

# W271-2 – Spring 2016 – Lab 3

Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song

April 22, 2016

## Contents

<b>Part 1</b>	<b>2</b>
Modeling House Values . . . . .	2
<b>Part 2</b>	<b>20</b>
Modeling and Forecasting a Real-World Macroeconomic / Financial time series . . . . .	20
<b>Part 3</b>	<b>21</b>
Forecast the Web Search Activity for global Warming . . . . .	21
<b>Part 4</b>	<b>22</b>
Forecast Inflation-Adjusted Gas Price . . . . .	22

---

## Instructions

- Thoroughly analyze the given dataset or data series. Detect any anomalies in each of the variables. Examine if any of the variables that may appear to be top- or bottom-coded.
- Your report needs to include a comprehensive graphical analysis
- Your analysis needs to be accompanied by detailed narrative. Just printing a bunch of graphs and econometric results will likely receive a very low score.
- Your analysis needs to show that your models are valid (in statistical sense).
- Your rationale of using certain metrics to choose models need to be provided. Explain the validity / pros / cons of the metric you use to choose your “best” model.
- Your rationale of any decisions made in your modeling needs to be explained and supported with empirical evidence.
- All the steps to arrive at your final model need to be shown and explained clearly.
- All of the assumptions of your final model need to be thoroughly tested and explained and shown to be valid. Don’t just write something like, “the plot looks reasonable”, or “the plot looks good”, as different people interpret vague terms like “reasonable” or “good” differently.

## Part 1

### Modeling House Values

In Part 1, you will use the data set `houseValue.csv` to build a linear regression model, which includes the possible use of the instrumental variable approach, to answer a set of questions interested by a philanthropist group. You will also need to test hypotheses using these questions.

The philanthropist group hires a think tank to examine the relationship between the house values and neighborhood characteristics. For instance, they are interested in the extent to which houses in neighborhood with desirable features command higher values. They are specifically interested in environmental features, such as proximity to water body (i.e. lake, river, or ocean) or air quality of a region.

The think tank has collected information from tens of thousands of neighborhoods throughout the United States. They hire your group as contractors, and you are given a small sample and selected variables of the original data set collected to conduct an initial, proof-of-concept analysis. Many variables, in their original form or transformed forms, that can explain the house values are included in the dataset. Analyze each of these variables as well as different combinations of them very carefully and use them (or a subset of them), in its original or transformed version, to build a linear regression model and test hypotheses to address the questions. Also address potential (statistical) issues that may be caused by omitted variables.

---

First, we will load the data and conduct an exploratory analysis

```
# Loading the Data -----
# setwd('./HW8/data')

ex1df <- read.csv("houseValueData.csv")
```

The data consists of 400 observations of 11 numeric variables related to the value of each house and characteristics that describe it and its surrounding neighborhood. A numeric summary of each variable is provided below.

Table 1: Summary Statistics of Wage Data

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Median	Pctl(75)	Max
crimeRate_pc	400	3.763	8.872	0.006	0.083	0.266	3.675	88.976
nonRetailBusiness	400	0.112	0.070	0.007	0.051	0.097	0.181	0.277
withWater	400	0.068	0.251	0	0	0	0	1
ageHouse	400	68.932	27.977	2.900	45.675	77.950	94.150	100.000
distanceToCity	400	9.638	8.786	1.228	3.240	6.115	13.628	54.197
distanceToHighway	400	9.582	8.672	1	4	5	24	24
pupilTeacherRatio	400	21.391	2.168	15.600	19.900	21.900	23.200	25.000
pctLowIncome	400	15.795	9.341	2	8	14	21	49
homeValue	400	499,584.400	196,115.700	112,500	384,187.5	477,000	558,000	1,125,000
pollutionIndex	400	40.615	11.825	23.500	29.875	38.800	47.575	72.100
nBedRooms	400	4.266	0.719	1.561	3.883	4.193	4.582	6.780

The summary table shows that the dataset has no missing values and that many of the variables likely have a skewed distribution and may benefit from tranformation. Potential transformations will be discussed as the exploratory analysis proceeds.

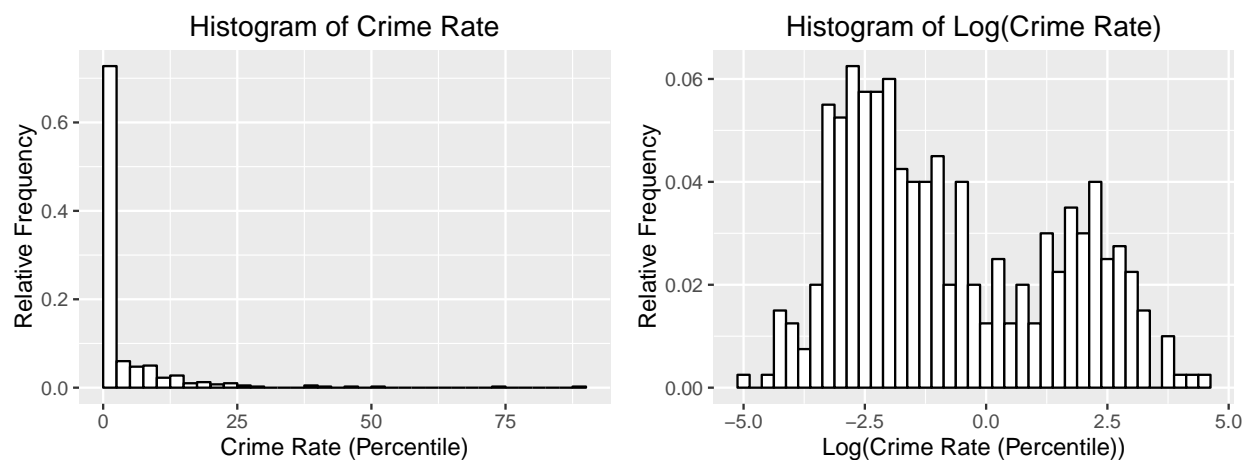


Figure 1: Histograms of the Crime Rate Variable

The crime rate variable is highly right-skewed, with most houses being in very low and low crime neighborhoods and a few houses ranging from moderate to very high crime neighborhoods. Log transformation of the variable reveals a roughly bimodal distribution, with a peak below zero representing low crime neighborhoods and a peak above zero representing moderate crime neighborhoods.

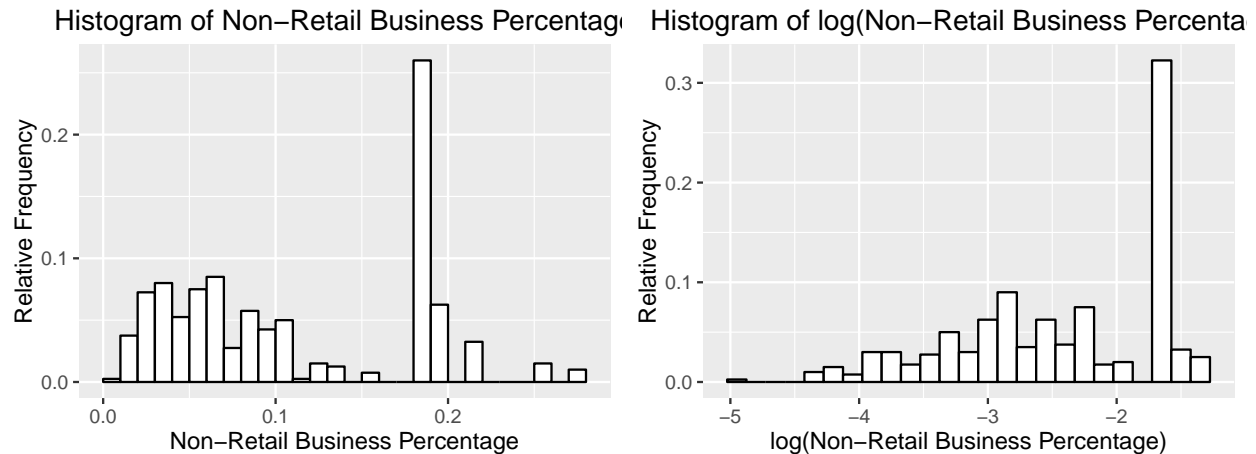


Figure 2: Histograms of Non-Retail Business Percentage Variable

The non-retail business percentage variable also shows multimodality, with one group of houses having a very low percentage of non-retail businesses. Another group of 104 houses has a non-retail business percentage of exactly 0.181. This large group with the same value suggests that a large portion of the sample may come from the same location, or perhaps that value-imputation was used to fill in missing values or infer values for this dataset. ???This variable may be a proxy for if the house is more urban or rural, so perhaps retail businesses are a sign of greater economic activity and thus we would expect these houses to have higher value.

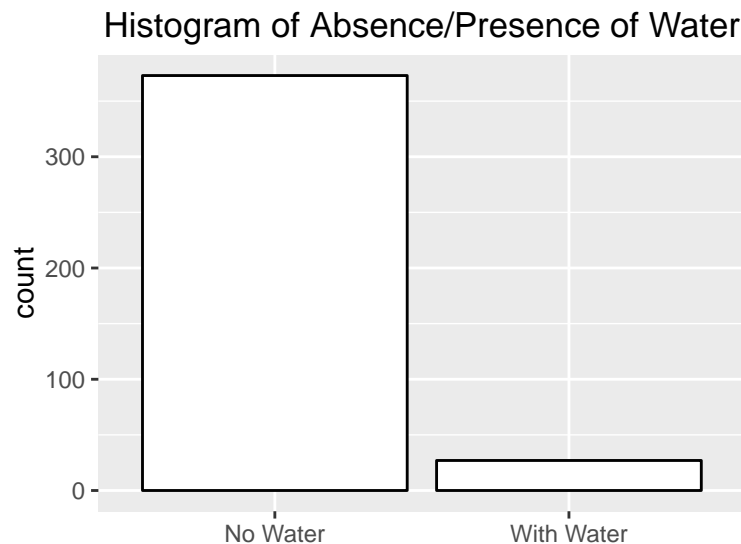


Figure 3: Histogram of Water Variable

The water variable shows that most houses are not located adjacent to water. The rarity of having water is likely a signal that having water is a strongly desirable trait for some home buyers and thus we would expect it to be correlated with higher home values.



Figure 4: Histogram of House Age

The age of houses in the same ranged from less than 5 years to about 100 years and was roughly uniformly distributed between 10 and 75 years and increasing thereafter. The majority of houses are over 50 years old, with a spike in home age at around 100 years. Log transformation of the house age variable produced a roughly monotonically increasing distribution and does not seem to be an appropriate transformation for purposes of modeling.

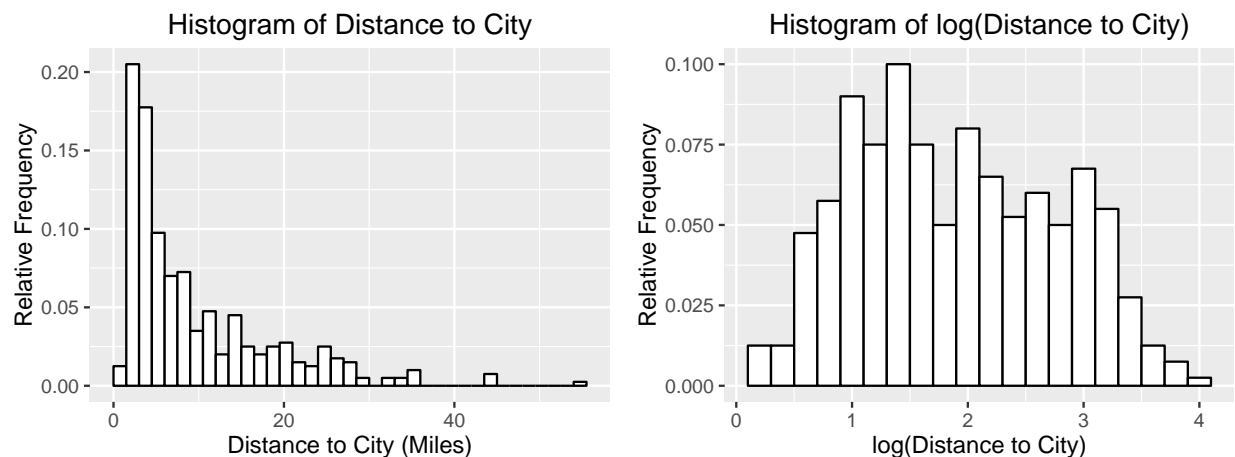


Figure 5: Histogram of Distance to City

The distance from a house to the nearest city had a right-tailed distribution, with most houses within 10 miles of a city and a minority of houses with greater than 25 miles to the nearest city. Log transformation of the city distance variable removed the skewness of the distribution and produced a more approximately normal distribution, although certainly still non-normal.

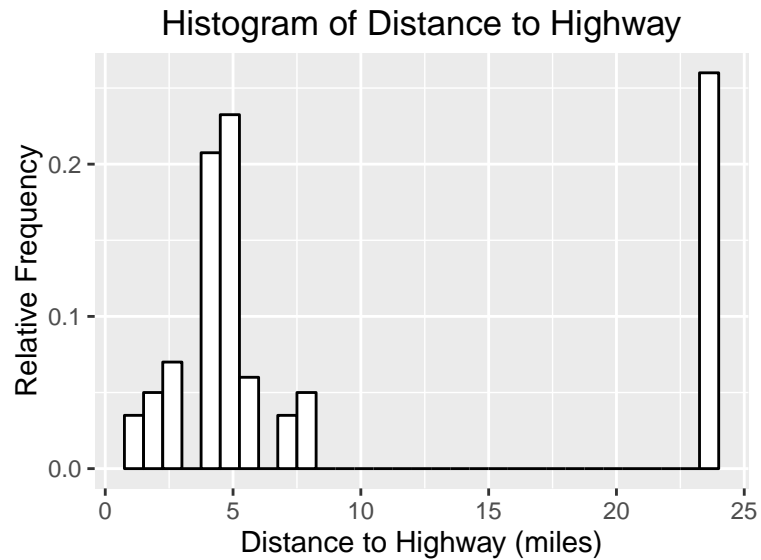


Figure 6: Histogram of Distance to Highway

The distance to highway values fall in to 8 bins. This suggests that these values may actually represent factors, or perhaps that the values were heavily rounded. The majority of houses were within 10 miles of a highway, with a large group of houses located 24 miles from the nearest highway.

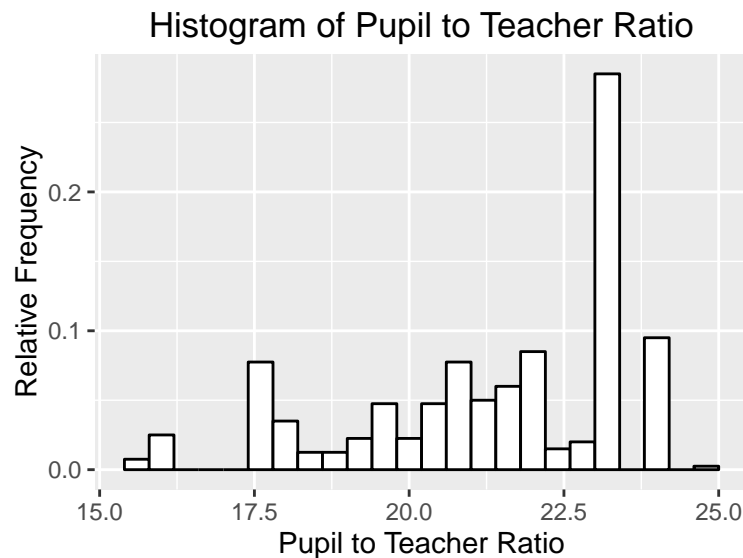


Figure 7: Histogram of Distance to Highway

The pupil to teacher ratio histogram shows a large portion of school districts have ratios between 17 and 23, and a small minority have ratios below 16.

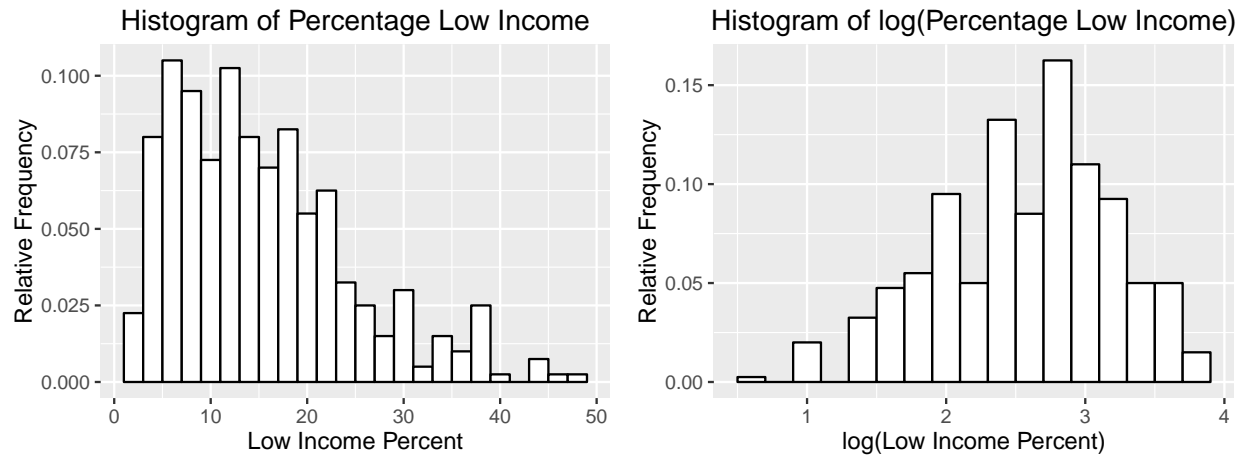


Figure 8: Histogram of Distance to Highway

The percentage of low income housing displayed a right-skewed distribution, with most values falling around the mean of 15.795 and then a long tail stretching to the maximum of nearly 50% low income housing. Log transformation produced a more normal looking distribution.

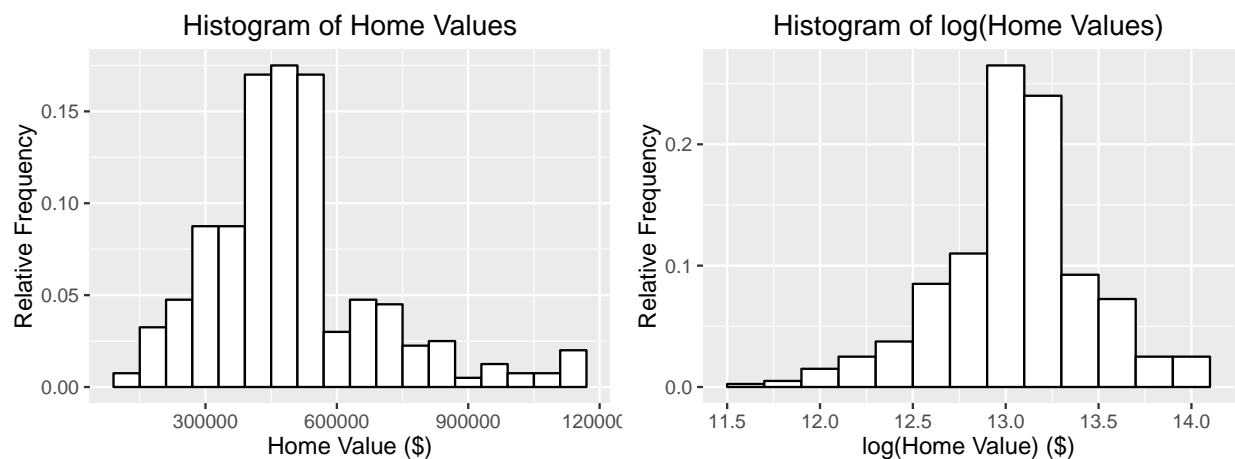


Figure 9: Histogram of Distance to Highway

The housing value plot was also right-skewed, with most values around the mean of 15.8 and a right tail extending to around \$1,200,000. Again, log transformation seems appropriate and produces a more normal looking distribution.

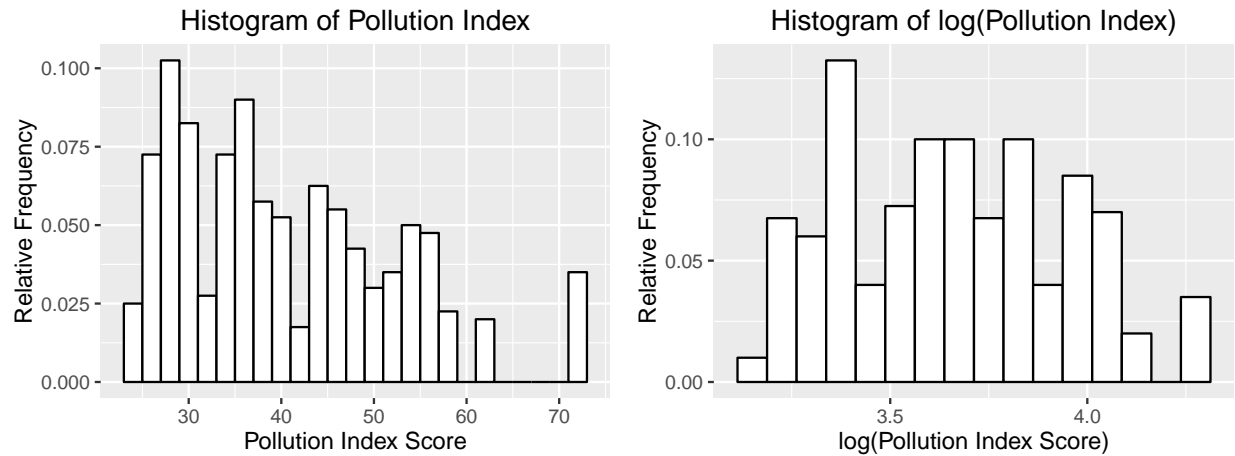


Figure 10: Histogram of Distance to Highway

The pollution index scores have a slightly-right tailed appearing distribution, with thin tails and evidence of multimodality. Log transformation of the pollution index reduced the right-skewness while still showing evidence of multimodality and thinner tails than a normal distribution.

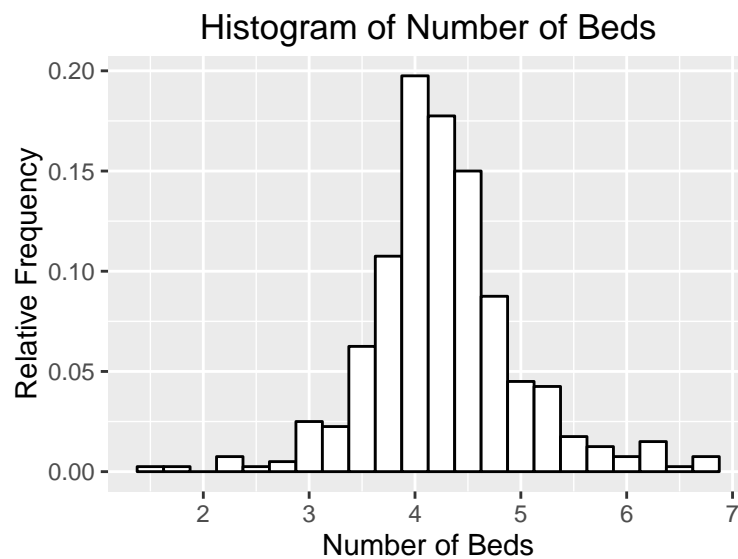


Figure 11: Histogram of Distance to Highway

The number of beds in houses had a roughly normal appearing distribution, but with considerably longer tails.

After visually inspect the distribution of each individual variable, we are also interested in how the variables relate, especially to the outcome variable of our model, housevalue. To gain an understanding of how each potential input variable relates to the outcome variable, we make a scatter plot and summarize the results of a univariate regression for each input variable against the output variable.



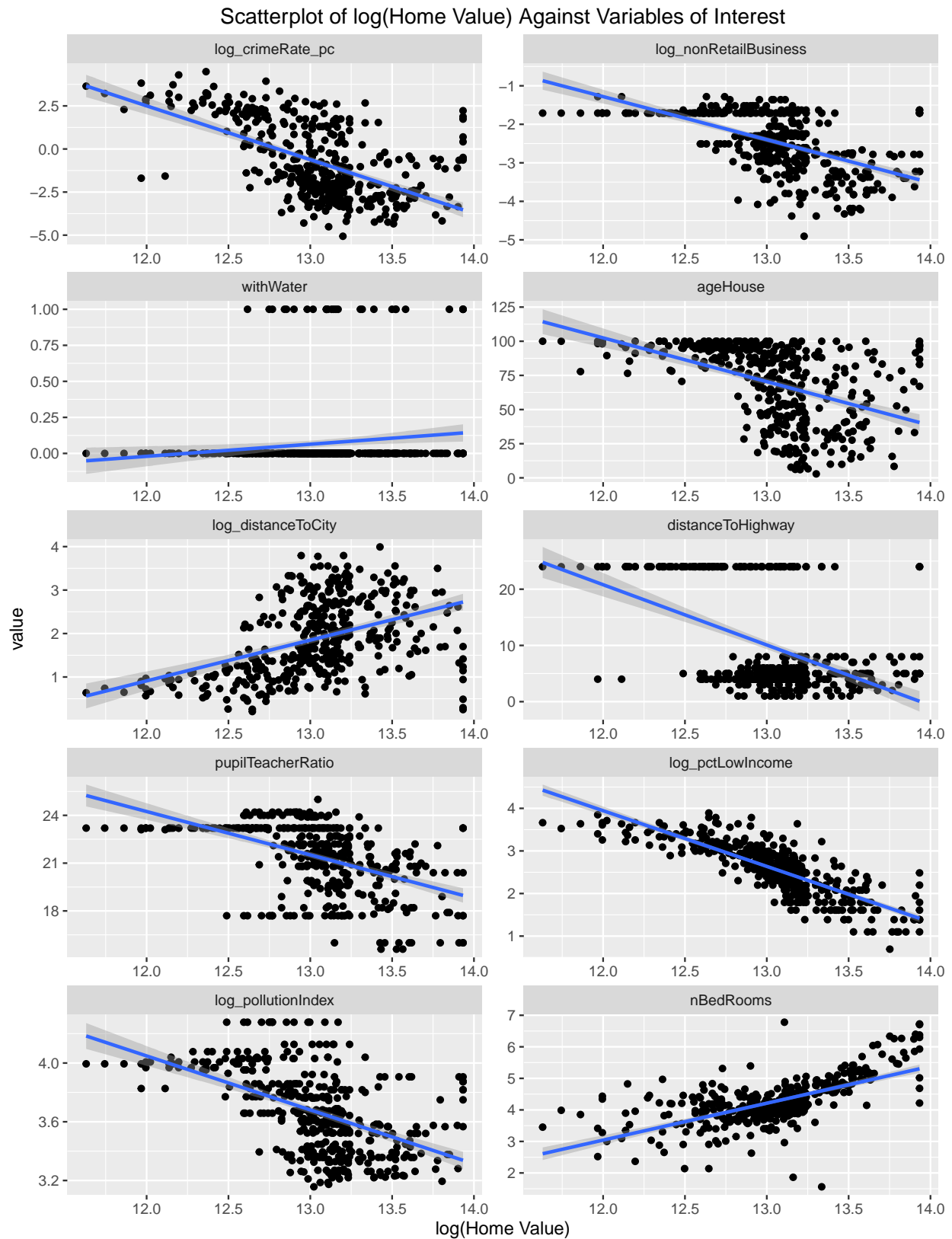


Figure 12: Scatter Plot of Home Value Against the Variables of Interest

Table 2: Univariate Regressions Against House Value

	Variable	Coefficient	P-value	R-Squared
1	log_crimeRate_pc	-0.105	0	0.328
2	log_nonRetailBusiness	-0.288	0	0.321
3	withWater	0.21	0.008	0.018
4	ageHouse	-0.006	0	0.207
5	log_distanceToCity	0.196	0	0.184
6	distanceToHighway	-0.022	0	0.241
7	pupilTeacherRatio	-0.091	0	0.249
8	log_pctLowIncome	-0.518	0	0.677
9	log_pollutionIndex	-0.728	0	0.268
10	nBedRooms	0.357	0	0.417

The scatterplots and regression summaries show that each of the variables, when not controlling for any other variables, is a better predictor of  $\log(\text{house price})$  than the mean house price. This fact may complicate any effort to select one of these variables as an instrument, as being un-related to the outcome variable is one condition of the exclusion restriction for an instrumental variable approach. However, this does not necessarily preclude all the variables from being used as an instrument, as some variables may not be significant when controlling for other variables.

Examining the scatter plot, it is noteworthy that many of the input variables seem to have a negative relationship with the outcome variable. To check for potential issues with multicollinearity, we can test the values in the correlation matrix for the dataset to see if they are over a threshold.

```
# Calculate correlation matrix
corrs <- cor(ex1df)
# Check if any variables other than the diagonals are perfectly correlated
length(corrs[corrs == 1])
```

```
## [1] 11
```

```
# Check if any variables have a correlation coefficient above .9
length(corrs[corrs > 0.9 & corrs != 1])
```

```
## [1] 0
```

Analysis of the correlation matrix shows that there are no variable pairs with a correlation coefficient above .9. While a good check against obvious issues of multicollinearity, this does not preclude the possibility that one variable is highly correlated with linear combinations of the other variables.

Since we are tasked with determining the impact of environmental variables on the value of homes, we will want to understand how those variables relate to the other variables in the dataset.

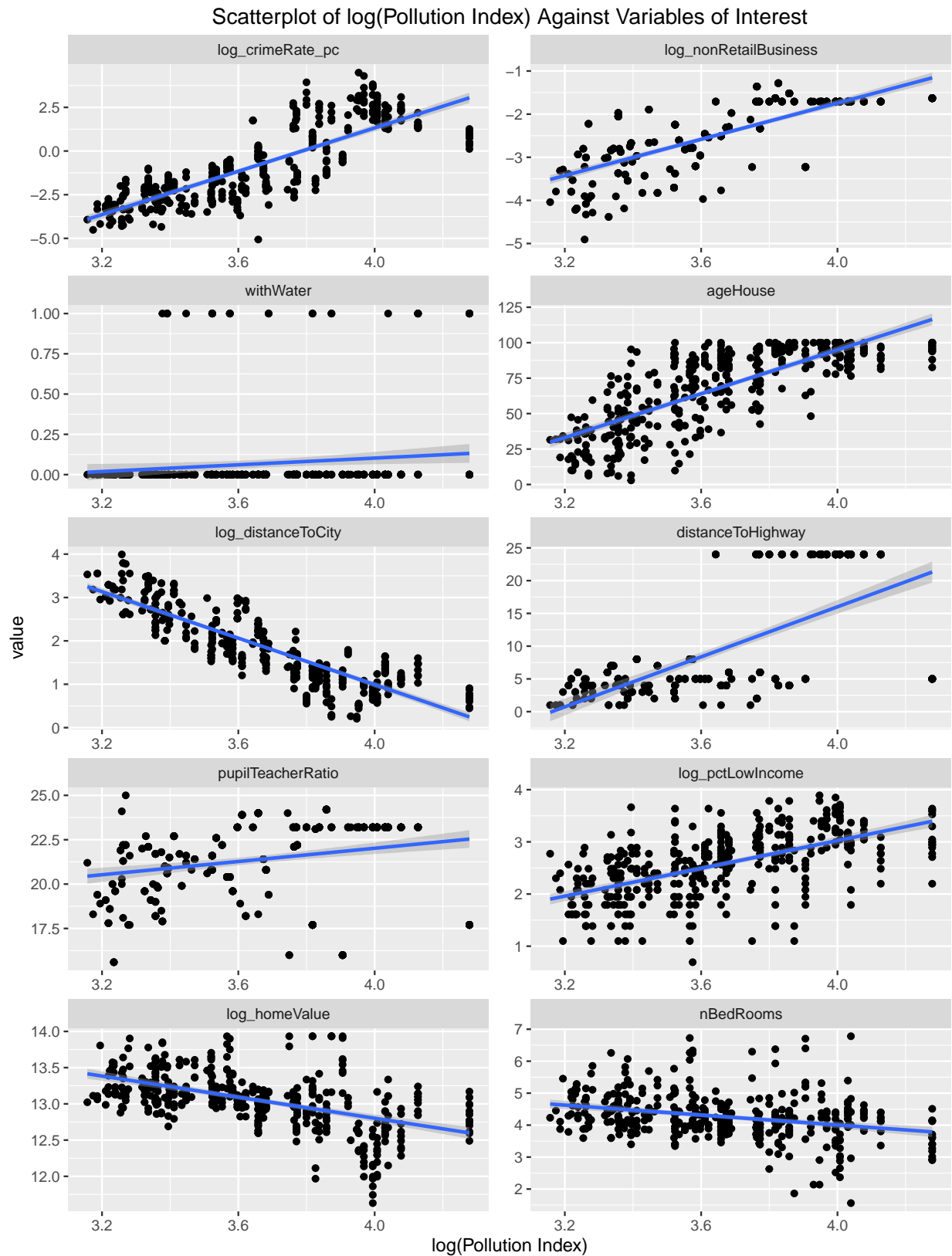


Figure 13: Scatter Plot of Pollution Index Against the Variables of Interest

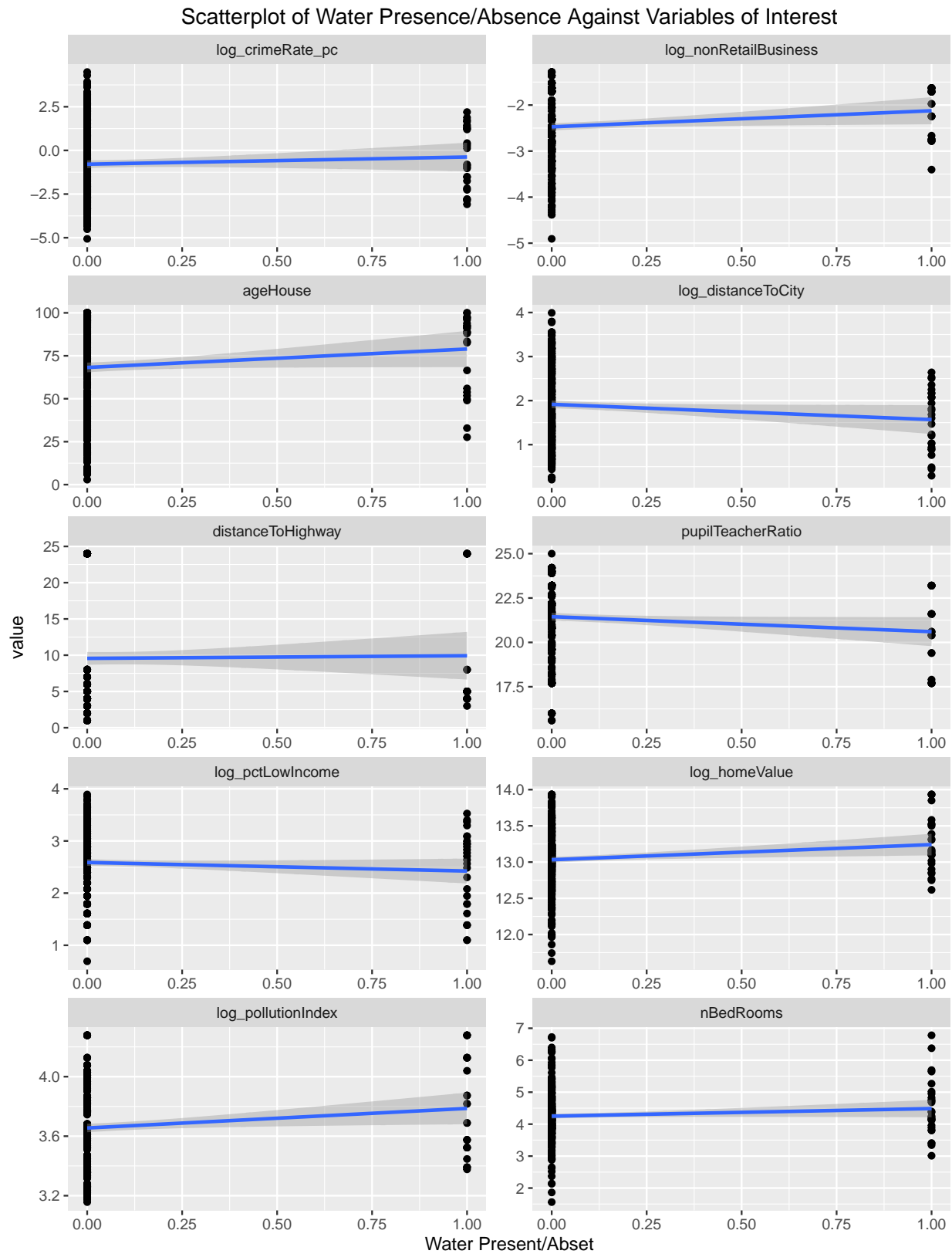


Figure 14: Scatter Plot of Water Presence/Absence Against the Variables of Interest

The scatter plots of the variables in the dataset against the pollution index reveal that there are many strong negative and positive correlations. This provides evidence that estimating the effects of pollution on home values requires controlling for a number of potential confounding variables. The scatter plot of adjacency to water with the variables reveals generally weaker relationships, which is not surprising given that the variable is a factor and that the occurrence of water is less influenced by social factors than pollution.

Before beginning the empirical process of model selection, we will stipulate that the variables for number of bedrooms and percentage of low income housing should be included in *any* regression model with house value because they have a well established relationship with the outcome variable.

Taking our variables of interest and the two variables included for theoretical reasons, the baseline model is:

$$\log(\text{homeValue}) = \beta_0 + \beta_1 \log(\text{pollutionIndex}) + \beta_2 \text{withWater} + \beta_3 n\text{BedRooms}$$

Table 3: Base Model Regression summary

	<i>Dependent variable:</i>
	log(home value)
log(PollutionIndex)	−0.126* (0.050)
Water Absence/Presence	0.136*** (0.038)
Number of Bedrooms	0.099** (0.031)
log(Percentage Low Income Housing	−0.405*** (0.036)
Constant	14.120*** (0.234)
F Statistic	163.005***
df	4; 395
Observations	400
R <sup>2</sup>	0.703
Adjusted R <sup>2</sup>	0.700
Residual Std. Error	0.217

·p<0.1; \*p<0.05; \*\*p<0.01; \*\*\*p<0.001

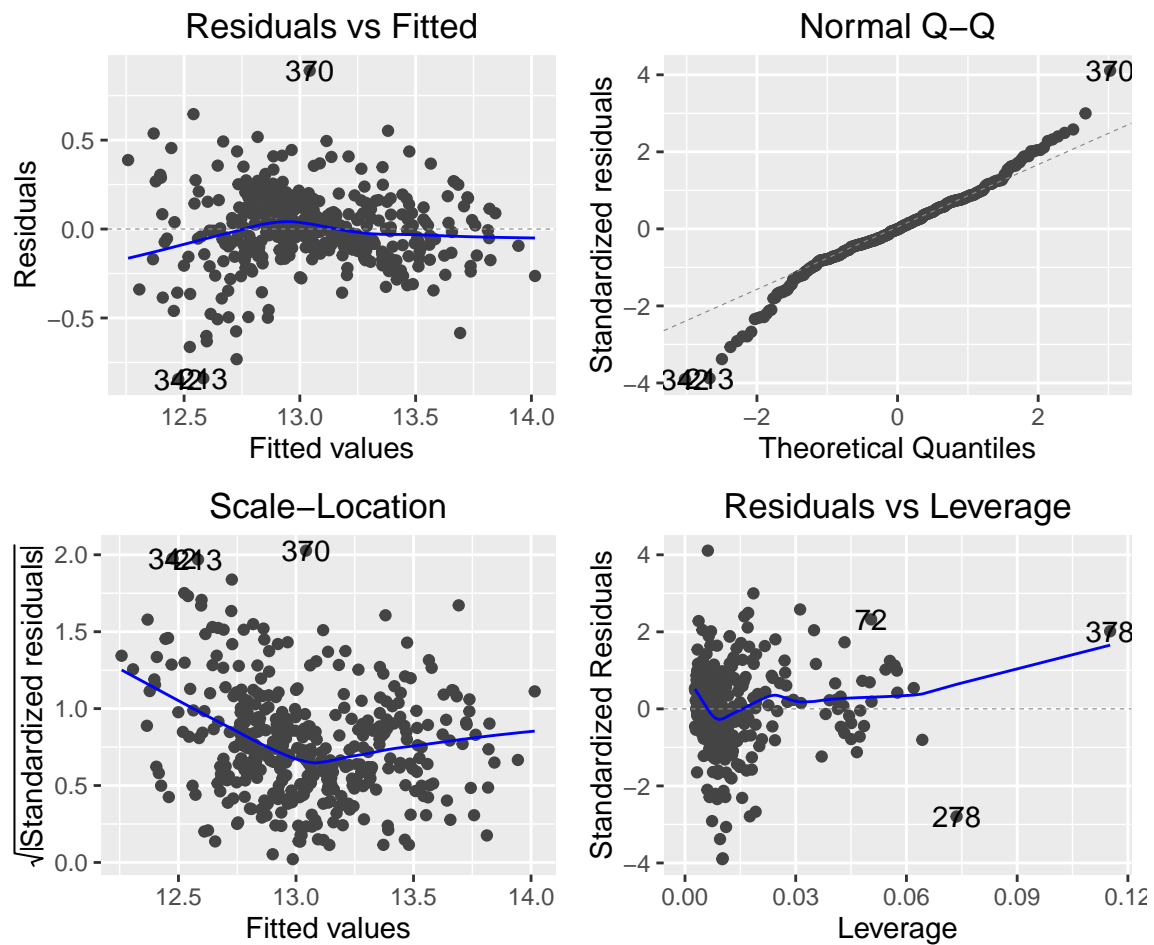


Figure 15: Diagnostic Plots for Base Regression Model

Looking at the results of the base regression, we see that all coefficients are significant, a very basic confirmation that these 3 variables are indeed variables of interest, and that the  $R^2$  value is 0.7029376. Looking at the diagnostic plots, we see in the residuals versus fitted plot that the band of values seems to widen in from left to right and then narrow, suggesting that the use of **heteroskedasticity robust standard errors** is appropriate.

Starting with our basemodel, we are interesting in selecting a model that 1) is a good fit of our sample data 2) can generalize to the larger dataset 3) provides insight into the effect of environmental factors on the value of homes. One approach we examine for selecting a final model is adding additional parameters while penalizing the model for increasing complexity using a measure like the AIC/BIC.

```
base_model <- lm(log_homeValue~log_pollutionIndex + withWater + nBedRooms +
                 log_pctLowIncome, data = s_df)
full <- lm(log_homeValue~., data=s_df)
step(base_model, scope=list(lower=base_model, upper=full), direction="both")
```

```
## Start:  AIC=-1215.67
## log_homeValue ~ log_pollutionIndex + withWater + nBedRooms +
##      log_pctLowIncome
##
##              Df Sum of Sq    RSS    AIC
## + pupilTeacherRatio    1    1.27809 17.399 -1242.0
## + distanceToHighway    1    1.19555 17.482 -1240.1
## + log_crimeRate_pc     1    0.93842 17.739 -1234.3
## + log_distanceToCity   1    0.50244 18.175 -1224.6
## + ageHouse             1    0.31582 18.361 -1220.5
## + log_nonRetailBusiness 1    0.12941 18.548 -1216.5
## <none>                  18.677 -1215.7
##
## Step:  AIC=-1242.02
## log_homeValue ~ log_pollutionIndex + withWater + nBedRooms +
##      log_pctLowIncome + pupilTeacherRatio
##
##              Df Sum of Sq    RSS    AIC
## + log_distanceToCity   1    0.57834 16.821 -1253.5
## + distanceToHighway    1    0.45576 16.943 -1250.6
## + log_crimeRate_pc     1    0.42263 16.976 -1249.9
## + ageHouse             1    0.41612 16.983 -1249.7
## <none>                  17.399 -1242.0
## + log_nonRetailBusiness 1    0.00007 17.399 -1240.0
## - pupilTeacherRatio    1    1.27809 18.677 -1215.7
##
## Step:  AIC=-1253.54
## log_homeValue ~ log_pollutionIndex + withWater + nBedRooms +
##      log_pctLowIncome + pupilTeacherRatio + log_distanceToCity
##
##              Df Sum of Sq    RSS    AIC
## + log_crimeRate_pc     1    0.56562 16.255 -1265.2
## + distanceToHighway    1    0.46512 16.355 -1262.8
## + ageHouse             1    0.15403 16.667 -1255.2
## <none>                  16.821 -1253.5
## + log_nonRetailBusiness 1    0.03313 16.788 -1252.3
## - log_distanceToCity   1    0.57834 17.399 -1242.0
## - pupilTeacherRatio    1    1.35399 18.175 -1224.6
```

```
##
## Step: AIC=-1265.22
## log_homeValue ~ log_pollutionIndex + withWater + nBedRooms +
##   log_pctLowIncome + pupilTeacherRatio + log_distanceToCity +
##   log_crimeRate_pc
##
##           Df Sum of Sq   RSS   AIC
## + ageHouse      1    0.12855 16.127 -1266.4
## <none>                16.255 -1265.2
## + distanceToHighway 1    0.03559 16.219 -1264.1
## + log_nonRetailBusiness 1    0.00126 16.254 -1263.3
## - log_crimeRate_pc      1    0.56562 16.821 -1253.5
## - log_distanceToCity    1    0.72132 16.976 -1249.9
## - pupilTeacherRatio     1    0.76782 17.023 -1248.8
##
## Step: AIC=-1266.4
## log_homeValue ~ log_pollutionIndex + withWater + nBedRooms +
##   log_pctLowIncome + pupilTeacherRatio + log_distanceToCity +
##   log_crimeRate_pc + ageHouse
##
##           Df Sum of Sq   RSS   AIC
## <none>                16.127 -1266.4
## - ageHouse      1    0.12855 16.255 -1265.2
## + distanceToHighway 1    0.01379 16.113 -1264.7
## + log_nonRetailBusiness 1    0.00001 16.127 -1264.4
## - log_distanceToCity    1    0.43001 16.556 -1257.9
## - log_crimeRate_pc      1    0.54014 16.667 -1255.2
## - pupilTeacherRatio     1    0.80898 16.936 -1248.8
##
## Call:
## lm(formula = log_homeValue ~ log_pollutionIndex + withWater +
##     nBedRooms + log_pctLowIncome + pupilTeacherRatio + log_distanceToCity +
##     log_crimeRate_pc + ageHouse, data = s_df)
##
## Coefficients:
##      (Intercept)  log_pollutionIndex      withWater
##          15.105338          -0.243462           0.109052
##          nBedRooms  log_pctLowIncome  pupilTeacherRatio
##           0.087657          -0.377696          -0.024465
## log_distanceToCity  log_crimeRate_pc      ageHouse
##          -0.084006          -0.031751           0.001202

stepModel <- lm(formula = log_homeValue ~ log_pollutionIndex + withWater +
                 nBedRooms + log_pctLowIncome + pupilTeacherRatio + log_distanceToCity +
                 log_crimeRate_pc + ageHouse, data = s_df)
```

The model chosen by step-wise addition of variables and AIC penalization has 4 additional parameters. One obvious drawback to this approach is it tends to under-state the confidence intervals and thus the resulting p-values for the parameters are too low. Indeed, in the resulting model, the final parameter's coefficient is not statistically different from zero using heteroskedasticity-robust standard errors.

Since we are estimating a model using only a portion of the data, we should also be concerned with the out of sample fit for our proposed model. Here, we can test the accuracy of predictions made with increasingly complex models using a subsample of withheld data.



```

# Split data into training and testing portions
set.seed(1099)
train <- sample_frac(s_df, 0.8)
r_id <- as.numeric(rownames(train))
test <- s_df[-r_id,]

# Define the base model and add parameters
base_params <- colnames(s_df)[c(10, 3, 11, 8)]
base <- lm(log_homeValue~log_pollutionIndex + withWater + nBedRooms +
           log_pctLowIncome, data = train)
plus_one_params <- colnames(s_df)[c(10, 3, 11, 8, 7)]
plus_one <- lm(log_homeValue~log_pollutionIndex + withWater + nBedRooms +
              log_pctLowIncome
              + pupilTeacherRatio, data = train)
plus_two_params <- colnames(s_df)[c(10, 3, 11, 8, 7, 5)]
plus_two <- lm(log_homeValue~log_pollutionIndex + withWater + nBedRooms +
              log_pctLowIncome +
              pupilTeacherRatio + log_distanceToCity, data = train)
plus_three_params <- colnames(s_df)[c(10, 3, 11, 8, 7, 5, 1)]
plus_three <- lm(log_homeValue~log_pollutionIndex + withWater + nBedRooms +
              log_pctLowIncome +
              pupilTeacherRatio + log_distanceToCity + log_crimeRate_pc,
              data = train)
plus_four_params <- colnames(s_df)[c(10, 3, 11, 8, 7, 5, 1, 4)]
plus_four <- lm(log_homeValue~log_pollutionIndex + withWater + nBedRooms +
              log_pctLowIncome +
              pupilTeacherRatio + log_distanceToCity + log_crimeRate_pc +
              ageHouse, data = train)

# Use each model to predict the outcome variable in the test data
base_preds <- predict(base, test[,base_params], interval = "prediction")
plus_one_preds <- predict(plus_one, test[, plus_one_params],
                        interval = "prediction")
plus_two_preds <- predict(plus_two, test[, plus_two_params],
                        interval = "prediction")
plus_three_preds <- predict(plus_three, test[, plus_three_params],
                        interval = "prediction")
plus_four_preds <- predict(plus_four, test[, plus_four_params],
                        interval = "prediction")

# Summarize the model predictions and AIC, BIC in a dataframe
pred_df <- data.frame(rbind(accuracy(base_preds[,1],
                                   test$log_homeValue)[,c("RMSE", "MAE")],
                        accuracy(plus_one_preds[,1], test$log_homeValue)[,c("RMSE", "MAE")],
                        accuracy(plus_two_preds[,1], test$log_homeValue)[,c("RMSE", "MAE")],
                        accuracy(plus_three_preds[,1], test$log_homeValue)[,c("RMSE", "MAE")],
                        accuracy(plus_four_preds[,1], test$log_homeValue)[,c("RMSE", "MAE")])))

pred_df$model <- c("base", "plus_one", "plus_two", "plus_three", "plus_four")
diag_df <- data.frame(
  cbind(BIC(base, plus_one, plus_two, plus_three, plus_four),
        AIC=AIC(base, plus_one, plus_two, plus_three, plus_four)[, 2]))
rownames(diag_df) <- c(1:5)

```

```
model_df <- data.frame(cbind(model =pred_df[,3],diag_df[, 1:3], pred_df[,c(1,2)]))
```

The results of out-of-sample fitting are summarized in the table below:

Table 4: Summary of Model Diagnostics and Out-Of-Sample Fit

	model	df	BIC	AIC	RMSE	MAE
1	base	6	-56.235	-78.844	0.240	0.172
2	plus_one	7	-73.034	-99.412	0.232	0.165
3	plus_two	8	-75.289	-105.436	0.225	0.161
4	plus_three	9	-81.502	-115.417	0.223	0.158
5	plus_four	10	-78.715	-116.398	0.222	0.152

Using the criterion of the RMES or MAE, the model chosen would include 4 additional parameters, while the AIC would favor either the model with 3 or 4 additional parameters, and the BIC would favor the model with 3 additional parameters.

Taking the results of step-wise addition and out-of-sample forecasting together, one could persuasively argue for both the model with 3 additional parameters and the model with 4. Given the similarity of the out of sample fit and the preference for models with less complexity, **I would select a model with the fewest additional parameters, in this case 3**

Table 5: Selected Model Regression summary

	<i>Dependent variable:</i> log(home value)
log(PollutionIndex)	-0.211* (0.099)
Water Absence/Presence	0.113** (0.040)
Number of Bedrooms	0.097** (0.030)
log(Percentage Low Income Housing	-0.359*** (0.038)
Pupil to Teacher Ratio	-0.024*** (0.005)
log(Distance to City	-0.101*** (0.027)
log(Average Crime Rate	-0.032** (0.011)
Constant	14.997*** (0.485)
F Statistic	133.939***
df	7; 392
Observations	400
R <sup>2</sup>	0.741
Adjusted R <sup>2</sup>	0.737
Residual Std. Error	0.204

·p<0.1; \*p<0.05; \*\*p<0.01; \*\*\*p<0.001

According to the model summary, there is in fact an effect of environmental factors on home values. For pollution index value, an increase of 1 in the log value of pollution index score results in a decrease of -0.211 (-0.042, -0.381),  $p = 0.015$  in the log home value. For a house with median value, going from the 25th percentile on the pollution index to the median value would result in a decrease in home value of \$12499.53. For locations adjacent to water, going from not being adjacent to water to being adjacent results in an increase in log home value of 0.113 (0.195, 0.032),  $p = 0.007$ . For a house with median value, being adjacent to water would result in an increase of home value of \$57064.43

## Part 2

### Modeling and Forecasting a Real-World Macroeconomic / Financial time series

Build a time-series model for the series in `lab3_series02.csv`, which is extracted from a real-world macroeconomic/financial time series, and use it to perform a 36-step ahead forecast. The periodicity of the series is purposely not provided. Possible models include AR, MA, ARMA, ARIMA, Seasonal ARIMA, GARCH, ARIMA-GARCH, or Seasonal ARIMA-GARCH models.

---

## Part 3

### Forecast the Web Search Activity for global Warming

Imagine that your group is part of a data science team in an apparel company. One of its recent products is Global-Warming T-shirts. The marketing director expects that the demand for the t-shirts tends to increase when global warming issues are reported in the news. As such, the director asks your group to forecast the level of interest in global warming in the news. The dataset given to your group captures the relative web search activity for the phrase, “global warming” over time. For the purpose of this exercise, ignore the units reported in the data as they are unimportant and irrelevant. Your task is to produce the weekly forecast for the *next 3 months* for the relative web search activity for global warming. For the purpose of this exercise, treat it as a *12-step ahead forecast*.

The dataset for this exercise is provided in `globalWarming.csv`. Use only models and techniques covered in the course (up to lecture 13). Note that one of the modeling issues you may have to consider is whether or not to use the entire series provided in the data set. Your choice will have to be clearly explained and supported with empirical evidence. As in other parts of the lab, the general instructions in the *Instruction Section* apply.

---

## Part 4

### Forecast Inflation-Adjusted Gas Price

During 2013 amid high gas prices, the Associated Press (AP) published an article about the U.S. inflation-adjusted price of gasoline and U.S. oil production. The article claims that there is “*evidence of no statistical correlation*” between oil production and gas prices. The data was not made publicly available, but comparable data was created using data from the Energy Information Administration. The workspace and data frame `gasOil.Rdata` contains the U.S. oil production (in millions of barrels of oil) and the inflation-adjusted average gas prices (in dollars) over the date range the article indicates.

In support of their conclusion, the AP reported a single p-value. You have two tasks for this exercise, and both tasks need the use of the data set `gasOil.Rdata`.

Your first task is to recreate the analysis that the AP likely used to reach their conclusion. Thoroughly discuss all of the errors the AP made in their analysis and conclusion.

Your second task is to create a more statistically-sound model that can be used to predict/forecast inflation-adjusted gas prices. Use your model to forecast the inflation-adjusted gas prices from 2012 to 2016.

---