W271-2 – Spring 2016 – **Lab 3**

Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song

April 22, 2016

Contents

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 neighbhorhood 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 transfomed 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 casued by omitted variables.

Exploratory

Based on the information in `homeValueData_VariableDescription.txt`, the variables and their meaning are:

- `crimeRate_pc`: crime rate per capital, measured by number of crimes per 1000 residents in neighborhood.
- `nonRetailBusiness`: the proportion of non-retail business acres per neighborhood.
- `withWater`: the neighborhood within 5 miles of a water body (lake, river, etc); 1 if true and 0 otherwise.
- `ageHouse`: proportion of house built before 1950.
- `distanceToCity`: distances to the nearest city (measured in miles).
- `pupilTeacherRatio`: average pupil-teacher ratio in all the schools in the neighborhood.
- `pctLowIncome`: percentage of low income household in the neighborhood
- `homeValue`: median price of single-family house in the neighborhood (measured in dollar).
- `pollutionIndex`: pollution index, scaled between 0 and 100, with 0 being the best and 100 being the worst (i.e. uninhabitable).
- `nBedRooms`: average number of bed rooms in the single family houses in the neighborhood.

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

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

```
##            crimeRate_pc nonRetailBusiness withWater ageHouse
## nbr.val         400.000           400.000   400.000  400.000
## nbr.na            0.000             0.000     0.000    0.000
## skewness          4.962             0.288     3.435   -0.614
## kurtosis         33.982            -1.274     9.823   -0.947
## normtest.p        0.000             0.000     0.000    0.000
##            distanceToCity distanceToHighway pupilTeacherRatio pctLowIncome
## nbr.val           400.000           400.000           400.000      400.000
## nbr.na              0.000             0.000             0.000        0.000
## skewness            1.629             1.002            -0.772        0.967
## kurtosis            2.868            -0.871            -0.348        0.610
## normtest.p          0.000             0.000             0.000        0.000
##            homeValue pollutionIndex nBedRooms
## nbr.val      400.000        400.000   400.000
## nbr.na         0.000          0.000     0.000
## skewness       1.057          0.718     0.369
## kurtosis       1.545         -0.134     2.041
## normtest.p     0.000          0.000     0.000
```

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

The data consists of 400 observations (with no missing values) of 11 numeric variables: the ones mentioned above (median price of single-family houses in different neighborhoods and characteristics about those houses and neighborhoods) plus an additional one, not mentioned in the `txt` file:

- `distanceToHighway`: self-explanatory (and probably measured in miles, same as `distanceToCity`).

Based on the kurtosis, skewness (all of them far from zero to a greater or lesser extent) and the p-values of a normality test (all highly significant), none of the variables in the sample is normally distributed. That

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

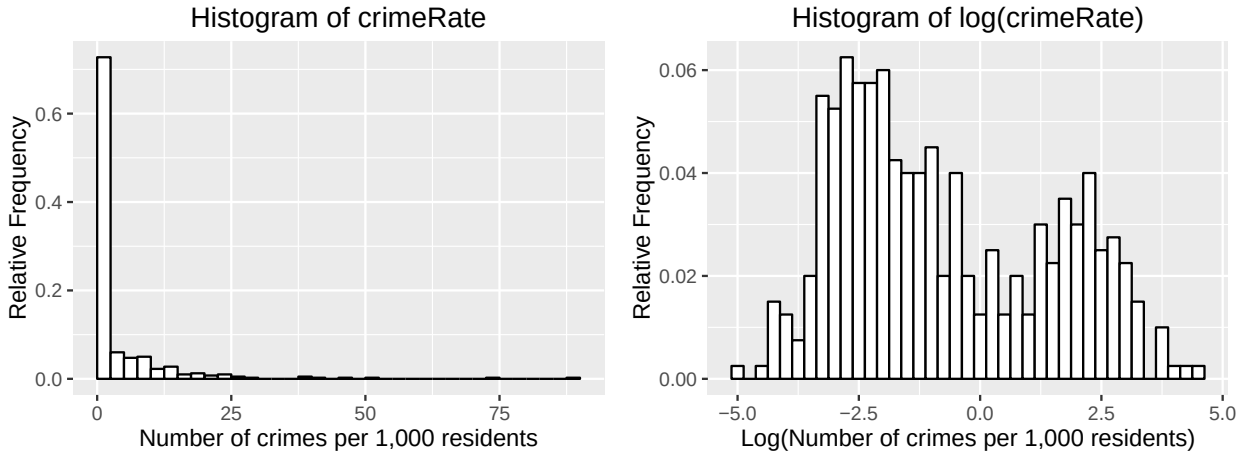

Figure 1: Histograms of the Crime Rate Variable

means they might benefit from transformation (potential transformations will be discussed as the exploratory analysis proceeds).

As shown in the 1st Figure in the following page, the crime rate variable is highly right-skewed, with most neighborhoods having a very low number of crimes per 1,000 residents, and a few having a high number. Using the log does not normalize that variable (the distribution is bimodal).

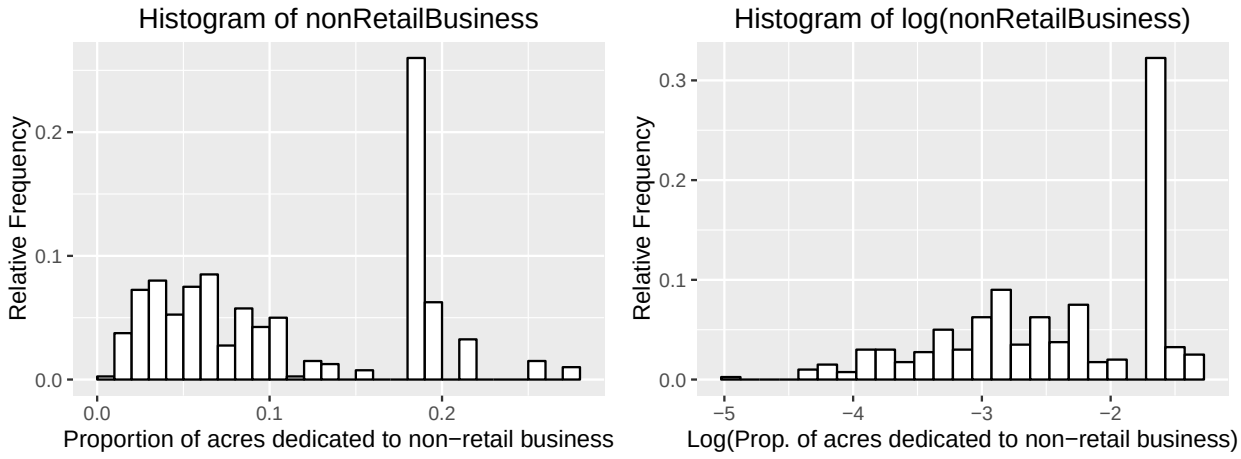

Figure 2: Histograms of Non-Retail Business Percentage Variable

As for the proportion of non-retail business acres per neighborhood, a high proportion of neighborhoods have non-retail business covering about 18% of their area, and most of the rest have much fewer non-retail businesses. A log transformation does not help to normalize this variable either.

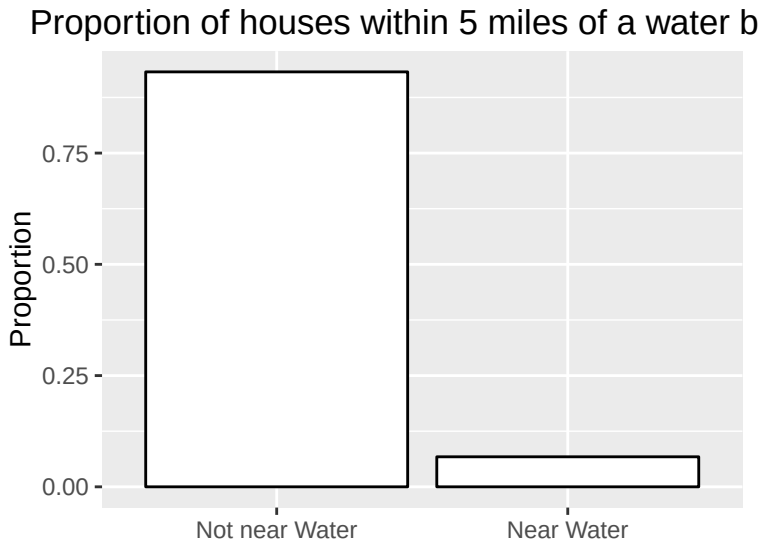

Figure 3: Histogram of Water Variable

Most neighborhoods are not located within 5 miles to a water body. Being near a lake or a river seems highly desirable, in principle, so it's a good candidate to have an effect on home values.

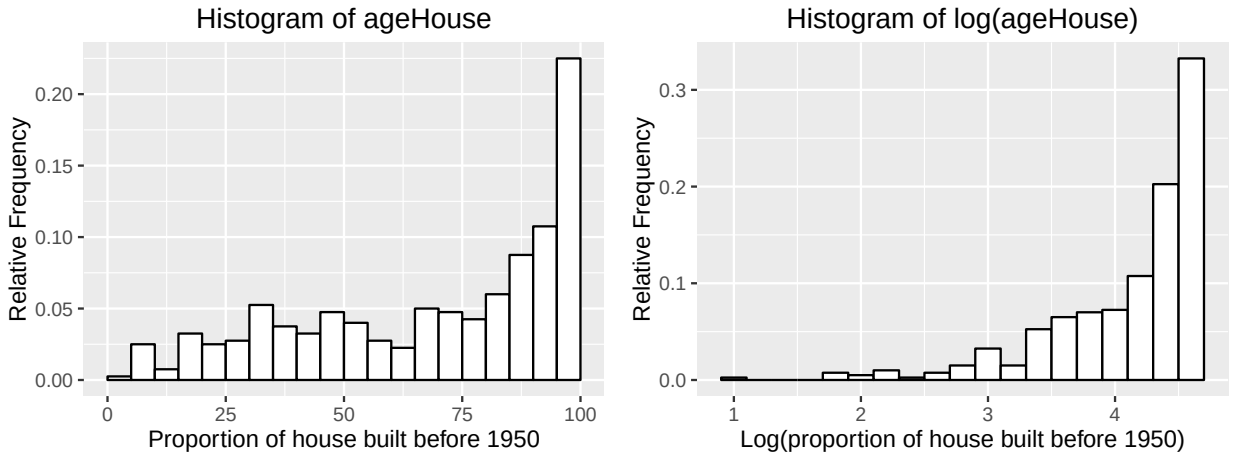

Figure 4: Histogram of House Age

In almost 15% (13.75%) of the neighborhoods, more than 97.5% of the houses were built before 1950. If we lower that percentage of "hold houses" to 75%, that occurs in more than half of the neighborhoods (52.25%). In less than 10% of the neighborhoods (9.25%) only 25% of the houses or less are "old"". Once again, a log transformation does not help to normalize the data.

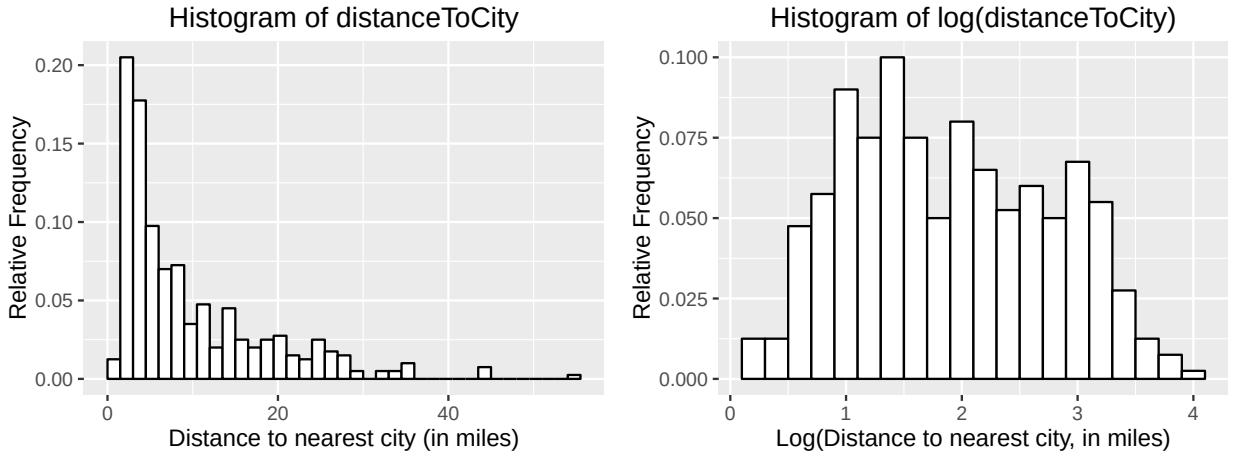

Figure 5: Histogram of Distance to City

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

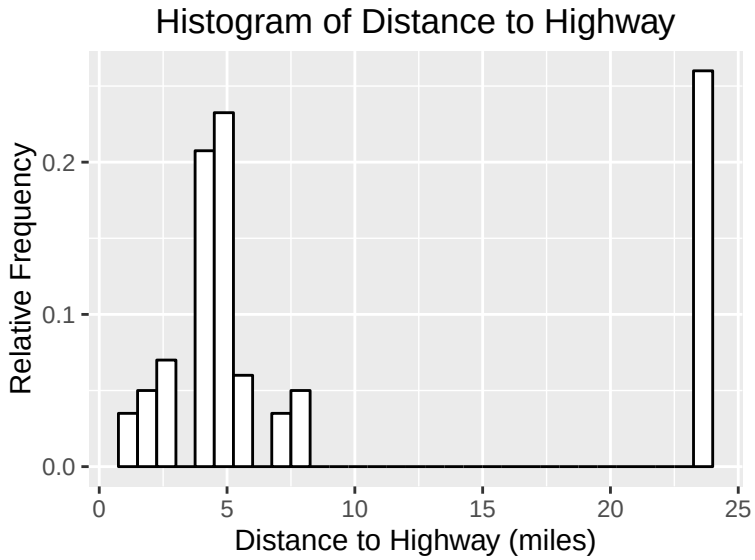

Figure 6: Histogram of Distance to Highway

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.

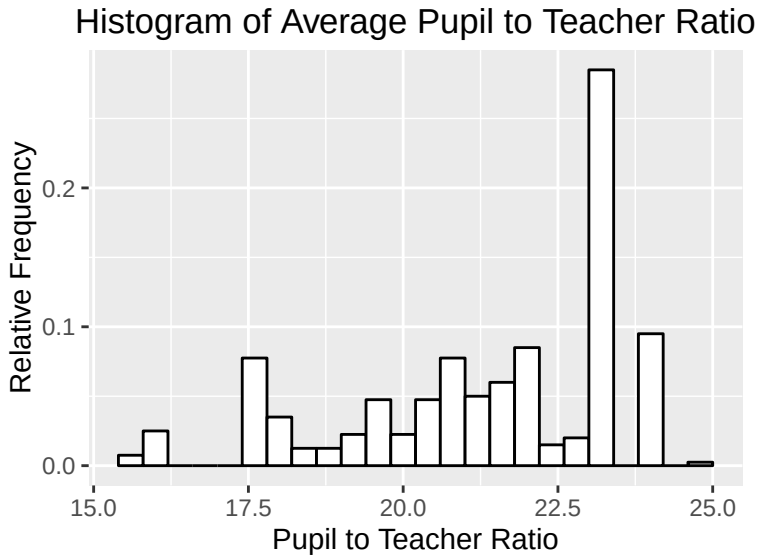

Figure 7: Histogram of Distance to Highway

The pupil to teacher ratio histogram shows a large portion of neighborhoods have average pupil-to-teacher ratios between 17.5 and 23, with a gap between 17.5 and 16, and a small minority that have ratios below 16.

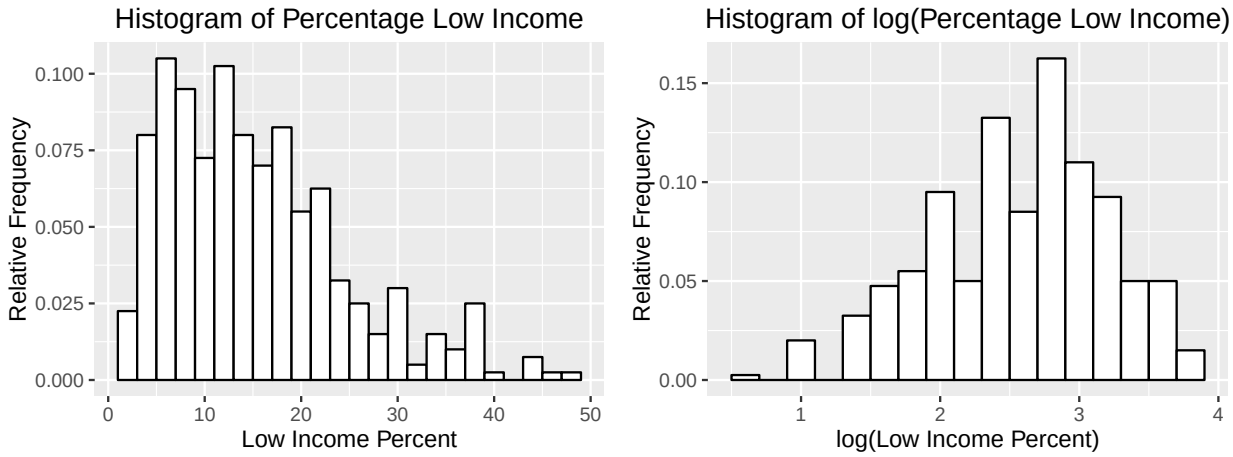

Figure 8: Histogram of Pupil to Teacher Ratio

The percentage of low income housing in a neighborhood 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.

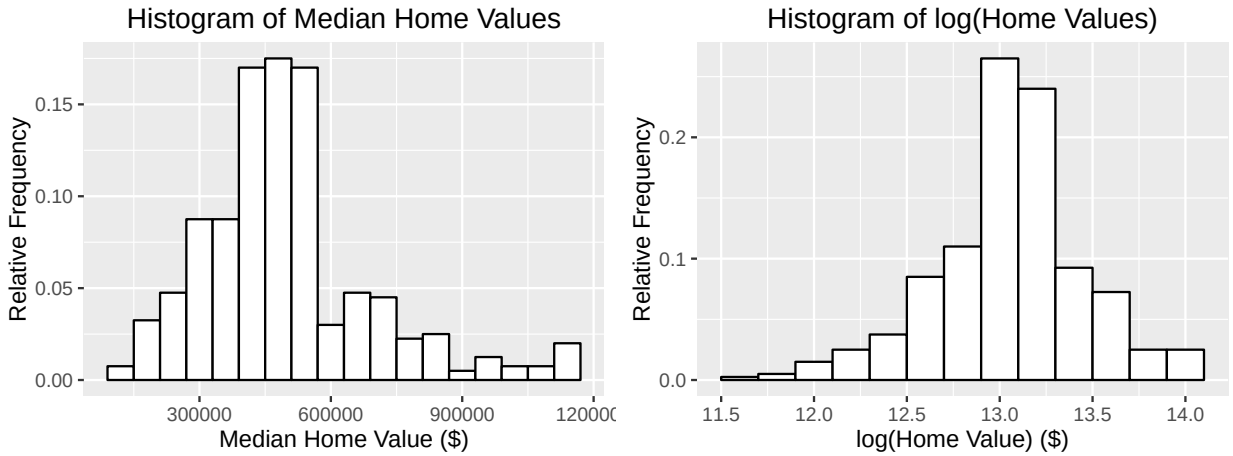

Figure 9: Histogram of Percentage Low Income Housing

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.

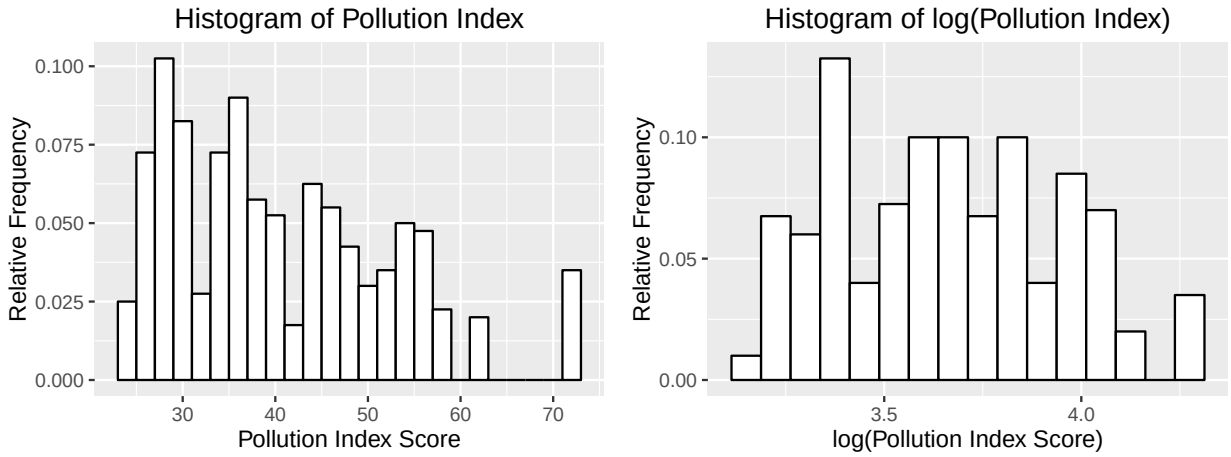

Figure 10: Histogram of Housing Value

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.

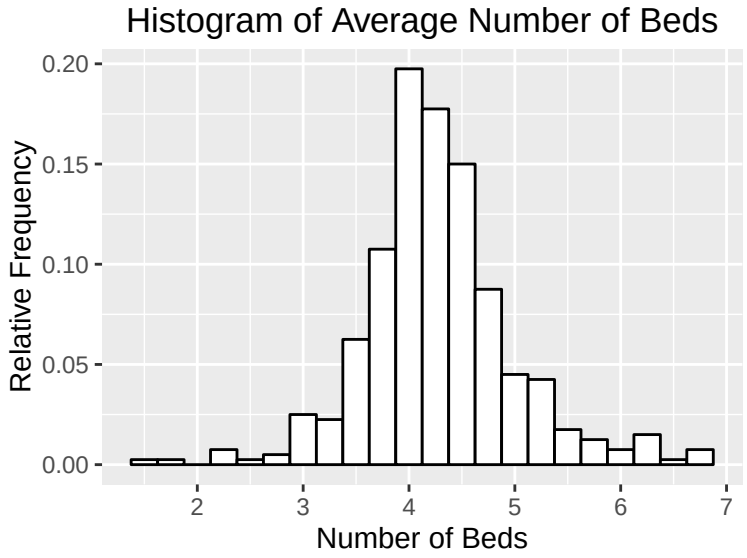

Figure 11: Histogram of Pollution Index

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

After visually inspecting the distribution of each individual variable, we are also interested in how the variables relate, especially to the outcome variable of our model, homeValue. 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.

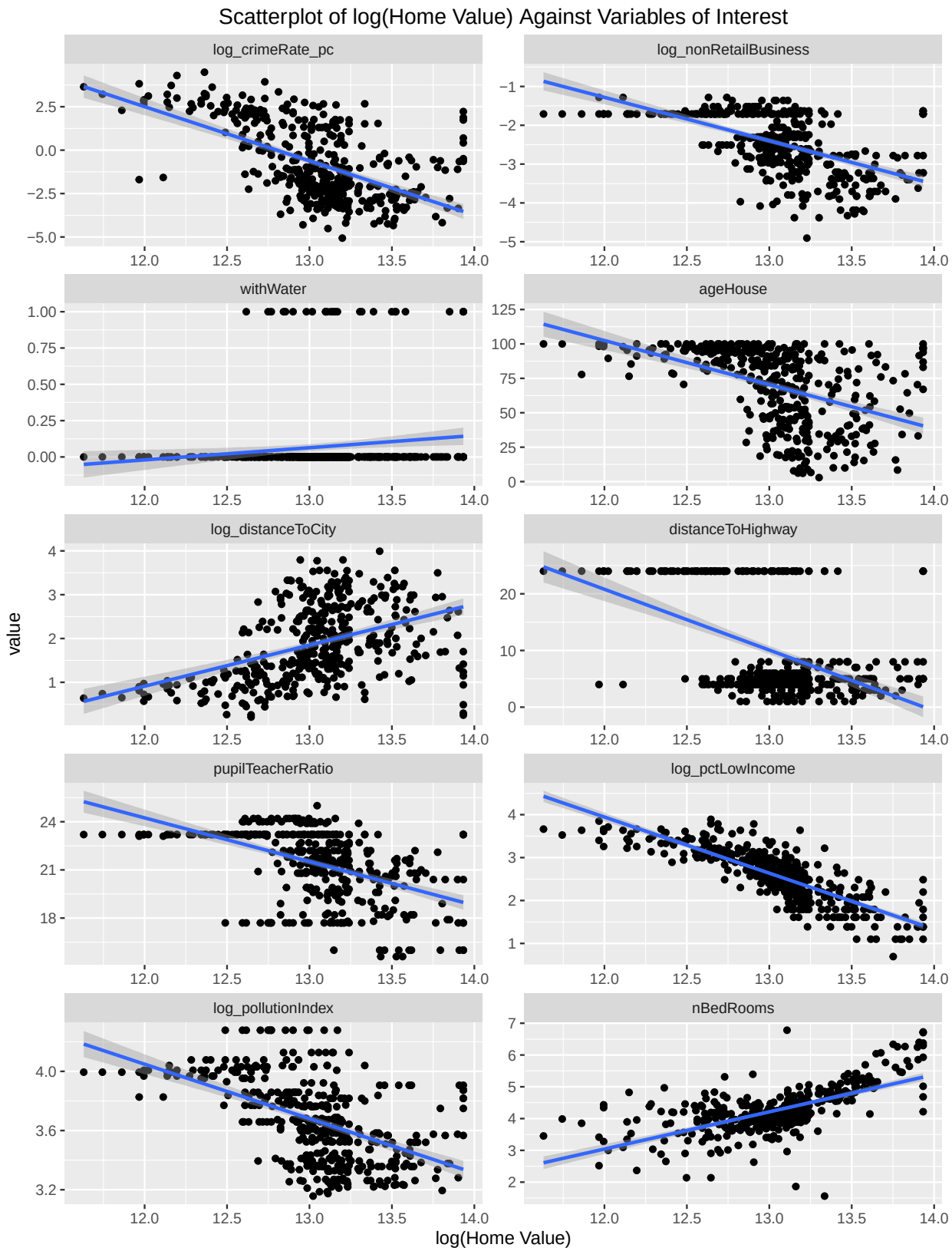

Figure 12: Scatter Plot of Home Value Against the Variables of Interest

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(median house price) than the mean. This fact may complicate our 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 multicollinarity, 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 multicolliniarity, 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 also want to understand how those variables relate to the other variables in the dataset.

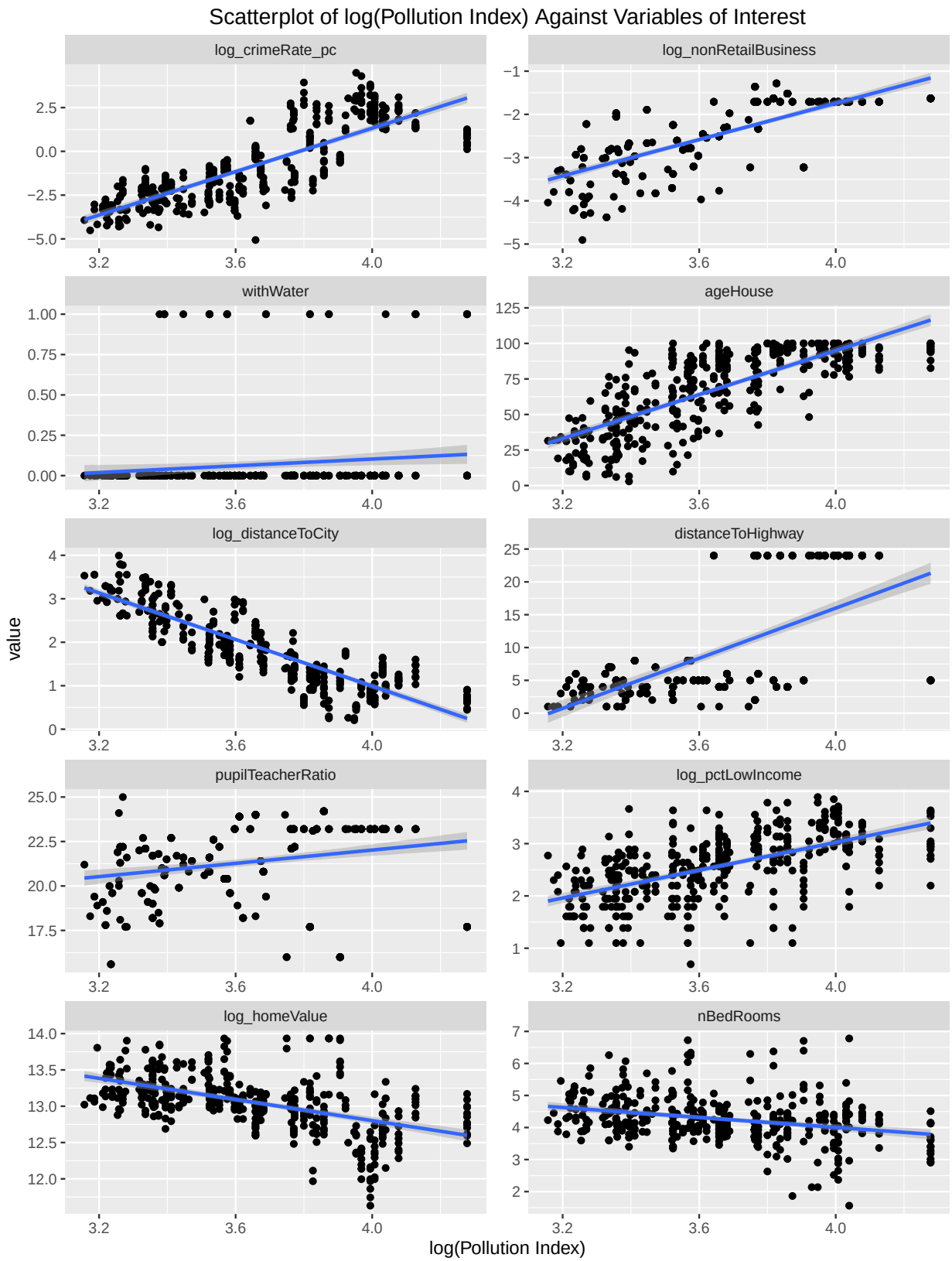

Figure 13: Scatter Plot of Pollution Index Against the Variables of Interest

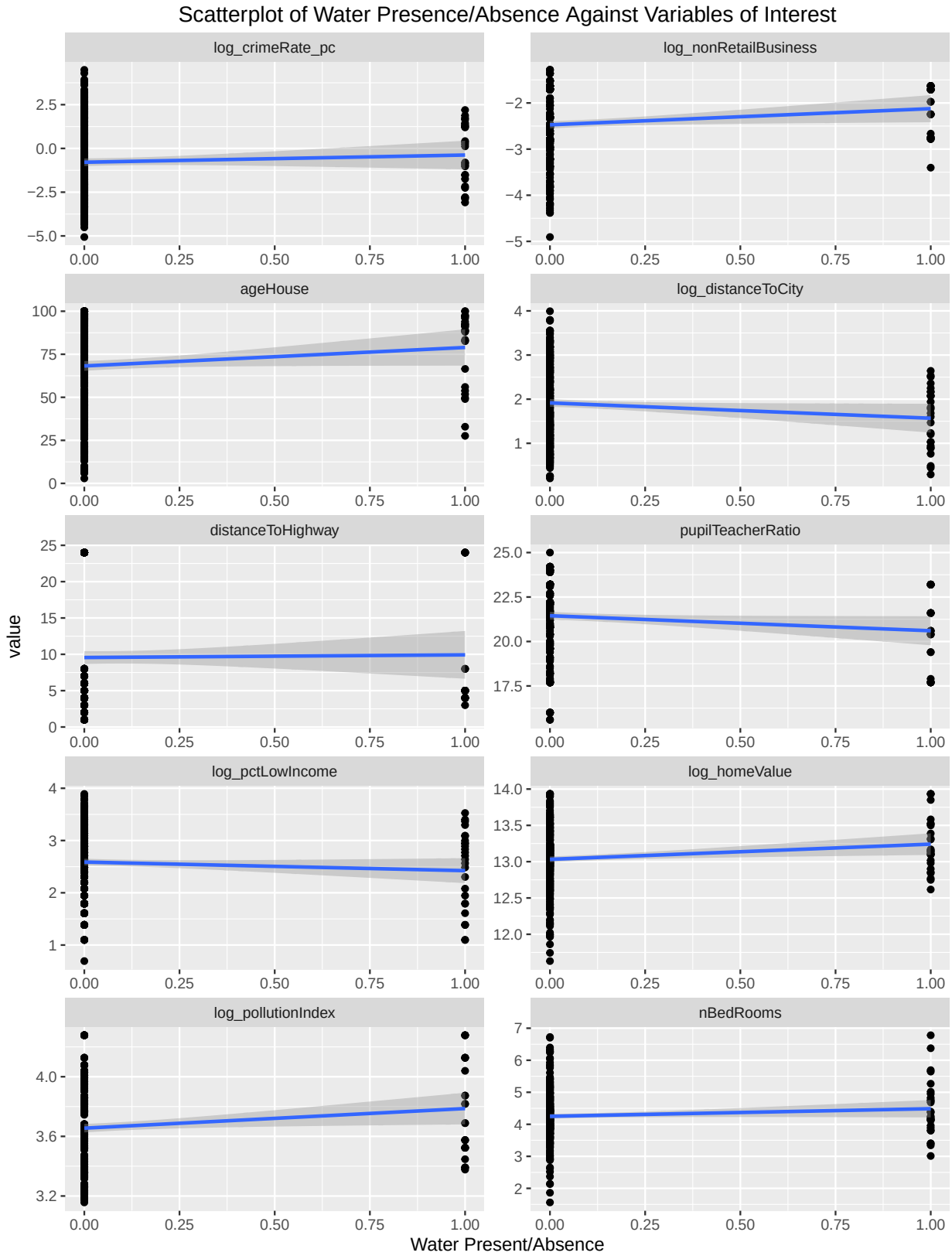

Figure 14: Scatter Plot of Water Presence/Absence Against the Variables of Interest

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 and demographic factors than pollution.

Model selection

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 median home 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:

$$log(homeValue) = \beta_0 + \beta_1 log(pollutionIndex) + \beta_2 withWater + \beta_3 nBedRooms + \epsilon$$

Table 3: Base Model Regression summary

	Dependent variable:
	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^2	0.703
Adjusted R^2	0.700
Residual Std. Error	0.217

·p<0.1; *p<0.05; **p<0.01; ***p<0.001

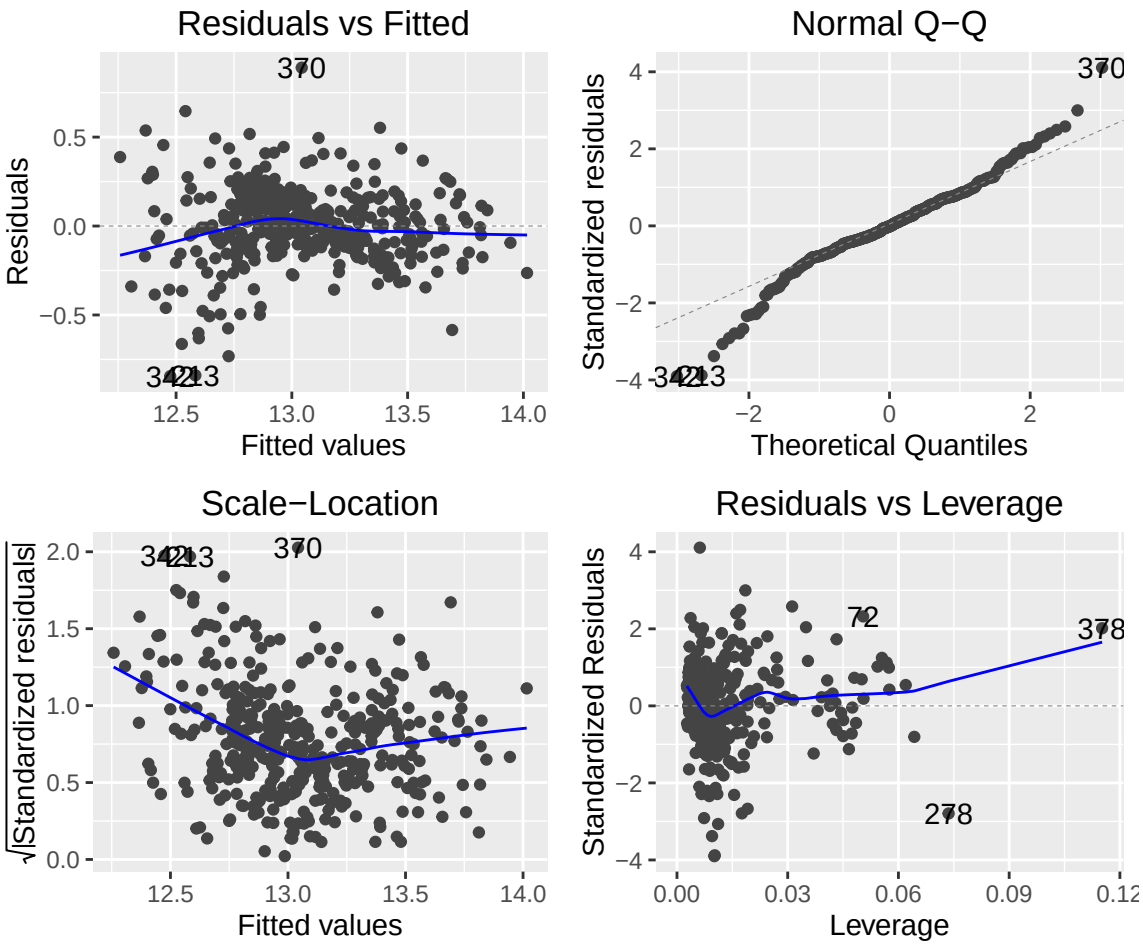

Figure 15: Diagnoistic Plots for Base Regression Model

Looking at the results of the base regression, we see that all coefficients are significant and the R^2 value is 0.7029376, a very basic confirmation that these 4 variables are indeed variables of interest. Looking at the diagnostic plots, we see in the residuals versus fitted plot that the band of values seems to widen out from left to right and then narrow, suggesting that the use of **heteroskadasticity-robust standard errors** is appropriate. The residuals vs leverage plot suggests that some points may represent outliers, although they do not exert leverage on the model that is outside of the norm. The presence of outliers is not surprising in a dataset like this, as we know the values of homes in some neighborhoods are simply much higher or lower due to conditions that may not be perfectly captured by the dataset. **Since the presence of outliers is simply a reality in datasets relating to housing, we do not argue in favor of any outlier removal procedure.**

Starting with our base model, we are interested in selecting a final 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 that, due to multiple comparisons, 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 sub-sample 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 Diagnoistics 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, **the model we select for predicting home values is the base model with three additional paramaters: pupil to teacher ratio, log(distance to city), and log(crime rate per capita)** . This model has the form:

$$log(homeValue) = \beta_0 + \beta_1 log(pollutionIndex) + \beta_2 withWater + \beta_3 nBedRooms +$$

$$\beta_4 pupilTeacherRatio + \beta_5 log(distanceToCity) + \beta_6 log(crimeRate_pc) + \epsilon$$

Table 5: Perfered Model Regression summary

	Dependent variable:
	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^2	0.741
Adjusted R^2	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. Compared to our original base model, controlling for the additional variables has resulted is a large (negative) increase in the coefficient for pollution index and a small decrease in the coefficient for absence/presence of water. 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.

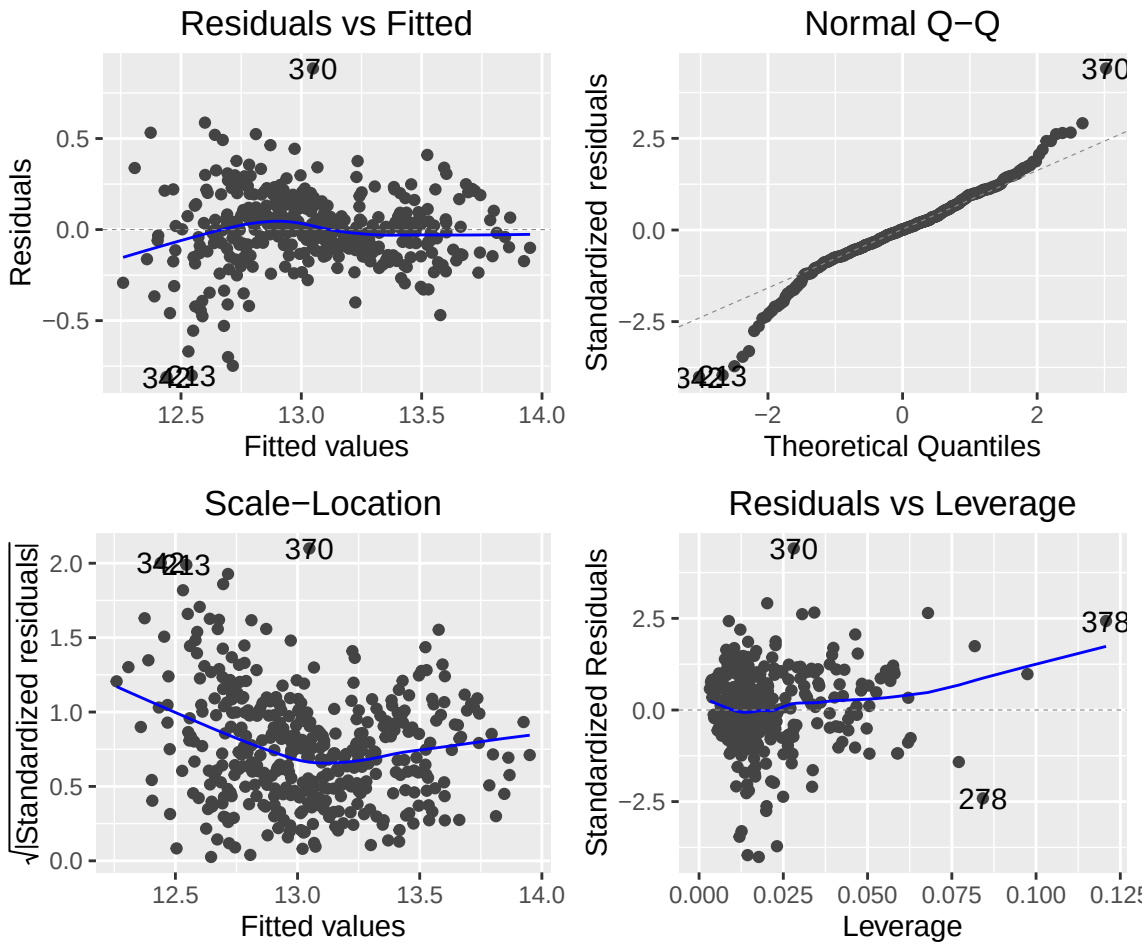

Figure 16: Diagnoistic Plots for Selected Regression Model

The residual vs fitted plot for our selected model shows a distinct tunneling of values from left to right, suggesting heteroskedasticity and the need to use heteroskedasticity-robust standard errors.

The primary issue in estimating home value using this dataset is the likelihood of omitted variable bias. For instance, we could imagine that the variable of interest, pollutionIndex, has a causal path to homeValue via intermediary variables. Perhaps total green space is an important variable for influencing people's beliefs about the environmental quality of a neighborhood, and people 'estimate' the pollutionIndex of a place based on the degree of green space. Then the true effect of pollution on home values is mediated through green space. In this case, our estimates of the coefficient for pollutionIndex would be biased because we are not controlling for green space. Given the likely complexity of the casual pathway between neighborhood desirability and home value, it is highly likely that there are variables relating to the 'desirability' of a neighborhood that have an effect on the value of homes that are not captured by this dataset.

Using IV to estimate coefficient for pollution index

Given that we are interested in how environmental factors influence home prices, we may worry that the coefficient for pollution index is biased due to considerations of omitted variables outlined above. We argue that the the proportion of non-retail businesses acres is a source of random variation with regards to home values (controlling for covariates), and thus represents a possible instrument for pollution index.

In order for non-retail business acreage to be a suitable IV for pollution index score, two conditions must

hold: * Non-retail business acreage must be correlated with pollution index * Non-retail business acreage must be uncorrelated with the error term in the regression model.

Given that non-retail businesses include industrial operations like manufacturing, which are sources of pollution, it is not surprising that the two variables are positively correlated (See scatterplot above). For the IV approach to be causal, we need to argue persuasively that non-retail business percentage is not endogenous to the model, that is to say that home values do not depend directly on non-retail business acreage, conditional on the covariates in the model. Here, we argue that the main proxy of non-retail business would be urban/rural, given that rural areas tend to have more non-retail businesses, both in terms of total number of business and the proportion of acreage occupied by those businesses, and since we are controlling for distance to city, then, on average, non-retail business percentage is not a determinant of house valuation. This assumption is generally not testable, and can be seen as more or less persuasive depending on the IV proposed. In this case, one can imagine enough situations where the closeness to certain kinds of non-retail business (heavy manufacturing, open-pit mining) would have an impact on the valuation of a home that I wouldn't find non-retail business percentage to be a completely persuasive instrument for pollution index.

Performing a two-step regression process using non-retail business percentage as an instrument for pollution index score produces the coefficients presented in the table below:

Table 6: Instrumental Variable Regression summary

	Dependent variable:
	log(home value)
log(PollutionIndex)	-0.052
	(0.083)
Water Absence/Presence	0.107^{*}
	(0.043)
Number of Bedrooms	0.096^{**}
	(0.031)
log(Percentage Low Income Housing)	-0.367^{***}
	(0.038)
Pupil to Teacher Ratio	-0.021^{***}
	(0.005)
log(Distance to City)	-0.068^{**}
	(0.021)
log(Average Crime Rate)	-0.041^{***}
	(0.010)
Constant	14.306^{***}
	(0.410)
F Statistic	115.442^{***}
df	7; 392
Observations	400
R^2	0.738
Adjusted R^2	0.733
Residual Std. Error	0.205

·p<0.1; *p<0.05; **p<0.01; ***p<0.001

The coefficient for pollution index using an instrumental variable approach is considerably smaller than when using straightforward OLS regression. Depending on how you interpret the strength of the instrument, this result could be used to argue in favor of the hypothesis that people are generally not aware of the pollution index of an area and it has a small effect on home valuation, while omitted variables may give the appearance of greater importance for pollution index scores. However, I think there are strong arguments to be made against this particular instrument, and when combined with the large standard errors of the estimate, **we**

can be fairly confident that there is a negative effect of pollution index on home values, but the magnitude of the effect remains uncertain.

One hypothesis that may be of particular interest to the think-tank is whether or not the effect of pollution is the same for neighborhoods with a lot of low income housing and neighborhoods with very little low income housing. Given the bi-model distribution of low income housing noted in the exploratory analysis, this dataset seems well suited to testing this hypothesis. Formally, we are interested in the interaction term ($\beta_7 = log(PollutionIdex) \cdot log(PercentageLowIncomeHousing)$) and the hypothesis:

$$H_0 : \beta_7 = 0$$
$$H_A : \beta_7 \neq 0$$

Adding the interaction term β_7 to our preferred model yields:

Table 7: Interaction Effect Regression summary

	Dependent variable:
	log(home value)
log(PollutionIndex)	1.049***
	(0.230)
Water Absence/Presence	0.078*
	(0.038)
Number of Bedrooms	0.089***
	(0.027)
log(Percentage Low Income Housing)	1.385***
	(0.270)
Pupil to Teacher Ratio	−0.029***
	(0.005)
log(Distance to City)	−0.088***
	(0.026)
log(Average Crime Rate)	−0.023*
	(0.010)
log_PI x log_LIH	−0.481***
	(0.077)
Constant	10.601***
	(0.876)
F Statistic	146.257***
df	8; 391
Observations	400
R^2	0.776
Adjusted R^2	0.771
Residual Std. Error	0.190

·p<0.1; *p<0.05; **p<0.01; ***p<0.001

The coefficient for the interaction effect represents a practically and significantly difference in the effect of pollution for neighborhoods with little low income housing and neighborhoods with a lot of low income housing ($\beta_7 = -0.481$, s.e. $= 0.062$, p $= 0$). For neighborhoods with a high proportion of low income housing, pollution has a considerably larger negative effect on home values, representing about a 5% decrease relative to neighborhoods with little low income housing.

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 you group is part of a data science team in an appreal 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 exericse, 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.
