number 97 is an outlier, which has a residual of 5.08, but it is not important because it is just one outlier.

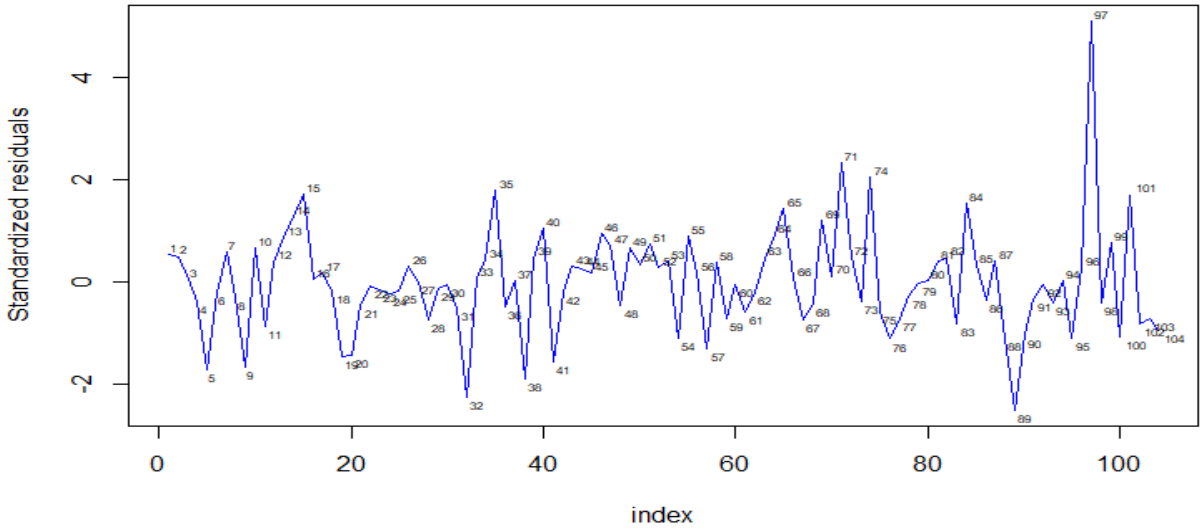

Figure 9: Plot of the standardized residuals against index vector

Additionally, it was used another diagnostic test called Box-Cox regression. In Figure 10, it can be observed the Box-Cox, that suggests doing the transformation of the response and explanatory.

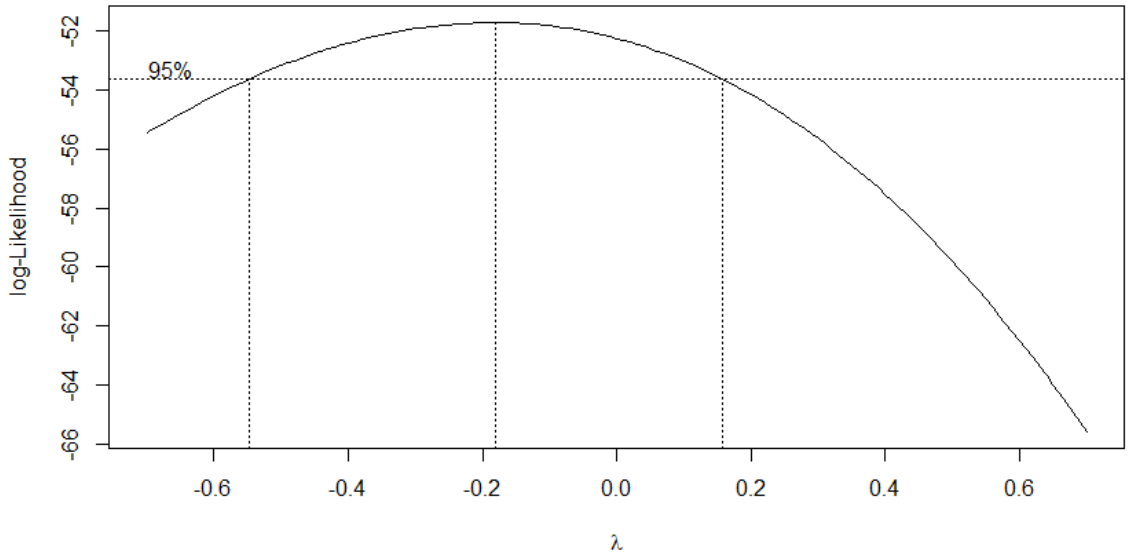

Figure 10: Box Cox

Conclusion

The analysis shows that the distance to the trail has little effect on the houses prices, so there is no evidence that rail trails in Northampton benefit the economy for people who have their house close to the rail trail. However, variables such as number of full baths, square footage of interior space, and score of walk friendliness have a significant effect on the houses prices. And, as it was found in the exploratory analysis, those houses with more rooms, full baths and garage space are close to the trail. So people could think that the distance to the trail has an impact to the houses prices, but in fact the final model does not show that.

Reference

Hartenian, E. and Horton, N. J. (2015) 'Journal of Statistics Education, Volume 23, Number 2, (2015)', 23(2).

Appendix

```
#Set directory
library(car)
library(calibrate)
library("ggplot2")

setwd("C:/Users/anaid/Box Sync/Lancaster Data Science
Master/Modules/Statistical methods/Generalised Linear
Models/Coursework/coursework for glm course-20171103")
hoPric=read.csv("houseprices.csv", header=T)
names(hoPric)
#Question 1.
plot(hoPric$distance,hoPric$price2014, pch=20,cex=1.3, col="red", main=
"Simple linear regression", ylab="House prices", xlab="Distance to the
trail") #forma bolas
#Fit a simple linear regression and to minimise the residual sum of
squares
model1=lm(price2014 ~ distance, data=hoPric)
summary(model1)
lines(hoPric$distance, fitted(model1))

#Question 2
#Multiple linear regression
hoPric$zipF <- factor(hoPric$zip)
#hoPric$zipF2 <- factor(c(1,2), lab = c("1060", "1062"))
reg1 <- lm(price2014 ~ acre + bedgroup + bikescore + distance +
garage_spaces
           + no_full_baths + no_half_baths + no_rooms + squarefeet +
walkscore
           + zipF , data = hoPric)
summary(reg1)


#Question 3
#Step #1
rega1 <- lm(price2014 ~ acre + bedgroup + bikescore + distance +
garage_spaces + no_full_baths + no_half_baths + no_rooms + squarefeet +
walkscore
           + zipF , data = hoPric)
Anova(rega1)#all variables
#Step #2 without acre
rega2 <- lm(price2014 ~ bedgroup + bikescore + distance + garage_spaces +
no_full_baths + no_half_baths + no_rooms + squarefeet + walkscore
            + zipF , data = hoPric)
Anova(rega2)#without acre

#Step #3 without distance
rega3 <- lm(price2014 ~ bedgroup + bikescore + garage_spaces +
no_full_baths + no_half_baths + no_rooms + squarefeet + walkscore
            + zipF , data = hoPric)
Anova(rega3)#without distance

#Step #4 without bedgroup
rega4 <- lm(price2014 ~ bikescore + garage_spaces + no_full_baths +
no_half_baths + no_rooms + squarefeet + walkscore
            + zipF , data = hoPric)
```

```
Anova(rega4)#without bedgroup

#Step #5 without no_rooms
rega5 <- lm(price2014 ~ bikescore + garage_spaces + no_full_baths +
no_half_baths + squarefeet + walkscore
            + zipF , data = hoPric)
Anova(rega5)#without no_rooms

#Step #6 without no_half_baths
rega6 <- lm(price2014 ~ bikescore + garage_spaces + no_full_baths +
squarefeet + walkscore
            + zipF , data = hoPric)
Anova(rega6)#without no_half_baths

#Step #7 without bikescore
rega7 <- lm(price2014 ~ garage_spaces + no_full_baths + squarefeet +
walkscore
            + zipF , data = hoPric)
Anova(rega7)#without bikescore

#Step #8 without garage_spaces
rega8 <- lm(price2014 ~ no_full_baths + squarefeet + walkscore
            + zipF , data = hoPric)
Anova(rega8)#without garage_spaces
#
#
#The last model
#Step #9 without zipF
rega9 <- lm(price2014 ~ no_full_baths + squarefeet + walkscore, data =
hoPric)
Anova(rega9)#without zipF
summary(rega9)

#Question 4
#Using standardized residuals
sresid <- rstandard(rega9)
qqnorm(sresid, ylab="Standardized residuals",
xlab="Theoretical Quantiles", main="Q-Q Plot - Prices houses", col =
"blue")
#textxy(sresid, sresid, labs = hoPric$housenum)
qqline(sresid)

hist(sresid)


#Checking the normality of the residuals
resid<- residuals(rega9)
qqnorm(resid, ylab="Sample Quantiles",
xlab="Theoretical Quantiles", main="Q-Q Plot - Prices houses")
qqline(resid)
hist(resid)

#-----Aberrant and influential points
#to identify the outliers
library(calibrate)
rstandard=rstandard(rega9)
index=seq(1,length(hoPric$price2014),1)
```

```
plot(index,rstandard, 'l', ylab="Standardized residuals", col = "blue")
textxy(index,rstandard,index)


####------BOX-COX reg1-------
library(MASS)
boxc<- boxcox(rega1)

sss= seq(-0.7,0.7,0.01)
boxcox(rega1,lambda=sss)

bcfit=boxcox(rega1,lambda=sss)
cbind(bcfit$x, bcfit$y)

logdno_full_baths=log(hoPric$no_full_baths)
logsquarefeet=log(hoPric$squarefeet)
logwalkscore=log(hoPric$walkscore)

boxcox(hoPric$price2014~logdno_full_baths+logsquarefeet+logwalkscore,
lambda=seq(-1,1,0.01))

boxcox(hoPric$price2014 ~ log(hoPric$acre) + log(hoPric$bikescore) +
log(hoPric$distance) + log(hoPric$garage_spaces)
+ log(hoPric$no_full_baths) + log(hoPric$no_half_baths) +
log(hoPric$no_rooms) + log(hoPric$squarefeet) + log(hoPric$walkscore))


##########---------DATA EXPLORATION------################
mean(hoPric$walkscore)
min(hoPric$walkscore)
max(hoPric$walkscore)

#Used in the report
par(mfrow=c(2,2))
plot(hoPric$price2014, hoPric$distance,  ylab="Distance to the trail",
xlab = "House prices", col="red")
cor(hoPric$price2014, hoPric$distance)
plot(hoPric$price2014, hoPric$squarefeet, ylab="Square footage", xlab =
"House prices", col="green")
cor(hoPric$price2014, hoPric$squarefeet)
plot(hoPric$price2014, hoPric$no_full_baths, ylab="Number of full baths",
xlab = "House prices", col="blue", ylim=c(1, 5))
cor(hoPric$price2014, hoPric$no_full_baths)
plot(hoPric$price2014, hoPric$walkscore, ylab="Walkscore", xlab = "House
prices", col="brown")
cor(hoPric$price2014, hoPric$walkscore)

par(mfrow=c(2,2))
plot(hoPric$distance, hoPric$no_rooms, ylab="Number of rooms", xlab =
"Distance to the trail", col="purple", ylim=c(3, 15))
cor(hoPric$no_rooms, hoPric$distance)

plot(hoPric$distance, hoPric$no_full_baths, ylab="Number of full baths",
xlab = "Distance to the trail", col="orange", ylim=c(1, 5))
cor(hoPric$no_full_baths, hoPric$distance)
```

```
plot(hoPric$distance, hoPric$squarefeet, ylab="Square footage", xlab =
"Distance to the trail", col="red")
cor(hoPric$squarefeet, hoPric$distance)

plot(hoPric$distance, hoPric$walkscore, ylab="Walkscore", xlab =
"Distance to the trail", col="blue", ylim=c(0, 100))
cor(hoPric$walkscore, hoPric$distance)

#Other
par(mfrow=c(2,2))
hist(hoPric$distance, xlab="Distance", main="Histogram of distance to the
trail")
hist(hoPric$squarefeet, xlab="Squarefeet", main="Histogram of square
footage of interior space")
hist(hoPric$price2014, xlab="Price", main="Histogram of house prices")
hist(hoPric$no_rooms, xlab="Number of room", main="Histogram of number of
room")

ggplot(hoPric, aes(x=hoPric$price2014,y=hoPric$walkscore ,colour=
hoPric$squarefeet)) +
  geom_point(position=position_jitter(height=0.03, width=0)) +
  xlab("Prices") + ylab("walkscore") + facet_grid(. ~ no_full_baths,
labeller = "label_value", as.table = TRUE) +
  scale_y_discrete(limits = c(0, 1))

plot(hoPric$price2014, hoPric$no_rooms, ylab="Number of rooms", xlab =
"House prices", col="blue")
cor(hoPric$price2014, hoPric$no_rooms)

plot(hoPric$distance, hoPric$garage_spaces, ylab="Number of garage
space", xlab = "Distance to the trail", col="blue", ylim=c(0, 5))
cor(hoPric$garage_spaces, hoPric$distance)
```