Homework 5

Pranab

April 06, 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)
```

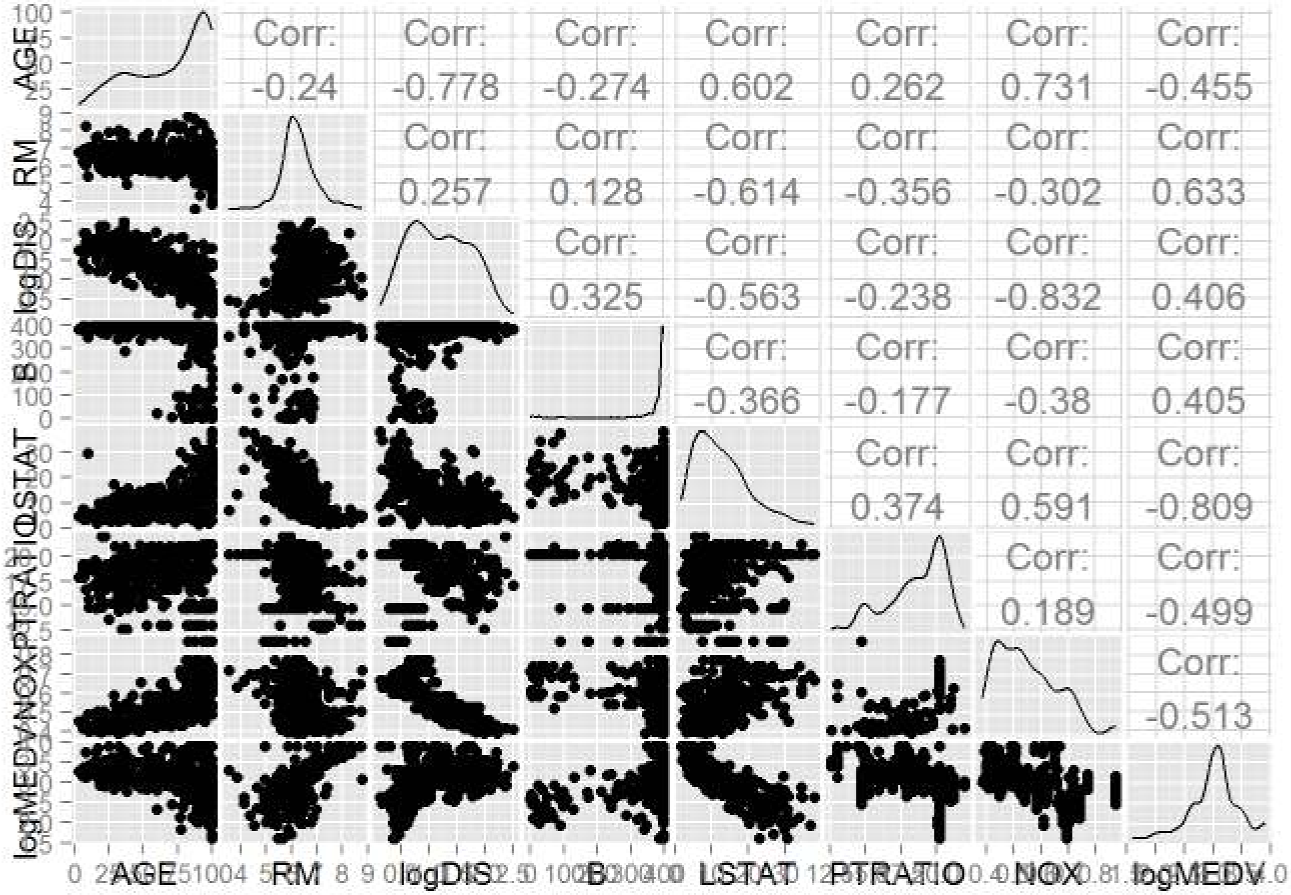
```
##         TOWN TOWNNO TRACT      LON     LAT MEDV CMEDV    CRIM ZN INDUS
## 1     Nahant      0  2011 -70.9550 42.2550 24.0  24.0 0.00632 18  2.31
## 2 Swampscott      1  2021 -70.9500 42.2875 21.6  21.6 0.02731  0  7.07
## 3 Swampscott      1  2022 -70.9360 42.2830 34.7  34.7 0.02729  0  7.07
## 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)
```

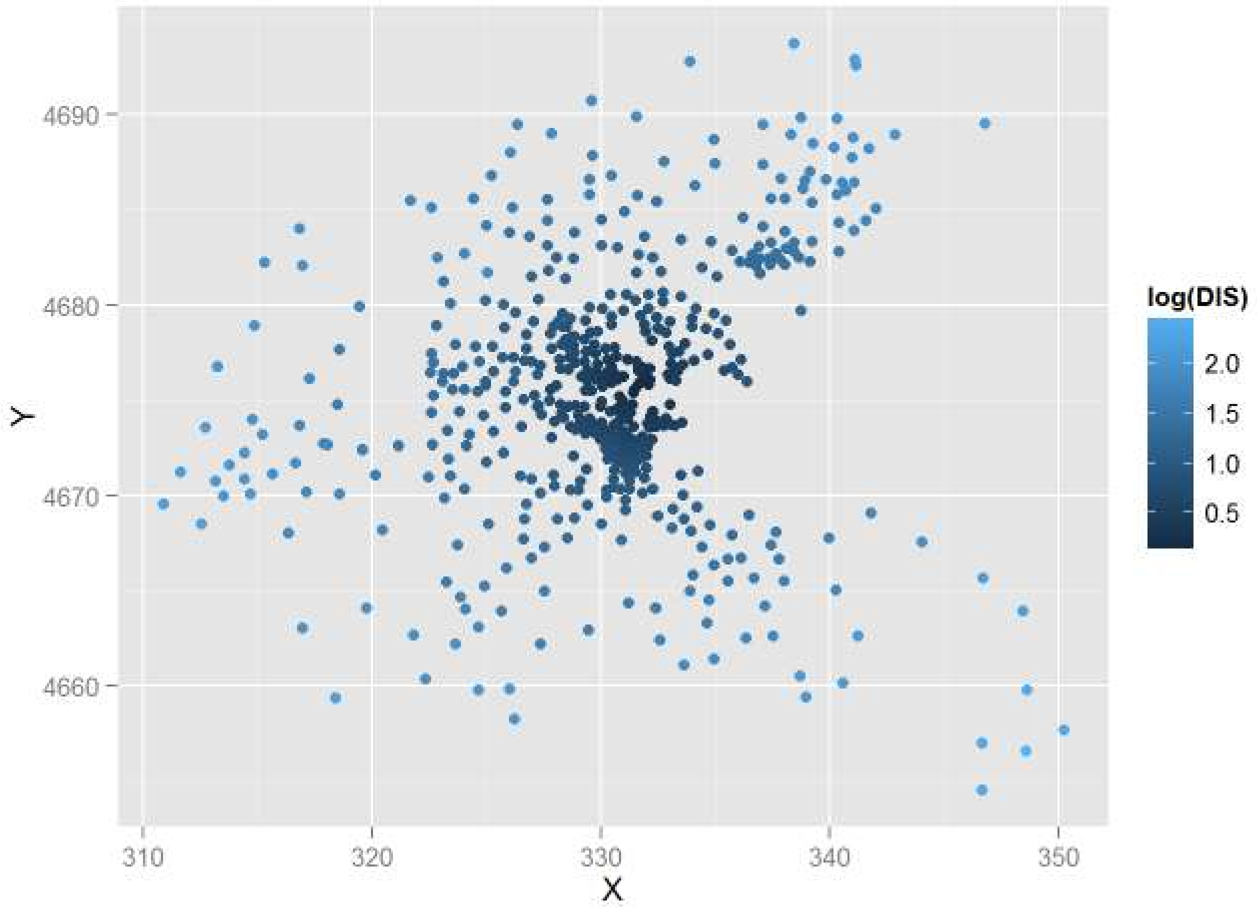
```
## Warning: package 'GGally' was built under R version 3.1.3
```

```
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`),

a. Describe the correlates of house price.

Answer:

From the scatterplot matrix the house price is found to be negatively correlated with median house prices. The correlation coefficient reported is -0.455. This could be interpreted as other price determining variables being equal, aged houses will be sold for lower prices.

b. Describe the correlates of NOX.

Anser:

The NOX is again found to be negatively correlated with the median house price with a correlation coefficient of -0.513. This signifies that buyers will pay a lesser price for houses located in areas with higher correlation, holding all other price determinig variables equal.

2. 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.

Answer:

Here we try to predict the way NOX, AGE, log(DIS), RM, CRIM, PTRATIO, B, LSTAT and CHAS affect log(MEDV) in a multivariate regression set up. In other words, how the median house values are affected by each of these factors one at a time, holding the other ones equal/fixed.

NOX~log(CMEDV): Now if all other factors are held fixed, a buyer would pay less price for a house which is situated in a polluted zone (where NOX concentration is higher). Thus, all other factors being equal, the NOX will have a negative regression coefficient.

AGE~log(CMEDV): If we assume other factors are fixed, then an aged house will sell for a lower price, as it would have maintainance issues and the buyer generally would have to spend additionl money on the upkeep of the house after purchase. Thus, if other factors are fixed, AGE will have a negative correlation coefficient.

log(DIS)~log(CMEDV): All other factors being fixed, the relationship between the distance and house price are negative. The farther it is from the city center, lower the value is.

RM~log(CMEDV): If other factors are assumed to play equal parts, a house with more number of rooms will have higher selling price. Thus the RM would then have a positive regression coefficient.

CRIM~log(CMEDV): IF we were to assume other factors played equal role in price determination, a house in a high crime infested locality will bring in low selling pricecompared to the ones located in low crime areas. Thus, other factor held fixed, CRIM will have a negative regression coefficient.

PTRATIO~log(CMEDV): If we consider all other factors are equal, then a pupil teacher ratio will signify a better educational environment and will push th emedian house prices higher. Thus there will be a positive relationship.

B~log(CMEDV): Given all other factors held fixed with equal effect and given the historical era the data pertains to, I fear these two variables may be negatively correlated or there could be no relationship between these two at all.

LSTAT~log(CMEDV): In general, holding all other factors equal, higher proportion lower status population will result in lower house prices. Thus, there will be a negative relationship.

CHAS~log(CMEDV): We know that Charles River is a part of industrial areas of the town. Thus being close to this place would mean staying close to job, which is a good thing for house owners. Thus with increasing distance from this place, the house prices will go down, signifying a negative relationship.

```
reg2 <<- lm(log(CMEDV)~NOX + AGE + log(DIS) + RM + CRIM + PTRATIO + B + LSTAT + CHAS, data=boston.c
)
summary(reg2)
```

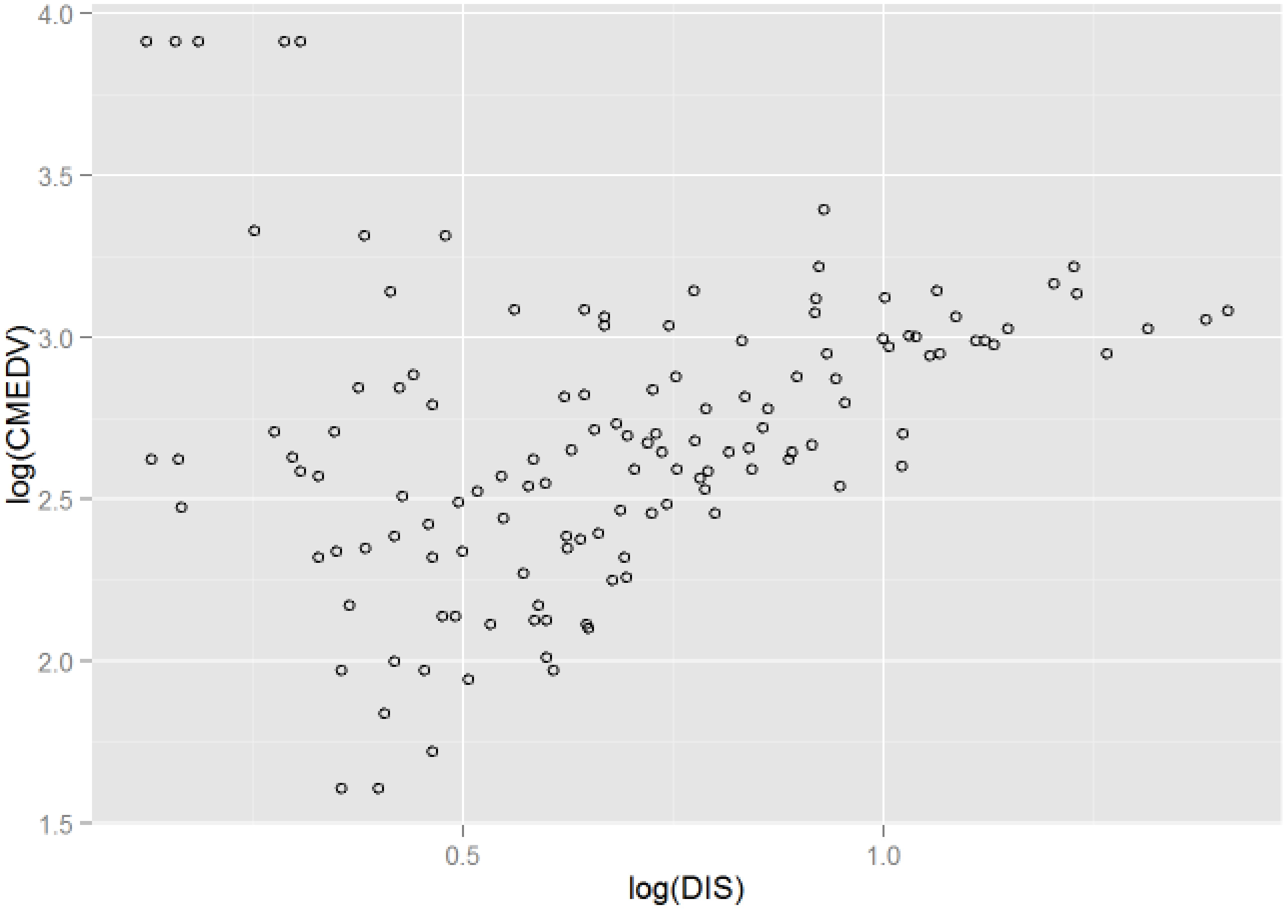
```
##
## Call:
## lm(formula = log(CMEDV) ~ NOX + AGE + log(DIS) + RM + CRIM +
##     PTRATIO + B + LSTAT + CHAS, data = boston.c)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.72962 -0.09850 -0.00867  0.09295  0.86452
##
## Coefficients:
##               Estimate Std. Error 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
```

3. 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

a. Report the slope of this regression and interpret it’s value.

Answer:

```
boston <- boston.c[grep("Boston", boston.c$TOWN), ]
ggplot(boston, aes(x=log(DIS), y=log(CMEDV))) + geom_point(shape=1)
```



```
reg1 <- lm(log(CMEDV)~log(DIS), data=boston)
summary(reg1)
```

```
##
## Call:
## lm(formula = log(CMEDV) ~ log(DIS), data = boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.97862 -0.23330 -0.01213  0.17116  1.42414
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.44375    0.09856  24.795  < 2e-16 ***
## log(DIS)     0.36211    0.13338   2.715  0.00753 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4364 on 130 degrees of freedom
## Multiple R-squared:  0.05366,    Adjusted R-squared:  0.04638
## F-statistic: 7.371 on 1 and 130 DF,  p-value: 0.00753
```

The bivariate regression run between log(CMEDV) and log(DIS) results in a weak positive relationship having regression coefficient of 0.05366. The slope of the line is 0.36211. This suggests that with increase of 1 unit of diatance from city core, the resultant median house prices will increase by 0.36211 units.

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

Answer:

```
confint(reg1, 'log(DIS)', level=0.95)
```

```
##                2.5 %    97.5 %
## log(DIS) 0.09823669 0.6259905
```

The approximate 95% confidence intervals for the slope is 0.09823669 and 0.6259905. This could also be calculated as the intercept (0.36211) +/- two times the standard error (0.13338).

4. Fit the linear regression from question 2.

```
reg3 <<- lm(log(CMEDV)~NOX + AGE + log(DIS) + RM + CRIM + PTRATIO + B + LSTAT + CHAS, data=boston.c
)
summary(reg3)
```

```
##
## 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
```

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

Answer:

The coefficient of log Distance from the regression summary is -0.2499594. This means with increase of an extra unit of diastance, the median house value will decrease by -0.2499594 units. This highlights the negative relationship between the distance from city core and median house prices given all pther factors are fixed.

```
confint(reg3, 'log(DIS)', level=0.95)
```

```
##               2.5 %     97.5 %
## log(DIS) -0.3133072 -0.1866116
```

The 95% confidence interval is -0.3133072 and -0.1866116.

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

Answer:

We got a negative relation (-0.2499594) while conducting the multivariate regression but in a bivariate regression between the distance and median gouse price we got weak positive relationship with a coefficient of 0.36211.

This could be described in the light of the fact that during multivariate regression, the regrassion model assumes that all other factors are acting equally at all distances while detrmining the coefficient for log(DIS). This results in the relationship following the general convention that all other things being equal, the house price decreases with increasing dostance from city core. No assumption here is being made to analyse other factors that vary with distance (like pollution, crime rate, new development, schools etc) and play an important role to determine price.

But while doing the bivariate analysis, we are just considering a data that reflects house prices at varying distances. Since we are not considering other factors into analysis, we are not holding them fixed as well and that effectively reverses the relationship to positive.

c. 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.

Answer:

From the multivariate regression output as shown at the begining of th ethis question, we get a coefficient of -0.9250979 for NOX. This suggests that assuming all other factors being equal, every unit of increase in NOX associates with a fall of -0.9250979 unit in median house prices. From this measure, we may se a very strong negative association between house prices and air quality.

The relationship beyween the NOX and log(CMEDV) is found to be statistically significant (p value much lesser than 0.05), thus we may assume that the air quality is a significant determinant of house value.

```
confint(reg3, 'NOX', level=0.95)
```

```
##         2.5 %     97.5 %
## NOX -1.202763 -0.6474324
```

The value of the relationship is -0.9250979 and the range of values at 95% confidence level lies between -1.202763 and -0.6474324.