Applied Statistics Assignment 1

CID: 01389741

October 25, 2021

1 Exploratory Data Analysis

In this section we explore the Boston Housing Dataset provided from the University of Toronto, originally published by Harrison et al. [1]. We begin with the exploratory analysis by looking at some key characteristics of the dataset:

```
> dim(dfm)
[1] 506  13
```

The dataset includes 506 observations and 13 variables. We can explore the head of the dataset to understand the general structure:

```
> head(dfm)
     crim zn indus chas   nox    rm  age    dis rad tax ptratio lstat medv
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3  4.98 24.0
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8  9.14 21.6
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8  4.03 34.7
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7  2.94 33.4
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7  5.33 36.2
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7  5.21 28.7
```

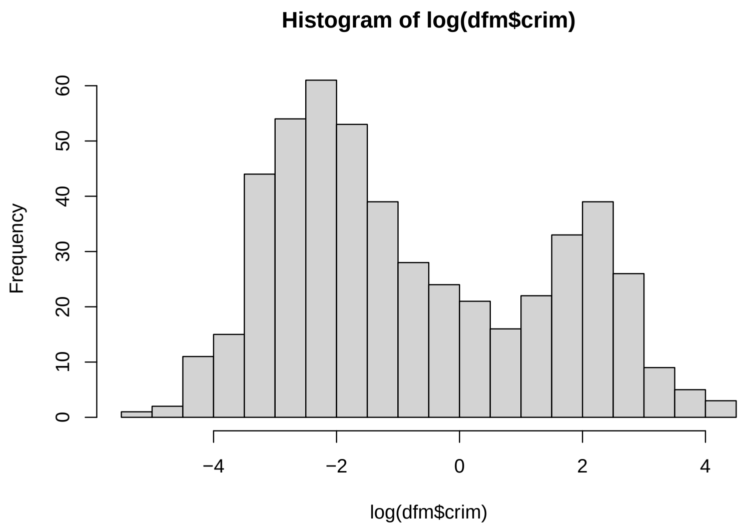
Insightful statistics of the dataset can be seen from the summary statistics of each of the variables:

```
> summary(dfm[,1:5])
      crim                zn             indus       chas         nox
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   0:471   Min.   :0.3850
 1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1: 35   1st Qu.:0.4490
 Median : 0.25651   Median :  0.00   Median : 9.69           Median :0.5380
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14           Mean   :0.5547
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10           3rd Qu.:0.6240
 Max.   :88.97620   Max.   :100.00   Max.   :27.74           Max.   :0.8710

> summary(dfm[,6:9])
       rm             age              dis              rad
 Min.   :3.561   Min.   :  2.90   Min.   : 1.130   Min.   : 1.000
 1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100   1st Qu.: 4.000
 Median :6.208   Median : 77.50   Median : 3.207   Median : 5.000
 Mean   :6.285   Mean   : 68.57   Mean   : 3.795   Mean   : 9.549
 3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188   3rd Qu.:24.000
 Max.   :8.780   Max.   :100.00   Max.   :12.127   Max.   :24.000

> summary(dfm[,11:13])
    ptratio          lstat            medv
```

```
 Min.   :12.60   Min.   : 1.73   Min.   : 5.00
 1st Qu.:17.40   1st Qu.: 6.95   1st Qu.:17.02
 Median :19.05   Median :11.36   Median :21.20
 Mean   :18.46   Mean   :12.65   Mean   :22.53
 3rd Qu.:20.20   3rd Qu.:16.95   3rd Qu.:25.00
 Max.   :22.00   Max.   :37.97   Max.   :50.00

> summary(dfm[,10:13])
      tax           ptratio          lstat            medv
 Min.   :187.0   Min.   :12.60   Min.   : 1.73   Min.   : 5.00
 1st Qu.:279.0   1st Qu.:17.40   1st Qu.: 6.95   1st Qu.:17.02
 Median :330.0   Median :19.05   Median :11.36   Median :21.20
 Mean   :408.2   Mean   :18.46   Mean   :12.65   Mean   :22.53
 3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:16.95   3rd Qu.:25.00
 Max.   :711.0   Max.   :22.00   Max.   :37.97   Max.   :50.00
```

These tables contain a lot of information. We can observe that the crime variable should include at least one outlier since the mean is around the value 0 with upper and lower quantiles of 0.08 and 3.61 but the maximum value of the dataset is 88.97 which is either an outlier, or the data is heavily skewed. Observing the histogram of the logarithm of this variable shows two peaks. This should be taken in account if it will be used in a model.

```
hist(log(dfm$crim), breaks=30)
```

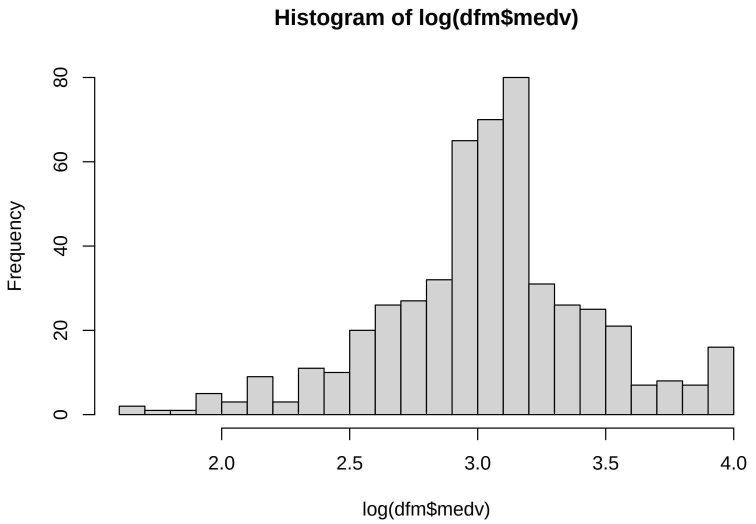


The `medv` variable which corresponds to house prices seems to have similar statistics so again a plot of the histogram was made to conclude that this variable is also heavily skewed so we take the logarithm and observe the histogram:

```
> hist(log(dfm$medv), breaks=30)
> mean(log(dfm$medv))
[1] 3.034513
> sd(log(dfm$medv))
[1] 0.4087569
```

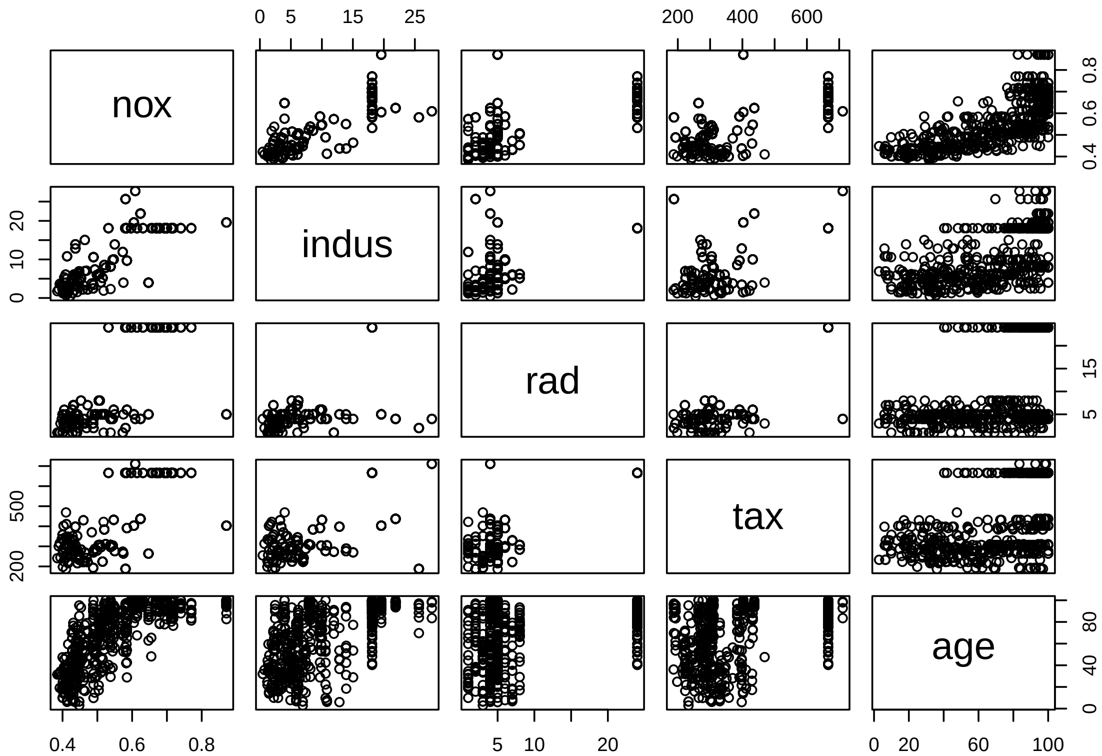
So this plot suggests that the house price follows a normal distribution with mean 3.03 and standard deviation of 0.41.

Now looking at the pairplot of the variables we aim to use regression on later on could give us an insight in important relationships:




```
> pairs(nox ~ indus + rad + tax + age, data=dfm )
```

The pairplot suggests that the variables `rad` and `tax` have a high correlation. This would suggest that it would not be a good idea for both to be included in a model since their columns might be linearly dependent. The `nox` variable which is the dependent variable seems to have a linear correlation with `indus`, a linear correlation with age (with a lot of deviation) and some suspicious points when plotted against `rad` and `tax`. Namely, it seems that many of points have the same value of `rad` and `tax` but different values of `nox`. Given that these two independent variables are highly correlated, we will definetely have to look at this in more depth in the analysis to come.



2 Linear Regression

We consider a Normal Linear Regression model for this dataset with the depended variable the nitric oxide concentration (`nox`). The form of the model is:

$$Y_i = \beta_1 + \beta_2 x_i + \epsilon_i, \quad i = 1, \ldots, n \tag{1}$$

where Y_i are the responses for the data y_i ($i = 1, \ldots n$) with covariates x_i and $\epsilon_i \overset{\text{iid}}{\sim} N\left(0, \sigma^2\right)$. In this model, we expect to see points of y_is and x_is follow a linear trend. Note that we do not actually know β_1 and β_2 before making the fit, we will actually predict them later on. The expected line will not perfectly fit all of the datapoints, instead the vertical differences are expected to follow $N\left(0, \sigma^2\right)$. The assumptions made in this model are the following:

- Linearity of the mean (possible non-linear trend of responses with covariates)
- Erros are Normally Distributed
- Mean of errors is zero
- Variances of the errors are the same
- Off-diagonal elements of the error matrix are zero. Namely, there is no covariance in the model.

In our dataset the exact equation of the regression fit will have the following form:

$$y_i^{\text{nox}} = \beta_0 + \beta_1 x_i^{\text{indus}} + \beta_2 x_i^{\text{rad}} + \beta_3 x_i^{\text{tax}} + \beta_4 x_i^{\text{age}} + \epsilon_i, \quad i = 1, \ldots, n, \quad \epsilon \sim \mathcal{N}\left(0, \sigma^2 \mathrm{I}\right) \tag{2}$$

The βs will be estimated using a likelihood function to obtain:

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y \in \mathbb{R}^p, \quad p = 5 \;\text{ including the intercept term} \tag{3}$$

where

$$X = \left[1 \left| x_i^{\text{indus}} \right| x_i^{\text{rad}} \left| x_i^{\text{tax}} \right| x_i^{\text{age}}\right] \in \mathbb{R}^{n \times p}, \quad i = 1, \ldots, n$$

Implementing this model in `R` we get the following summary statistics:

```
> model1 <- lm(formula = nox ~ indus + rad + tax + age, data = dfm)
> summary(model1)

Call:
lm(formula = nox ~ indus + rad + tax + age, data = dfm)

Residuals:
      Min        1Q    Median        3Q       Max
-0.142896 -0.035140 -0.009734  0.024423  0.249569

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.401e-01  1.205e-02  28.230  < 2e-16 ***
indus       6.488e-03  6.860e-04   9.457  < 2e-16 ***
rad         2.227e-03  7.996e-04   2.786  0.00554 **
tax         3.002e-05  4.771e-05   0.629  0.52953
age         1.586e-03  1.314e-04  12.077  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06301 on 501 degrees of freedom
```

```
Multiple R-squared:  0.7067,        Adjusted R-squared:  0.7044
F-statistic: 301.8 on 4 and 501 DF,  p-value: < 2.2e-16
```

The summary provides a few insights towards the fit that was done. The test performed to check the impact of the variables on the fit was a t-test which is based on the t-statistic:

$$t = \frac{\hat{\beta}_j}{\text{se}(\hat{\beta})}, \quad \text{where} \quad \text{se}(\hat{\beta}) = \sqrt{\frac{e^T e}{n-p} (X^T X)^{-1}}$$

and e are the residuals.

The summary suggests that tax was not significant in the fit of the regression. We can consider a hypothesis test for β_3 – the coefficient for tax. The null hypothesis is $H_0 : \beta_3 = 0$ against the alternative hypothesis $H_1 : \beta_3 \neq 0$ (in presence of the other variables). The t-statistic value is 0.629 and comparing this t-statistic value to a Student's t-distribution with $n-p$ degrees of freedom we get a large p-value of 0.529, giving us insufficient evidence to reject the null hypothesis that tax is statistically significant in modelling the nitric oxide concentration when the other variables are present in the model at a 1% significance level. Looking at the other variables, all of them present p-values very close to zero and we can conclude under the same hypothesis test as above for their respective βs that there is sufficient evidence to reject the null hypothesis that βs for variables `indus`, `rad`, `age` and intercept are 0 at a 1% significance level and should therefore be included in the model.

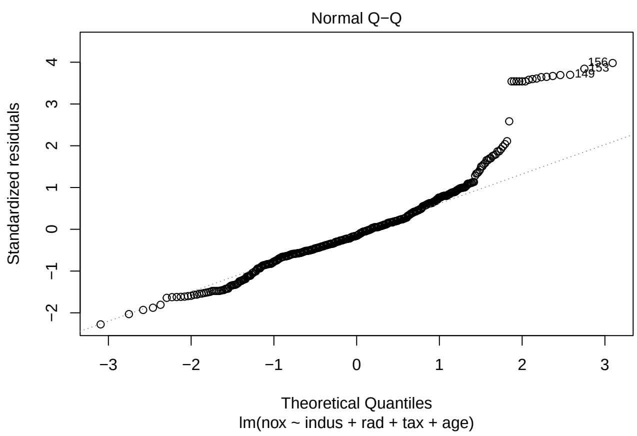
We can also look at the adjusted R^2 value defined as:

$$R^2_{\text{adj}} = 1 - \frac{\text{RSS}/(n-p)}{\left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right)/(n-1)}$$

where RSS is the residual sum of squares, the y_is are the random variables and $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$. This value lies between the values 0 and 1 and it shows how strong the correlation is between the covariates and the dependent variable is. The value of 1 corresponds to the perfect model. We get a value of 0.70 which shows a fairly good correlation but could definitely be improved. Note that we decided to include the adjusted R^2 rather than just R^2, since it takes into account the degrees of freedom of the model.

Now we look at the Quantile-Quantile Plot which creates a graph of the quantiles of the two distributions against one another. In the normal linear model as stated above, the errors are assumed to be normally distributed independently, and the residuals and standardised residuals are also assumed to be normally distributed for large n. If they are indeed normally distributed, the pattern should follow the line $y = x$. [2].

```
> plot(model1,2)
```



Clearly there is a problem with the residuals. They follow the expected pattern until the value of approximately 1 for theoretical quantiles but then there is a jump which diverges from the line for the rest of the points. This shows that the suspicious points do not follow $N\left(0, \sigma^2\right)$ and should be investigated later on.

3 Analysis of Variances

We can look at the variance more closely. Consider two nested models, where we fit a model with the same variables as above with the difference that in the one model we have the tax variable at the beginning of the equation and the other model with the tax at the end. We can look at a significance test known as ANOVA (analysis of variance). Note that comparing the sum of squared residuals under two models follows an F-distribution [2]. So the F statistic value seen in the ANOVA tables is the value of significance of each estimator for the response.

```
> anova(lm(formula = nox ~ tax + indus + rad  + age, data = dfm))
Analysis of Variance Table

Response: nox
           Df  Sum Sq Mean Sq F value    Pr(>F)
tax         1 3.02604 3.02604  762.29 < 2.2e-16 ***
indus       1 1.12358 1.12358  283.04 < 2.2e-16 ***
rad         1 0.06352 0.06352   16.00 7.288e-05 ***
age         1 0.57903 0.57903  145.86 < 2.2e-16 ***
Residuals 501 1.98879 0.00397
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(lm(formula = nox ~ indus + rad  + age + tax, data = dfm))
Analysis of Variance Table

Response: nox
           Df Sum Sq Mean Sq  F value    Pr(>F)
indus       1 3.9544  3.9544 996.1619 < 2.2e-16 ***
rad         1 0.2587  0.2587  65.1713 5.138e-15 ***
age         1 0.5775  0.5775 145.4743 < 2.2e-16 ***
tax         1 0.0016  0.0016   0.3958    0.5295
Residuals 501 1.9888  0.0040
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

It is clear from the above tables that the ordering of the tax variable has a significant change the analysis on their variances. Namely, when it is included in the model in the first position there is sufficient evidence to reject the null hypothesis (p-value $= 10^{-6}$) that the estimator for the tax is 0, whereas when the tax is last in the equation, then there is not enough evidence for the null hypothesis to be rejected. This is a strange result and it shows that the observational data is unbalanced. We can see this because ANOVA uses Type I significance testing where testing is done in the order the variables are specified in our model. It tests how much variance can be explained from the first variable, and then tests the how much of the remaining variance can be explained by the second variable and so on. This could be avoided by using a Type II or Type III test [3].

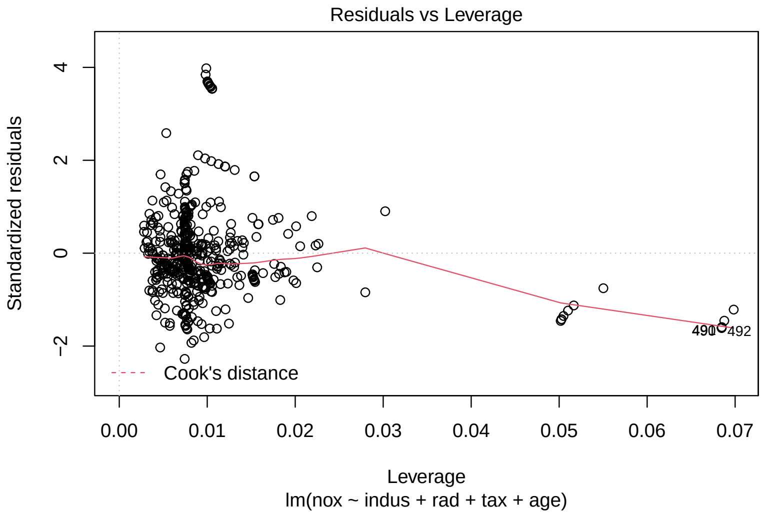
4 Cook's Distance

Cook's Distance is a way to measure the *influence* it has on the estimator β. It is defined as:

$$C_i = \frac{\left(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}}\right)^T X^T X \left(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}}\right)}{p\text{RSS}/(n-p)} \tag{4}$$

where $\widehat{\boldsymbol{\beta}}$ is the estimator calculated *without* using the ith observation [4]. We can plot the Standardized residuals against the leverage, defined as a value that measures the potential influence of a certain observation in the regression fit. It solely depends on the covariates.

```
plot(model1, which=5)
```



The plot shows that there are a few points which have a very high leverage value which influence the model. They have leverage values of around 0.07 and we are going to consider these points as outliers and see how the model changes. The threshold we will use will be Cook's distance (as defined in 4) of 0.02. That is, we will remove any points with Cook's distance larger than 0.02.

The following code was adapted from [5].

```
> influential <- as.numeric(names
(cooks.distance(model1))[(cooks.distance(model1) > 0.02)])
> length(influential)
[1] 24
> new_dfm <- dfm[-influential, ]
> dim(new_dfm)
[1] 482  13
> model2 <- lm(formula = nox ~ indus + rad + tax + age, data = new_dfm)
> summary(model2)

Call:
lm(formula = nox ~ indus + rad + tax + age, data = new_dfm)

Residuals:
      Min        1Q    Median        3Q       Max
-0.141139 -0.024327 -0.001127  0.022885  0.161647

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.511e-01  9.958e-03  35.262  < 2e-16 ***
indus       4.228e-03  5.471e-04   7.728 6.53e-14 ***
rad         3.926e-03  7.112e-04   5.520 5.58e-08 ***
```

```
tax          3.517e-05  4.112e-05   0.855    0.393
age          1.409e-03  9.539e-05  14.776  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04551 on 477 degrees of freedom
Multiple R-squared:  0.8063,	Adjusted R-squared:  0.8047
F-statistic: 496.5 on 4 and 477 DF,  p-value: < 2.2e-16
```

We identify that there are 24 outliers in our model. Removing these and storing the cleaned data under the name `new_dfm`, we can see that now we have a total of 482 observations. The parameter estimates of this new model are similar to the previous model (`model1`) except the fact that we can see that the p-value for the `rad` estimate is smaller than before (order of 10^{-8} in comparison with order of 10^-2) so there is stronger evidence to reject the null hypothesis of not including the `rad` observation. Another change in this model is that the adjusted R^2 has increased to 80%, showing that the correlation now is better between the variables, reinforcing our claim that the points removed indeed negatively influenced our model.

5 Prediction

Now we wish to make some predictions with respect to the model we have fitted. We will use the $\hat{\beta}$s defined in 3 to make a prediction of what the response $y_\star$ should be given $x_\star$. We define this response as

$$y_\star = x_\star \cdot \beta + \epsilon_\star \approx x_\star \cdot \hat{\beta} + \epsilon_\star \tag{5}$$

The prediction intervals for these $y_\star$ are given by:

$$\widehat{y}_\star \pm t_{n-p}^{(\alpha/2)} \widehat{\sigma} \sqrt{1 + \boldsymbol{x}_\star^T \left(X^T X\right)^{-1} \boldsymbol{x}_\star} \tag{6}$$

where $P\left(T \leq t_{n-p}^{(\alpha/2)}\right) = 1 - \alpha/2, T \sim t_{n-p}$, t_{n-p} is the Student's t-distribution with $n-p$ degrees of freedom and $\widehat{\sigma}$ is the *estimate* of the standard deviation. Note that this is the $100(1-\alpha)\%$ confidence interval for a *single future response*. Using `R` to make this prediction:

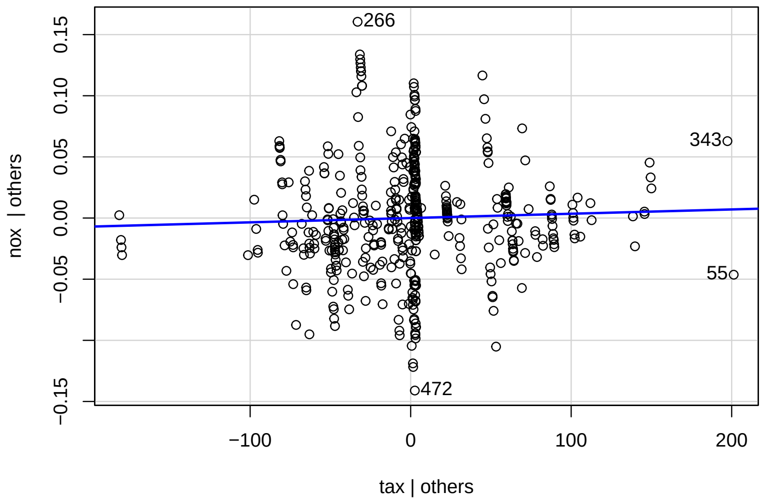
```
> trial_xs <- c(median(new_dfm[['indus']]), median(new_dfm[['rad']]),
+               median(new_dfm[['tax']]), median(new_dfm[['age']]))
>
> predict(model2,
+         newdata = data.frame(indus = trial_xs[1], rad = trial_xs[2],
+                                           tax = trial_xs[3], age = trial_xs[4]),
+         interval = 'prediction', level = 0.99)
        fit       lwr       upr
1 0.5237466 0.4058262 0.6416671
```

which means that $y_\star = 0.52$ with 99% confidence intervals of 0.41 and 0.64. The interval shows that when predicting $y_\star$ with the $\hat{\beta}$s, 99% of the time the value will be within the range of 0.41 and 0.64.

6 Added Variable Plot

Now we make an added variable plot for the variable `tax` with respect to the other predictors `indus, rad, age`. The reason for this is that we want to know if it is sensible to include the `tax` variable in the model or not. This plot will be the residuals of a model which has dependent variable `tax` and independent variables `indus, rad` and `age`, against the residuals of the model of `nox` against `tax, indus, rad` and `age`. In order to do this plot we use the function `avPlots` from the library `car`.

```
> library(car)
> res <- avPlots(model2,layout=c(1,1))
> lsfit(res$tax[,1], res$tax[,2])$coefficients
   Intercept             X
1.514637e-18 3.516851e-05
```



We can see that the blue line which is the line of best fit of this plot has a gradient of nearly zero (10^{-5}). This suggests that there is a correlation of the tax variable with some other variable and this clearly affects the model. This was foreshadowed in the EDA section 1, since the `tax` and `rad` observations were highly correlated. The gradient of this line suggests excluding the `tax` variable from the model.

7 Revised Model

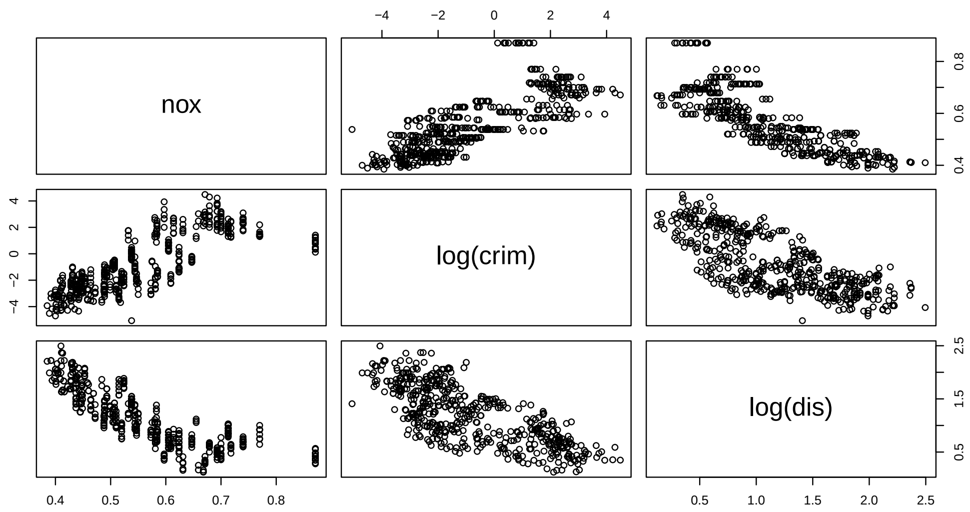
Now we look for adding new parameters in the model. First we plotted the variables excluded from the initial model against `nox` to look for a linear relationship. It seemed like none of the variables had a linear relationship with `nox` but `crim` and `dis` seemed to have some kind of relationship. Taking the logarithm of both observations eluded a relatively linear relationship between `nox` and those two observations. This can be seen on the following pairplot

```
pairs(nox ~ log(crim) + log(dis), data =dfm)
```

This linear relationship suggested that we should include them in the model. We created two new models (`m2` and `m3`) with the same observations as before (`m1`), and adding `log(crim)` to the first one and `log(dis)` to the second one.

Now there are multiple ways with which we can determine the goodness of a fit. We could look for the lowest RSS value although usually the model with more parameters typically has the lowest RSS, we could look for the highest R^2 value, look at the analysis of the variances of the models or divide the data into a training set and a test set and find the best predictive performance. A combination of the above methods is optimal which is what we did for this report.

Table 1 shows the values of RSS and Adjusted R^2 we found by running all 3 models. The lowest RSS value is given by `m3` but as explained before this might be because it has more parameters, and the highest Adjusted R^2 value is given again by `m3`.



Parameters	m1	m2	m3
RSS	0.063	0.059	0.057
Adjusted R^2	0.704	0.740	0.756

Table 1: Summary of models m1, m2, m3

Looking at the ANOVA tables, and testing the Null Hypothesis of having the model `m1` against the alternative of having model `m2`, through an F-test there is evidence to reject the null hypothesis with a p-value $= 10^{-5}$. The same test with alternative hypothesis on having model `m3` again presents sufficient evidence to reject the Null Hypothesis with p value $= 10^{-6}$.

So combining the observed RSS, Adjusted R^2 and ANOVA analysis we know that both `m2` and `m3` are better than `m1`. To decide between the two models we are going to use prediction, split the dataset into training and testing set, compute the mean squared error (MSE) $\sum_{i=1}^{M} (\widehat{y}_i - y_i)^2$ where M is the number of sample in the test and repeat this process enough times to avoid biases. This will be done for all 3 models and the one with the lowest MSE will be the best one.

Splitting the dataset into 80% training and 20% testing and we replicated 1,000 simulations for different training the testing (this was essential since we knew there were outliers in our dataset from previous sections but we are confident that these 1,000 simulations were enough to decrease the bias). Taking the average of the simulations, the results are shown in table 2. The code of this part was long and therefore it was included in the Appendix for the interest of the reader.

m1	m2	m3
0.0040(8)	0.0035(6)	0.0033(6)

Table 2: Summary of models m1, m2, m3

We conclude that `m3`, the model with the same observations as the original model but with added term the logarithm of `dis` (weighted distances to five Boston employment centres), is the best model out of the 3 based on RSS, adjusted R^2, conducted F-tests and predictive performance.

References

[1] David Harrison and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, 1978. ISSN 0095-0696. doi: https://doi.org/10.1016/0095-0696(78)90006-2. URL https://www.sciencedirect.com/science/article/pii/0095069678900062.

[2] Nick Heard. Normal linear model. *Imperial College London (Lecture)*, page 35, October 2021.

[3] EdM. Why do p-values change in significance when changing the order of covariates in the aov model? Cross Validated. URL https://stats.stackexchange.com/q/212700.

[4] Din-Houn Lau. Normal linear model. *Imperial College London (Lecture)*, page 11, October 2020.

[5] user3459010. Removing outliers based on cook39;s distance in r language. Cross Validated. URL https://stats.stackexchange.com/q/164099.

A Code for Model Comparison

```
> set.seed(111)
> MSE_func <- function(){
+     split <- sample(seq_len(nrow(dfm)), size = floor(0.80*nrow(dfm)))
+     train <- dfm[split, ]
+     test <- dfm[-split, ]
+
+     m1 <- lm(formula = nox ~ indus + rad + tax + age, data = train)
+     m2 <- lm(formula = nox ~ indus + rad + tax + age + log(crim), data = train)
+     m3 <- lm(formula = nox ~ indus + rad + tax + age + log(dis), data = train)
+     y_hat_m1 <- predict(m1, test, level = 0.99)
+     y_hat_m2 <- predict(m2, test, level = 0.99)
+     y_hat_m3 <- predict(m3, test, level = 0.99)
+
+     y_actual <- test['nox']
+     data_all <- data.frame(actual=y_actual, pred_m1 = y_hat_m1,
+                            pred_m2 = y_hat_m2, pred_m3 = y_hat_m3)
+
+     MSE_m1 <- mean((data_all$nox - data_all$pred_m1)^2)
+     MSE_m2 <- mean((data_all$nox - data_all$pred_m2)^2)
+     MSE_m3 <- mean((data_all$nox - data_all$pred_m3)^2)
+
+     return(c(MSE_m1, MSE_m2, MSE_m3))
+ }
>
> tr <- replicate(1000, MSE_func())
> mean(tr[1,])
[1] 0.004006903
> mean(tr[2,])
[1] 0.003531534
> mean(tr[3,])
[1] 0.003312969
```