CASE WESTERN RESERVE UNIVERSITY

College of Arts and Sciences
Department of Mathematics, Applied Mathematics, and Statistics

---

MODELING PREDICTED UNITED STATES HOUSING PRICE :

Multiple Linear Regression Analysis

Author
Carl Conti

Advisors
Paula FitzGibbon
Joon Woo Bae

27 April 2020

INTRODUCTION

The United States Census Bureau collects housing data on approximately 1% of every American home (single-family, multi-family, group homes, apartments, etc.) each year. These statistics are published in the annual American Community Survey (ACS) which lists information on both the building itself as well as the people living within. Each year's dataset contains over 300,000,000 data points covering an exhaustive list of variables.

The variables to be used in this study will include: state where the property is listed, dollar value of the property, lot size in acres, dollar sales of agricultural products from the land, number of bedrooms, number of bathrooms, number of rooms, dollar cost of electricity, dollar cost of gas, dollar cost of fuel (non-electricity or gas), dollar cost of water, type of heating, vehicles available, year built, complete kitchen facilities (Y/N), complete plumbing facilities (Y/N), years living in the house, and dollar cost in yearly property taxes. A model will be built to accurately predict the dollar value of each property.

DATA PREPARATION

Although there was a large amount of data available, not all of it could be used. Of the 213 descriptive variables provided by the Census, a selection of 17 were chosen that best captured the overall status of the building and its property. Next, the data was cleansed of any observations that did not have a complete set of all required variables as all would be needed in order to develop the most complete model. Finally, under the law of large numbers it is understood that as the sample size increases, it will more and more closely resemble the total population. Because of this it is unnecessary to include all 200,000 observations. Instead a sample size of 2,500 was used as testing data, on which the complete dataset would be tested later. This size was chosen as to provide a 95% confidence interval and 2% margin of error on the larger sampling population.

EXPLORATORY DATA ANALYSIS

This section is intended to show us how certain regression models look and tell us if any transformation must be made. Looking at Figure 1, the scatterplot matrix of the quantitative variables, we can see that there is linear relationship between many of the variables. However, there is a slight upward curve on a few of the relationships which indicates that an exponential transformation may be the best course. A boxcox transformation would be the best solution to this problem.

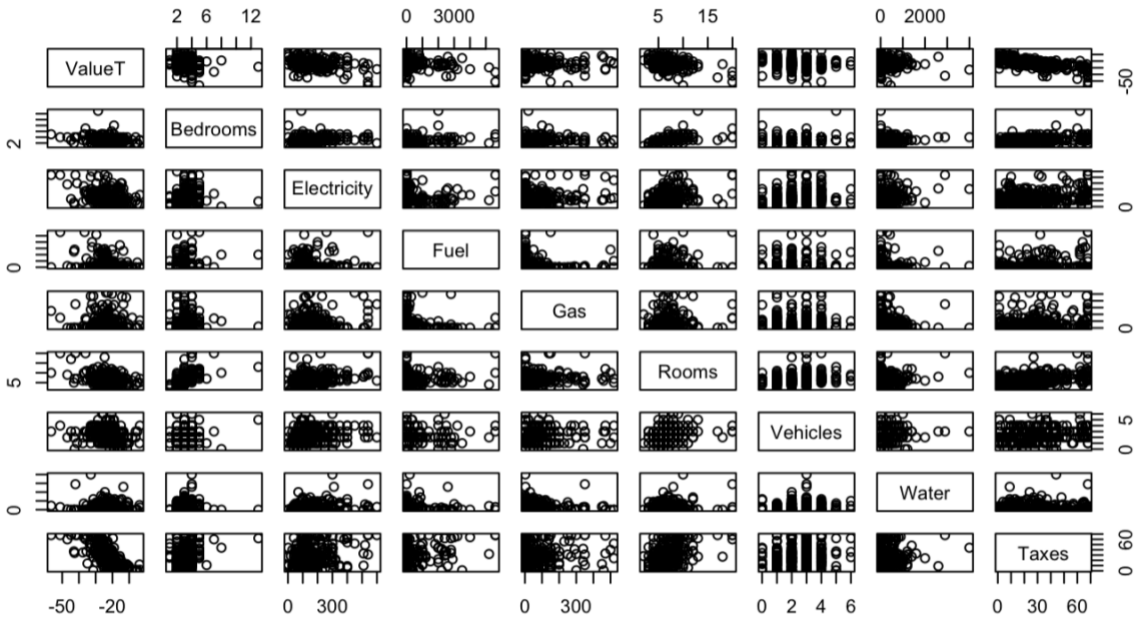

Figure 2: Scatterplot matric of all quantitative variables in the data set

From the diagnostic plots show in Figure 2, we can see that not all of the conditions for normality. The residual plot shows an upwards skew and the normal probability plot is skewed upwards in the same manner as in Figure 1. The information indicates that a boxcox transformation is the correct path to pursue.

The boxcox transformation on the full sample set yields a lambda of 0.2626. After applying the transformation, the standard diagnostic plots show a good fit, seen below in Figure 3. The new residual plot shows a consistent variance while the normal probability plot follows a nearly linear path.

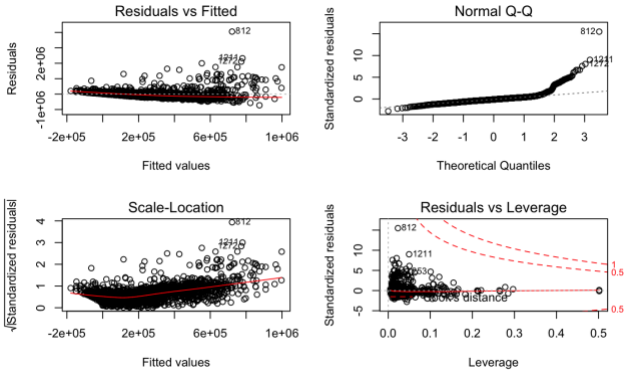

Figure 2: Standard Diagnostic Plots for un-transformed data

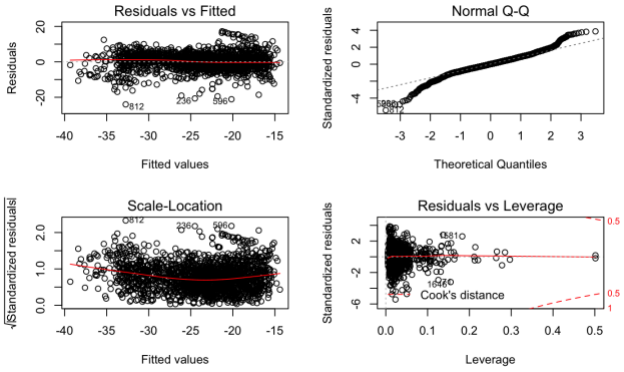

Figure 3: Standard Diagnostic Plots for boxcox-transformed data

The boxcox transformation on the full sample set yields a lambda of 0.2626. After applying the transformation, the standard diagnostic plots show a good fit, seen above in Figure 3. The new residual plot shows a consistent variance while the normal probability plot follows a nearly linear path.

The next step is to determine the which variables to include in the final product. Figure 4 shows the scatterplot matrix of the new boxcox transformed variables. It is clear to see that there are several variables with linear relationships on the transformed Value factor. However, there appears to be little to no relationship between cost of water bill and transformed dollar value. Conversely, it appears as though number of vehicles available does not have a significant relationship with any other variables. There may also be collinearity between electricity and gas costs compared to fuel as they are all slightly different measures of the same variable. These are all possible changes to the model that may need to be implemented later.

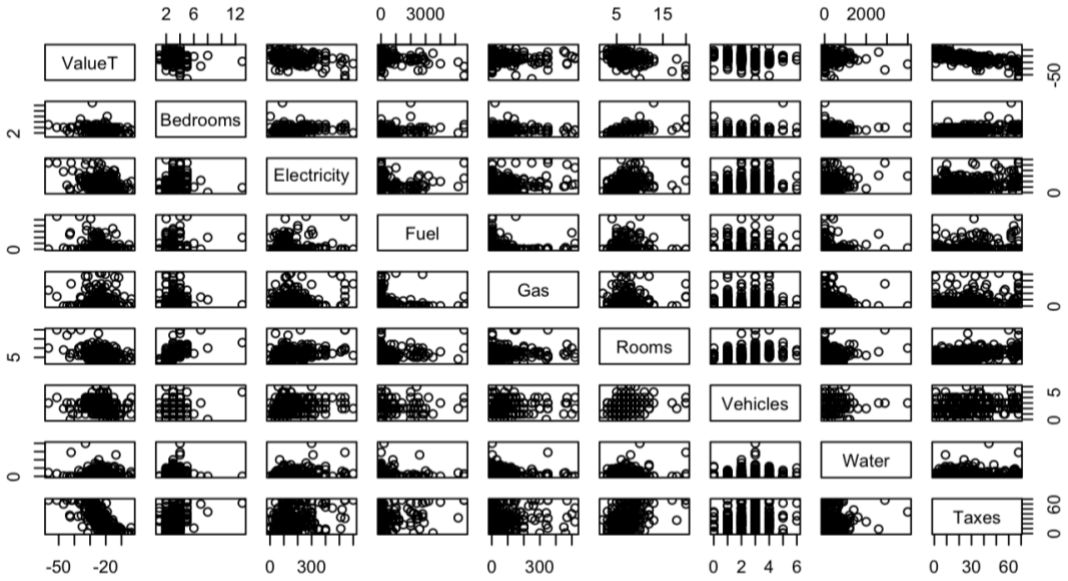

Figure 4: Scatterplot Matric of transformed quantitative variables

Figure 5 below shows the strength of the variables Electricity, Fuel, Gas, Vehicles and Water. This is a way to measure the added value of each factor to the overall model. We can use this figure to determine which of those listed to keep moving forward. In terms of collinearity, it appears that Fuel has a limited effect on the model compared to rest of the heating cost factors, so it will be removed moving forward. The number of vehicles available did not show that it added enough to the model, so it will be removed from the model as well.

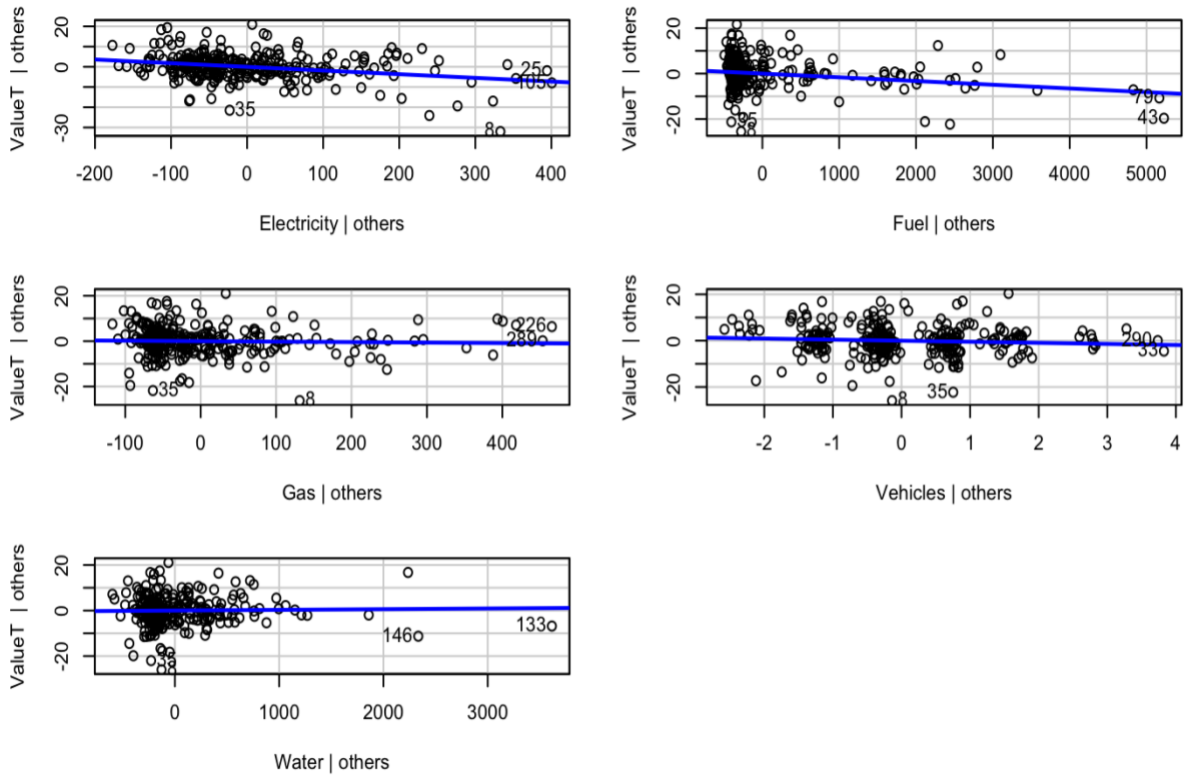

Figure 5: Add-Variable plot comparing fuel costs and vehicles

The qualitative variables must be analyzed with equal scrutiny to understand their significance in the overall model. To do this a boxplot for each variable is created to compare them to the boxcox transformed price, shown in Figure 6. The boxplot comparing transformed value to lot size as well as plumbing and heating type show that there is little change to the transformed value of a home when it is on a large or small plot, type of plumbing or the type of heating in the home, respectively. This tells us that these three variables can be removed from the model as dummy variables. All other qualitative variables showed enough variance to be included.

There is now a consensus on which variables and transformations are necessary to create an accurate model on the testing data. The factors to be used for predicting home value are sales of agricultural products, number of bedrooms, cost of electricity, cost of gas, number of rooms, cost of water, year built, years living in the home, and property taxes.

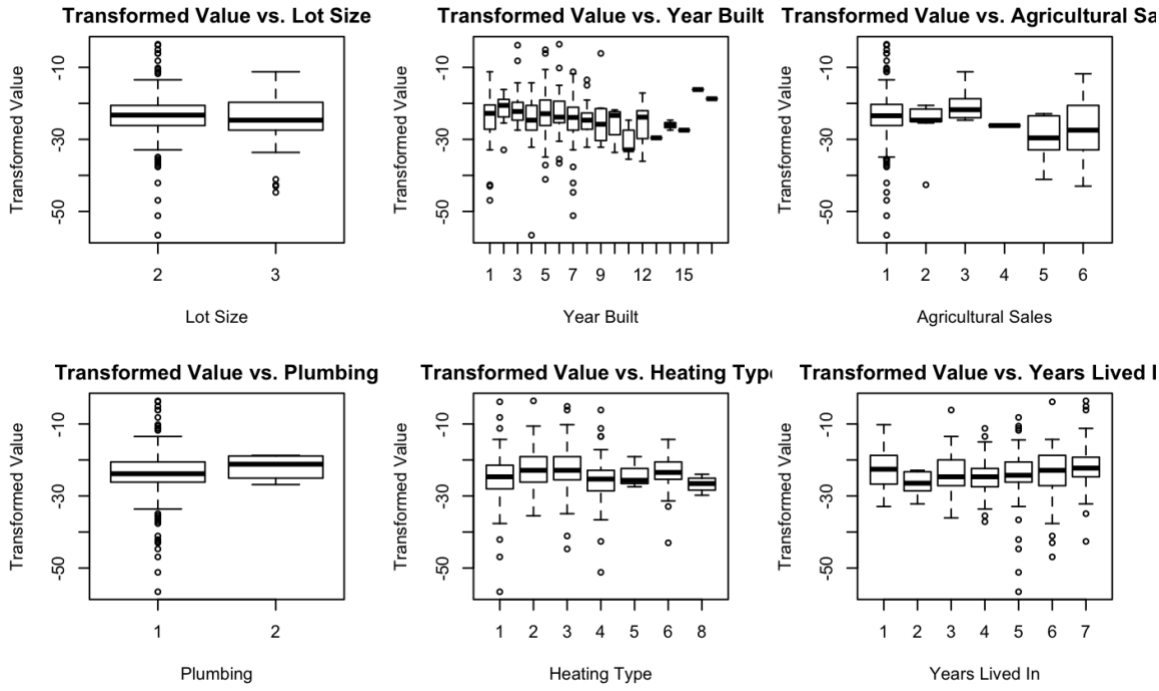

Figure 6: Boxplot of each qualitative variable compared to transformed price

MODEL BUILDING

The model has now been transformed to meet conditions with collinearity and dummy variables have been removed. It is important at this stage to continue to remove explanatory variables. Figure 7 shows the R_2 values, BIC values, and CP/AIC values for all the remaining factors. These plots show which combinations of variables best fits to the data. The BIC and CP/AIC plots are very similar to each other in terms of that variables each prefers in their respective best models. Clearly the weakest factor in each is Years Lived In which never appears in any model, followed closely by Water Costs and Bedrooms. The strongest factors appear to be Rooms, Electricity and Taxes which are included in nearly every top combination. The R_2 plot shows the same summary as the other two by including Electricity, Rooms and Taxes in nearly all models and disregarding the impact of Years Lived In, Water and Bedrooms. The only differences between the three are the inclusion of other variables in the intermediate-strength combination sets.

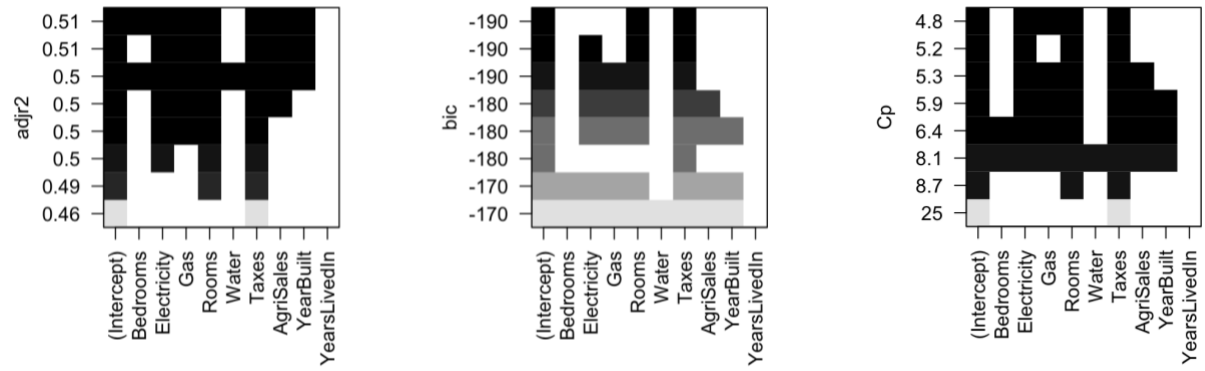

Figure 7: Ordered possible subsets of variables based on AIC, BIC, and R_2

The five best models for each method are as follows:

ADJR:

1. Bedrooms, Electricity, Gas, Rooms, Taxes, Agricultural Sales, Year Built
2. Electricity, Gas, Rooms, Taxes, Agricultural Sales, Year Built
3. Bedrooms, Electricity, Gas, Rooms, Water, Taxes, Agricultural Sales, Year Built
4. Electricity, Gas, Rooms, Taxes, Agricultural Sales
5. Electricity, Gas, Rooms, Taxes

BIC

1. Rooms, Taxes
2. Electricity, Rooms, Taxes
3. Electricity, Gas, Rooms, Taxes
4. Electricity, Gas, Rooms, Taxes, Agricultural Sales
5. Electricity, Gas, Rooms, Taxes, Agricultural Sales, Year Built

CP/AIC

1. Electricity, Gas, Rooms, Taxes
2. Electricity, Rooms, Taxes
3. Electricity, Gas, Rooms, Taxes, Agricultural Sales
4. Electricity, Gas, Rooms, Taxes, Agricultural Sales, Year Built
5. Bedrooms, Electricity, Gas, Rooms, Taxes, Agricultural Sales, Year Built

From the stepwise regression of both directions we attain a model that includes Electricity, Gas Rooms, Taxes, Agricultural Sales and Year built. This is second best model in ADJR, fourth best model in BIC and fourth best model in CP/AIC. Two models based on the inclusion of these factors will be used, one which includes Year Built and one which does not.

MODEL DIAGNOSTICS

In the summaries of Model 1 and Model 2 it turns out that only the coefficients of Rooms, Taxes and Electricity are statistically significant at the 95% confidence level, with Gas being very close. Agricultural Sales were not close with a p-value 20%. We will now refine the model once again and no longer include Agricultural Sales. Model 3 will contain Rooms, Taxes, Electricity, Year Built and Gas while Model 4 will contain the same factors except for Gas.

The summaries of Model 3 and Model 4 are seen in Figures 8 and 9, respectively. They show that every variable in each model is statistically significant at the 95% confidence level, where before it was not. The only minor difference between the two is the inclusion of Gas, which although it is significant, it does not pass the 5% p-value threshold by much. This will need to be looked at in more detail later.

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.0071920  0.2042702 -44.095  < 2e-16 ***
## Electricity -0.0026237  0.0005846  -4.488 7.59e-06 ***
## Gas         -0.0012863  0.0006308  -2.039   0.0416 *
## Rooms       -0.1484024  0.0233454  -6.357 2.55e-10 ***
## Taxes       -0.1131209  0.0030636 -36.924  < 2e-16 ***
## YearBuilt   -0.0911356  0.0191029  -4.771 1.97e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.484 on 1994 degrees of freedom
## Multiple R-squared:  0.5002, Adjusted R-squared:  0.499
```

Figure 8: Summary of Model 3 (including Year Built)

```
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -9.0650146  0.2024525 -44.776  < 2e-16 ***
## Electricity -0.0025902  0.0005848  -4.429 9.96e-06 ***
## Rooms       -0.1520283  0.0232960  -6.526 8.54e-11 ***
## Taxes       -0.1135409  0.0030591 -37.116  < 2e-16 ***
## YearBuilt   -0.0881065  0.0190601  -4.623 4.03e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.486 on 1995 degrees of freedom
## Multiple R-squared:  0.4992, Adjusted R-squared:  0.4982
```

Figure 9: Summary of Model 4 (not including Year Built)

The standard diagnostic plots for each model show that each model has a constant variance and are generally normally distributed. However, as both plots match the assumption, we cannot determine with is a better for out test data. The next step is to determine which points count as outliers. Figure 12 and Figure 13 are residual plots for each model using studentized and semi-studentized residuals. As can be seen in the figures, they are quite similar and thus tell us there are not any outliers that significantly affect the models.

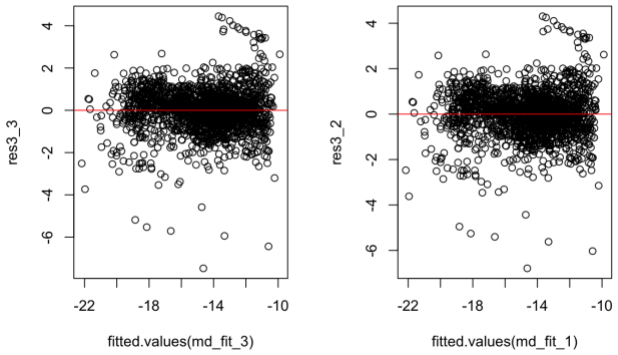

Figure 12: Residual plots for Model 3 (including Year Built)

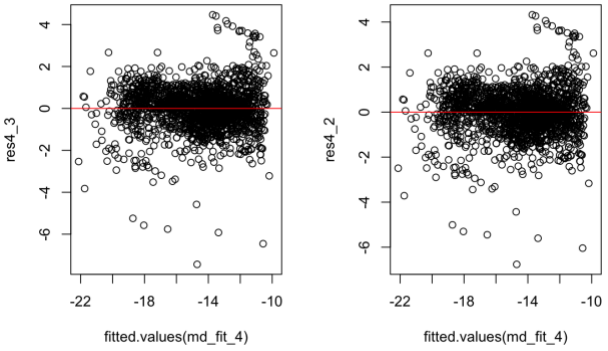

Figure 13: Residual plots for Model 4 (not including Year Built)

Figure 14, below, plots the Cook's distance for each observation and plots them against the case number. They are consistent with our current information as they show that no observations are outlier, so there is no need to delete any observations.

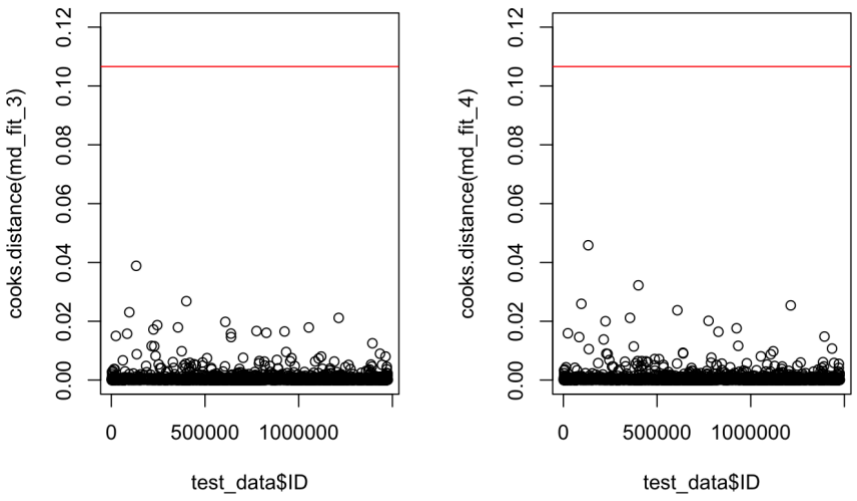

Figure 14: Cook's Distances plotted for Model 3 (left) and Model 4 (right)

MODEL VALIDATION

The model at this point has been refined and all significant outliers have been removed. Now the model will be run against another data set to ensure that it is accurate. To see if each model fits against the validation data we obtain and compare the MSPR to the MSE. For Model 3 we have a MSPR of 6712.45 and an MSE of 6.00. For Model 4 we have a MSPR of 6710.59 and an MSE of 6.00. These statistics show that each model does a very good job at modeling the data, but neither is significantly better than the other.

```
##                    2.5 %        97.5 %
## (Intercept) -9.0834038547 -8.9265889818
## Electricity -0.0027545598 -0.0022997695
## Gas         -0.0008460398 -0.0003807405
## Rooms       -0.1923038472 -0.1743584405
## Taxes       -0.1115698790 -0.1091965333
## YearBuilt   -0.0819602722 -0.0667962230
```

Figure 15: 97.5% confidence interval for Model 3 factors

```
##                    2.5 %       97.5 %
## (Intercept) -9.107975801 -8.952284234
## Electricity -0.002729389 -0.002274878
## Rooms       -0.194259887 -0.176373194
## Taxes       -0.111807525 -0.109440615
## YearBuilt   -0.080959822 -0.065810601
```

Figure 16: 97.5% confidence interval for Model 4 factors

```
##                    2.5 %        97.5 %
## (Intercept) -9.1141426357 -9.0277094961
## Electricity -0.0025268208 -0.0022766424
## Gas         -0.0005258832 -0.0002682076
## Rooms       -0.1857104863 -0.1758469527
## Taxes       -0.1118986315 -0.1105915841
## YearBuilt   -0.0750687087 -0.0667150051
```

Figure 17: 97.5% confidence intervals for Predicted Model 3 factors

```
##                    2.5 %       97.5 %
## (Intercept) -9.130467397 -9.044703911
## Electricity -0.002510808 -0.002260816
## Rooms       -0.186866995 -0.177031705
## Taxes       -0.112055725 -0.110752621
## YearBuilt   -0.074474191 -0.066128386
```

Figure 18: 97.5% confidence intervals for Predicted Model 4 factors

The above tables show confidence intervals for Model 3 and Model 4 alongside Predicted Model 3 and Predicted Model 4, which tests against the rest of the available data. From this we see that the estimates for Model 3 and its validation model are contained within their confidence interval. The same can be said for Model 4 and its validation model. Because the R_2 value for Model 3 (50.14%) is greater than Model 4 (50.12%) as well as the validation models (49.96% vs. 49.95%) were all so close it is impossible to tell from them which model is better. However, when you include the largest differences in MSPR for each and that Model 3 included an extra variable that could still be an explanatory variable, Model 4 is the better model.

CONCLUSION

We can conclude that the best model for predicting home and property value is:

$$\text{VALUE}^{-0.2222} = -9.056 - 0.002431 \cdot \text{ELECTRICITY} - 0.0004477 \cdot \text{GAS} - 0.1814 \cdot \text{ROOMS} - 0.1110 \cdot \text{TAXES} - 0.07169 \cdot \text{YEAR BUILT} + \epsilon$$

This model was constructed with 97.5% confidence based on the sampling size as well as in model validation. The margin of error was 2%. This model explains 49.96% of the variation by the regression line.

```
## Coefficients:
##               Estimate Std. Error  t value Pr(>|t|)
## (Intercept) -9.056e+00  1.931e-02 -468.956  < 2e-16 ***
## Electricity -2.431e-03  5.592e-05  -43.476  < 2e-16 ***
## Gas         -4.477e-04  5.751e-05   -7.786 6.96e-15 ***
## Rooms       -1.814e-01  2.205e-03  -82.246  < 2e-16 ***
## Taxes       -1.110e-01  2.921e-04 -380.210  < 2e-16 ***
## YearBuilt   -7.169e-02  1.867e-03  -38.409  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.447 on 214868 degrees of freedom
## Multiple R-squared:  0.4996, Adjusted R-squared:  0.4996
## F-statistic: 4.291e+04 on 5 and 214868 DF,  p-value: < 2.2e-16
```

Figure 19: Statistical Summary of Final Model

R CODE

```
## Setting Working Directory

setwd("/Users/carlconti/Downloads")


library(MASS)

library(vctrs)

library(readxl)

library(car)

library(leaps)

# Establishing two halves of data

file1 <- read_xlsx("/Users/carlconti/Downloads/capstone data.xlsx", sheet = "
1-28")

file2 <- read_xlsx("/Users/carlconti/Downloads/capstone data.xlsx", sheet = "
29-56")

names <- c("State", "HousingType", "LotSize", "AgriSales", "Tub/Shower", "Bed
rooms", "Electricity", "Fuel", "Gas", "HeatingType", "Rooms", "Value", "Vehic
les", "Water", "YearBuilt", "KitchenApps", "YearsLivedIn", "Plumbing", "Taxes
", "ID")


## combining datasets (exluding NA's and non-HousingTypes #1's)

all_data <- rbind(file1, file2)

all_data["ID"] <- c(1:nrow(all_data))

all_data <- all_data[complete.cases(all_data), ]

names(all_data) <- names

all_data <- subset(all_data, HousingType  == 1)


## test data

n <- nrow(all_data)

obs <- sample(1:n, size=2500)

test_data <- all_data[obs, ]


## extra data

extra_data <- all_data[-obs, ]

## FIT 1
```

```r
## linear fit of all factors

full_fit <- lm(formula = Value ~ factor(LotSize) + factor(YearBuilt) + factor
(AgriSales) + factor(Plumbing) + factor(HeatingType) + factor(YearsLivedIn) +
Bedrooms + Electricity + Fuel + Gas + Rooms + Vehicles + Water + KitchenApps
+ Taxes, data = test_data)


## scatterplot matric of all factors

pairs(~ Value + Bedrooms + Electricity + Fuel + Gas + Rooms + Vehicles + Wate
r + Taxes, data = test_data, main = "Quantitative Variable Scatter Matrix")


## residual plots

par(mfrow = c(2, 2))

plot(full_fit)

## FIT 2 (boxcox)


## boxcox transformation

boxcox <- boxcox(full_fit)

lambda <- boxcox$x[which.max(boxcox$y)]


## applying lambda

test_data$ValueT <- - test_data$Value^lambda


## transformed fit

bc_fit <- lm(formula = ValueT ~ factor(LotSize) + factor(YearBuilt) + factor(
AgriSales) + factor(Plumbing) + factor(HeatingType) + factor(YearsLivedIn) +
Bedrooms + Electricity + Fuel + Gas + Rooms + Vehicles + Water + KitchenApps
+ Taxes, data = test_data)


## plot fit

par(mfrow = c(2, 2))

plot(bc_fit)

## FIT 3 (removing variables)


## scatterplot matrix of quantitative variables

pairs(~ ValueT + Bedrooms + Electricity + Fuel + Gas + Rooms + Vehicles + Wat
er + Taxes, data = test_data, main = "Quantitative Variable Scatter Matrix")


## added-variable plots to fuels costs and vehicles
```

```r
par(mfrow = c(2, 3))

avPlots(lm(ValueT~Electricity+Fuel+Gas+Vehicles+Water,data=test_data))


## boxplot on all qualitative varibles

par(mfrow = c(2, 3))

boxplot(ValueT ~ LotSize, data = test_data, main = "Transformed Value vs. Lot Size", xlab = "Lot Size", ylab = "Transformed Value")

boxplot(ValueT ~ YearBuilt, data = test_data, main = "Transformed Value vs. Year Built", xlab = "Year Built", ylab = "Transformed Value")

boxplot(ValueT ~ AgriSales, data = test_data, main = "Transformed Value vs. Agricultural Sales", xlab = "Agricultural Sales", ylab = "Transformed Value")

boxplot(ValueT ~ Plumbing, data = test_data, main = "Transformed Value vs. Plumbing", xlab = "Plumbing", ylab = "Transformed Value")

boxplot(ValueT ~ HeatingType, data = test_data, main = "Transformed Value vs. Heating Type", xlab = "Heating Type", ylab = "Transformed Value")

boxplot(ValueT ~ YearsLivedIn, data = test_data, main = "Transformed Value vs. Years Lived In", xlab = "Years Lived In", ylab = "Transformed Value")


rv_fit <- lm(ValueT ~ factor(AgriSales) + factor(YearBuilt) + factor(YearsLivedIn) + Bedrooms + Electricity + Gas + Rooms + Water + Taxes, data = test_data)

## FIT 4 (explanatory variables)


## best subset of remaining varibles

attach(test_data)

leaps <- regsubsets(ValueT ~ Bedrooms + Electricity + Gas + Rooms + Water + Taxes + AgriSales + YearBuilt + YearsLivedIn, data = test_data, nbest = 1, method = "backward", intercept = TRUE, nvmax = 8)

summary(leaps)


par(mfrow = c(1, 1))

subsets(leaps, statistic = "adjr2", legend = FALSE, cex.subsets = 0.4, xlim = c(1, 9))

subsets(leaps, statistic = "bic", legend = FALSE, cex.subsets = 0.4, xlim = c(1, 9))

subsets(leaps, statistic = "cp", legend = FALSE, cex.subsets = 0.4, xlim = c(1, 9))


par(mfrow = c(1,2))
```

```r
plot(leaps, scale = "adjr2")

plot(leaps, type = 'b')

plot(leaps, scale = 'Cp')


step_fit <- lm(ValueT ~ Bedrooms + Electricity + Gas + Rooms + Water + Taxes
+ AgriSales + YearBuilt + YearsLivedIn, data = test_data)

step(step_fit, data = test_data, scope = bc_fit, direction = "both", test = "
F")


## two final models

par(mfrow = c(2,2))

md_fit_1 <- lm(ValueT ~ Electricity + Gas + Rooms + Taxes + AgriSales + YearB
uilt, data = test_data)

md_fit_2 <- lm(ValueT ~ Electricity + Gas + Rooms + Taxes + AgriSales, data =
test_data)

md_fit_3 <- lm(ValueT ~ Electricity + Gas + Rooms + Taxes + YearBuilt, data =
test_data)

md_fit_4 <- lm(ValueT ~ Electricity + Rooms + Taxes + YearBuilt, data = test_
data)


par(mfrow = c(1,1))

summary(md_fit_1)

summary(md_fit_2)

summary(md_fit_3)

summary(md_fit_4)

## MODEL DIAGNOSTICS


## residual plots for fit 3

par(mfrow = c(1, 2))

yhat_3 <- fitted(md_fit_3)

mse3_1 <- summary(md_fit_3)$sigma^2

h_3 <- hatvalues(md_fit_3)

res3_1 <- residuals(md_fit_3)

res3_2 <- res3_1 / sqrt(mse3_1)

mse3_2 <- ( (2494) * mse3_1 - (res3_1^2)/(1-h_3) ) / (2495)

res3_3 <- res3_1 / sqrt(mse3_2 * (1-h_3))

plot(fitted.values(md_fit_3), res3_3)
```

```r
abline(a = 0, b = 0, col = "red")
plot(fitted.values(md_fit_1), res3_2)
abline(a = 0, b = 0, col = "red")

## residual plots for fit 4
par(mfrow = c(1, 2))
yhat_4 <- fitted(md_fit_4)
mse4_1 <- summary(md_fit_4)$sigma^2
h_4 <- hatvalues(md_fit_4)
res4_1 <- residuals(md_fit_4)
res4_2 <- res4_1 / sqrt(mse4_1)
mse4_2 <- ( (2494) * mse4_1 - (res4_1^2)/(1-h_4) ) / (2494)
res4_3 <- res4_1 / sqrt(mse4_2 * (1-h_4))
plot(fitted.values(md_fit_4), res4_3)
abline(a = 0, b = 0, col = "red")
plot(fitted.values(md_fit_4), res4_2)
abline(a = 0, b = 0, col = "red")

## standard diagnostic plots for models 3 and 4
par(mfrow = c(1,2))
plot(md_fit_3)
plot(md_fit_4)

## cook's distance model 3
par(mfrow = c(1,2))
plot(test_data$ID, cooks.distance(md_fit_3), ylim = c(0,0.12))
abline(h = 1 - qf(0.5, 6, 2494), col = "red")

## cook's distance model 4
plot(test_data$ID, cooks.distance(md_fit_4), ylim = c(0,0.12))
abline(h = 1 - qf(0.5, 6, 2494), col = "red")
## Model Validation
attach(test_data)
```

```r
## MSE and MSPR of model 3
pred_3 <- predict(md_fit_3, newdata = extra_data, interval = "prediction")
mspr_3 <- sum( (ValueT - pred_3)^2 ) / 2494
anova(md_fit_3)

## MSE and MSPR of model 4
pred_4 <- predict(md_fit_4, newdata = extra_data, interval = "prediction")
mspr_4 <- sum( (ValueT - pred_4)^2 ) / 2495
anova(md_fit_4)

## model 3 and model 4 on prediction data
extra_data$ValueT <- - extra_data$Value^lambda
pred_fit_3 <- lm(ValueT ~ Electricity + Gas + Rooms + Taxes + YearBuilt, data
= extra_data)
pred_fit_4 <- lm(ValueT ~ Electricity + Rooms + Taxes + YearBuilt, data = ext
ra_data)

confint(md_fit_3)
summary(md_fit_3)

confint(md_fit_4)
summary(md_fit_4)

confint(pred_fit_3)
summary(pred_fit_3)

confint(pred_fit_4)
summary(pred_fit_4)

all_data$ValueT <- - all_data$Value^lambda
final_fit <- lm(ValueT ~ Electricity + Gas + Rooms + Taxes + YearBuilt, data
= all_data)
summary(final_fit)
```

DATA SET

Data provided by United States Census Bureau as part of their annual American Community Survey (ACS), this project specifically used the year-2013 data. Dataset was posted to Kaggle by the official U.S. Census account.

https://www.kaggle.com/census/2013-american-community-survey