Boston Housing Data Labs

Aaron Swoboda

This document describes a series of labs I developed for my undergraduate senior seminar in the economics of housing at Carleton College. The goal of the labs is to give students experience working with empirical data to prepare them for their own senior empirical project. Labs 1 and 2 can be completed with nothing more than a working version of `R` and the appropriate packages. Labs 3 and 4 require access to ArcMap GIS software and extra data files (provided in the `BostonGISLab.zip` file).

Course Background

These labs were developed as part of ECON 395: Advanced Topics in the Economics of Housing. This course is typically taken during the fall term by senior economics majors at Carleton College as part one of the two term senior Comprehensive Exercise. During the senior seminar 10-15 students read and discuss primary literature related to the seminar topic and ultimately propose an individual empirical research project to be completed in the subsequent term.

The primary goal of the seminar is to help students write a research prospectus containing:

- a tractible research question,
- a description of an appropriate and accessible dataset,
- a proposed analysis methodology and identification strategy,
- and, a knowledge of how the proposed work fits within the scholarly literature.

This is the first course in the major in which Econometrics is a prerequisite. Therefore, this is typically the first course for which students can apply their econometric tools to the task of reading primary literature. As such, they often struggle understanding the myriad steps involved "behind the scenes" that are necessary to construct the dataset described in the paper (for instance, merging datasets from multiple sources). They commonly struggle to understand what is feasible as they propose their own projects and often find tehmselves in more challenging circumstances than expected.

Abstract from Harrison and Rubinfeld (1978)

This paper investigates the methodological problems associated with the use of housing market data to measure the willingness to pay for clean air. With the use of a hedonic housing price model and data for the Boston metropolitan area, quantitative estimates of the willingness to pay for clean air improvements are generated. Marginal air pollution damages (as revealed in the housing market) are found to increase with the level of air pollution and with household income. The results are relatively sensitive to the specification of the hedonic housing price equation, but insensitive to the specification of the air quality demand.

Accessing the Data

The dataset consisting of 14 variables across 506 census tracts is available in at least three different ways.

1. Belsley, Kuh, and Welsch (1980) (a book)
2. The UCI Machine Learning Repository
3. Within the Statistical Software R (for instance, as the `Boston` dataset in the `MASS` package).

Overview of the Labs

The sequence of labs is based on the primary results of Harrison and Rubinfeld (1978). Our first lab took place after reading and discussing the paper as a class.

- Lab 1 asks students to use the basic dataset from Harrison and Rubinfeld (1978) to reproduce the basic summary statistics and OLS regression results. The primary challenge in this lab is to rescale some of the variables.
- Lab 2 asks students to use the same dataset to update the results of Harrison and Rubinfeld (1978) as presented in O. Gilley and Pace (1996). Students must fix some typos in the original dataset and estimate a new regression technique to account for the censored nature of the dependent variable.
- Lab 3 asks students to construct the Harrison and Rubinfeld (1978) from smaller component datasets. Students perform one-to-one and many-to-one merges as well as use GIS tools to perform spatial joins and construct an indicator variable based on spatial location.
- Lab 4 extends the spatial thinking of Lab 3 to try and incorporate the spatial nature of the data to replicate the results of R. K. Pace and Gilley (1997).

Lab 1: Replicating the Basic Results of Harrison and Rubinfeld (1978)

Below are images of two tables of results from the Harrison and Rubinfeld (1978) paper that can be replicated using the available data:

- The summary statistics from Table V, and
- The "basic equation" OLS regression results from Table VII:

Summary Statistics

The raw dataset does not immediately reproduce the summary statistics shown in the paper.

```
library(MASS)
data("Boston")
library(stargazer)
stargazer(Boston, type = "text")
```

```
##
## ================================================
## Statistic  N   Mean   St. Dev.  Min     Max
## ------------------------------------------------
## crim      506  3.614   8.602   0.006  88.976
## zn        506 11.364   23.322  0.000  100.000
## indus     506 11.137   6.860   0.460  27.740
## chas      506  0.069   0.254     0       1
## nox       506  0.555   0.116   0.385   0.871
## rm        506  6.285   0.703   3.561   8.780
## age       506 68.575   28.149  2.900  100.000
## dis       506  3.795   2.106   1.130  12.127
## rad       506  9.549   8.707     1      24
## tax       506 408.237 168.537   187     711
## ptratio   506 18.456   2.165   12.600 22.000
## black     506 356.674  91.295  0.320  396.900
## lstat     506 12.653   7.141   1.730  37.970
## medv      506 22.533   9.197   5.000  50.000
## ------------------------------------------------
```

TABLE V

Summary Statistics for Housing Value Equation Variables

Variable	Mean	SD
MV	22,532	9,197
RM	6.28	0.70
AGE	68.6	28.1
B	0.06	0.18
LSTAT	0.13	0.07
CRIM	3.61	8.60
ZN	11.36	23.32
INDUS	11.13	6.86
TAX	408.2	168.5
PTRATIO	18.5	2.16
DIS	3.79	2.10
RAD	9.55	8.70
NOX	5.55	1.16
PART	6.31	1.50

Figure 1:

A comparison of Table V and the results above shows that some of the variables are not in the approporiate units and must be rescaled.

```
library(dplyr)
# copy the data frame
boston.df.1978 = Boston
# change the names to uppercase for consistency with H&R (1978)
names(boston.df.1978) = toupper(names(boston.df.1978))

# now rescale and reorder for consistency
boston.df.1978 = boston.df.1978 %>%
  # rescale
  mutate(MV = MEDV*1000,
         Btransformed = BLACK/1000,
         LSTAT = LSTAT/100,
         NOX = NOX*10) %>%
  # and reorder
  select(MV, RM, AGE, Btransformed, LSTAT, CRIM, ZN,
         INDUS, TAX, PTRATIO, DIS, RAD, NOX, CHAS)

stargazer(boston.df.1978,
          type = "text",
          digits = 2)
```

```
##
## ================================================
## Statistic     N    Mean     St. Dev.  Min    Max
## ------------------------------------------------
## MV           506 22,532.81 9,197.10 5,000  50,000
## RM           506   6.28      0.70    3.56   8.78
## AGE          506   68.57    28.15    2.90  100.00
```

```
## Btransformed 506   0.36      0.09   0.0003  0.40
## LSTAT        506   0.13      0.07    0.02   0.38
## CRIM         506   3.61      8.60    0.01  88.98
## ZN           506   11.36    23.32    0.00  100.00
## INDUS        506   11.14     6.86    0.46  27.74
## TAX          506  408.24    168.54   187    711
## PTRATIO      506   18.46     2.16   12.60  22.00
## DIS          506   3.80      2.11    1.13  12.13
## RAD          506   9.55      8.71     1     24
## NOX          506   5.55      1.16    3.85   8.71
## CHAS         506   0.07      0.25     0      1
## ----------------------------------------------
```

That's better.[1]

Regression Results

Now let's estimate the OLS equation shown in TABLE VII (being careful to transform the variables appropriately).

```
HR.lm = lm(log(MV) ~ I(RM^2) + AGE + log(DIS) +
                     log(RAD) + TAX + PTRATIO + Btransformed +
                     log(LSTAT) + CRIM + ZN + INDUS + CHAS +
                     I(NOX^2),
                   data = boston.df.1978)
stargazer(HR.lm, type = "text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                log(MV)
## ------------------------------------------------
## I(RM2)                        0.006***
##                                (0.001)
##
## AGE                            0.0001
##                                (0.001)
##
## log(DIS)                      -0.191***
##                                (0.033)
##
## log(RAD)                      0.096***
##                                (0.019)
##
## TAX                          -0.0004***
##                                (0.0001)
##
## PTRATIO                       -0.031***
##                                (0.005)
##
## Btransformed                  0.364***
##                                (0.103)
```

[1]The original "Black" variable cannot be recovered from the these data, which have been transformed using the equation $(Black - 0.63)^2$.

Variable	"Basic equation" Equation 1
Dependent	Log (MV)
Constant	9.76
	(65.22)
RM^2	0.0063
	(4.83)
AGE	8.98×10^{-5}
	(1.7)
Log (DIS)	−0.19
	(−5.73)
Log (RAD)	0.096
	(5.00)
TAX	-4.20×10^{-4}
	(−3.43)
$PTRATIO$	−0.031
	(−6.21)
$(B - 0.63)^2$	0.36
	(3.53)
Log ($STAT$)	−0.37
	(−14.84)
$CRIM$	−0.012
	(−9.53)
ZN	8.03×10^{-5}
	(0.16)
$INDUS$	2.41×10^{-4}
	(0.10)
$CHAS$	0.088
	(2.75)
NOX^2	−0.0064
	(−5.64)
P	2
$PART^{PP}$	
PP	
R^2	0.81

* t statistics are in parentheses.

Figure 2:

```
##
## log(LSTAT)                    -0.371***
##                                (0.025)
##
## CRIM                          -0.012***
##                                (0.001)
##
## ZN                             0.0001
##                                (0.001)
##
## INDUS                          0.0002
##                                (0.002)
##
## CHAS                          0.091***
##                                (0.033)
##
## I(NOX2)                       -0.006***
##                                (0.001)
##
## Constant                      9.756***
##                                (0.150)
##
## ------------------------------------------------
## Observations                     506
## R2                              0.806
## Adjusted R2                     0.801
## Residual Std. Error       0.182 (df = 492)
## F Statistic           157.128*** (df = 13; 492)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Tweaking a few settings[2] to customize our table shows just how similar our results are to the reported values.

```
stargazer(HR.lm, type = "text",
          intercept.bottom = FALSE,
          keep.stat = "rsq",
          omit.table.layout = "n",
          report = "vct",
          digits = 5,
          no.space = TRUE)
```

```
##
## ================================
##              Dependent variable:
##              -------------------
##                    log(MV)
## --------------------------------
## Constant           9.75629
##                  t = 65.22075
## I(RM2)             0.00633
##                   t = 4.82256
## AGE                0.00009
##                   t = 0.17242
## log(DIS)          -0.19126
```

[2]Check out this cheatsheet for the **stargazer** package: http://jakeruss.com/cheatsheets/stargazer.html

```
##                 t = -5.72750
## log(RAD)           0.09571
##                  t = 5.00207
## TAX                -0.00042
##                 t = -3.42613
## PTRATIO            -0.03112
##                 t = -6.20808
## Btransformed        0.36370
##                  t = 3.52708
## log(LSTAT)         -0.37116
##                 t = -14.84063
## CRIM               -0.01186
##                 t = -9.53205
## ZN                 0.00008
##                  t = 0.15853
## INDUS              0.00024
##                  t = 0.10134
## CHAS               0.09140
##                  t = 2.75268
## I(NOX2)           -0.00638
##                 t = -5.63930
## --------------------------------
## R2                 0.80589
## ================================
```

Commentary on Lab 1

I gave this lab to my students without much instruction or advice. Many of them expected it to be easy. How hard could it be to reproduce some simple numbers from a paper with the same dataset?

My students struggled to keep track of the changes they made (many worked with the data in Excel) as well as whether or not a variable had been logged, squared, etc. At the end of the hour-long lab, most students were able to replicate the main results via help from their peers or myself.

I intentionally did not require students to turn in anything related to this lab. Most students did not save their notes nor their code for producing the results.

Lab 2: Replicating the Results of O. Gilley and Pace (1996)

After Lab 1, I asked students to read O. Gilley and Pace (1996), which revealed that the underlying data used in the original Harrison and Rubinfeld (1978) analysis had two significant problems.

- There were data entry mistakes (eight dependent values appeared to have been entered incorrectly)
- Some dependent variable values were censored (census tracts with median home values above $50,000 were simply recorded as $50,000)

For Lab 2, I then asked students to replicate the results from O. Gilley and Pace (1996): correcting data entry errors[3] and updating the regression method. I told them that this lab should be simpler than the first lab because they knew how to rescale the provided data and run the initial regression. The vast majority of my students did not save their work from the previous lab and had to start over.

About half-way through the lab I provided my code that successfully completed the task of transforming and running the OLS regression. I also introduced the students to R Markdown via RStudio and encouraged

[3]One of my students pointed out that there is a typo in the O. Gilley and Pace (1996) TABLE I. They believe observation 191 (rather than 119) should have a corrected median value of 33.0 (from an incorrect value of 37.0). The results shown in O. Gilley and Pace (1996) can be replicated when observation 191 is corrected but not when observation 119 is "corrected."

them to place all of their notes and code (and therefore results) in one place to be continually updated and available for future reference.

TABLE I
Miscoded Dependent Variable Observations

Observation and tract number	Median value	Corrected median value
8-2042	27.1	22.1
39-2084	24.7	24.2
119-3585	37.0	33.0
241-3823	22.0	27.0
438-0905	8.7	8.2
443-0911	18.4	14.8
455-0923	14.9	14.4
506-1805	11.9	19.0

Figure 3:

Replicating the O. Gilley and Pace (1996) Uncorrected OLS Results

Inspection of the first column of results shows nearly identical values from our earlier regression. However, the regression intercepts do not match.

The answer is scaling (again!). Notably, O. Gilley and Pace (1996) present the median house value variable in $1,000s (i.e. 27.1 instead of $27,100). Let's rescale the dependent variable (and two others) and rerun the regression.

```
GP.lm.Uncorrected = lm(log(MV/1000) ~  CRIM + ZN + INDUS + CHAS +
                       I(NOX^2/100) + I(RM^2) + AGE +
                       log(DIS) + log(RAD) + TAX + PTRATIO +
                       I(Btransformed*1000) + log(LSTAT),
                       data = boston.df.1978)
```

Replicating the O. Gilley and Pace (1996) Corrected OLS Results

Let's create a new variable, `CMV` for the corrected median values and see if we can replicate the other two columns from Table 3.

```
obs.incorrect = c(8, 39, 191, 241, 438, 443, 455, 506) # note 191 instead of 119
correct.values = c(22.1, 24.2, 33, 27, 8.2, 14.8, 14.4, 19)

boston.df.1978 = boston.df.1978 %>%
  mutate(CMV = MV/1000)
boston.df.1978$CMV[obs.incorrect] = correct.values
```

We now should be able to replicate the results from Column 2 in Table III simply by re-estimating the regression from column 1 but using `CMV` as the dependent variable.

```
GP.lm.Corrected = lm(log(CMV) ~ CRIM + ZN + INDUS +
                         CHAS + I(NOX^2/100) + I(RM^2) + AGE +
                         log(DIS) + log(RAD) + TAX + PTRATIO +
                         I(Btransformed*1000) + log(LSTAT),
                       data = boston.df.1978)
```

TABLE III

Estimation Results for the Harrison and Rubinfeld Data

Variable	Uncorrected OLS	Corrected OLS	TOBIT
Constant	2.84853	2.83601	1.10758
	(19.04)	(19.22)	(7.42)
CRIM	−0.01186	−0.01177	−0.01170
	(−9.53)	(−9.59)	(−9.45)
ZN	0.00008	.00009	0.00014
	(0.15)	(0.18)	(0.27)
INDUS	0.00024	0.00018	0.00101
	(0.10)	(0.08)	(0.43)
CHAS	0.09139	0.09213	0.10540
	(2.75)	(2.81)	(3.12)
NOX^2	−0.63805	−0.63724	−0.66618
	(−5.64)	(−5.71)	(−5.91)
RM^2	0.00633	0.00625	0.00666
	(4.82)	(4.83)	(5.01)
AGE	0.00009	0.00007	0.00024
	(0.17)	(0.14)	(0.45)
LDIS	−0.19125	−0.19784	−0.20454
	(−5.73)	(−6.01)	(−6.13)
LRAD	0.09571	0.08957	0.08937
	(5.00)	(4.75)	(4.69)
TAX	−0.00042	−0.00042	−0.00041
	(−3.43)	(−3.46)	(−3.38)
PTRATIO	−0.03112	−0.02960	−0.03096
	(−6.21)	(−5.99)	(−6.18)
B	0.00036	0.00036	0.00036
	(3.53)	(3.55)	(3.53)
LSTAT	−0.37116	−0.37489	−0.39122
	(−14.84)	(−15.20)	(−15.23)
σ			−0.1813
R^2	0.806	0.811	
Log-likelihood	149.955	156.979	125.532

Figure 4:

Replicating the O. Gilley and Pace (1996) Corrected Tobit Results

Let's just go ahead and estimate the Tobit model as well and present all three regressions in one table using `stargazer`.

```
library(AER)
GP.lm.Tobit = tobit(log(CMV) ~  CRIM + ZN + INDUS +
                      CHAS + I(NOX^2/100) + I(RM^2) +
                      AGE + log(DIS) + log(RAD) + TAX + PTRATIO +
                      I(Btransformed*1000) + log(LSTAT),
                    right = log(50),
                    data = boston.df.1978)
```

```
stargazer(GP.lm.Uncorrected, GP.lm.Corrected, GP.lm.Tobit,
          type = "text", digits = 5,
          intercept.bottom = FALSE,
          keep.stat = "rsq",
          omit.table.layout = "n",
          report = "vct",
          no.space = TRUE)
```

```
##
## ===============================================================
##                                  Dependent variable:
##                          ---------------------------------------
##                          log(MV/1000)           log(CMV)
##                               OLS           OLS          Tobit
##                               (1)           (2)           (3)
## ---------------------------------------------------------------
## Constant                    2.84853       2.83601       2.80445
##                           t = 19.04243  t = 19.22372  t = 18.79884
## CRIM                        -0.01186      -0.01177      -0.01170
##                           t = -9.53205  t = -9.59008  t = -9.44951
## ZN                           0.00008       0.00009       0.00014
##                            t = 0.15853   t = 0.18401   t = 0.27166
## INDUS                        0.00024       0.00018       0.00101
##                            t = 0.10134   t = 0.07676   t = 0.42624
## CHAS                         0.09140       0.09213       0.10541
##                            t = 2.75268   t = 2.81347   t = 3.12409
## I(NOX2/100)                 -0.63805      -0.63724      -0.66621
##                           t = -5.63930  t = -5.71086  t = -5.91071
## I(RM2)                       0.00633       0.00626       0.00666
##                            t = 4.82256   t = 4.83322   t = 5.01024
## AGE                          0.00009       0.00007       0.00024
##                            t = 0.17242   t = 0.13679   t = 0.45037
## log(DIS)                    -0.19126      -0.19784      -0.20455
##                           t = -5.72750  t = -6.00743  t = -6.13354
## log(RAD)                     0.09571       0.08957       0.08937
##                            t = 5.00207   t = 4.74638   t = 4.68565
## TAX                         -0.00042      -0.00042      -0.00041
##                           t = -3.42613  t = -3.46406  t = -3.37752
## PTRATIO                     -0.03112      -0.02960      -0.03096
##                           t = -6.20808  t = -5.98646  t = -6.17677
## I(Btransformed * 1000)       0.00036       0.00036       0.00036
##                            t = 3.52708   t = 3.55082   t = 3.53534
```

```
## log(LSTAT)               -0.37116      -0.37489      -0.39123
##                        t = -14.84063 t = -15.19961 t = -15.23336
## ------------------------------------------------------------------
## R2                        0.80589       0.81076
## ==================================================================
```

Lab 3: Using GIS to Reconstruct the Harrison and Rubinfeld (1978) Data

This lab focuses less on replicating the same numbers shown in Harrison and Rubinfeld (1978) and more on the underlying process by which the final dataset might have been constructed.

The Harrison and Rubinfeld (1978) dataset uses census tract as the unit of observation, but many of the variables are not measured via the census and have to be merged with census data. Some of the variables are measured at the town level, others contain spatial information (adjacency to the Charles River). These represented opportunities to help the students understand several ways of merging data and creating new variables.

I "deconstructed" the final dataset that had been used in Labs 1 and 2 into split the dataset into four parts with the help of Wei-Hsin Fu, Carleton College's GIS Specialist.

1. CensusData1 contains TRACT, MEDV, RM, AGE, and TOWN
2. CensusData2 contains TRACT, DIS, CMEDV, CRIM, B, LSTAT
3. TownData contains TOWN-level variables: "INDUS", "PTRATIO", "RAD", "TAX", and "ZN"
4. PollutionMonitors contains the NOX variable as well as X, Y coordinates

The BostonLab3Instructions.docx file contains step-by-step instructions for how to use ArcMap GIS software to perform the processes needed to reconstruct the dataset.

1. Perform a one-to-one join of the two datasets containing census data using the census tract ID as the key variable.
2. Perform a many-to-one join of the census data and the town data using the town name as the key variable.
3. Spatially project the pollution monitor data and the census tract data and perform a spatial join to add the NOX data to the census tract dataset.
4. "Select by Location"" those census tracts that intersect with the Charles River to create the indicator variable CHAS.

After reconstructing the dataset, students are able to estimate the OLS regression with the new data and replicate the earlier results produced with the original dataset.

Lab 4: Using GIS and GeoDa to Replicate R. K. Pace and Gilley (1997)

Between Lab 3 and 4 students read R. K. Pace and Gilley (1997), which, after correcting the data entry errors noted in O. Gilley and Pace (1996), reestimated the relationship between census tract median house values and NOX levels while accounting for the spatial nature of the underlying data.

In this lab we pick up where we left off with Lab 3. We discuss the topic of spatial autocorrelation among the regression residuals and how, if present, this is a violation of the Gauss-Markov assumptions and prevent OLS from being the Best Linear Unbiased Estimator for our model.

The BostonLab4Instructions.docx file provides step-by-step instructions for using the data produced from Lab 3 to:

1. Check for spatial autocorrelation in the OLS residual estimates by calculating Moran's-I statistic
2. Estimate spatially explicit regressions (spatial error and spatial lag models)

References

Belsley, David A., Edwin Kuh, and Roy E. Welsch. 1980. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.* John Wiley & Sons.

Gilley, Otis, and R. Kelley Pace. 1996. "On the Harrison and Rubinfeld Data." *Journal of Environmental Economics and Management* 31 (3): 403–5. doi:10.1006/jeem.1996.0052.

Harrison, D, and D Rubinfeld. 1978. "Hedonic housing prices and the demand for clean air." *Journal of Environmental Economics and Management* 5 (1): 81–102.

Pace, R K, and O W Gilley. 1997. "Using the Spatial Configuration of the Data to Improve Estimation." *Journal of Real Estate Finance and Economics* 14 (3): 333–40.