Homework 5

Nicholas Nagle

March 30, 2015

For this assigment you will use a dataset of housing prices in Boston. These data were used in an early publication in environmental economics to study the effect of air quality on housing price. You can get a copy of the data in the spdep R package. Don't forget to use install.packages if you need to!

library(spdep)

## Warning: package 'spdep' was built under R version 3.1.3

## Warning: package 'sp' was built under R version 3.1.3

data(boston)

There is a codebook in the help file for this dataset

help(boston)

## starting httpd help server ... done

head(boston.c)

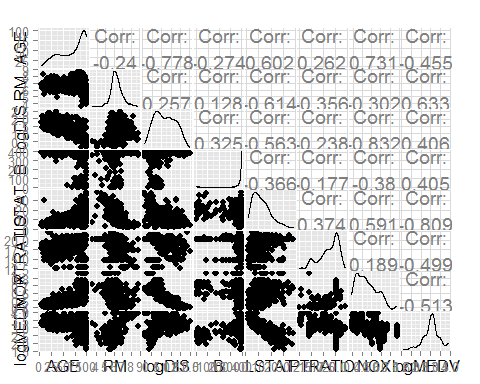
## TOWN TOWNNO TRACT LON LAT MEDV CMEDV CRIM ZN INDUS  
## 1 Nahant 0 2011 -70.9550 42.2550 24.0 24.0 0.00632 18 2.31  
## 2 Swampscott 1 2021 -70.9500 42.2875 21.6 21.6 0.02731 0 7.07  
## 3 Swampscott 1 2022 -70.9360 42.2830 34.7 34.7 0.02729 0 7.07  
## 4 Marblehead 2 2031 -70.9280 42.2930 33.4 33.4 0.03237 0 2.18  
## 5 Marblehead 2 2032 -70.9220 42.2980 36.2 36.2 0.06905 0 2.18  
## 6 Marblehead 2 2033 -70.9165 42.3040 28.7 28.7 0.02985 0 2.18  
## CHAS NOX RM AGE DIS RAD TAX PTRATIO B LSTAT  
## 1 0 0.538 6.575 65.2 4.0900 1 296 15.3 396.90 4.98  
## 2 0 0.469 6.421 78.9 4.9671 2 242 17.8 396.90 9.14  
## 3 0 0.469 7.185 61.1 4.9671 2 242 17.8 392.83 4.03  
## 4 0 0.458 6.998 45.8 6.0622 3 222 18.7 394.63 2.94  
## 5 0 0.458 7.147 54.2 6.0622 3 222 18.7 396.90 5.33  
## 6 0 0.458 6.430 58.7 6.0622 3 222 18.7 394.12 5.21

Most of these variables were selected because Economic theory suggests that each should impact median value. A scatterplot matrix is a helpful to quickly visualize many bivariate relations. I like the scatterpot matrix function in the GGally package called ggpairs. Sorry it looks so bad printed out. It's better on a big screen.

library(ggplot2)  
library(GGally)

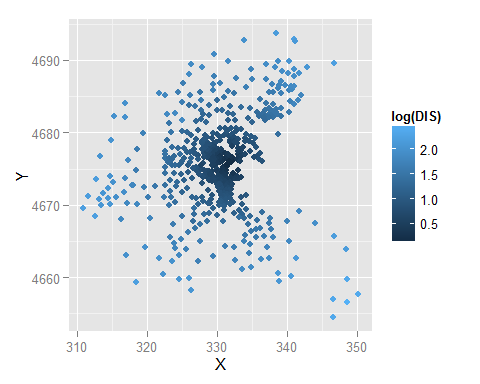
## Warning: package 'GGally' was built under R version 3.1.3

library(dplyr)  
boston.c %>% mutate(logMEDV = log(CMEDV), logDIS=log(DIS)) %>%   
 select(AGE, RM, logDIS, B, LSTAT, PTRATIO, NOX, logMEDV) %>% ggpairs()



We are trying to understand the various determinants of house price, including air pollution. One of the most important aspects of house price in the US is suburbanization. From the scatterplots, we see a significant relationship between value and distance. It may be helpful to map this out. You could use Latitude and Longitude to map it out, but it is better to use projected coordinates, which are in units of meters, rather than in units of geographic degrees. Fortunately, these have already been calculated for you.  
(Note, a GIS course would teach you more about projections.  
You could do the projection using GIS software like Quantum GIS or ArcGIS, or you could do it in R using the spTransform function in the sp package.)

boston.c$X <- boston.utm[,1]  
boston.c$Y <- boston.utm[,2]  
ggplot(boston.c) + geom\_point(aes(x=X, y=Y, color=log(DIS)))



# Homework Assignment:

1. Using the scatterplot matrix (ggpairs),
2. Describe the correlates of house price.

According to the scatterplot matrix, house prices have negative correlations with age, lower status populations, pupil-teacher ratios, and nitric oxides concentration. House prices have positive correlations with average number of rooms per house, log-distance, and the numeric vector of black population,

1. Describe the correlates of NOX.

Nitric oxide counts have a (quite strong) positive correlation with age, and otherwise positive correlations with lower-status populations and a slightly positive correlation with pupil-teacher ratio. Nitric oxide counts have negative correlations with rooms per house, distances from employment centers, and the numeric vector of black population.

1. Imagine the multivariate regression of log(CMEDV) on NOX, AGE, log(DIS), RM, CRIM, PTRATIO, B, LSTAT, and CHAS. DO NOT RUN THE REGRESSION YET. For each of these variables, predict whether you think the regression coefficient will be positive or negative, and why. Remember, the multivariate regression relationship is the relationship AFTER you hold the other values fixed. So, for instance to think about the relationship between Distance and value, you should think like: "Imagine two houses that have the same age, same number of rooms, same racial and ethnic neighborhood, same tax rate, etc. Now move one of those houses farther from workplaces. Should that change increase or decrease housing value." Answer: Economic theory suggest that everything else equal, being far from work is a bad thing. Note, the Charles River is a particulurly industrial part of town.

If we were to hold the other values all fixed, then I would imagine the regression coefficient for NOX would be negative, the coefficient for AGE would be negative, log(DIS) would be negative, the RM would be positive, the CRIM would be negative, PTRATIO would be negative, B would be negative, LSTAT would be negative, and CHAS would probably, based upon the industrial element, also be negative. So, the only factor that would drive up house prices (all things considered) would be RM (number of rooms per dwelling). All other factors, based upon generalities, negatively affect housing prices, especially in a city with a history of racial intolerance like Boston.

1. One of the relationships is a negative relationship between Distance from Work (primarily Boston) and House Value. Fit a bivariate regression between log CMEDV and log DIS

Call: lm(formula = MEDV ~ log(CMEDV) + log(DIS), data = boston.c)

Residuals: Min 1Q Median 3Q Max -9.1250 -1.6884 -0.7287 0.5194 12.9573

Coefficients: Estimate Std. Error t value Pr(>|t|)  
(Intercept) -43.3338 0.9008 -48.105 < 2e-16  ***log(CMEDV) 22.4540 0.3197 70.245 < 2e-16***  log(DIS) -1.9120 0.2419 -7.905 1.71e-14  ***--- Signif. codes: 0 ‘***’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.68 on 503 degrees of freedom Multiple R-squared: 0.9154, Adjusted R-squared: 0.9151 F-statistic: 2721 on 2 and 503 DF, p-value: < 2.2e-16

1. Report the slope of this regression and interpret it's value.

This slope is estimated for log Distance at -1.9, meaning that for every unit that the distance changes, it affects the Median Value approximately 2 units.

1. Report approximate 95% confidence intervals for the slope.

The Standard error of the log(DIS) is 0.24, so a 95% confidence interval would have the approximate amount of 0.48.

1. Fit the linear regression from question 2.

Call: lm(formula = log(CMEDV) ~ NOX + AGE + log(DIS) + RM + CRIM + PTRATIO + B + LSTAT + CHAS, data = boston.c)

Residuals: Min 1Q Median 3Q Max -0.72962 -0.09850 -0.00867 0.09295 0.86452

Coefficients: Estimate Std. Error t value Pr(>|t|)  
(Intercept) 4.0719345 0.1997748 20.383 < 2e-16  ***NOX -0.9250979 0.1413230 -6.546 1.48e-10***  AGE -0.0004881 0.0005235 -0.932 0.351613  
log(DIS) -0.2499594 0.0322420 -7.753 5.14e-14  ***RM 0.1113468 0.0159920 6.963 1.06e-11***  CRIM -0.0094586 0.0011974 -7.899 1.82e-14  ***PTRATIO -0.0338335 0.0043792 -7.726 6.20e-14***  B 0.0003547 0.0001053 3.368 0.000817  ***LSTAT -0.0292276 0.0020030 -14.592 < 2e-16***  CHAS1 0.1202545 0.0339289 3.544 0.000431  ***--- Signif. codes: 0 ‘***’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1887 on 496 degrees of freedom Multiple R-squared: 0.7902, Adjusted R-squared: 0.7864 F-statistic: 207.5 on 9 and 496 DF, p-value: < 2.2e-16

1. Report the coefficient of log Distance. Interpret it's value and report it's 95% confidence interval.

The coefficient of log Distance is approximately -0.25. This means that for every unit of distance farther away, it results in .25 less value on the property. It's standard error is .032, so a 95% confidence interval would be within .064 of that amount per unit.

1. Explain why the coefficient on log Distance changed so dramatically from in question 3.

It changed dramatically because in question 3, it only considered the log(DIS) against one other factor. It didn't factor in any of the other myriad factors that can affect housing prices, particularly in Boston. When weighed against everything from crime levels, building age, location on the Charles, etc. the impact of the distance can highly fluctuate.

1. Come to a conclusion regarding the relationship of air quality (measured by NOX). Is there evidence that NOX has a relationship on house value? Be sure to describe both the value of that relationship and the possible range of values.

The coefficient of NOX is -0.92, meaning that air quality has a fairly steady negative correlation on housing value. However, NOX is not independent from other influencing factors, like the heavily industrialized Charles River banks, poverty, and others, so it cannot drive housing prices alone. But a negative correlation of nearly 1/1 says it is, based around these pretenses, a good indicator of rising or sinking housing values.