Final_Sampathirao_A

Anvita Sampathirao

8/15/2019

```r
#install.packages("Ecdat")
library(Ecdat)
```

```
## Warning: package 'Ecdat' was built under R version 3.6.1
```

```
## Loading required package: Ecfun
```

```
## Warning: package 'Ecfun' was built under R version 3.6.1
```

```
##
## Attaching package: 'Ecfun'
```

```
## The following object is masked from 'package:base':
##
##     sign
```

```
##
## Attaching package: 'Ecdat'
```

```
## The following object is masked from 'package:datasets':
##
##     Orange
```

```r
data(Housing)
Housing$driveway <- ifelse(Housing$driveway == "yes", 1, 0)
Housing$recroom <- ifelse(Housing$recroom == "yes", 1, 0)
Housing$fullbase <- ifelse(Housing$fullbase == "yes", 1, 0)
Housing$gashw <- ifelse(Housing$gashw == "yes", 1, 0)
Housing$airco <- ifelse(Housing$airco == "yes", 1, 0)
Housing$prefarea <- ifelse(Housing$prefarea == "yes", 1, 0)
library(stargazer)
```

```
##
## Please cite as:
```

```
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
```

```
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```r
library(ggplot2)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

```r
suppressMessages(attach(Housing))
library(psych)
```

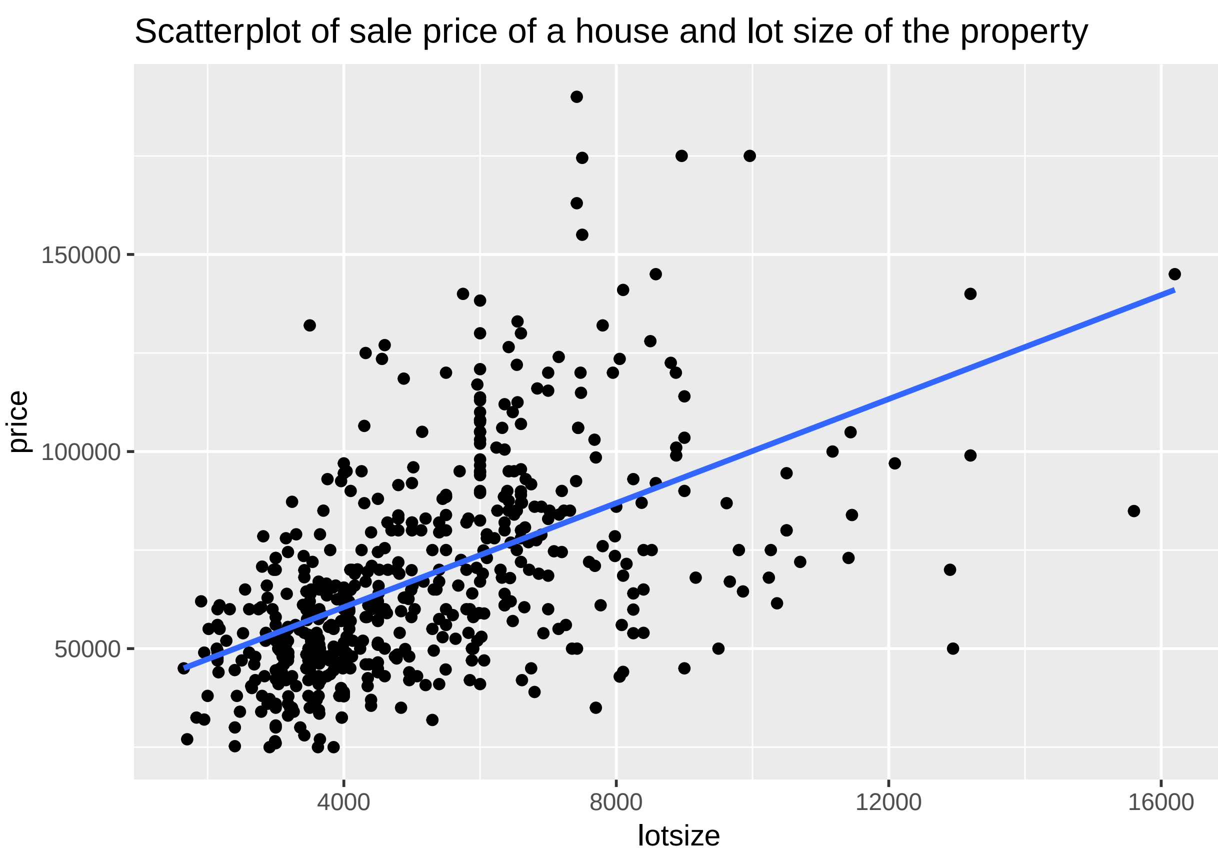
```
## Warning: package 'psych' was built under R version 3.6.1
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

#1

```r
g1 <- ggplot(data = Housing, aes(x = lotsize, y = price)) + geom_point() +
  ggtitle("Scatterplot of sale price of a house and lot size of the property")
g2 <- g1 + geom_smooth(method = "lm", formula = y~x, se = FALSE)
g2
```



The relationship between sale price of a house and the lot size of the property seems to be positively correlated, i.e., lower values of lotsize correspond to lower values of sale price of the house and higher values of the lotsize correspond to higher values of sale price of the house.

Also, Correlation does not imply causation. Therefore, we cannot conclude that lot size of the property causes the sale price of the house.

#2

```r
BV <- lm(price ~ lotsize)
stargazer(BV, type = "latex",
          header = FALSE,
          title = "Bivariate Regression Summary")
```

Table 1: Bivariate Regression Summary

	Dependent variable:
	price
lotsize	6.599***
	(0.446)
Constant	34,136.190***
	(2,491.064)
Observations	546
R^2	0.287
Adjusted R^2	0.286
Residual Std. Error	22,567.050 (df = 544)
F Statistic	219.056*** (df = 1; 544)
Note:	*p<0.1; **p<0.05; ***p<0.01

beta0: The average sale price of a house is 34,136.19 units if the lot size of the property is not taken into consideration, i.e. lotsize=0.

beta1: When the lot size of the property increases by a unit, on average, sale price of the house increases by 6.599 units.

R^2: 28.6% of the variation is sale price of the house can be explained by lot size of the property.

#3

```r
corData <- cor(Housing)
corData <- corData[, colnames(corData) %in% c("price", "lotsize")]
corData
```

```
          price      lotsize
```

price 1.00000000 0.535795672 lotsize 0.53579567 1.000000000 bedrooms 0.36644736 0.151851492 bathrms 0.51671925 0.193833484 stories 0.42119023 0.083674995 driveway 0.29716682 0.288777751 recroom 0.25495955 0.140327323 fullbase 0.18621767 0.047486731 gashw 0.09283654 -0.009200907 airco 0.45334656 0.221764888 garagepl 0.38330199 0.352871658 prefarea 0.32907432 0.234782230

```r
a<- corData[,1] * corData[,2]
sort(a, decreasing = TRUE)
```

```
    price       lotsize      garagepl         airco       bathrms
```

0.5357956724 0.5357956724 0.1352564095 0.1005363493 0.1001574933 driveway prefarea bedrooms recroom stories 0.0858151653 0.0772608029 0.0556455782 0.0357777908 0.0352430904 fullbase gashw 0.0088428685 -0.0008541804

```r
MV1 <- lm(price ~ lotsize + garagepl)
MV2 <- lm(price ~ lotsize + garagepl + airco)
MV3 <- lm(price ~ lotsize + garagepl + airco +
            bathrms)
MV4 <- lm(price ~ lotsize + garagepl + airco +
            bathrms + driveway)
MV5 <- lm(price ~ lotsize + garagepl + airco +
            bathrms + driveway + prefarea)
MV6 <- lm(price ~ lotsize + garagepl + airco +
            bathrms + driveway + prefarea + bedrooms)
MV7 <- lm(price ~ lotsize + garagepl + airco +
            bathrms + driveway + prefarea + bedrooms + recroom)
MV8 <- lm(price ~ lotsize + garagepl + airco +
            bathrms + driveway + prefarea + bedrooms +
            recroom + stories)
MV9 <- lm(price ~ lotsize + garagepl + airco +
            bathrms + driveway + prefarea + bedrooms +
            recroom + stories + fullbase)
MV10 <- lm(price ~ lotsize + garagepl + airco +
             bathrms + driveway + prefarea + bedrooms +
             recroom + stories + fullbase + gashw)

MVRegs1 <- list(MV1, MV2, MV3, MV4, MV5)
MVRegs2 <- list(MV6, MV7, MV8, MV9, MV10)

stargazer(MVRegs1, type = "latex",
          title = "(1/2)", intercept.bottom = FALSE, df = FALSE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sat, Aug 17, 2019 - 5:31:10 PM

```r
stargazer(MVRegs2, type = "latex",
          title = "(2/2)", intercept.bottom = FALSE, df = FALSE)
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Sat, Aug 17, 2019 - 5:31:10 PM

Looking at the results from the regression model, it seems that there is evidence that previously estimated parater in Q2 for lotsize was biased. After controlling for other factors, the estimated parameter for lotsize changes from 6.599 (in Q2) to 3.546 (in Model 10), which is approximately 53.7% reduction in magnitude of the estimated parameter.

Also, the R square value has improved from 28.6% (in Bivariate Model) to 66.6% (in Model 10). The variation is sale price of the house can be explained 66.6% by lot size of the property, number of garage

Table 2: (1/2)

	Dependent variable:				
	price				
	(1)	(2)	(3)	(4)	(5)
Constant	34,340.150***	32,934.040***	12,364.070***	5,781.983**	7,157.513**
	(2,417.072)	(2,222.856)	(2,551.472)	(2,895.059)	(2,809.440)
lotsize	5.635***	4.847***	4.287***	3.885***	3.496***
	(0.462)	(0.431)	(0.382)	(0.386)	(0.379)
garagepl	6,878.237***	5,946.030***	4,651.574***	4,168.203***	4,236.784***
	(1,163.740)	(1,072.100)	(949.676)	(938.945)	(908.370)
airco		19,268.380***	16,298.270***	15,993.610***	15,402.600***
		(1,902.763)	(1,692.126)	(1,663.580)	(1,612.137)
bathrms			19,671.880***	19,911.410***	19,782.560***
			(1,565.171)	(1,538.420)	(1,488.359)
driveway				10,220.790***	8,328.955***
				(2,249.915)	(2,197.989)
prefarea					10,911.740***
					(1,768.942)
Observations	546	546	546	546	546
R^2	0.330	0.437	0.564	0.580	0.608
Adjusted R^2	0.328	0.434	0.561	0.576	0.603
Residual Std. Error	21,894.510	20,095.930	17,696.170	17,383.490	16,816.170
F Statistic	133.827***	140.085***	174.983***	149.195***	139.201***
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 3: (2/2)

	Dependent variable:				
	price				
	(1)	(2)	(3)	(4)	(5)
Constant	−3,556.794	−3,049.459	−3,187.969	−4,115.163	−4,038.350
	(3,637.783)	(3,606.332)	(3,481.065)	(3,456.578)	(3,409.471)
lotsize	3.400***	3.316***	3.440***	3.536***	3.546***
	(0.373)	(0.370)	(0.358)	(0.355)	(0.350)
garagepl	4,009.048***	4,121.579***	4,559.991***	4,512.089***	4,244.829***
	(893.837)	(885.966)	(857.955)	(849.458)	(840.544)
airco	14,814.050***	14,349.720***	11,906.240***	11,693.250***	12,632.890***
	(1,589.164)	(1,580.061)	(1,572.891)	(1,558.328)	(1,555.021)
bathrms	17,433.230***	17,013.920***	15,175.730***	14,677.400***	14,335.560***
	(1,551.794)	(1,542.058)	(1,516.323)	(1,508.028)	(1,489.921)
driveway	9,048.952***	8,759.434***	6,840.416***	6,638.478***	6,687.779***
	(2,165.250)	(2,146.379)	(2,093.685)	(2,073.499)	(2,045.246)
prefarea	10,554.680***	9,854.542***	10,115.210***	9,007.644***	9,369.513***
	(1,739.668)	(1,735.582)	(1,675.765)	(1,689.676)	(1,669.091)
bedrooms	4,734.667***	4,671.830***	2,440.751**	1,919.541*	1,832.003*
	(1,047.159)	(1,037.370)	(1,061.108)	(1,061.250)	(1,047.000)
recroom		6,364.557***	6,846.258***	4,519.340**	4,511.284**
		(1,886.790)	(1,822.794)	(1,926.238)	(1,899.958)
stories			5,781.239***	6,678.946***	6,556.946***
			(909.938)	(937.576)	(925.290)
fullbase				5,558.221***	5,452.386***
				(1,609.766)	(1,588.024)
gashw					12,831.410***
					(3,217.597)
Observations	546	546	546	546	546
R^2	0.622	0.630	0.656	0.663	0.673
Adjusted R^2	0.617	0.624	0.650	0.657	0.666
Residual Std. Error	16,520.830	16,363.750	15,795.040	15,636.530	15,423.190
F Statistic	126.540***	114.281***	113.515***	105.437***	99.968***
Note:					*p<0.1; **p<0.05; ***p<0.01

places, availability of air conditioning, number of full bathrooms , availability of a driveway, location in the preferred neighborhood, number of bedrooms, availability of recreational rooms, number of stories, availability of a full finished basement, and availability of gas for hot water heating. Therefore, Multivariate Regression- Model 10 is the least biased and our best model for further analyses.

#4

```
vif <- function(reg, data){
  XvarNames <- names(reg$coefficients)
  XvarNames <- XvarNames[!(XvarNames %in% "(Intercept)")]
  k <- length(XvarNames)
  vifs <- rep(0, k)
  for(i in 1:k){
    indVars <- paste(XvarNames[!(XvarNames %in% XvarNames[i])], collapse = " + " )
    strFormula <- paste(XvarNames[i], indVars, sep = "~")
    auxReg <- lm(as.formula(strFormula), data = data)
    r2 <- summary(auxReg)$r.squared
    vifs[i] <- 1/(1-r2)
  }
  return(vifs)
}
multiTable <- data.frame(severe = logical(1),
moderate = logical(1))
multiTable$severe <- ifelse(any(vif(MV10, Housing) >= 10), TRUE, FALSE)
multiTable$moderate <- ifelse(any(vif(MV10, Housing) >= 5), TRUE, FALSE)
stargazer(multiTable, type = "latex", summary = FALSE, rownames = FALSE, header = FALSE,
title ="Multicollinearity Tests")
```

Table 4: Multicollinearity Tests

severe	moderate
FALSE	FALSE

Model 10 is the model with the largest amount of independent variables. Therefore, checking for multicollinearity for Model 10 produces the result that the model does not suffer from multicollinearity and we can trust the precision of the estimated standard errors and hypothesis tests. Thus, we can also conclude that there is no multicollinearity in the other models with fewer independent variables.

#5

```
#5.a- Adding a quadratic term
MV10Q <- lm(price ~ lotsize + I(lotsize^2) + garagepl +
              airco + bathrms + driveway + prefarea +
              bedrooms + recroom + stories + fullbase + gashw)
#5.b- Adding a cubic term
MV10C <- lm(price ~ lotsize + I(lotsize^2) + I(lotsize^3) +
              garagepl + airco + bathrms + driveway + prefarea +
              bedrooms + recroom + stories + fullbase + gashw)
NLRegs <- list(MV10, MV10Q, MV10C)
stargazer(NLRegs, type = "latex", header = FALSE, intercept.bottom = FALSE, df = FALSE)
```

$$y = price$$

Table 5:

	Dependent variable:		
	price		
	(1)	(2)	(3)
Constant	−4,038.350	−9,730.712**	−22,635.440***
	(3,409.471)	(4,470.917)	(7,564.538)
lotsize	3.546***	5.857***	12.904***
	(0.350)	(1.229)	(3.556)
I(lotsize^2)		−0.0002*	−0.001**
		(0.0001)	(0.001)
I(lotsize^3)			0.00000**
			(0.00000)
garagepl	4,244.829***	4,101.499***	4,299.041***
	(840.544)	(841.494)	(843.981)
airco	12,632.890***	12,184.820***	11,908.400***
	(1,555.021)	(1,567.636)	(1,568.052)
bathrms	14,335.560***	14,289.530***	14,156.340***
	(1,489.921)	(1,486.152)	(1,482.698)
driveway	6,687.779***	6,086.083***	5,422.837***
	(2,045.246)	(2,062.766)	(2,079.970)
prefarea	9,369.513***	9,328.949***	9,972.037***
	(1,669.091)	(1,664.790)	(1,687.142)
bedrooms	1,832.003*	1,888.165*	1,759.889*
	(1,047.000)	(1,044.614)	(1,043.014)
recroom	4,511.284**	3,852.820**	3,609.536*
	(1,899.958)	(1,924.436)	(1,921.683)
stories	6,556.946***	6,446.463***	6,571.614***
	(925.290)	(924.553)	(923.473)
fullbase	5,452.386***	5,555.420***	5,816.975***
	(1,588.024)	(1,584.681)	(1,584.417)
gashw	12,831.410***	12,884.020***	12,776.510***
	(3,217.597)	(3,209.171)	(3,199.218)
Observations	546	546	546
R^2	0.673	0.675	0.678
Adjusted R^2	0.666	0.668	0.670
Residual Std. Error	15,423.190	15,382.260	15,332.610
F Statistic	99.968***	92.446***	86.231***
Note:			*p<0.1; **p<0.05; ***p<0.01

$$x_1 = mean(lotsize)$$

$$x_{1new} = mean(lotsize) + 1 * stdev(lotsize)$$

$$\Delta x = x_{1new} - x_1 = stdev(lotsize)$$

$$\text{Best Model (BM)} : y = \beta_0 + \beta_1 * x_1 + \sum_i^k \beta_i * x_k + \epsilon$$

$$\text{BM.a.} : y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_1^2 + \sum_i^k \beta_i * x_k + \epsilon$$

$$\text{BM.b.} : price = \beta_0 + \beta_1 * x_1 + \beta_2 * x_1^2 + \beta_3 * x_1^3 + \sum_i^k \beta_i * x_k + \epsilon$$

$$\Delta y_{BM} = \beta_1 * \Delta x$$

$$\Delta y_{BM.a.} = (\beta_1 + 2\beta_2 x_1) * \Delta x$$

$$\Delta y_{BM.b.} = (\beta_1 + 2\beta_2 x_1 + 3\beta_3 x_1^2) * \Delta x$$

```r
betasBM <- as.numeric(MV10$coefficients)
betasBMa <- as.numeric(MV10Q$coefficients)
betasBMb <- as.numeric(MV10C$coefficients)
x_1 <- mean(lotsize)
deltax1 <- sd(lotsize)
deltayBM <- betasBM[2]*deltax1
deltayBMa <- (betasBMa[2] + (2*betasBMa[3]*x_1))* deltax1
deltayBMb <- (betasBMb[2] + (2*betasBMb[3]*x_1) + (3*betasBMb[4]*x_1^2))* deltax1
deltays <- list(deltayBM, deltayBMa, deltayBMb)
Models <- c("Best Model", "Quadratic Model", "Cubic Model")
deltays <- cbind(Models, deltays)
stargazer(deltays, type = "latex",title ="Results", header = FALSE)
```

Table 6: Results

Models	deltays
Best Model	7688.94772689347
Quadratic Model	8928.64549962263
Cubic Model	8536.14448788785

In the Quadratic Model, we see that the estimated parameter for the quadratic term is negative, therefore the change in sale price of the house increases as lot size of the property grows (when compared to the best model which is linear).

In the Cubic Model, we see that the estimated parameter for the cubic term is positive, therefore the changes in the sale price of the house increases as the lot size of the property grows (when compared to the best model which is linear).

$$H0 : \beta_{lotsize^3} = 0$$

$$H0 : \beta_{lotsize^3} \neq 0$$

When considering the Cubic Model, we see that the parameter of the cubic term is significative at alpha = 0.95, we reject the hypothesis of linearity and quadratic. However the value of the estimated parameter is negligible and a very small value and it does not imply that the parameter is important in practical terms.

#6

$$H0 : \beta_{lotsize*prefarea} = 0$$

$$H1 : \beta_{lotsize*prefarea} \neq 0$$

```r
MV11 <- lm(price ~ lotsize + I(lotsize*prefarea) + garagepl +
             airco + bathrms + driveway + prefarea + bedrooms +
             recroom + stories + fullbase + gashw)
Regs <- list(MV10, MV11)
stargazer(Regs, type = "latex", header = FALSE, intercept.bottom = FALSE, df = FALSE)
```

```r
anova(MV10, MV11)
```

Analysis of Variance Table

Model 1: price ~ lotsize + garagepl + airco + bathrms + driveway + prefarea + bedrooms + recroom + stories + fullbase + gashw Model 2: price ~ lotsize + I(lotsize * prefarea) + garagepl + airco + bathrms + driveway + prefarea + bedrooms + recroom + stories + fullbase + gashw Res.Df RSS Df Sum of Sq F Pr(>F)
1 534 1.2703e+11
2 533 1.2635e+11 1 675749814 2.8506 0.09192 . — Signif. codes: 0 '*' ***0.001*** '*' *0.01* " 0.05 '.' 0.1 ' ' 1

From the regression results, the interaction parameter is not statistically significative at alpha=0.95 but is significative at alpha=0.90.

From the F test, we note that p-value is greater than 0.05 which means that we fail to reject the null hypothesis that the estimated parameter for the interaction term between lotsize and prefarea is 0. Then, we can reject that the effect of lot size on price is moderated by prefarea.

#7

```r
Housing1 <- scale(Housing)
covHousing <- cov(Housing1)
fact <- fa(Housing1, nfactors = 2)
```

```
## Loading required namespace: GPArotation
```

```r
fact1 <- fact$loadings[,1]
fact1[order(fact1)]
```

```
##       gashw     stories    bedrooms     bathrms       airco    driveway
## 0.007361468 0.100426460 0.232825159 0.351343719 0.361199509 0.381280336
##    fullbase     recroom    garagepl    prefarea     lotsize       price
## 0.391277143 0.400448034 0.423111330 0.461010507 0.608631576 0.906643383
```

```r
fact2 <- fact$loadings[,2]
fact2[order(fact2)]
```

```
##    fullbase    prefarea     recroom     lotsize    driveway    garagepl
## -0.35485404 -0.23539311 -0.21106894 -0.13008623 -0.12774952 -0.05936481
##       gashw       price       airco     bathrms    bedrooms     stories
##  0.05304230  0.14348784  0.15386980  0.31140542  0.39043088  0.70766815
```

Table 7:

	Dependent variable:	
	price	
	(1)	(2)
Constant	$-4{,}038.350$	$-2{,}899.316$
	(3,409.471)	(3,469.795)
lotsize	3.546^{***}	3.182^{***}
	(0.350)	(0.411)
I(lotsize *prefarea)		1.179^{*}
		(0.699)
garagepl	$4{,}244.829^{***}$	$4{,}142.113^{***}$
	(840.544)	(841.294)
airco	$12{,}632.890^{***}$	$12{,}590.640^{***}$
	(1,555.021)	(1,552.535)
bathrms	$14{,}335.560^{***}$	$14{,}390.840^{***}$
	(1,489.921)	(1,487.706)
driveway	$6{,}687.779^{***}$	$7{,}220.913^{***}$
	(2,045.246)	(2,065.985)
prefarea	$9{,}369.513^{***}$	2,601.531
	(1,669.091)	(4,341.065)
bedrooms	$1{,}832.003^{*}$	$1{,}892.736^{*}$
	(1,047.000)	(1,045.809)
recroom	$4{,}511.284^{**}$	$4{,}664.056^{**}$
	(1,899.958)	(1,898.831)
stories	$6{,}556.946^{***}$	$6{,}587.293^{***}$
	(925.290)	(923.866)
fullbase	$5{,}452.386^{***}$	$5{,}266.002^{***}$
	(1,588.024)	(1,589.118)
gashw	$12{,}831.410^{***}$	$12{,}975.440^{***}$
	(3,217.597)	(3,213.169)
Observations	546	546
R^2	0.673	0.675
Adjusted R^2	0.666	0.668
Residual Std. Error	15,423.190	15,396.530
F Statistic	99.968^{***}	92.192^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01	

According to the first factor loading, a possible categorization for a person shortlisting houses in the city of windsor: High average on luxury items & amenities criteria such as full basement, recreational rooms, garageplace, preferred neighborhood, lot size of the property and sale price of the house (which corresponds to valuation of the house).

Low average on essentials criteria such as gas for heating and hot water, number of stories, bedrooms,bathrooms, air conditioning and driveway availability.

#8

```
set.seed(1)
kout2 <- kmeans(Housing1, centers = 2, nstart = 25)
centroids2 <- kout2$centers
topvars_centroid21 <- centroids2[1,order(centroids2[1,])]
topvars_centroid22 <- centroids2[2,order(centroids2[2,])]
tail(topvars_centroid21)
```

```
##  garagepl   recroom     airco   bathrms   lotsize     price
## 0.5001346 0.5007542 0.5860177 0.6462950 0.7658282 0.9911868
```

```
tail(topvars_centroid22)
```

```
##    prefarea    bedrooms     stories    driveway    fullbase       gashw
## -0.27739213 -0.27735429 -0.27400490 -0.19218679 -0.17777251 -0.04182529
```

Using two centers divided the data into two groups.

One with garage place, recreational room, airconditioning, bathrooms, lotsize and price as one category which can be interpreted as a luxury criteria for people in Windsor with a higher average.

Another one with preferred neighborhood, bedrooms, stories, driveway, full basement and gas for heating & hot water which can be interpreted as an essential criteria for people in Windsor with a lower average.

Yet there are variables (such as availability of a full basement) in the second group which may not reflect how an average individual shortlists houses. Similarly there are variables (such as price) in the first group which may be an essential criterion for shortlisting houses.

Cluster Analysis seems to be identifying personal preferences in essentials category which suggests there might be another category that might group overlapping factors in a third category.