First Coursework - Applied Statistics

CID: 02091191

October 25, 2021

Abstract

In this document we are presenting the work conducted during the first Applied Statistics coursework of the year.

Introduction

The coursework focuses on a version of the Boston Housing data [1] relating median house prices in different suburbs of Boston in 1978 to some attributes. The dataset contains 506 observations of 13 attributes which are described as:

CRIM: per capita crime rate by town

ZN: proportion of residential land zoned for lots over 25,000 sq.ft.

INDUS: proportion of non-retail business acres per town

CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)

NOX: nitric oxides concentration (parts per 10 million)

RM: average number of rooms per dwelling

AGE: proportion of owner-occupied units built prior to 1940

DIS: weighted distances to five Boston employment centres

RAD: index of accessibility to radial highways

TAX: full-value property-tax rate per $10,000

PTRATIO: pupil-teacher ratio by town

LSTAT: % lower status of the population

MEDV: median value of owner-occupied homes in $1000's

1 Exploratory analysis of the data

1.1 Data types and ranges of values

First, we perform an exploratory analysis in **R** of the data as it is the first step before considering or fitting any model. Using the `summary` function in **R** provides the following results we brought together in Table 1 for the sake of readability.

	Min	1st Qu.	Median	Mean	3rd Qu.	Max
crim	0.00632	0.08205	0.25651	3.61352	3.67708	88.97620
zn	0.00	0.00	0.00	11.36	12.50	100.00
indus	0.46	5.19	9.69	11.14	18.10	27.74
chas	0.00	0.00	0.00	0.06917	0.00	1.00
nox	0.3850	0.4490	0.5380	0.5547	0.6240	0.8710
rm	3.561	5.886	6.208	6.285	6.623	8.780
age	2.90	45.02	77.50	68.57	94.08	100.00
dis	1.130	2.100	3.207	3.795	5.188	12.127
rad	1.000	4.000	5.000	9.549	24.000	24.000
tax	187.0	279.0	330.0	408.2	666.0	711.0
ptratio	12.60	17.40	19.05	18.46	20.20	22.00
lstat	1.73	6.95	11.36	12.65	16.95	37.97
medv	5.00	17.02	21.20	22.53	25.00	50.00

Table 1: Boston Housing data ranges and distributions

All these features are `numeric` except from `chas` variable which is `factor` (or Boolean). Also, there are no missing values in the data set. The `chas` variable is highly unbalanced with 471 FALSE and 35 TRUE. We need to be careful using this variable further in the coursework when fitting linear models.

We notice that `crim`, `zn`, `medv`, `nox` and `dis` variables have significantly large differences between their mean and median, which may result in highly skewed data and maybe outliers. In addition, the `rm` variable shows a high range of values with 1st and 3rd quartiles close to the mean and median, which may also result in the presence of outliers. Plotting the box plots would help us confirming these hypotheses.

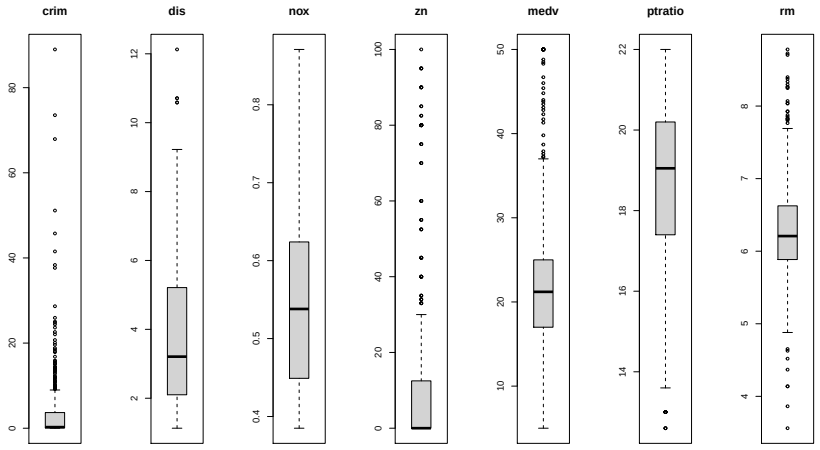
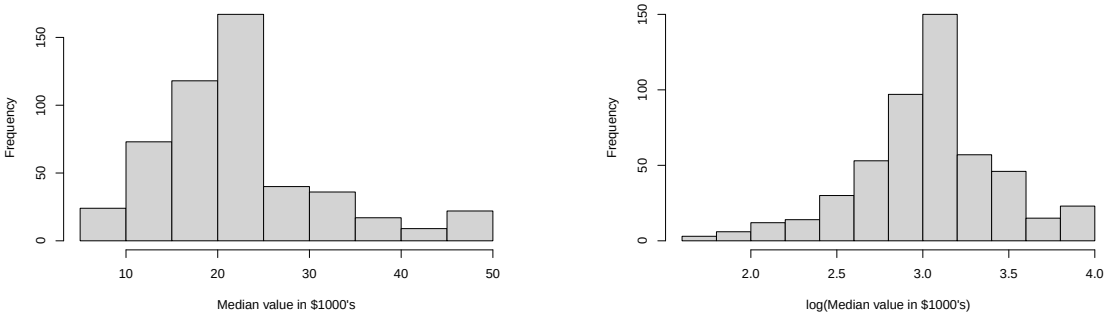


Figure 1.1: Box Plots of variables which may contain outliers

As Figure 1.1 shows, `crim`, `zn`, `rm` and `medv` contain lots of outliers. If we wanted to consider them in a linear regression model for example, it would be wise to transform them to make them more relevant, using a `log` transformation for right-skewed data `crim`, `zn` and maybe `medv`. The `dis` and `nox` variables are also right-skewed but don't seem to have as much outlier values as the precedent cited variables: applying a log transformation would be wise too. The `ptratio` variable seems in the contrary to be left-skewed: here, applying a square transformation would be also wise in order to make it normally distributed. The `rm` variable seems to look like the bell curve of a normal distribution with thick tails: it may for issue follow a Cauchy distribution

and it requires further investigation to find a relevant transformation.

Thus, some variables are highly skewed or contain extreme values. Considering these variables directly in a linear model would lead to non significant coefficients (in terms of range). Moreover, the housing prices data contained in `medv` variable are right-skewed and it seems consistent as a lot of houses are around the median value and few of them are very expensive.



(a) House pricing distributions before transformation

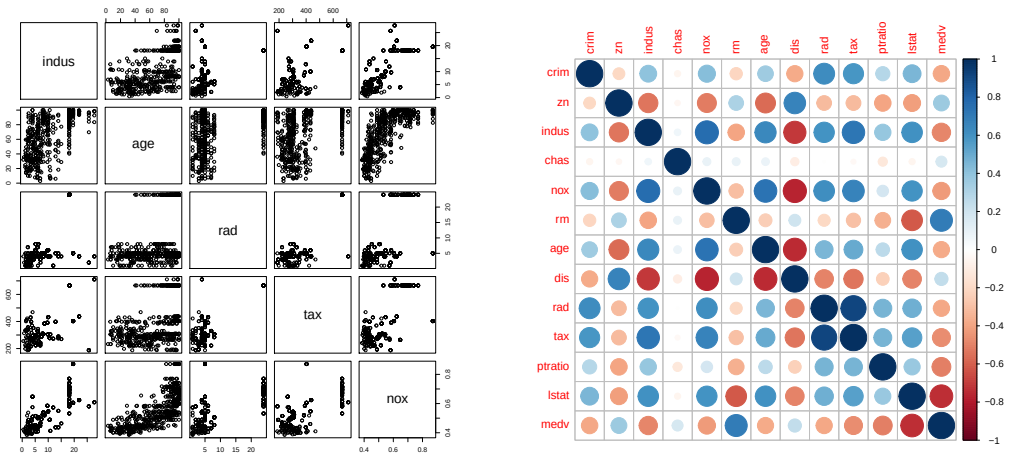
(b) House pricing distributions after `log` transformation

Figure 1.2: House pricing distributions

As shown on Figure 1.2 and discussed before, the house pricing distribution is right-skewed. However, it doesn't seem that high pricing values are extreme or real outliers. The distribution shape is the result of the market prices with a lot of "middle priced" housings and only a few "expensive" housings. Applying a `log` transformation would be wise here as shown on Figure 1.2b where the distribution looks more normal.

1.2 Data interactions and relationships

Now focus on relationships between data. Representing the whole pair plot would not have been relevant here as there are too many variables and we would not have seen anything. Thus, I chose the variables we will focus on in the next part of the coursework. We may focus on relationships between variables shown on a pair plot and on a correlation matrix.



(a) Pair plot between some variables

(b) Correlation matrix

Figure 1.3: Relationships between housing data

The pair plot on Figure 1.3a shows trends between variables. In particular the last row of this pair plot is interesting when considering a model for `nox` variable with all other variables as predictors. It makes sense to consider these variables as there are trends between them and the nitric oxyde concentration. However, we may first notice that sometimes data points are splitting in two groups: this may be caused by outlier values or integer (index) values. Also, some trends seem not to be strictly linear as we may for example observe a curve for the relationship between `age` and `nox` variables: this could be improved by transforming the data (here a log transformation for `nox` variable would be appropriate). These trends are confirmed on the right Figure 1.3b: all variables seem to be correlated to each other. Thus, considering linear models here between some variables (for example modelling the housing prices from all the other variables) would be consistent here.

2 Fitting a linear model

2.1 Specification of the normal linear model

Consider a normal linear model for nitric oxyde concentration `nox` with the following predictors: `indus`, `rad`, `tax`, `age` along with an intercept term.

$$\texttt{nox}_i = \beta_1 + \beta_2 \texttt{indus}_i + \beta_3 \texttt{rad}_i + \beta_4 \texttt{tax}_i + \beta_5 \texttt{age}_i + \epsilon_i \quad i = 1, \dots, n$$

where n is the number of observations and $\varepsilon_i \overset{iid}{\sim} N\left(0, \sigma^2\right)$. The matrix form of this equation is:

$$\underbrace{\begin{bmatrix} \texttt{nox}_1 \\ \vdots \\ \texttt{nox}_n \end{bmatrix}}_{\boldsymbol{Y}} = \underbrace{\begin{bmatrix} 1 & \texttt{indus}_1 & \texttt{rad}_1 & \texttt{tax}_1 & \texttt{age}_1 \\ \vdots & \vdots & \vdots & & \\ 1 & \texttt{indus}_n & \texttt{rad}_n & \texttt{tax}_n & \texttt{age}_n \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}}_{\boldsymbol{\varepsilon}}, \ \boldsymbol{\epsilon} \sim N\left(0, \sigma^2 I_n\right)$$

where

- $\boldsymbol{Y}$ is the n-dimensional random vector of observations, called the response
- $X \in \mathbb{R}^{n \times p}$ is the design matrix which contains the predictors
- $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown parameter vector
- ϵ is the n-variate unobserved error

As $\boldsymbol{\epsilon} \sim N\left(0, \sigma^2 I_n\right)$ for some $\sigma^2 > 0$, the normal linear model can also be written as

$$\boldsymbol{Y} \sim N\left(X\boldsymbol{\beta}, \sigma^2 I_n\right)$$

The assumptions of this model are:

- the linearity of the mean: if $\mathbb{E}[\boldsymbol{Y}] \neq X\boldsymbol{\beta}$, we would not expect any pattern between the response and the predictors.
- $\boldsymbol{\epsilon} \sim N\left(0, \sigma^2 I_n\right)$: explicitly the distribution of errors is normally distributed, with the expectation of the errors is equal to zero, the variance terms are all the same ($\text{Var}[Y_i] = \sigma^2$ for $i = 1, \dots, n$) and the off-diagonal elements are zero ($\text{Cov}[Y_i, Y_j] = 0$ for $i = 1, \dots, n$).

Fitting this model in **R** gives the following results:

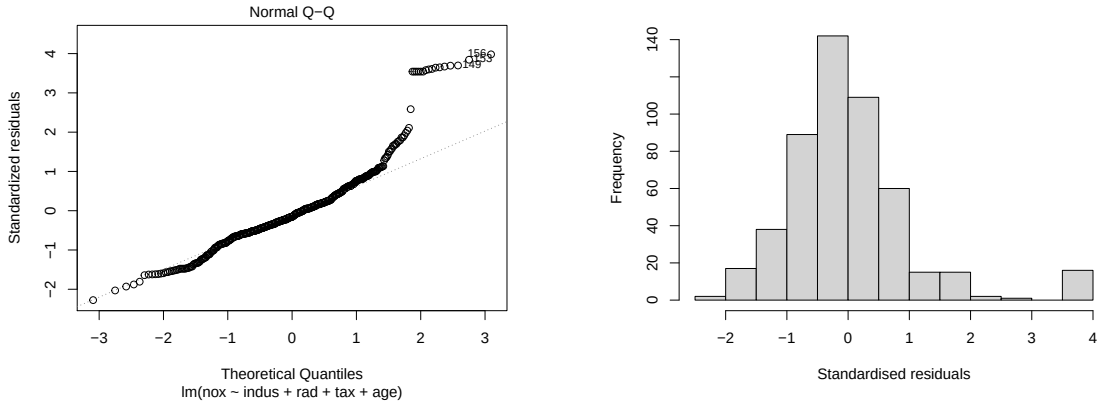
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.401e-01  1.205e-02  28.230  < 2e-16 ***
indus       6.488e-03  6.860e-04   9.457  < 2e-16 ***
rad         2.227e-03  7.996e-04   2.786  0.00554 **
tax         3.002e-05  4.771e-05   0.629  0.52953
age         1.586e-03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

For each $\boldsymbol{\beta}$ coefficient in the model, we proceed to a t-statistic test to test the null hypothesis $H_0 : \beta_j = 0$ against the alternative hypothesis $H_1 : \beta_j \neq 0$. Regarding to the results, there is no sufficient evidence to reject the null hypothesis (or we fail to reject the null hypothesis) that $\beta_4 = 0$, i.e. `tax` is not statistically significant in modelling `nox` when `indus`, `rad` and `age` are present in the model. In the contrary, the other t tests made here show enough evidence to reject the null hypothesises. Thus, the `tax` variable does not seem statistically significant in the model here, whereas `indus`, `rad` and `age` have significant regression coefficients.

We may furthermore notice that the regression coefficients associated to the significant variables here are yet close to zero. This is caused by the difference of range between `nox` values (close to zero) and the predictors' values (much higher). We could have re-scaled predictor values by dividing them by the same value (for example 100) in order to get higher regression coefficients, but it would not have changed the t-statistics test at all and we would fall to the same conclusions.

2.2 Quality of the fit

Consider the residuals of the model: $e = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$. The assumptions cited before require the residuals to be normally distributed. We can check this by plotting the Q-Q plot of the standardised residuals.



(a) Q-Q plot of standardised residuals

(b) Histogram of standardised residuals

Figure 2.1: Residuals distribution

Figure 2.1a shows that most of the standardised residuals are following a normal distribution. However, after approximately the 1.5 theoretical quantile, the standardised residuals jumps to extreme values: this would mean that the standardised residuals are not clearly following a normal distribution as the right tail is thicker than it should be. This can be confirmed plotting the histogram of the standardised residuals on Figure 2.1b: some residuals are too high and make the distribution non normal. We may perform a Kolmogorov-Smirnov test in **R** in order to test

the null hypothesis H_0: residuals are normally distributed against the alternative hypothesis H_1: residuals are not normally distributed. We obtain a `p-value = 1.49e-6` (< 0.05) so we reject the null hypothesis that the residuals are normally distributed. Also, we could have focused on the residuals vs fitted values plot.

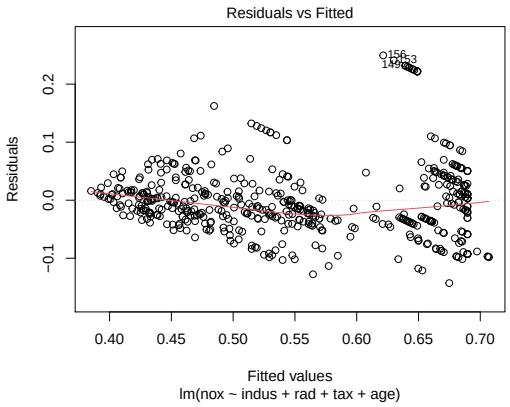


Figure 2.2: Residuals against fitted values plot

This scatter plot shows that the variance of the residuals does not seem constant along fitted values but increasing in terms of magnitude. This represents also an assumption violation (σ^2 should be constant). Thus, our model may present weakness such as outliers data points which are responsible of high variance in residuals and a non normal distribution of the standardised residuals. These outliers would be the data points associated to high residuals on Figure 2.1a. We may need further investigation in order to find these outliers and delete them from our model.

We now perform two analysis of variance for the above linear model. In the first model we include `tax` variable last and in the second one we include `tax` variable first. We proceed to an analysis of variance (ANOVA) in **R**.

	F value	**p-value**
`tax` first	762.29	$< 2.2e-16$
`tax` last	0.3958	0.5295

Table 2: Comparison of F values and p-values for `tax` variable performed by two ANOVA: one with `tax` variable first and one with `tax` variable last in the model

These commands lead to same residuals sum of squares (RSS) but slightly different results shown in Table 2. In fact, with `tax` first in the model, the ANOVA considers the `tax` variable to be relevant in the model in terms of variance with a `p-value < 2.2e-16`. However, when `tax` is the last variable in the model, the F-test (from Fischer distribution) gives a `p-value = 0.5295` which makes us think that the variable is not relevant by failing to reject the null. These different may seem unusual but in fact are explainable. However modifying the order of variables changes the design matrix, changing the order of variables should not change the fit process and we would obtain the same regression coefficients for each variable. To recall, the ANOVA purpose is to provide information about how much this new variable in the model would reduce the residual sum of squares. And this is why here when `tax` is the first variable, the ANOVA shows that `tax` helps to reduce the RSS, while when it is the last variable, it doesn't help to reduce the RSS because the other precedent variables already reduced the RSS and the new variable `tax` is not significantly important to still reduce the RSS. In a F-test, the order of variables matters whereas in a t-test it doesn't. Thus, the `tax` variable is not useless when used

first in the model, but doesn't give significantly more information on the response when at the end of the model.

2.3 Outliers detection and prediction

We now want to remove outliers from our data. We first plot standardised residuals and Cook's distance against leverage in **R**.

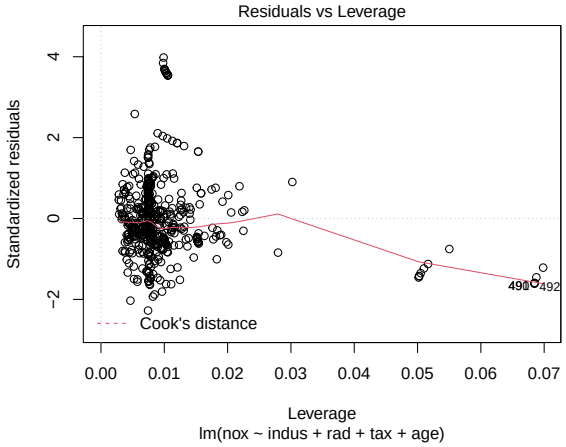


Figure 2.3: Standardised residuals and Cook's distance against leverage

As shown on Figure 2.3, some points in our data have higher leverage than most of the other points (with a leverage higher than 0.05). Also, some points have high standardised residuals (roughly 4 here). These points may be considered as outliers. In order to get rid of these data points that are distorting our linear model, we choose to delete all data points that have a leverage exceeding 0.05 or a Cook's distance exceeding 0.02. To recall, the Cook's distance is a measure which combines both leverage and standardised residual defined as:

$$C_i = r_i^2 \frac{h_{ii}}{(1 - h_{ii})r}$$

where r_i is the standardised residual, h_{ii} the leverage and $r = rank(X)$. After deleting points as described before we obtain a new data set containing 478 observations so we deleted 28 data points (roughly 5% of initial observations, which seems consistent). We then fit a new linear model to this new data set. Fitting this new model in **R** gives the following results:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.527e-01  1.039e-02  33.942  < 2e-16 ***
indus       4.392e-03  6.249e-04   7.029 7.32e-12 ***
rad         3.993e-03  7.238e-04   5.516 5.71e-08 ***
tax         2.630e-05  4.415e-05   0.596    0.552
age         1.406e-03  9.617e-05  14.624  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

	R squared	Adjusted R squared	RSE
Old model	0.7067	0.7044	0.06301
New model	0.8064	0.8047	0.04568

Table 3: Comparison between the same model on both filtered (new model) an unfiltered (old model) data sets on R squared, adjusted R squared and residual standard error (RSE)

We may notice the model on new data (without outliers) has approximately the same parameters' estimations as the previous one. Again, the regression coefficient of `tax` is not statistically significant. As shown in Table 3, the new model has significantly improved the R squared and the adjusted R squared. Moreover, the RSE is higher for the previous model on unfiltered data compared to on the new one on filtered data. This confirms our choice to remove the outlier was not a bad choice. For further investigation, we could have splitted the data in both a train set and a test set and evaluated the performances of both models in terms of prediction.

Using this new model, we are able to predict the response for a new data point, i.e. predict the nitric oxyde concentration knowing all the predictors. We can illustrate this by computing a prediction with the four predictors variables equal to their median values from the full data set.

```
> predict(mylm.2, newdata=data.frame(indus=9.69, rad=5.000, tax=330.0, age=77.50),
        interval="prediction", level=0.99)
        fit       lwr       upr
1 0.5328794 0.4145148 0.6512439
```

Hence, in a such suburb of Boston, we expect the nitric oxyde concentration to be equal to 0.533 approximately, and between 0.414 and 0.651 at a 99% confidence level. This confidence interval is defined as follows:

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

where $P(T < t_{n-p}^{(\alpha/2)}) = 1 - \alpha/2$ and $T \sim t_{n-p}$ and $\boldsymbol{x}_\star$ is the new predictors data. In fact, this interval takes into account the variance of $\boldsymbol{\epsilon}$ as it is a single response prediction, i.e. the nitric oxyde concentration $\widehat{y}_\star$ for the predictors data $\boldsymbol{x}_\star$ will be $\boldsymbol{x}_\star\boldsymbol{\beta} + \boldsymbol{\epsilon}$ whereas in mean it only will be $\boldsymbol{x}_\star\boldsymbol{\beta}$. To construct this interval, we should notice that the variance of a new measurement can be written

$$\begin{aligned}
\sigma_p^2 &= \mathrm{Var}[\epsilon] + \mathrm{Var}[\boldsymbol{x}_\star\boldsymbol{\beta}] \\
&= \sigma^2 + \sigma^2 \boldsymbol{x}_\star^T \left(X^T X\right)^{-1} \boldsymbol{x}_\star \\
&= \sigma^2 \left(1 + \boldsymbol{x}_\star^T \left(X^T X\right)^{-1} \boldsymbol{x}_\star\right)
\end{aligned}$$

Then,

$$\frac{y - \hat{y}_*}{\sigma_p^2} \sim t_{n-p}$$

And finally a $100(1-\alpha)\%$ confidence interval for a single future response is the one above. Note that we are replacing σ^2 which is unknown by its unbiased estimator $\hat{\sigma}^2 = \frac{RSS}{n-p}$.

2.4 Model checking and specification of a new linear model

We may wonder if the `tax` variable should be taken in account in our model. To do so, we are interested in the appropriateness of the linearity of the mean assumption in our linear model. We proceed as follows:

- First, we define a linear model with response `nox` and predictors `indus`, `rad` and `age`. We call it model one.
- Then, we define a second linear model with response `tax` and predictors `indus`, `rad` and `age`. We call it model two.
- Finally, we define a third linear model with response the residuals of model one and predictors the residuals of model two. We call it model three.
- If the slope of model 3 is significantly not equal to zero, then we should consider the `tax` variable in our model.

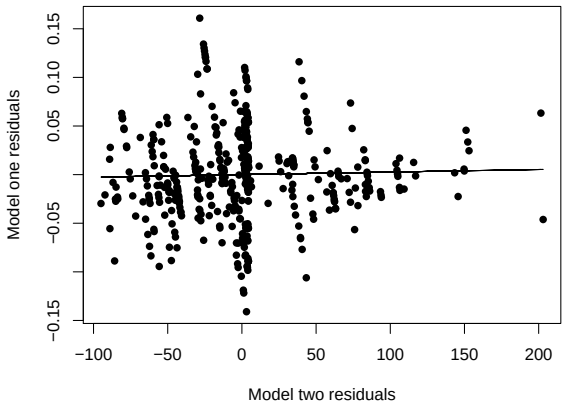


Figure 2.4: Residuals of model one against residuals of model two. Fitted line of model three

The result shown in Figure 2.4 and model three's summary in **R** suggest that the regression coefficient (or the slope) of this model is $2.6e-05$. The t-statistics test used to test if the slope is significantly equal to 0 or not has a p-value of 0.55: thus, we fail to reject the null hypothesis and we can affirm that the slope is equal to zero. Hence, we may remove the `tax` predictor from our data as it is not significantly useful to explain the response in the context of other predictors.

As we have shown that we could remove `tax` variable from the model, we may think about adding a new feature to replace it. As we want to model the nitric oxyde concentration which is directly related to air pollution, it seems to be wise to take into account as a new feature the weighted distances to five Boston employment centres (`dis` variable in **R**). Indeed, if the distance from the suburb to the centre of Boston is high we should expect the air pollution to be lower than if the distance from the suburb to the centre was low. In other terms, the farther you get from downtown Boston, the less polluted the air becomes. This can be confirmed on Figure 1.3b with a negative correlation coefficient. Moreover, as talked before and shown on Figure 1.1, the `dis` variable is right skewed: we would prefer to use a transformed version of `dis`: we will use `log(dis)` which seems to be normally distributed. Thus, we consider the new linear model with response `nox` and predictors `indus`, `rad`, `age` and `log(dis)`.

We can fit this new model in **R** and try to compare it to the old one we focused during this coursework.

- **RSS comparison:** The RSS of the previous linear model is 1.99 whereas the RSS of the new linear model is 1.64. Hence, in terms of RSS the new linear model is better than the old one.

- **Adjusted R squared:** The adjusted R squared of the old linear model was 0.70 whereas the adjusted R squared of the new linear model is 0.76. Thus, the new model has a slightly better adjusted R squared.
- **ANOVA:** Here, a rigorous ANOVA would not be possible as our two models have completely different covariates and we are not comparing a sub-model of an initial linear model.

Hence, considering the new feature `log(dis)` seems to improve significantly our initial model.

Conclusion

To conclude, this coursework allowed us to work on data exploratory, a necessary step before considering any linear regression model on these data. In addition, we were able to train and compare several linear regression models to model the nitric oxyde concentration in the air from several predictors. Finally, we were able to modify our initial model by removing outliers or adding a more relevant variable to improve performance. Due to time constraints, we were not able to separate the data into training and test datasets in order to compare the performance of the models in terms of prediction.

References

[1] **R** Documentation. Boston Housing Dataset description. `http://lib.stat.cmu.edu/datasets/boston`

[2] Dr Din-Houn Lau (Autumn 2020) *MATH70071 - Applied Statistics lecture notes*, Imperial College London MSc Statistics resources

[3] Dunn, P. & Smyth, G (2018) *Generalized Linear Models With Examples In R.* Springer Texts in Statistics. Springer

[4] Marina Evangelou (Oct. 2021) Introduction to LaTeX