Econ 104 Project 1 Group 12

Seungyeon Yoo, AJ Jeng, Youxue Xia, Kevin Carranza

2023-10-20

Contents

1. Introduction

```
library(AER)
```

(a) Loading the Dataset

```
library(AER)
data(HousePrices)
names(HousePrices) # Names of all the columns
```

```
##  [1] "price"      "lotsize"    "bedrooms"   "bathrooms"  "stories"
##  [6] "driveway"   "recreation" "fullbase"   "gasheat"    "aircon"
## [11] "garage"     "prefer"
```

```
# This will contain factor columns for plots
HousePrices1 <- HousePrices
# This will be converted to all-numeric-columns dataset for correlation matrix
HousePrices2 <- HousePrices
```

(b) Describing the question we are trying to answer

We are attempting to determine which of the variables in the HousePrices dataset influenced housing prices in Windsor, Canada in the late summer of 1987. To do so, we measure the effect of variables that are potentially related to the price a house was sold for. There are a total of 11 variables in question, starting with the lot size of the house. Next are the 4 variables that denote the number of bedrooms, bathrooms, stories, and garage spaces, respectively. Lastly are the 6 variables that each denote whether or not the property has a driveway, a recreational room, a full-finished basement, a gas heater for water, central air conditioning, or if the property is located in the preferred neighborhood.

2. Description of the Dataset – `HousePrices`

(a) Citation of of the Dataset

```r
# Citing AER package by which the HousePrices dataset was sourced
citation(package = "AER")
```

```
## To cite AER, please use:
##
##   Christian Kleiber and Achim Zeileis (2008). Applied Econometrics with
##   R. New York: Springer-Verlag. ISBN 978-0-387-77316-2. URL
##   https://CRAN.R-project.org/package=AER
##
## A BibTeX entry for LaTeX users is
##
##   @Book{,
##     title = {Applied Econometrics with {R}},
##     author = {Christian Kleiber and Achim Zeileis},
##     year = {2008},
##     publisher = {Springer-Verlag},
##     address = {New York},
##     note = {{ISBN} 978-0-387-77316-2},
##     url = {https://CRAN.R-project.org/package=AER},
##   }
```

(b) Summary of the Dataset

As described earlier in the introduction, these are more detailed explanations of the 11 variables that we are working with:

`lotsize` – Lot size of a property in square feet

`bedrooms` – Number of bedrooms

`bathrooms` – Number of full bathrooms

`stories` – Number of stories excluding basement.

`driveway` – Does the house have a driveway?

`recreation` – Does the house have a recreational room?

`fullbase` – Does the house have a full finished basement?

`gasheat` – Does the house use gas for hot water heating?

`aircon` – Is there central air conditioning?

`garage` – Number of garage places

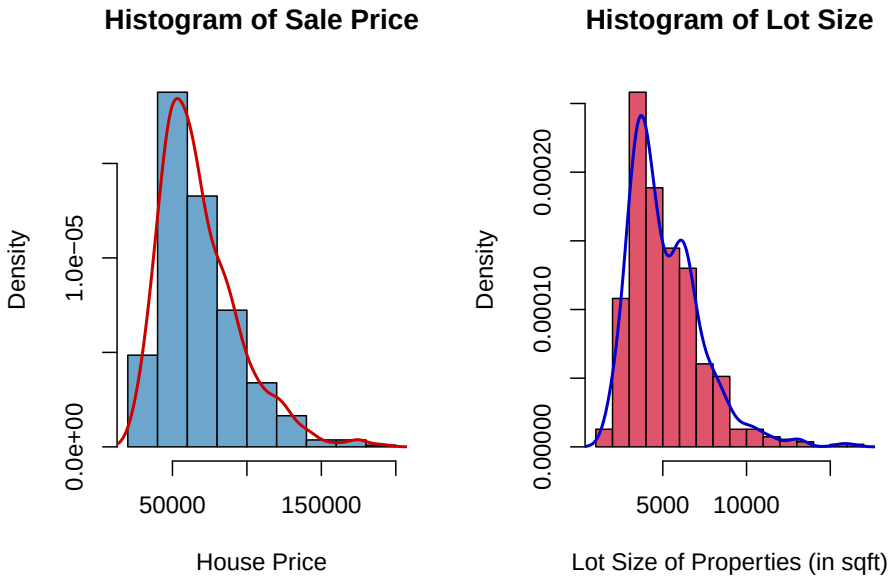
`prefer` – Is the house located in the preferred neighborhood of the city?

(c) Descriptive Analysis of the variables

```r
# Changing the data types of binary and discrete variables to logical and factors
HousePrices1$driveway <- ifelse(HousePrices$driveway == "yes", TRUE, FALSE)
HousePrices1$recreation <- ifelse(HousePrices$recreation == "yes", TRUE, FALSE)
HousePrices1$fullbase <- ifelse(HousePrices$fullbase == "yes", TRUE, FALSE)
HousePrices1$gasheat <- ifelse(HousePrices$gasheat == "yes", TRUE, FALSE)
HousePrices1$aircon <- ifelse(HousePrices$aircon == "yes", TRUE, FALSE)
HousePrices1$prefer <- ifelse(HousePrices$prefer == "yes", TRUE, FALSE)
HousePrices1$stories <- factor(HousePrices$stories)
HousePrices1$bedrooms <- factor(HousePrices$bedrooms)
HousePrices1$bathrooms <- factor(HousePrices$bathrooms)
HousePrices1$garage <- factor(HousePrices$garage)
```

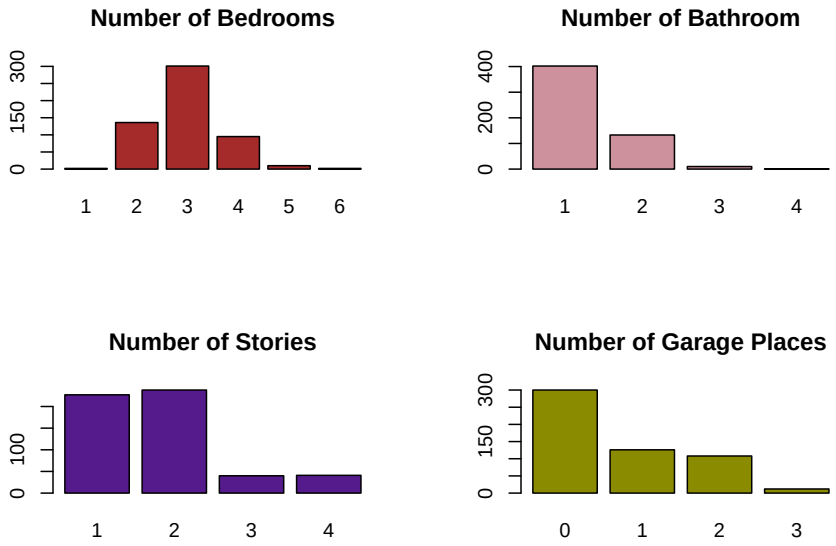
```r
# Histograms of continous variables and fitted distribution lines
par(mfrow = c(1,2))
hist(HousePrices1$price, prob = TRUE,
     xlab = "House Price",
     main = "Histogram of Sale Price",
     col = "skyblue3")
lines(density(HousePrices1$price), col = "red3",lwd = 2)

hist(HousePrices1$lotsize, prob = TRUE,
     xlab = "Lot Size of Properties (in sqft)",
     main = "Histogram of Lot Size",
     col = 2)
lines(density(HousePrices1$lotsize), col = "blue3", lwd = 2)
```

Histograms

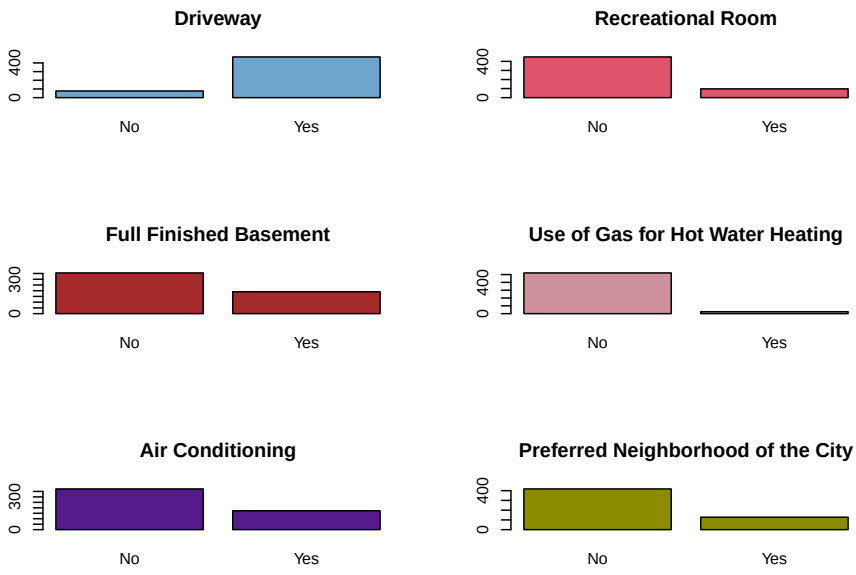


```r
# Creating barplots for the discrete variables
par(mfrow = c(2,2))
barplot(table(HousePrices1$bedrooms), main = "Number of Bedrooms", col = "brown")
barplot(table(HousePrices1$bathrooms), main = "Number of Bathroom", col = "pink3")
barplot(table(HousePrices1$stories), main = "Number of Stories", col = "purple4")
barplot(table(HousePrices1$garage), main = "Number of Garage Places", col = "yellow4")
```

Barplots




```r
# Creating barplots for the binary variables
par(mfrow = c(3,2))
barplot(table(HousePrices1$driveway), main = "Driveway",
        names.arg = c("No", "Yes"), col = "skyblue3")
barplot(table(HousePrices1$recreation), main = "Recreational Room",
        names.arg = c("No", "Yes"), col = 2)
barplot(table(HousePrices1$fullbase), main = "Full Finished Basement",
        names.arg = c("No", "Yes"), col = "brown")
barplot(table(HousePrices1$gasheat), main = "Use of Gas for Hot Water Heating",
        names.arg = c("No", "Yes"), col = "pink3")
barplot(table(HousePrices1$aircon), main = "Air Conditioning",
        names.arg = c("No", "Yes"), col = "purple4")
barplot(table(HousePrices1$prefer), main = "Preferred Neighborhood of the City",
        names.arg = c("No", "Yes"), col = "yellow4")
```



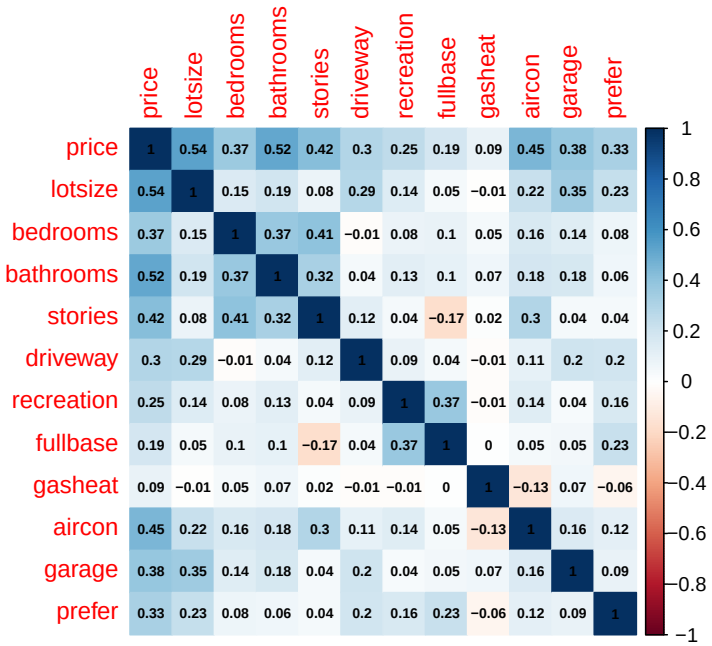
As we can see with the histograms and barplots, the `lotsize`, `bedrooms`, `bathrooms`, `stories`, and `garage` Variables all seem to be right-skewed, with a majority of observations being on the lower end of the spectrum.

For the 6 Boolean variables, only `driveway` has more observations with the feature than not. For the rest of the variables, the majority of observations do not have the feature.

```
# Changing all the values in the dataset to numerical values
HousePrices2 <- HousePrices

HousePrices2$driveway <- ifelse(HousePrices2$driveway == "yes", 1, 0)
HousePrices2$recreation <- ifelse(HousePrices2$recreation == "yes", 1,0)
HousePrices2$fullbase <- ifelse(HousePrices2$fullbase == "yes", 1,0)
HousePrices2$gasheat <- ifelse(HousePrices2$gasheat == "yes", 1,0)
HousePrices2$aircon <- ifelse(HousePrices2$aircon == "yes", 1,0)
HousePrices2$prefer <- ifelse(HousePrices2$prefer == "yes", 1,0)
HousePrices2$bedrooms <- as.numeric(HousePrices2$bedrooms)
HousePrices2$bathrooms <- as.numeric(HousePrices2$bathrooms)
HousePrices2$stories <- as.numeric(HousePrices2$stories)
HousePrices2$garage <- as.numeric(HousePrices2$garage)

cor_matrix <- cor(HousePrices2) # Correlation Matrix
```

```
# Creating a correlation matrix
library(corrplot)
corrplot(cor_matrix, method = "color", addCoef.col="black", number.cex = 0.6)
```

Correlation matrix



As seen in the correlation matrix, all of the variables have a positive correlation with the House Price, meaning that increasing the value of the variable or having the feature (for the Boolean variables) equates to a higher price.

```
# 5 number summary
summary(HousePrices1)
```

Five Number summary

```
##      price           lotsize      bedrooms bathrooms stories  driveway
##  Min.   : 25000   Min.   : 1650   1:  2    1:402     1:227   Mode :logical
##  1st Qu.: 49125   1st Qu.: 3600   2:136    2:133     2:238   FALSE:77
##  Median : 62000   Median : 4600   3:301    3: 10     3: 40   TRUE :469
##  Mean   : 68122   Mean   : 5150   4: 95    4:  1     4: 41
##  3rd Qu.: 82000   3rd Qu.: 6360   5: 10
##  Max.   :190000   Max.   :16200   6:  2
##  recreation       fullbase        gasheat          aircon        garage
##  Mode :logical   Mode :logical   Mode :logical   Mode :logical   0:300
##  FALSE:449       FALSE:355       FALSE:521       FALSE:373       1:126
##  TRUE :97        TRUE :191       TRUE :25        TRUE :173       2:108
##                                                                  3: 12
##
##
##    prefer
##  Mode :logical
##  FALSE:418
##  TRUE :128
##
##
##
```

The 5 number summary reflects what was discussed for the fitted histograms, with most observations falling on the lower half of the spectrum for their respective variables. With the exception of `driveway`, the other 5 Boolean variables are mostly FALSE, meaning the observations do not have the respective feature.

(d) Possible Violation of the Regression Assumptions

According to the data visualization and summary statistics, we can see a few possible violations of the Gauss Markov assumptions listed below:

- Independence – We currently assume independence due to lack of knowledge of the methods of data collection for this data set.
- Homoskedasticity – Intuitively, there may be heteroskedasticity present due to the presence of Boolean variables.
- Multicollinearity – We do not yet know if some variables may be perfectly collinear with one another. For example, the number of bedrooms and number of bathrooms, may be a 2 to 1 ratio.

3. Estimating a Multiple Linear Regression Model

```r
# Creating a model that includes all the possible variables
model_all <- lm(price ~ ., data = HousePrices2)
summary(model_all) # Summary statistics of model_all
```

```
##
## Call:
## lm(formula = price ~ ., data = HousePrices2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -41389  -9307   -591   7353  74875
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4038.3504  3409.4713  -1.184 0.236762
## lotsize         3.5463     0.3503  10.124  < 2e-16 ***
## bedrooms     1832.0035  1047.0002   1.750 0.080733 .
## bathrooms   14335.5585  1489.9209   9.622  < 2e-16 ***
## stories      6556.9457   925.2899   7.086 4.37e-12 ***
## driveway     6687.7789  2045.2458   3.270 0.001145 **
## recreation   4511.2838  1899.9577   2.374 0.017929 *
## fullbase     5452.3855  1588.0239   3.433 0.000642 ***
## gasheat     12831.4063  3217.5971   3.988 7.60e-05 ***
## aircon      12632.8904  1555.0211   8.124 3.15e-15 ***
## garage       4244.8290   840.5442   5.050 6.07e-07 ***
## prefer       9369.5132  1669.0907   5.614 3.19e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15420 on 534 degrees of freedom
## Multiple R-squared:  0.6731, Adjusted R-squared:  0.6664
## F-statistic: 99.97 on 11 and 534 DF,  p-value: < 2.2e-16
```

(a) Statistical and Economic Significance of the Estimates

According to the p-values of each coefficients, `bedrooms` appears to be **statistically not significant** at the 10% significance level. However, this may seem odd because the number of bedrooms in a house is intuitively a large determinant of the house price. Thus, it may be reasonable to suspect that the significance of the particular variable has multicollinearity with some other variable. ***To determine whether to remove `bedrooms` in the model, practical reasoning and further exploration of data may be required.***

Our model at this time suggests that all variables have an effect on the house price. Intuitively, this makes sense as additional features to a house would logically make the house price increase. Economically, this suggests that homeowners adding additional features to their house will make the selling price of their property increase.

(b) Overall Fit of the Model

```r
# RMSE -- Root Mean Squared Error
# RMSE = {[(1/n) * sigma(actual values - fitted values)^2]}^(1/2)
actual <- HousePrices$price # the actual values
fitted <- model_all$fitted.values # the fitted values
RMSE <- sqrt((sum((actual - fitted)^2)) / length(actual))
RMSE
```

```
## [1] 15252.76
```

```r
median(actual)
```

```
## [1] 62000
```

```r
RMSE / median(actual)
```

```
## [1] 0.2460122
```

This means that the model has an error of 15252.76 units (the same unit as the price variable). When we compare this with the median of the price from the data, we find that if we were to predict a house price with this model, our prediction may be off by about 24%.

4. Testing for Multicollinearity (VIF Test)

```
vif(lm(price ~ ., data = HousePrices2)) # Running the VIF test
```

```
##    lotsize   bedrooms  bathrooms    stories   driveway recreation   fullbase
##   1.321632   1.365633   1.282494   1.478584   1.163091   1.210501   1.316543
##    gasheat     aircon     garage     prefer
##   1.038246   1.201397   1.200839   1.147639
```

```
# all variables are in between 1-5 scale
```

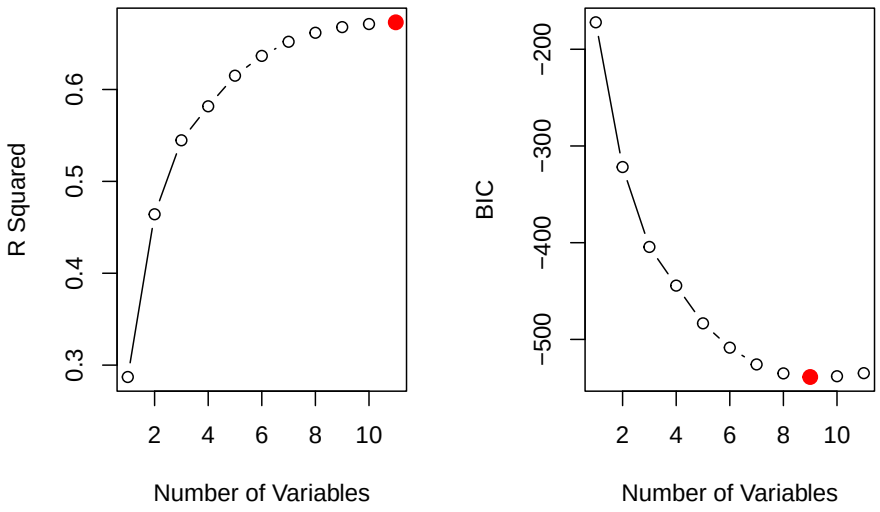
All independent variables are close to 1, which means that the provided variables are not highly correlated with each other. Thus, we do not have enough evidence to justify removing any variables based on collinearity at this point.

5. Determining which variables to keep or remove (AIC / BIC)

```r
library(leaps) # Loading leaps package

# Using regsubsets() function to see which variables are significant
# when limiting the number of variables to include
reg.sub <- regsubsets(price ~ ., data = HousePrices, nvmax = 11)
reg.sub.sum <- summary(reg.sub)
reg.sub.sum
```

```
## Subset selection object
## Call: regsubsets.formula(price ~ ., data = HousePrices, nvmax = 11)
## 11 Variables  (and intercept)
##               Forced in Forced out
## lotsize           FALSE      FALSE
## bedrooms          FALSE      FALSE
## bathrooms         FALSE      FALSE
## stories           FALSE      FALSE
## drivewayyes       FALSE      FALSE
## recreationyes     FALSE      FALSE
## fullbaseyes       FALSE      FALSE
## gasheatyes        FALSE      FALSE
## airconyes         FALSE      FALSE
## garage            FALSE      FALSE
## preferyes         FALSE      FALSE
## 1 subsets of each size up to 11
## Selection Algorithm: exhaustive
##          lotsize bedrooms bathrooms stories drivewayyes recreationyes
## 1  ( 1 ) "*"     " "      " "       " "     " "          " "
## 2  ( 1 ) "*"     " "      "*"       " "     " "          " "
## 3  ( 1 ) "*"     " "      "*"       " "     " "          " "
## 4  ( 1 ) "*"     " "      "*"       "*"     " "          " "
## 5  ( 1 ) "*"     " "      "*"       "*"     " "          " "
## 6  ( 1 ) "*"     " "      "*"       "*"     " "          " "
## 7  ( 1 ) "*"     " "      "*"       "*"     " "          " "
## 8  ( 1 ) "*"     " "      "*"       "*"     " "          " "
## 9  ( 1 ) "*"     " "      "*"       "*"     "*"          " "
## 10  ( 1 ) "*"     " "      "*"       "*"     "*"          "*"
## 11  ( 1 ) "*"     "*"      "*"       "*"     "*"          "*"
##          fullbaseyes gasheatyes airconyes garage preferyes
## 1  ( 1 ) " "         " "        " "       " "    " "
## 2  ( 1 ) " "         " "        " "       " "    " "
## 3  ( 1 ) " "         " "        "*"       " "    " "
## 4  ( 1 ) " "         " "        "*"       " "    " "
## 5  ( 1 ) " "         " "        "*"       " "    "*"
## 6  ( 1 ) " "         " "        "*"       "*"    "*"
## 7  ( 1 ) "*"         " "        "*"       "*"    "*"
## 8  ( 1 ) "*"         "*"        "*"       "*"    "*"
## 9  ( 1 ) "*"         "*"        "*"       "*"    "*"
## 10  ( 1 ) "*"         "*"        "*"       "*"    "*"
## 11  ( 1 ) "*"         "*"        "*"       "*"    "*"
```

```r
library(leaps)
reg.sub <- regsubsets(price ~ ., data = HousePrices, nvmax = 11)
reg.sub.sum <- summary(reg.sub)

par(mfrow = c(1, 2))

plot(reg.sub.sum$rsq, xlab = "Number of Variables", ylab = "R Squared", type = 'b')
best_rsq <- which.max(reg.sub.sum$rsq)
points(best_rsq, reg.sub.sum$rsq[best_rsq], col = "red", cex = 2, pch = 20)

plot(reg.sub.sum$bic, xlab = "Number of Variables", ylab = "BIC", type = 'b')
best_bic <- which.min(reg.sub.sum$bic)
points(best_bic, reg.sub.sum$bic[best_bic], col = "red", cex = 2, pch = 20)
```




```r
# Include everything
var11_mod <- lm(price ~ lotsize + bedrooms + bathrooms + stories +
                  driveway + recreation + fullbase + gasheat +
                  aircon + garage + prefer , data = HousePrices)
# exclude bedrooms
var10_mod <- lm(price ~ lotsize + bathrooms + stories + driveway +
                  recreation + fullbase + gasheat + aircon +
                  garage + prefer , data = HousePrices)
# 9 variables included model
var9_mod <- lm(price ~ lotsize + bathrooms + stories + driveway +
                 fullbase + gasheat + aircon + garage +
                 prefer , data = HousePrices)
# 8 variables included model
var8_mod <- lm(price ~ lotsize + bathrooms + stories + fullbase +
                 gasheat + aircon + garage + prefer , data = HousePrices)

aic_bic <- data.frame(AIC = AIC(var8_mod, var9_mod, var10_mod, var11_mod)[,2],
                      BIC = BIC(var8_mod, var9_mod, var10_mod, var11_mod)[,2])
```

```r
rownames(aic_bic) <- c("var8_mod", "var9_mod", "var10_mod", "var11_mod")
aic_bic
```

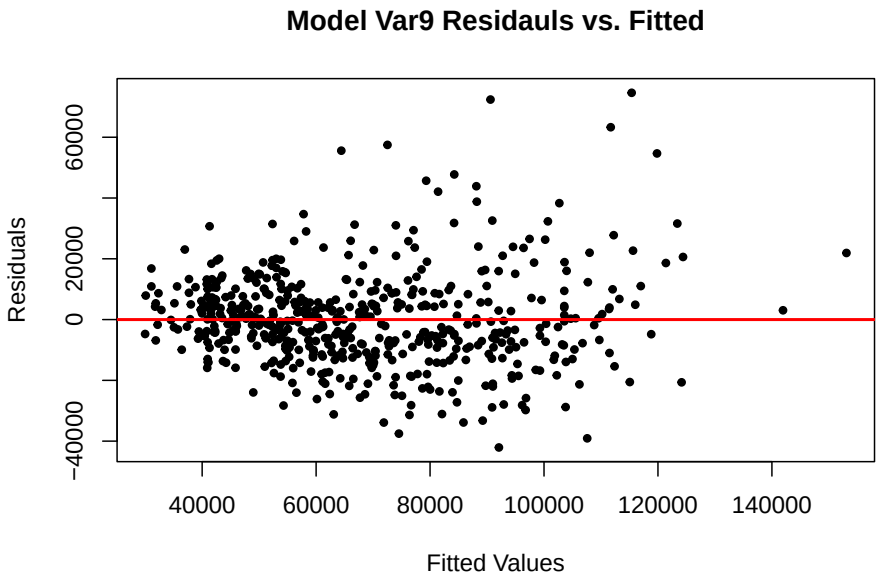
```
##                AIC      BIC
## var8_mod  12106.84 12149.87
## var9_mod  12098.84 12146.17
## var10_mod 12095.31 12146.94
## var11_mod 12094.19 12150.12
```

Using the regsubsets function, we can figure out which of the variables are best for any given number of variables. For example, if we want to limit our model to 10 variables (of the 11), the regsubsets function tells us that we can exclude the `bedrooms` variable. From the plots comparing the R^2 and BIC of the various models, we can see that the two plots disagree on the best number of variables. As such, we decided to calculate the AIC and BIC values for the models that included the 8 best variables to the model that includes all 11 to further investigate.

By AIC metrics, var11_mod has the lowest value. However, by BIC, var9_mod has the lowest value. Due to the higher restrictions in the BIC criteria, we have decided that this is sufficient evidence to use the var9_mod as our new model. This removes the following variables: `bedrooms`, `recreation`, and `driveway`.

6. Residuals vs. Fitted Values

```
# Model with 9 variables (bedrooms and recreation variables removed)
plot(var9_mod$fitted.values, var9_mod$residuals,
     main = "Model Var9 Residauls vs. Fitted",
     xlab = "Fitted Values",
     ylab = "Residuals",
     pch = 20)
abline(h = 0, col = "red", lwd = 2)
```



We can see clear signs that heteroskedasticity may be present based on the variance in the later half of the spectrum. We will need to do further testing to fix this.

7. RESET test on the model from (5)

```
library(lmtest)
resettest(var9_mod) # Reject --- we might be missing some higher term variables
```

```
##
##  RESET test
##
## data:  var9_mod
## RESET = 11.2, df1 = 2, df2 = 534, p-value = 1.719e-05
```

Based on the p-value of 1.719e-05, we reject the null and can conclude that adding higher powers or interaction terms to our model may improve it.

8. Testing for Heteroskedasticity

As seen in (6), there is grounds for heteroskedasticity based on visual inspection of the residuals.

```
bptest(var9_mod) # Reject H0 -- Heteroskedasticity present
```

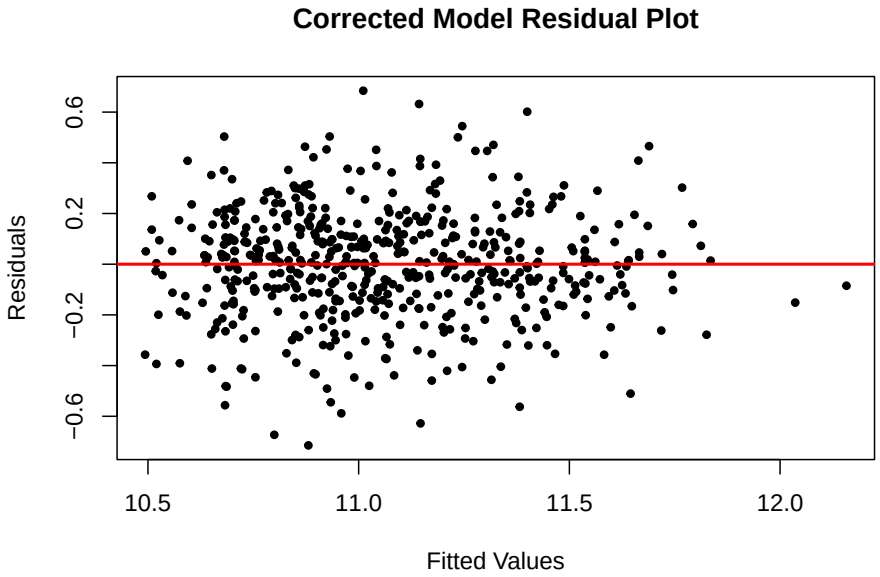
BP Test

```
##
##  studentized Breusch-Pagan test
##
## data:  var9_mod
## BP = 57.53, df = 9, p-value = 4e-09
```

Based on the BP test, we conclude that heteroskedasticity is in fact present.

```
# Taking log() on the dependent variable
corrected_mod1 <- lm(log(price) ~ lotsize + bathrooms + stories + driveway + fullbase +
                       gasheat + aircon + garage + prefer , data = HousePrices)
plot(corrected_mod1$fitted.values, corrected_mod1$residuals, pch = 20,
     xlab = "Fitted Values", ylab = "Residuals",
     main = "Corrected Model Residual Plot")
abline(h = 0, col = "red", lwd = 2)
```

Correcting the Model



```r
# BP Test for corrected model
bptest(corrected_mod1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  corrected_mod1
## BP = 18.621, df = 9, p-value = 0.02862
```

In order to fix for heteroskedasticity, we log the dependent variable, which upon visual inspection, is much cleaner. After running the BP test on the corrected model, we have insufficient evidence to conclude that heteroskedasticity is still present.

9. Adding Interaction Terms / Higher Power Terms

```
# Adding the higher terms
log_lin <- lm(log(price) ~ lotsize + bathrooms + stories + driveway +
                fullbase + gasheat + aircon + garage + prefer , data = HousePrices)
log_log <- lm(log(price) ~ log(lotsize) + bathrooms + stories +
                driveway + fullbase + gasheat + aircon + garage +
                prefer, data = HousePrices)


aci_bic <- data.frame(ACI = AIC(log_lin, log_log)[,2],
                      BIC = BIC(log_lin, log_log)[,2])
rownames(aci_bic) <- c("log_lin", "log_log")
aci_bic
```

```
##                ACI       BIC
## log_lin -112.5899 -65.26111
## log_log -132.1830 -84.85422
```

Based on the RESET test in (7), the model can benefit from changing the interaction terms. Intuitively, the square footage of a house/property does not significantly increase in price when a house has 1,000,000 ft^2 vs 1,000,001 ft^2. Knowing this, we decided to log the `lotsize` variable.

Next we tested the AIC and BIC values to compare the modified model with the original. We see that for both AIC and BIC, the modified model (log_log), has a smaller value.

```
# Running RESET test to see if there is more to be added
resettest(log_log, data = HousePrices)
```

```
##
##  RESET test
##
## data:  log_log
## RESET = 0.10956, df1 = 2, df2 = 534, p-value = 0.8963
```

Lastly, we quickly ran a RESET test to see if we needed to add any additional terms. Based on the results, we can conclude that the model does not need anymore interaction terms or higher power terms.

10. Conclusion

To conclude, we have found that not all of the variables in the data set are statistically significant. In the course of our testing, we removed `bedrooms`, `recreation`, and `driveway`. Our model was further improved by log-transforming the dependent variable and the `lotsize` variable to correct our model for heteroskedasticity and improve the performance of the model, respectively. One interesting finding was that the `bedrooms` variable was removed. Despite our pre-conceived importance of bedrooms to housing price, the variable was removed, much to our surprise. One possible explanation is that the number of bedrooms is less as important as the amount of space in each bedroom. For example, a house with four 100 ft^2 bedrooms may be similar in price to a house with two 200 ft^2 bedrooms. As such, the number of bedrooms does not necessarily mean more space/land for a house nor does it mean more value. We concluded that our model is a proper predictor of housing price in Windsor, Canada in 1987, that has good fit. With more data from other cities, perhaps this model can be expanded to predict housing prices more broadly like in developed nations with moderate climates in the late 1980's.