# Lab 3

*Gosuddin Siddiqi*

*February 9, 2017*

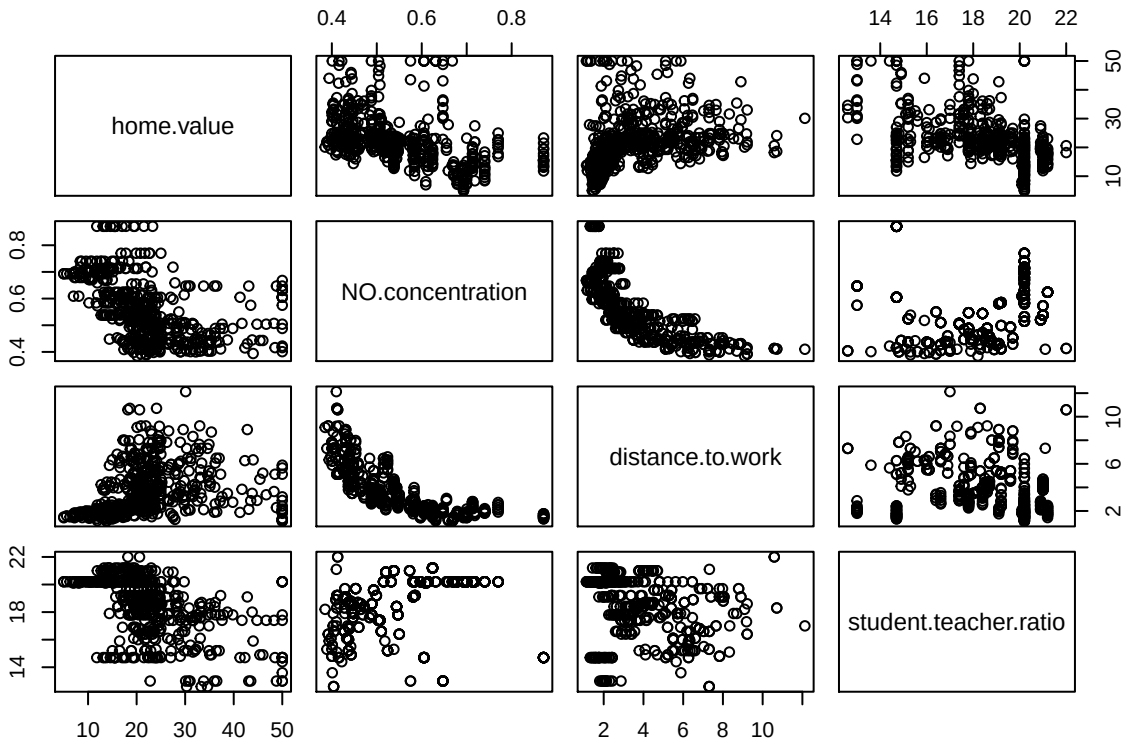
The data we are examining is housing values in suburbs of Boston and we wish to examine the various factors influencing the value of homes. Data come from: Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air. J. Environ. Economics and Management 5, 81-102. Dataset description: Table below gives the variables in the dataset. home.value: median value of owner-occupied homes in $1000s. NO.concentration: nitrogen oxides concentration (parts per 10 million) distance.to.work: weighted mean of distances to five Boston employment centers student.teacher.ratio: Student-teacher ratio by town

```
load("BostonData.Rdat")
```

Research questions: we wish to examine the effects of NOx, distance to employment, and education quality (as assessed through the student-teacher ratio) on home value. We will test them individually and then adjust for all factors.

0) Plot a scatterplot matrix of the data and describe what you see. Are any variables tightly related?

```
plot(boston)
```




On observing the scatterplot, the NO concentration decreases with the increase in distance to work. The correlation observed here is not tight but we can say that there exists negative correlation.

1) Fit single linear regressions with home.value as the outcome and each of the predictors:

```r
mod1 <- lm(home.value ~ NO.concentration, data = boston)
summary(mod1)
```

```
##
## Call:
## lm(formula = home.value ~ NO.concentration, data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.691  -5.121  -2.161   2.959  31.310
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       41.346      1.811   22.83   <2e-16 ***
## NO.concentration  -33.916      3.196  -10.61   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.323 on 504 degrees of freedom
## Multiple R-squared:  0.1826, Adjusted R-squared:  0.181
## F-statistic: 112.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
mod2 <- lm(home.value ~ distance.to.work, data = boston)
summary(mod2)
```

```
##
## Call:
## lm(formula = home.value ~ distance.to.work, data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.016  -5.556  -1.865   2.288  30.377
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       18.3901     0.8174  22.499  < 2e-16 ***
## distance.to.work   1.0916     0.1884   5.795 1.21e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.914 on 504 degrees of freedom
## Multiple R-squared:  0.06246,    Adjusted R-squared:  0.0606
## F-statistic: 33.58 on 1 and 504 DF,  p-value: 1.207e-08
```

```r
mod3 <- lm(home.value ~ student.teacher.ratio, data = boston)
summary(mod3)
```

```
##
## Call:
## lm(formula = home.value ~ student.teacher.ratio, data = boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.8342  -4.8262  -0.6426   3.1571  31.2303
##
```

```
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             62.345      3.029   20.58   <2e-16 ***
## student.teacher.ratio   -2.157      0.163  -13.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.931 on 504 degrees of freedom
## Multiple R-squared:  0.2578, Adjusted R-squared:  0.2564
## F-statistic: 175.1 on 1 and 504 DF,  p-value: < 2.2e-16
```

What are the associations of NOx, distance to employment, and education with home value?

Association of NOX with home value:

The model can be interpreted as with an increase in a unit of NOX concentration there is a decrease in the home value by $33k

Association of Distance to work with home value:

The model can be interpreted as with an increase in a unit of the distance to work (Boston employment centres)the home value increases by $1.09K

Association of education:

The model can be interpreted as with an increase in a unit of student-teacher ratio, the home value decreases by $2.15k

2) Compare the adjusted r^2 values for each model. Which predictor explains the data the best?

The R-Squared value for student-teacher ratio as a predictor, is maximum when compared to other predictors. Thus, we can say that Student-Teacher ratio accounts the most for variability in the data as compared to other predictors.

3) Run a multiple linear regression with home.value as the outcome and the other three variables as the predictors:

```
mod.full <- lm(home.value ~ distance.to.work + NO.concentration +
student.teacher.ratio, data = boston)
summary(mod.full)
```

```
##
## Call:
## lm(formula = home.value ~ distance.to.work + NO.concentration +
##     student.teacher.ratio, data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.434  -4.931  -1.270   2.951  32.606
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            89.0255     4.2358  21.017  < 2e-16 ***
## distance.to.work       -1.2803     0.2374  -5.393 1.07e-07 ***
## NO.concentration      -44.7740     4.2729 -10.479  < 2e-16 ***
## student.teacher.ratio  -1.9939     0.1503 -13.270  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.109 on 502 degrees of freedom
```

```
## Multiple R-squared:  0.4061, Adjusted R-squared:  0.4026
## F-statistic: 114.4 on 3 and 502 DF,  p-value: < 2.2e-16
```

Again interpret the associations. Do you see anything surprising (hint: maybe distance to work)? What could explain this discrepancy? Remember you are now adjusting each predictor for the other values.

The slopes have changed. The increasing effect of an unit increase has been reversed.

The model can be interpreted as: For every increase in a unit to distance to work, the home value decreases by $1.3k, keeping the other two predictors constant For every increase in a unit of NOX concentration, the home value decreases by $44.7k, keeping the other two predictors constant For every increase in a unit Student Teacher ratio, the home value decreases by $2k, keeping the other two predictors constant

There could be counfounding variables that explains the effect of distance in simple linear regression model.

4) Compare the adjusted r2 value for this multivariate model to the single linear regression values, which model fits the data the best?

For the model which takes into account all the 3 predictors has the best adjusted R-Squared value. Thus, the model with 3 predictors fits the best.

5) Predict and find the prediction interval the median home value of a home 3 km from work, with a NOx concentration of 0.35, and a student-teacher ratio of 10.

```
predict(mod.full, newdata=data.frame("distance.to.work" = 3,
"NO.concentration" = 0.35, "student.teacher.ratio" = 10),
interval="prediction")
```

```
##        fit      lwr      upr
## 1 49.57499 35.22737 63.92261
```

For the given values of the predictors, the median home value is lies between $35.27k and $63.92k