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 certian 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.

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.

```r
houseValue <- read.csv('houseValueData.csv', header = TRUE)
```

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

Table 1: Summary statistics of house values and features

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

Before plotting the distribution of each variable, we run a regression of the price value on all the other variables (after standardizing all to better compare their effects). This 1st regression model may not be the most appropriate one (data are not transformed, we won't check residulas, there may be multicollinearity…), but for the moment we just want to check if all the relationships make sense.

```
houseValue.std <- houseValue %>% mutate_each(funs(scale))
model.1 <- lm(homeValue ~ ., houseValue.std)
names(sort(abs(model.1$coefficients), decreasing = T))
```

```
##  [1] "pctLowIncome"      "nBedRooms"         "pupilTeacherRatio"
##  [4] "pollutionIndex"    "distanceToCity"    "distanceToHighway"
##  [7] "crimeRate_pc"      "withWater"         "nonRetailBusiness"
## [10] "ageHouse"          "(Intercept)"
```

As shown in the table in the next page, the variables that have a stronger effect on the house value are `pctLowIncome` (a one standard deviation increase in it—which translates in a 9.3 point increase in the percentage of low income household in the neighborhood—decreases price by 0.37 standard deviation), then `nBedRooms` (a one standard deviation increase in it—which corresponds to 0.72 additional bedroomm, on average—increases price by 0.34 standard deviation), and so on. The variable that has a lower effect on the house value is the `ageHouse` (the roportion of houses built before 1950): though it may seem surprising that the effect is positive, it is not significant(ly different from zero), that could make sense: it's the age of each individual house, and not the average in the neighborhood, which should affect the price (and the age is not necessarily a bad feature: mansions of the 19th century are certainly more valuable than low-priced small houses, no matter how new they may be). The two variables that are significant only at the 10% level are `nonRetailBusiness` and `distanceToHighway`. `withWater` is significant at the 5% level, and `crimeRatio` at the 5% level; all these variables have the lowest effects (and the rest are significant at the 1% level. But what we matter most, at this early stage, it's the sign of the coefficients; and all of them make sense:

- A higher crime rate,
- more acres dedicated to non-retail businesses,
- farther distances to the nearest city,
- higher pupil-teacher ratios,
- a higher percentage of low-income households, and
- more pollution

all lead (other factors being equal) to lower house values. Similarly,

- closeness to a water body,
- a higher proportion of houses built before 1950 (already explained), and
- farther distance to the nearest highway

decrease the house value, on average.

As usual, all the standard errors are heteroskedasticity-robust.

Table 2: Regression summary (with standardized variables)

	Dependent variable:
	Median price ($) of single-family house
Crime rate per capita	-0.103^{**}
	(0.032)
Proportion of non-retail business acres	$-0.075^{\cdot}$
	(0.045)
Water body less than 5 miles away	0.080^{*}
	(0.040)
Proportion of houses built before 1950	0.026
	(0.059)
Distance (miles) to nearest city	-0.210^{***}
	(0.045)
Distance (miles) to nearest highway	$0.106^{\cdot}$
	(0.057)
Average pupil-teacher ratio	-0.237^{***}
	(0.032)
Percentage of low-income households	-0.369^{***}
	(0.090)
Pollution index (0-100)	-0.220^{***}
	(0.059)
Average number of bedrooms	0.345^{***}
	(0.075)
Constant (intercept)	0.000
	(0.027)
F Statistic	74.444^{***}
df	10; 389
Observations	400
R^2	0.724
Adjusted R^2	0.717
Residual Std. Error	0.532

·p<0.1; *p<0.05; **p<0.01; ***p<0.001

Next we plot the histogram (or bar chart) of the variables (and, in many cases, their log; using the log of a variable may make sense to narrow its range, satisfy the CLM assumptions more closely—e.g., reducing the skewness of the residuals—, model a non-linear—e.g., exponential—relationship, etc.). In principle, we don't care if the distribution of the regressors is normal; it's the distribution of the residuals which has to be (and that's not the strongest CLM assumption).

We begin with the dependent variable: `homeValue` is slightly right-skewed, with most values around the mean of approximately $500,000 and the right tail extending to the maximum of $1,125,000. A log transformation produces a distribution closer to normal, and we'll use it (besides, it makes a lot of sense for this regressand: the meaning of the coefficients—if not too high—will be a percentage change in the value).

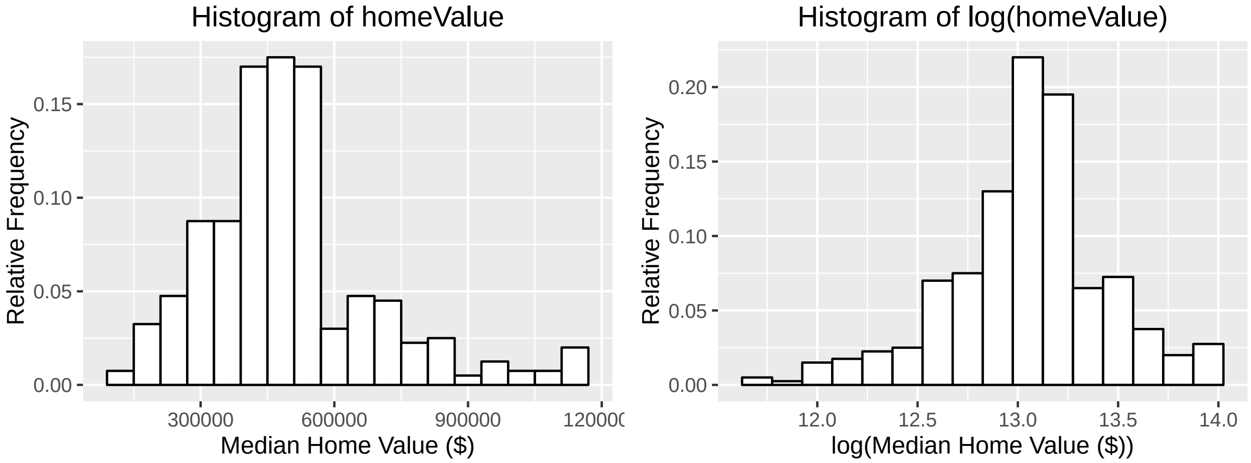

Figure 1: Histograms of home Value and its log

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 perfectly normalize that variable (the distribution is bimodal), but does a good enough job to use the log in this case.

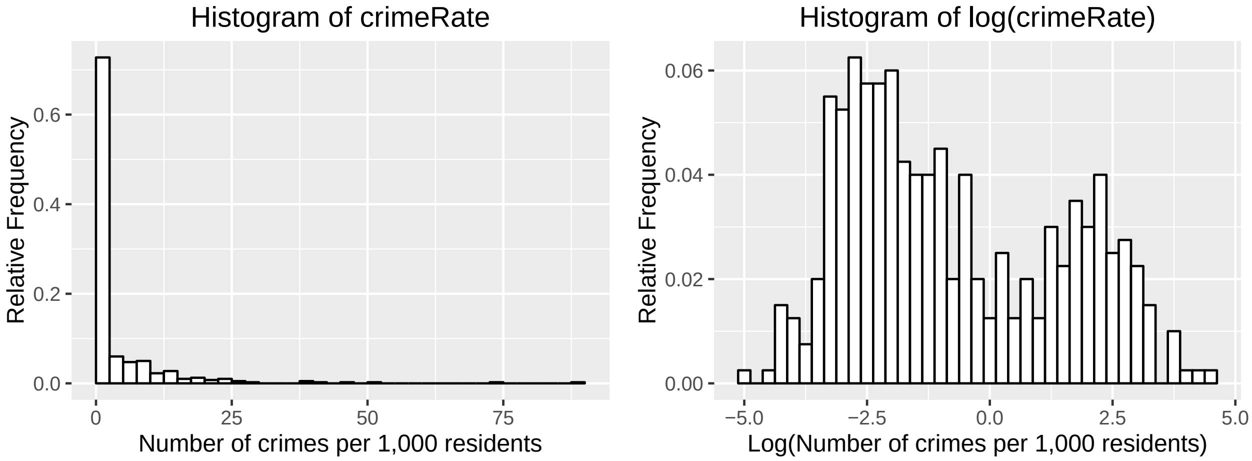

Figure 2: Histograms of Crime Rate and its log

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.

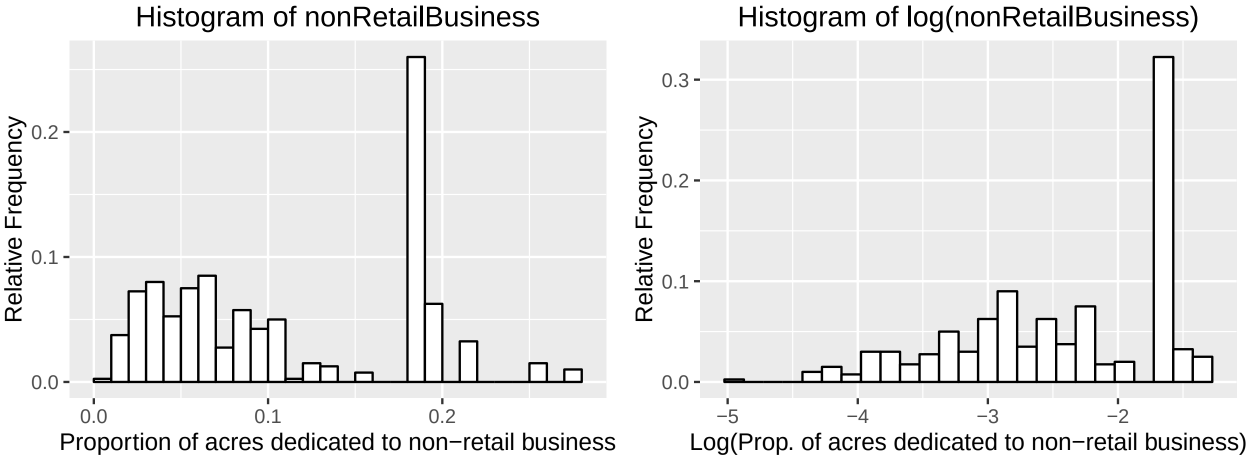

Figure 3: Histograms of non-retail Business acres and its log

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.

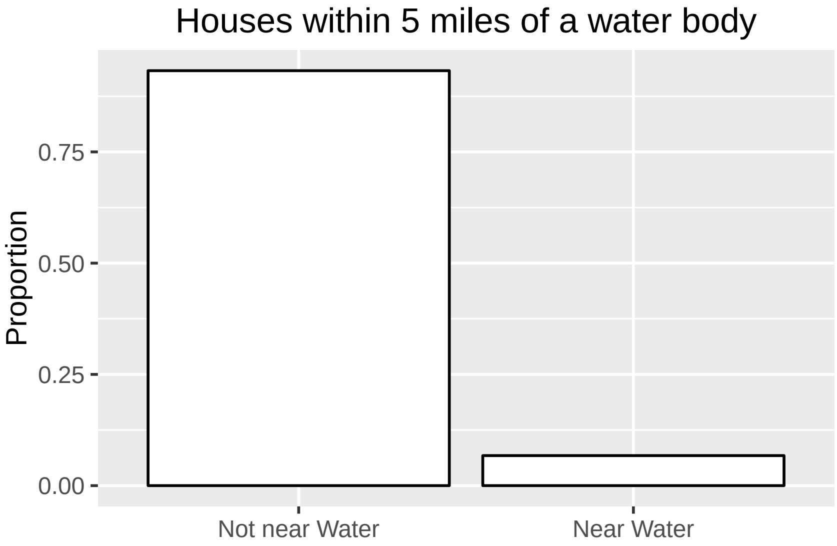

Figure 4: Char bart of proportion of houses within 5 miles of a water body

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.

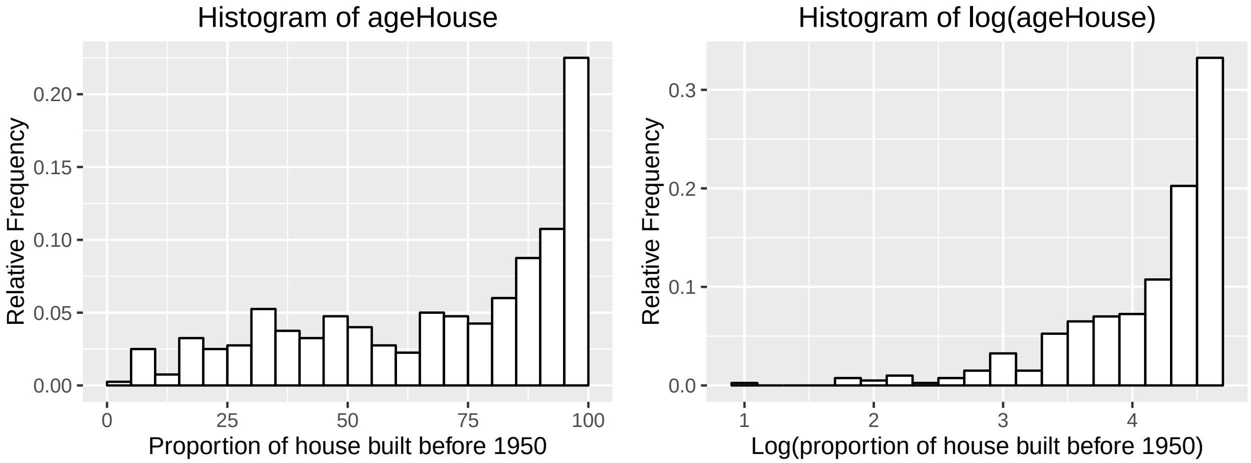

Figure 5: Histograms of Proportion of houses built before 1950 and its log

The distance from a neighborhood to nearby cities has a right-tailed distribution, with more than half of the neighborhoods (66.5%) within 10 miles of a city, and just 1% of them more than 40 miles away. Log transformation of this variable removed the skewness of the distribution and produced a more approximately normal distribution.

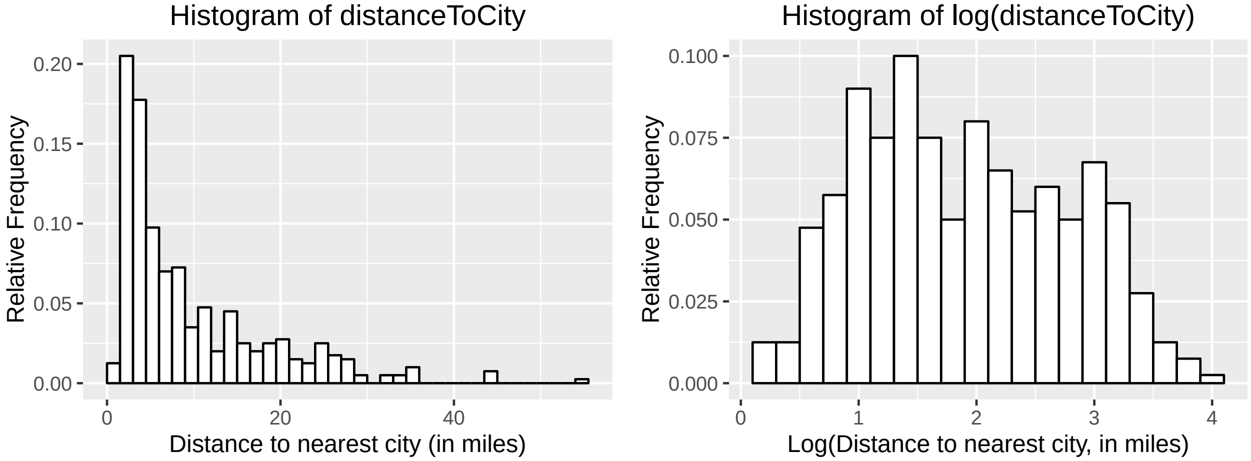

Figure 6: Histogram of distance to nearest city and its log

Since we'll use the log for this variable, it seems appropriate to also use it for the next one (though the log does not normalize the data, as explained in the next page).

26% of the neighborhoods were exactly 24 miles from the nearest highway. There are only 9 unique values (the other possible distances go from 1 to 8 miles), which suggests us thinkg that this variable was probably rounded and factorized (losing part of its explanatory value). That makes the distribution to be strongly bimodal…even if it the log of the variable is used.

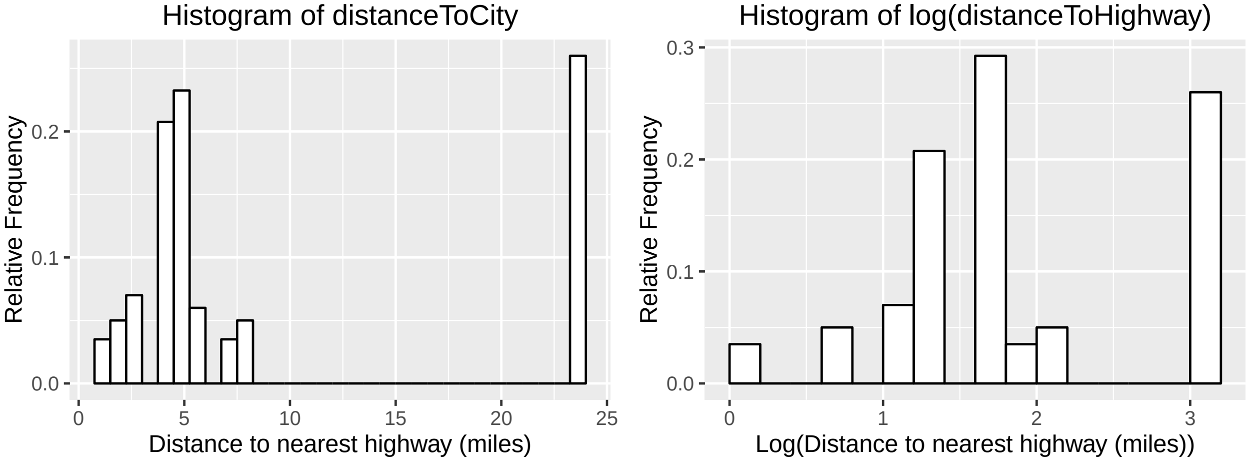

Figure 7: Histogram of distance to nearest highway and its log

The histogram of the average pupil-teacher ratio is left-skewed (and still is after using the log): many neighborhoods (27.5% of them has a ratio of 23.2 pupils per teacher; the other values (ranging from 15.6 to 25; and almost all lower) are approximately uniformly distributed). The distribution of the log of this variable looks pretty much the same.

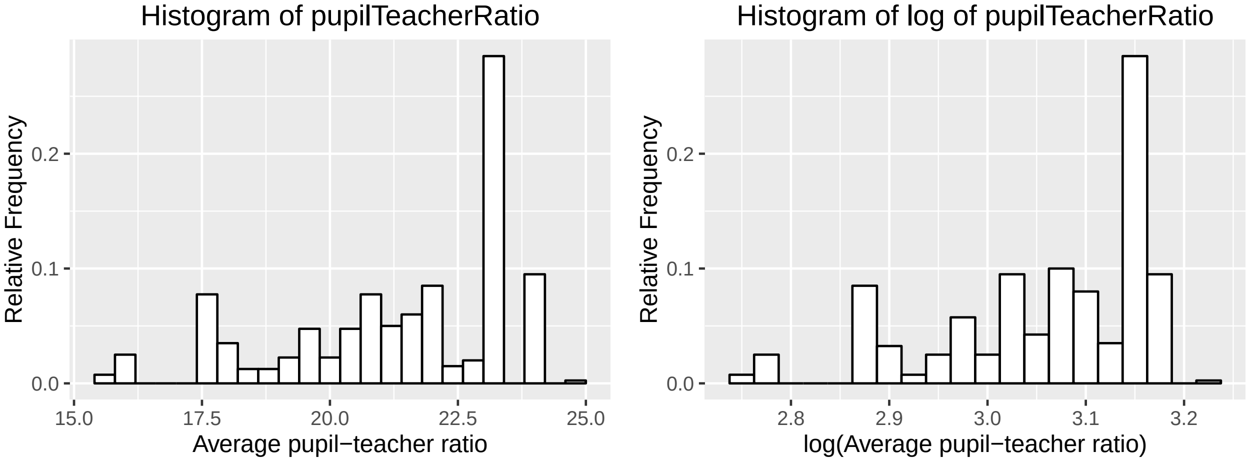

Figure 8: Histograms of Average pupil-teacher ratio and its log

The percentage of low-income households in a neighborhood displays a slightly right-skewed distribution. Since this is a percentage, keeping the data untransformed maintains the meaning of the regression coefficient: a unit increase means a 1% increase, which will result in a $100 \cdot beta_i$ increase (or decrease) in the home value (since we'll use the log of it).

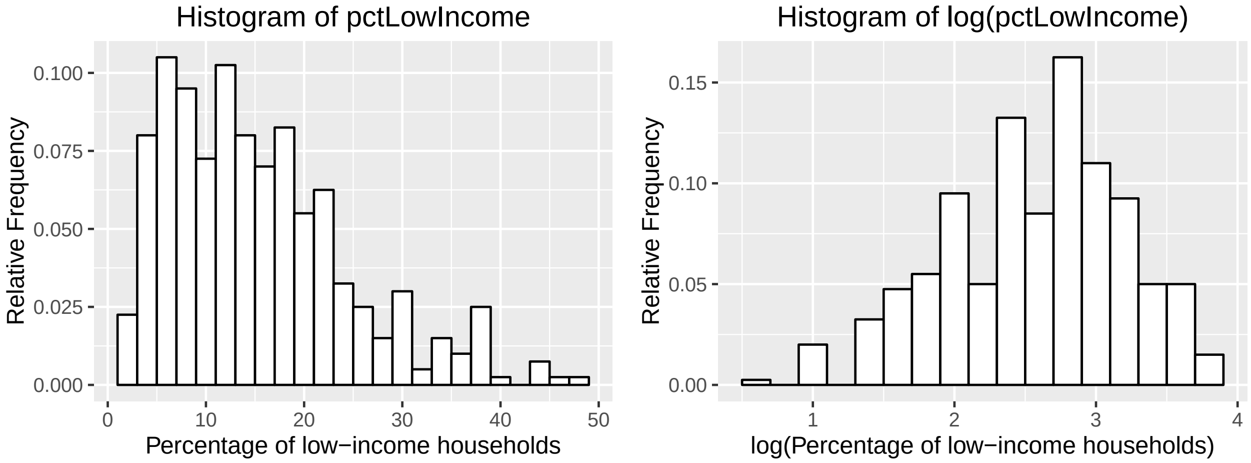

Figure 9: Histogram of percentage of low-income households and its log

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.

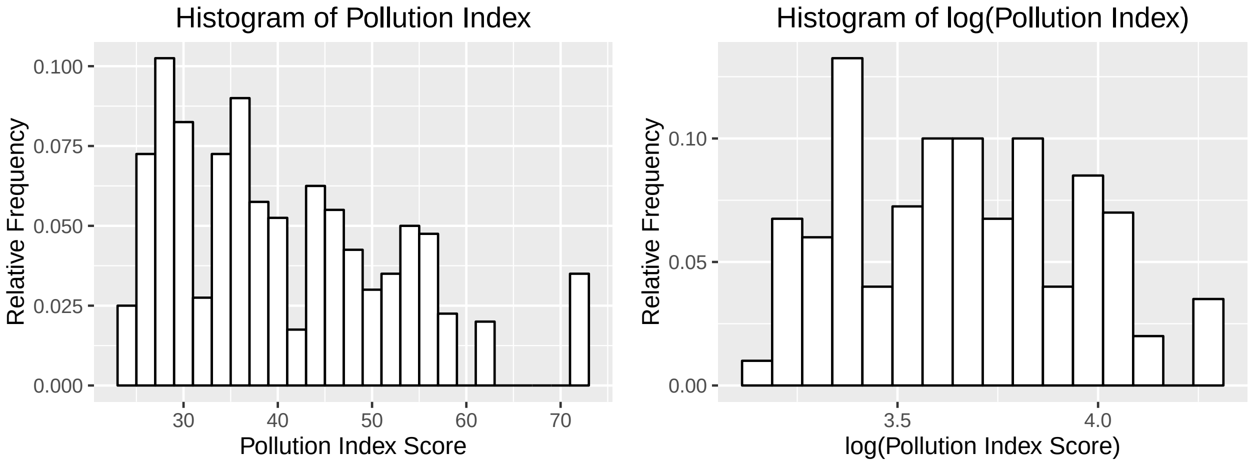

Figure 10: Histogram of percentage of low-income households and its log

The average number of bedrooms (with a mean of 4.3) is approximately normally distributed.

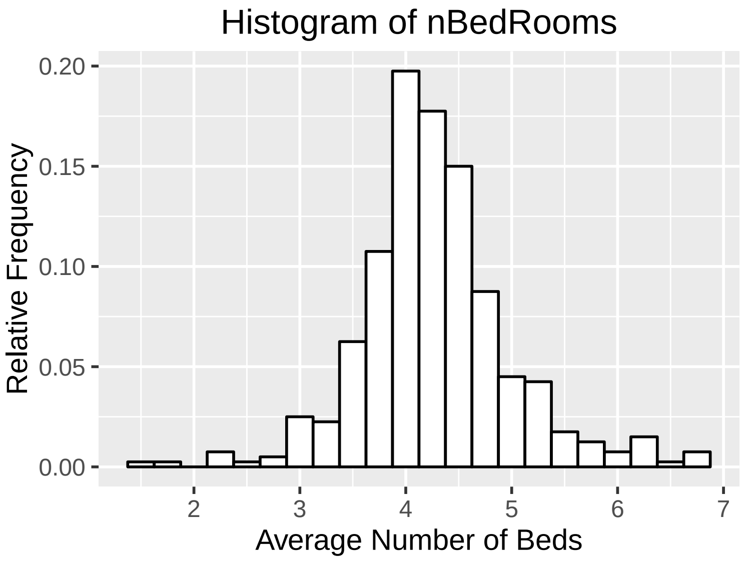

Figure 11: Histogram of average number of bedrooms

After visually inspecting each individual variable, we are also interested in how the other variables relate to the variable of interest, `homeValue`. First we apply the log to the (4) variables we previously mentioned (and change their names accordingly) and then we build a scatterplot matrix and run a simple regression of all the independent variables on `log_homeValue`.

```
vars_to_log <- c("homeValue", "crimeRate_pc", "distanceToCity",
                 "distanceToHighway")
houseValue.2 <- houseValue %>% mutate_each_(funs(log), vars_to_log) %>%
  setNames(c(paste0("log_", names(.)[1]), names(.)[2:4],
             paste0("log_", names(.)[5:6]), names(.)[7:8],
             paste0("log_", names(.)[9]), names(.)[10:11]))
```

```
regressors <- names(houseValue.2)[c(1:8, 10:11)]
model.list <- lapply(1:length(regressors), function(i)
  lm(as.formula(paste("log_homeValue ~", regressors[i])), houseValue.2))
```

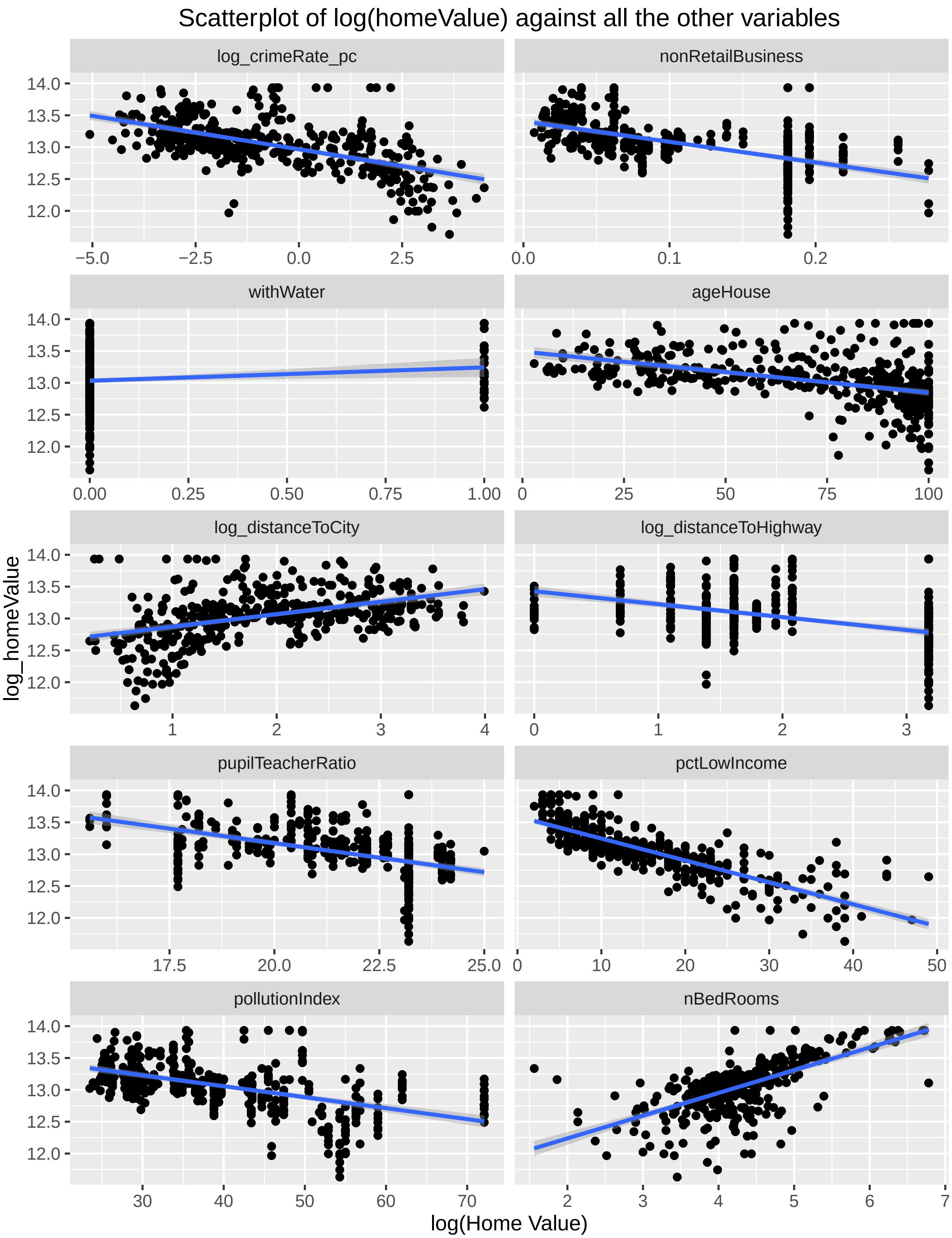

Figure 12: Scatter Plot of log(homeValue) against all the other variables

Table 3: Simple regression summary of log(homeValue)

	Dependent variable:									
	log(Median price ($) of single-family house)									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
log(crime rate)	−0.105***									
	(0.008)									
Prop. NR business		−3.230***								
		(0.253)								
Water<5 miles			0.210**							
			(0.077)							
Prop. houses<1950				−0.006***						
				(0.001)						
log(dist. city)					0.196***					
					(0.023)					
log(dist. highway)						−0.203***				
						(0.023)				
Avg p-t ratio							−0.091***			
							(0.008)			
% low-inc. house								−0.034***		
								(0.002)		
Pollution									−0.017***	
									(0.002)	
no. bedrooms										0.357***
										(0.029)
Constant	12.966***	13.406***	13.032***	13.491***	12.675***	13.427***	14.998***	13.590***	13.745***	11.524***
	(0.019)	(0.026)	(0.020)	(0.035)	(0.054)	(0.042)	(0.174)	(0.029)	(0.057)	(0.129)
F Statistic	169.087***	162.378***	7.469**	121.735***	73.192***	79.045***	125.722***	306.988***	126.672***	148.118***
df	1; 398	1; 398	1; 398	1; 398	1; 398	1; 398	1; 398	1; 398	1; 398	1; 398
Observations	400	400	400	400	400	400	400	400	400	400
R^2	0.328	0.321	0.018	0.207	0.184	0.195	0.249	0.656	0.263	0.417
Adjusted R^2	0.326	0.319	0.015	0.205	0.182	0.193	0.247	0.655	0.261	0.416
Residual Std. Error	0.326	0.328	0.394	0.354	0.359	0.357	0.345	0.233	0.341	0.303

·p<0.1; *p<0.05; **p<0.01; ***p<0.001

The Table in the previous page shows that each variable, when not controlling for any other, is a good predictor of the log of the median house value (much better than just the mean of it). This fact may complicate our effort to select one of these variables as an instrument, as being unrelated to the outcome variable is one condition of the exclusion restriction for an IV 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.

The only coefficients that change their sign when running a simple regression (as opposed to a multiple regression when all independent variables are used) are the distance to the nearest city and highway, respectively. As a result, if we don't control for other factors, a further distance to the nearest city increases the median value of a house, and the opposite occurs with the nearest highway. This is not due to the use of logarithms (the same happens if we don't apply them to either the home value or the distance) but because of the inclusion of other variables, some of them which may be related to those 2 variables.

Since we are particularly interested on 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. As shown in the following 4 pages, `pollutionIndex` is highly related (positively or negatively) with all the other independent variables, while `withWater` is only related with a few of them (`pollutionIndex` itself, `ageHouse`, and `nonRetailBusiness`, at the 5% level). The former fact provides evidence that estimating the effects of pollution on home values requires controlling for a number of potential confounding variables.

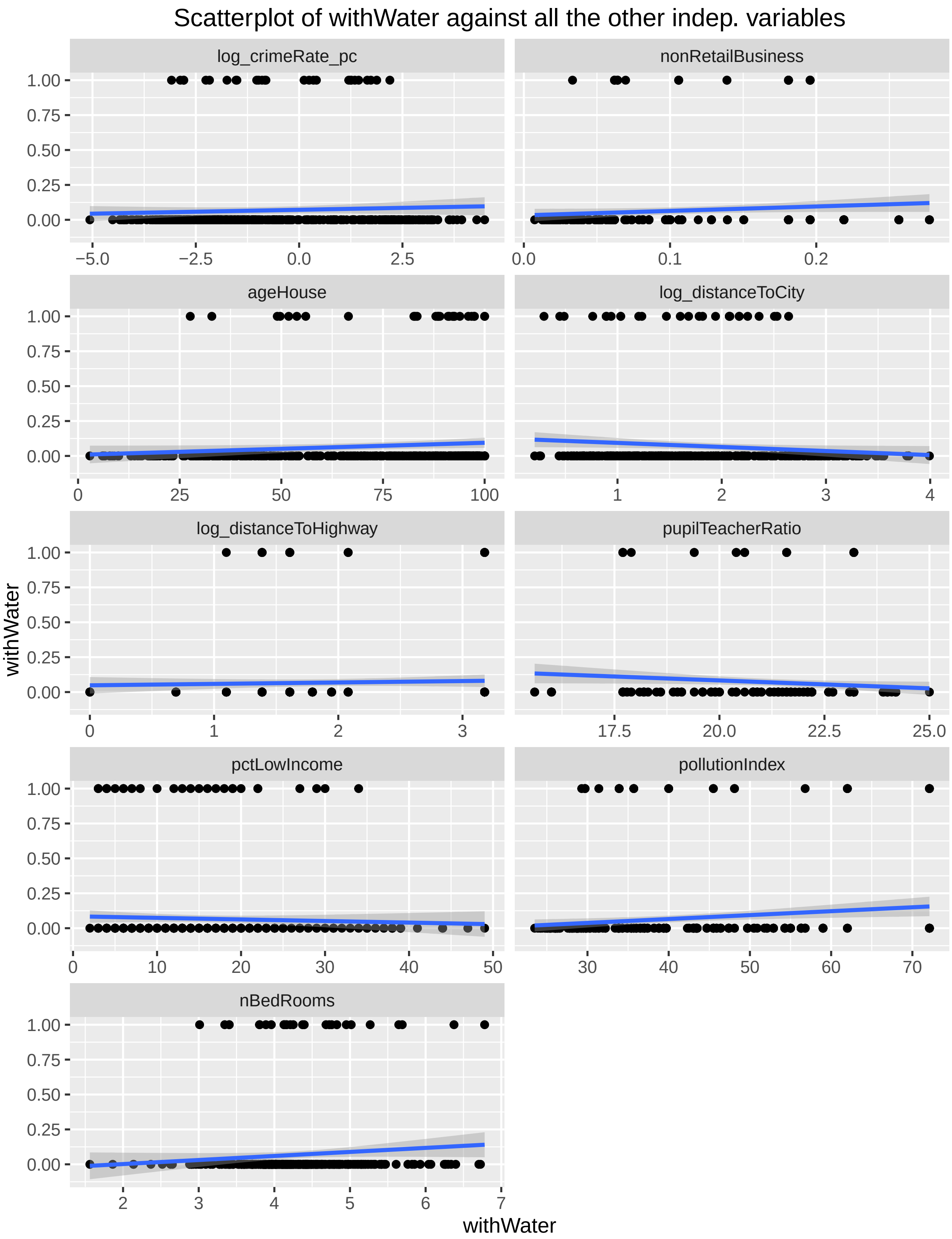

Figure 13: Scatter Plot of withWater against all the other independent variables

Table 4: Simple regression summary of withWater

	Dependent variable:								
	Proximity (< 5 miles) to a of a water body								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
log(crime rate)	0.006								
	(0.005)								
Prop. NR business		0.318*							
		(0.161)							
Prop. houses<1950			0.001*						
			(0.0004)						
log(dist. city)				−0.029*					
				(0.013)					
log(dist. highway)					0.010				
					(0.013)				
Avg p-t ratio						−0.011$^{\cdot}$			
						(0.006)			
% low-inc. house							−0.001		
							(0.001)		
Pollution								0.003*	
								(0.001)	
no. bedrooms									0.029
									(0.022)
Constant	0.072***	0.032$^{\cdot}$	0.008	0.122***	0.048$^{\cdot}$	0.310*	0.085***	−0.046	−0.056
	(0.014)	(0.018)	(0.024)	(0.031)	(0.025)	(0.132)	(0.026)	(0.053)	(0.091)
F Statistic	1.383	3.895*	5.207*	5.358*	0.642	3.662$^{\cdot}$	0.756	4.094*	1.782
df	1; 398	1; 398	1; 398	1; 398	1; 398	1; 398	1; 398	1; 398	1; 398
Observations	400	400	400	400	400	400	400	400	400
R^2	0.002	0.008	0.009	0.010	0.001	0.010	0.002	0.017	0.007
Adjusted R^2	−0.0002	0.005	0.007	0.008	−0.001	0.007	−0.001	0.015	0.004
Residual Std. Error	0.251	0.251	0.250	0.250	0.251	0.250	0.251	0.249	0.251

$^{\cdot}$p<0.1; *p<0.05; **p<0.01; ***p<0.001

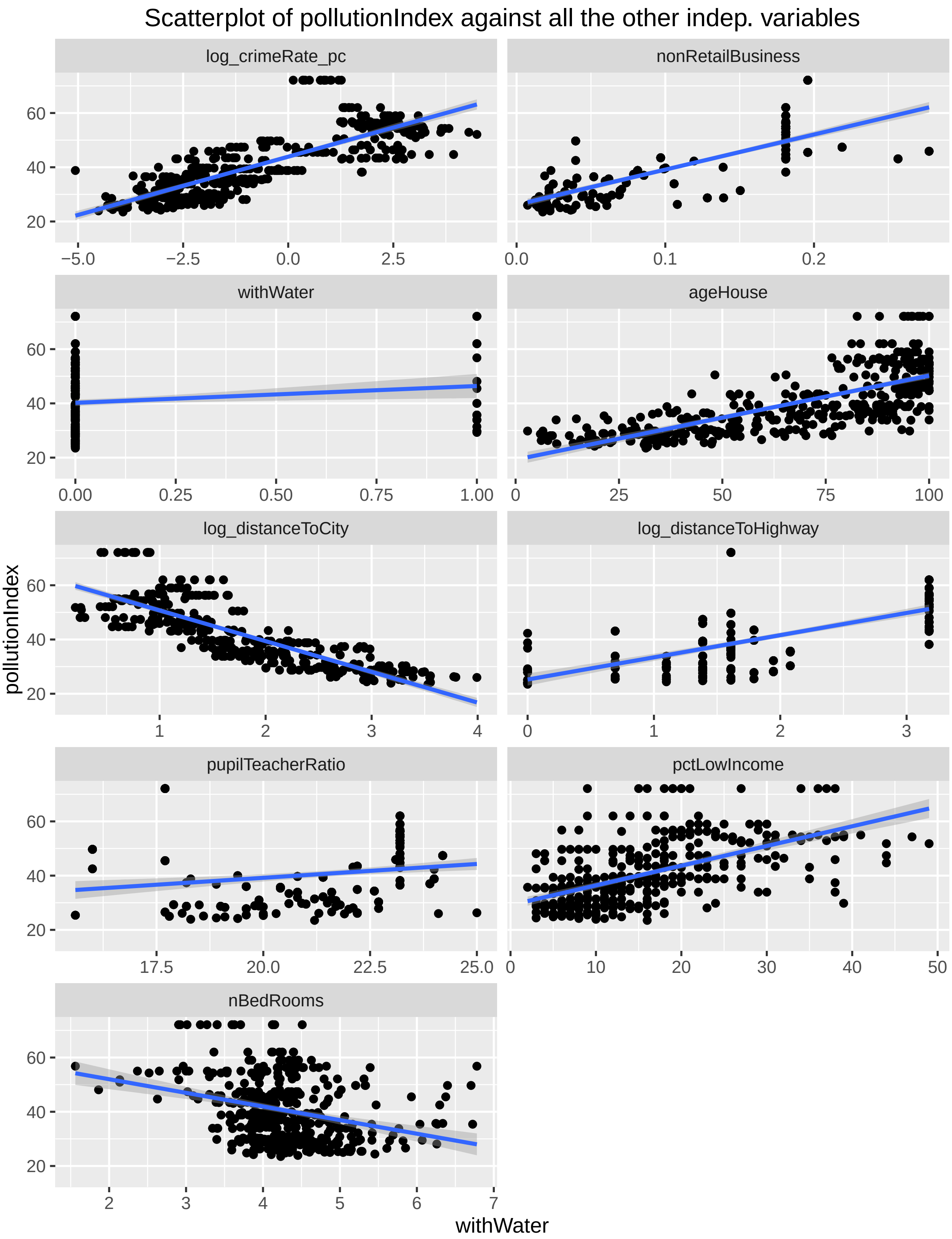

Figure 14: Scatter Plot of pollutionIndex against all the other independent variables

Table 5: Simple regression summary of pollutionIndex

	Dependent variable:								
	Pollution Index								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
log(crime rate)	4.295*** (0.162)								
Prop. NR business		129.499*** (6.727)							
Water<5 miles			6.207* (3.071)						
Prop. houses<1950				0.310*** (0.013)					
log(dist. city)					−11.335*** (0.394)				
log(dist. highway)						8.183*** (0.393)			
Avg p-t ratio							1.020** (0.346)		
% low-inc. house								0.726*** (0.055)	
no. bedrooms									−5.023*** (0.830)
Constant	43.890*** (0.437)	26.175*** (0.645)	40.196*** (0.593)	19.243*** (0.757)	62.062*** (0.967)	25.238*** (0.970)	18.786* (7.650)	29.141*** (0.901)	62.042*** (3.594)
F Statistic	701.991***	370.554***	4.087*	576.467***	829.270***	433.845***	8.710**	172.149***	36.639***
df	1; 398	1; 398	1; 398	1; 398	1; 398	1; 398	1; 398	1; 398	1; 398
Observations	400	400	400	400	400	400	400	400	400
R^2	0.618	0.581	0.017	0.538	0.692	0.358	0.035	0.329	0.093
Adjusted R^2	0.617	0.580	0.015	0.537	0.692	0.356	0.033	0.328	0.091
Residual Std. Error	7.320	7.667	11.737	8.047	6.568	9.489	11.631	9.697	11.275

·p<0.1; *p<0.05; **p<0.01; ***p<0.001

To select a candidate for a good model we'll make use of a stepwise selection (using the BIC rather than the AIC—though the output of R's `step` does not change the name of the criterion—to put a greater penalty in the number of regressors).

```
full.model <- lm(log_homeValue ~ ., data = houseValue.2)
null.model <- lm(log_homeValue ~ 1, data = houseValue.2)
# k=log(n) (instead of default k=2) to use  BIC rather than AIC
stepwise.reg <- step(null.model, scope = list(lower = null.model,
                                               upper=full.model),
                     direction="both", k = log(400))
```

```
## Start:  AIC=-734.15
## log_homeValue ~ 1
##
##                            Df Sum of Sq    RSS      AIC
## + pctLowIncome              1    41.239 21.634 -1154.90
## + nBedRooms                 1    26.243 36.629  -944.26
## + log_crimeRate_pc          1    20.617 42.255  -887.11
## + nonRetailBusiness         1    20.158 42.714  -882.79
## + pollutionIndex            1    16.514 46.359  -850.04
## + pupilTeacherRatio         1    15.624 47.249  -842.43
## + ageHouse                  1    13.002 49.870  -820.83
## + log_distanceToHighway     1    12.270 50.602  -815.01
## + log_distanceToCity        1    11.586 51.287  -809.63
## + withWater                 1     1.106 61.767  -735.26
## <none>                                  62.873  -734.15
##
## Step:  AIC=-1154.9
## log_homeValue ~ pctLowIncome
##
##                            Df Sum of Sq    RSS      AIC
## + pupilTeacherRatio         1     2.202 19.431 -1191.86
## + nBedRooms                 1     1.958 19.675 -1186.87
## + withWater                 1     0.611 21.023 -1160.37
## + nonRetailBusiness         1     0.490 21.144 -1158.08
## + log_crimeRate_pc          1     0.425 21.208 -1156.85
## + log_distanceToHighway     1     0.388 21.245 -1156.15
## <none>                                  21.634 -1154.90
## + pollutionIndex            1     0.214 21.420 -1152.89
## + log_distanceToCity        1     0.154 21.480 -1151.77
## + ageHouse                  1     0.075 21.559 -1150.30
## - pctLowIncome              1    41.239 62.873  -734.15
##
## Step:  AIC=-1191.86
## log_homeValue ~ pctLowIncome + pupilTeacherRatio
##
##                            Df Sum of Sq    RSS      AIC
## + nBedRooms                 1    1.5804 17.851 -1219.80
## + withWater                 1    0.4262 19.005 -1194.74
## + pollutionIndex            1    0.3051 19.126 -1192.20
## <none>                                  19.431 -1191.86
## + log_distanceToCity        1    0.1714 19.260 -1189.41
## + nonRetailBusiness         1    0.1576 19.274 -1189.12
## + ageHouse                  1    0.1322 19.299 -1188.60
## + log_crimeRate_pc       1    0.1187 19.313 -1188.32
## + log_distanceToHighway  1    0.0249 19.406 -1186.38
## - pupilTeacherRatio      1    2.2023 21.634 -1154.90
## - pctLowIncome           1   27.8174 47.249  -842.43
##
## Step:  AIC=-1219.8
## log_homeValue ~ pctLowIncome + pupilTeacherRatio + nBedRooms
##
##                         Df Sum of Sq    RSS     AIC
## + pollutionIndex         1    0.4309 17.420 -1223.6
## + log_crimeRate_pc       1    0.3607 17.490 -1222.0
## + withWater              1    0.3287 17.522 -1221.2
## <none>                               17.851 -1219.8
## + nonRetailBusiness      1    0.1658 17.685 -1217.5
## + log_distanceToHighway  1    0.1655 17.685 -1217.5
## + log_distanceToCity     1    0.0490 17.802 -1214.9
## + ageHouse               1    0.0047 17.846 -1213.9
## - nBedRooms              1    1.5804 19.431 -1191.9
## - pupilTeacherRatio      1    1.8245 19.675 -1186.9
## - pctLowIncome           1   12.9288 30.780 -1007.9
##
## Step:  AIC=-1223.58
## log_homeValue ~ pctLowIncome + pupilTeacherRatio + nBedRooms +
##     pollutionIndex
##
##                         Df Sum of Sq    RSS     AIC
## + log_distanceToCity     1    1.1615 16.259 -1245.2
## + withWater              1    0.4967 16.923 -1229.2
## + ageHouse               1    0.3357 17.084 -1225.4
## <none>                               17.420 -1223.6
## - pollutionIndex         1    0.4309 17.851 -1219.8
## + log_crimeRate_pc       1    0.0397 17.380 -1218.5
## + log_distanceToHighway  1    0.0080 17.412 -1217.8
## + nonRetailBusiness      1    0.0015 17.419 -1217.6
## - nBedRooms              1    1.7062 19.126 -1192.2
## - pupilTeacherRatio      1    1.9121 19.332 -1187.9
## - pctLowIncome           1    7.5976 25.018 -1084.8
##
## Step:  AIC=-1245.19
## log_homeValue ~ pctLowIncome + pupilTeacherRatio + nBedRooms +
##     pollutionIndex + log_distanceToCity
##
##                         Df Sum of Sq    RSS     AIC
## + withWater              1    0.4720 15.787 -1251.0
## <none>                               16.259 -1245.2
## + log_crimeRate_pc       1    0.1559 16.103 -1243.0
## + nonRetailBusiness      1    0.0966 16.162 -1241.6
## + log_distanceToHighway  1    0.0199 16.239 -1239.7
## + ageHouse               1    0.0152 16.243 -1239.6
## - log_distanceToCity     1    1.1615 17.420 -1223.6
## - nBedRooms              1    1.2952 17.554 -1220.5
## - pollutionIndex         1    1.5433 17.802 -1214.9
## - pupilTeacherRatio      1    2.2115 18.470 -1200.2
## - pctLowIncome           1    8.5804 24.839 -1081.7
```

```
##
## Step:  AIC=-1250.98
## log_homeValue ~ pctLowIncome + pupilTeacherRatio + nBedRooms +
##     pollutionIndex + log_distanceToCity + withWater
##
##                         Df Sum of Sq    RSS     AIC
## <none>                                15.787 -1251.0
## + log_crimeRate_pc       1    0.1394 15.647 -1248.5
## + nonRetailBusiness      1    0.1245 15.662 -1248.2
## + log_distanceToHighway  1    0.0181 15.768 -1245.5
## + ageHouse               1    0.0086 15.778 -1245.2
## - withWater              1    0.4720 16.259 -1245.2
## - log_distanceToCity     1    1.1368 16.923 -1229.2
## - nBedRooms              1    1.2195 17.006 -1227.2
## - pollutionIndex         1    1.7165 17.503 -1215.7
## - pupilTeacherRatio      1    2.0491 17.836 -1208.2
## - pctLowIncome           1    8.2783 24.065 -1088.3
```

Table 6: Regression model of log(homeValue)

	Dependent variable:
	log(Median price ($) of single-family house)
Percentage of low-income households	-0.025^{***}
	(0.003)
Average pupil-teacher ratio	-0.037^{***}
	(0.005)
Average number of bedrooms	0.101^{***}
	(0.027)
Pollution index (0-100)	-0.010^{***}
	(0.002)
log(Distance (miles) to nearest city)	-0.115^{***}
	(0.026)
Water body less than 5 miles away	0.140^{***}
	(0.040)
Constant	14.419^{***}
	(0.224)
F Statistic	140.988^{***}
df	6; 393
Observations	400
R^2	0.749
Adjusted R^2	0.745
Residual Std. Error	0.200

·p<0.1; *p<0.05; **p<0.01; ***p<0.001

Thus, the model is:

$$\log(homeValue) = \beta_0+\beta_1\cdot pctLowIncome+\beta_2\cdot pupilTeacherRatio+\beta_3\cdot nBedRooms+\beta_4\cdot pollutionIndex+\beta_5\cdot\log(distanceToC$$

And (because we've used the log of the dependent variable and one of the independent variables), the coefficients tell us that:

- when the percentage of low-income households in the neighborhood increases by 1, the median value of a house in that neighborhood decreases 2.5%.

- when the number of pupils per teahcer increases by 1, the median home value decreases by 3.7%.
- an additional bedroom increases the value of the house by approximately 10.1%.
- when the pollution index increases by 1, the median home value decreases by 1.0% (since the standard deviation of the Pollution Index is 11.8, an increase of 1 standard deviation in that index would decrease the value of the house by approximately 12.2%).
- when the distance to the nearest city increases by 1% (i.e., almost 0.1 mile, since the mean of that variable is 9.6 miles), the median home value decreases by 11.5%.
- the proximity to a water body increases the home value by 14.0%.

The previous F statistic tell us the overall significance of the regression model we've built. We can also test whether each individual regressor helps to explain the variability in home values (controlling for other factors; in other words, if it's worth keeping it in the model). As expected, all the included regressors are significantly different from zero.

```
hyp <- lapply(names(stepwise.reg$coefficients)[-1], function(x)
  linearHypothesis(stepwise.reg, x, vcov = vcovHC))
null.hyp_p <- unlist(lapply(c(1:(length(stepwise.reg$coefficients) - 1)),
                            function(i) (hyp[[i]])$`Pr(>F)`[2]))
names(null.hyp_p) <- names(stepwise.reg$coefficients)[-1]
null.hyp_p
```

```
##       pctLowIncome  pupilTeacherRatio          nBedRooms
##       2.302064e-15       6.550973e-12       2.147518e-04
##     pollutionIndex log_distanceToCity          withWater
##       1.009110e-07       1.635181e-05       5.168360e-04
```

But we can also test more complex hypotheses:

@knitr P1-model selection_3

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.

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.