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.

```
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

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).

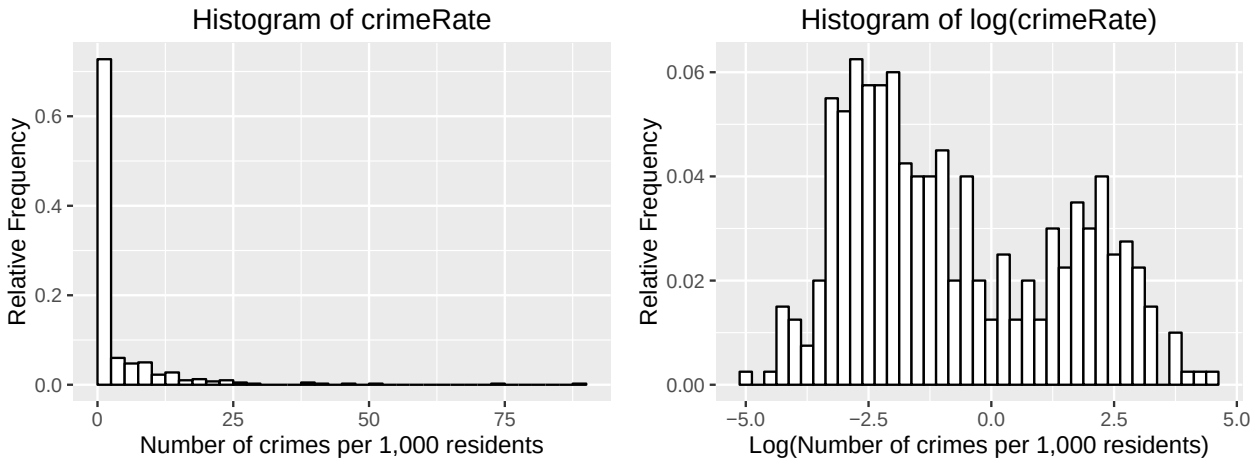

Figure 1: Histograms of Crime Rate and its log

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.

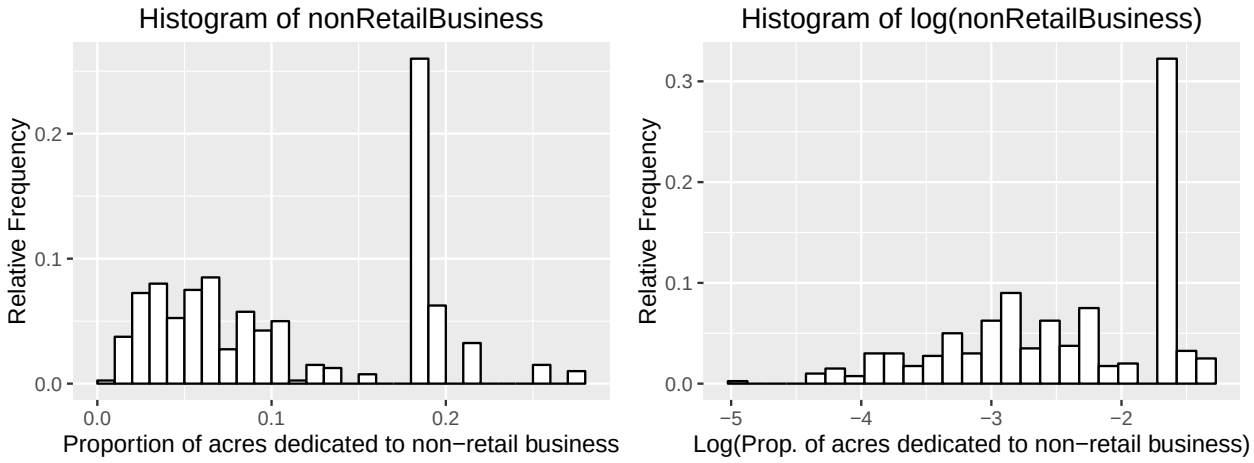

Figure 2: Histograms of non-retail Business acres and its log

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.

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.

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.

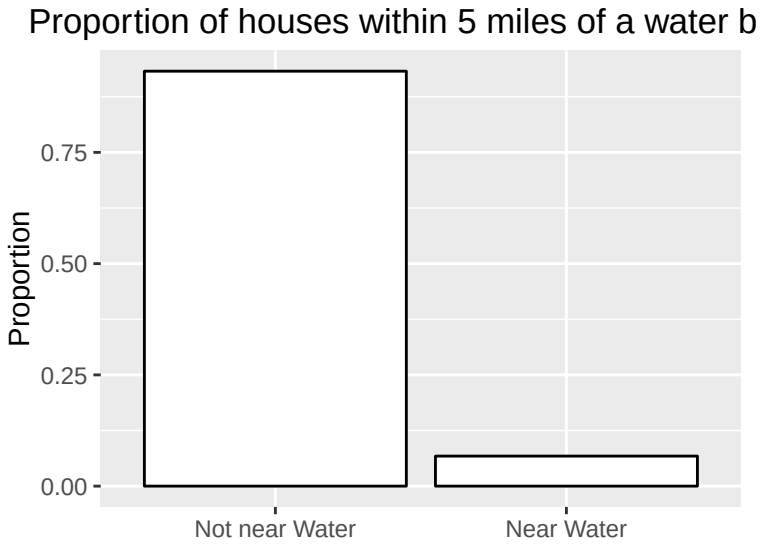

Figure 3: Proportion of of houses within 5 miles of a water body (or not)

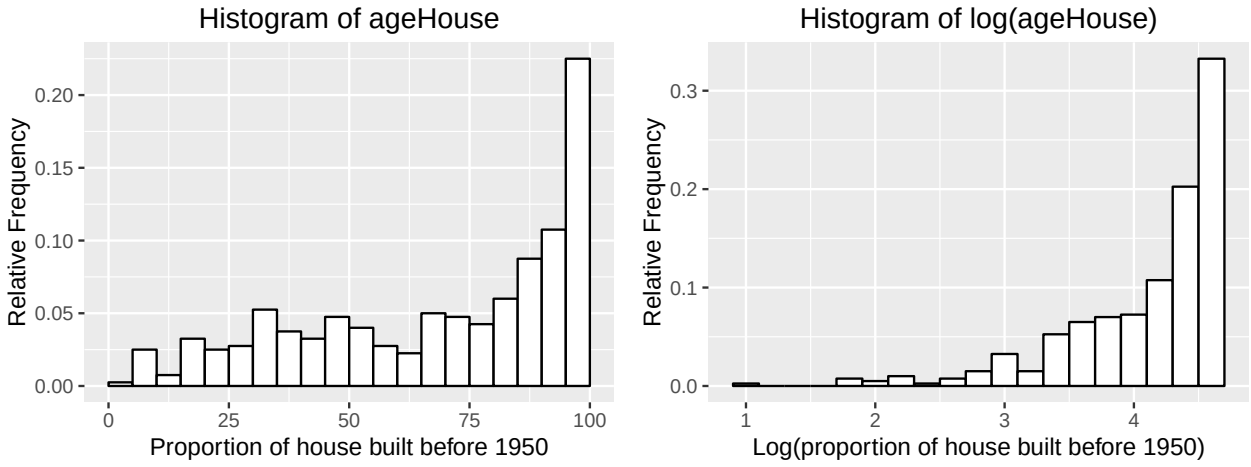

Figure 4: Histogram of House Age

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.