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.

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 it's 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 proceedes.

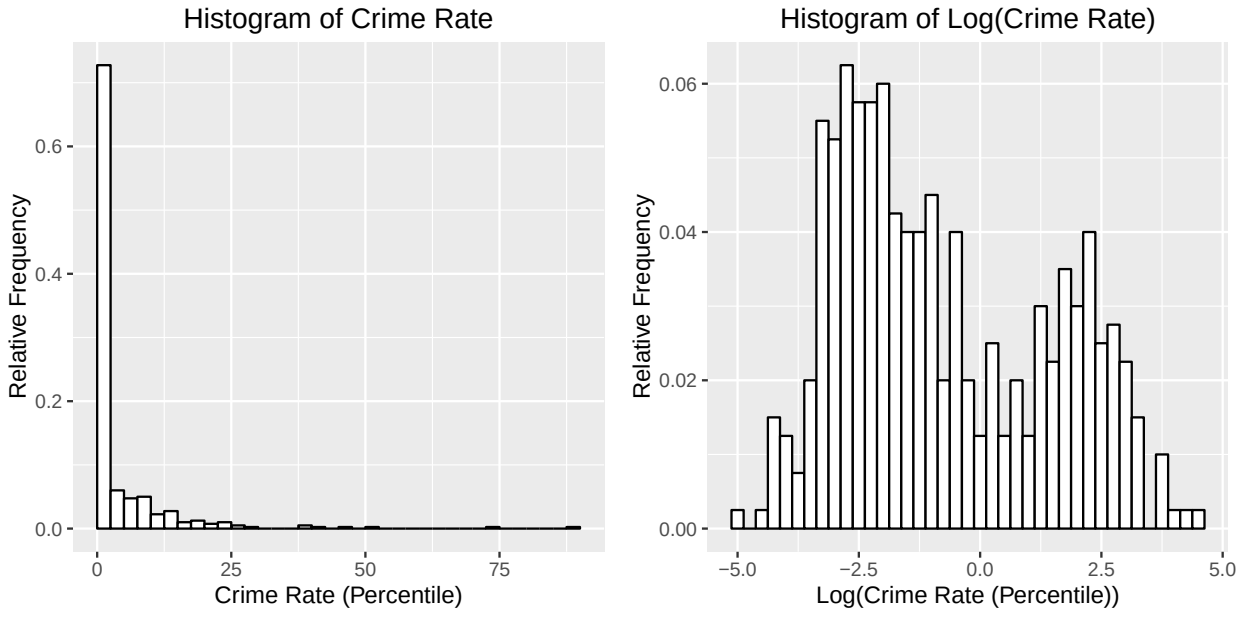

Figure 1: Histograms of the Crime Rate Variable

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 bimodel distribution, with a peak below zero representing low crime neighborhoods and a peak above zero representing moderate crime neighborhoods.

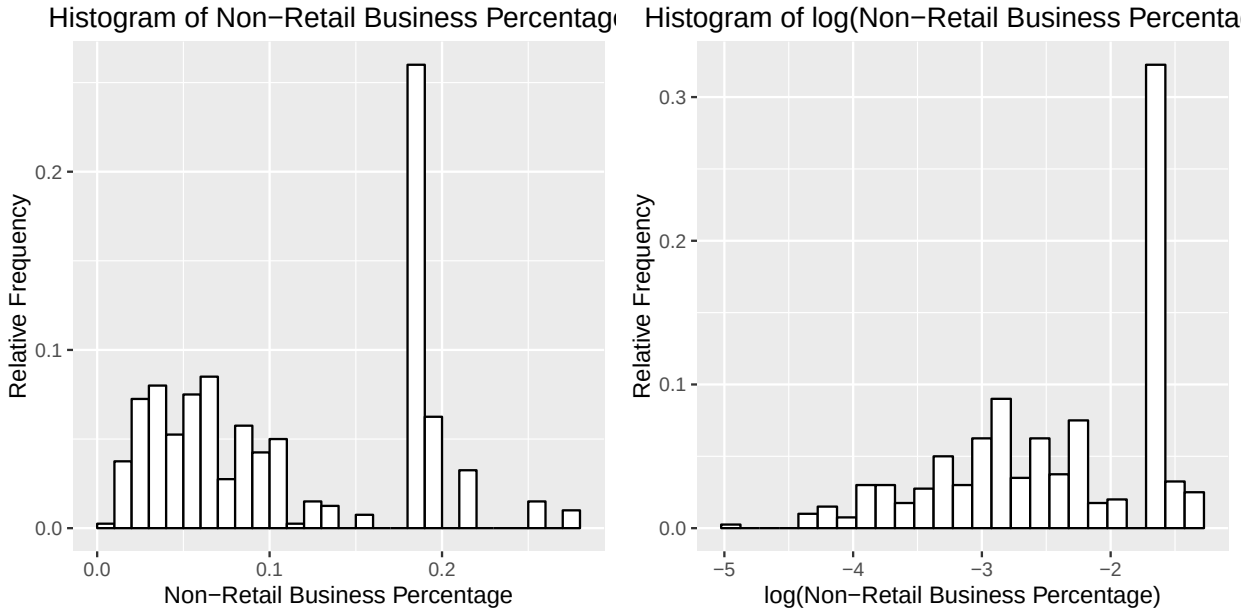

Figure 2: Histograms of Non-Retail Business Percentage Variable

The non-retail business percentage variabel also shows multimodiality, 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 to fill in missing values for this dataset.

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.
