MATH70071 Applied Statistics – Assessed Coursework

Coursework 1 Solutions

25 October 2021

Question 1

Obtaining the data:

```
library(mlbench)
data(BostonHousing)
dfm = as.data.frame(BostonHousing)
dfm = dfm[,!(names(dfm) == "b")] #Remove that covariate
dump("dfm",file="bos.R")
```

a)

Dimensions:

```
dim(dfm)
```

```
## [1] 506  13
```

```
sum(is.na(dfm))
```

```
## [1] 0
```

The data contain 506 observation vectors of 13 variables, with 0 missing values.

Data types:

```
sapply(dfm,class)
```

```
##      crim        zn     indus      chas       nox        rm       age       dis
## "numeric" "numeric" "numeric"  "factor" "numeric" "numeric" "numeric" "numeric"
##       rad       tax   ptratio     lstat      medv
## "numeric" "numeric" "numeric" "numeric" "numeric"
```

The variable chas is a binary factor variable (indicating whether the area borders the river) and the other 13 variables are numeric.

Summaries:

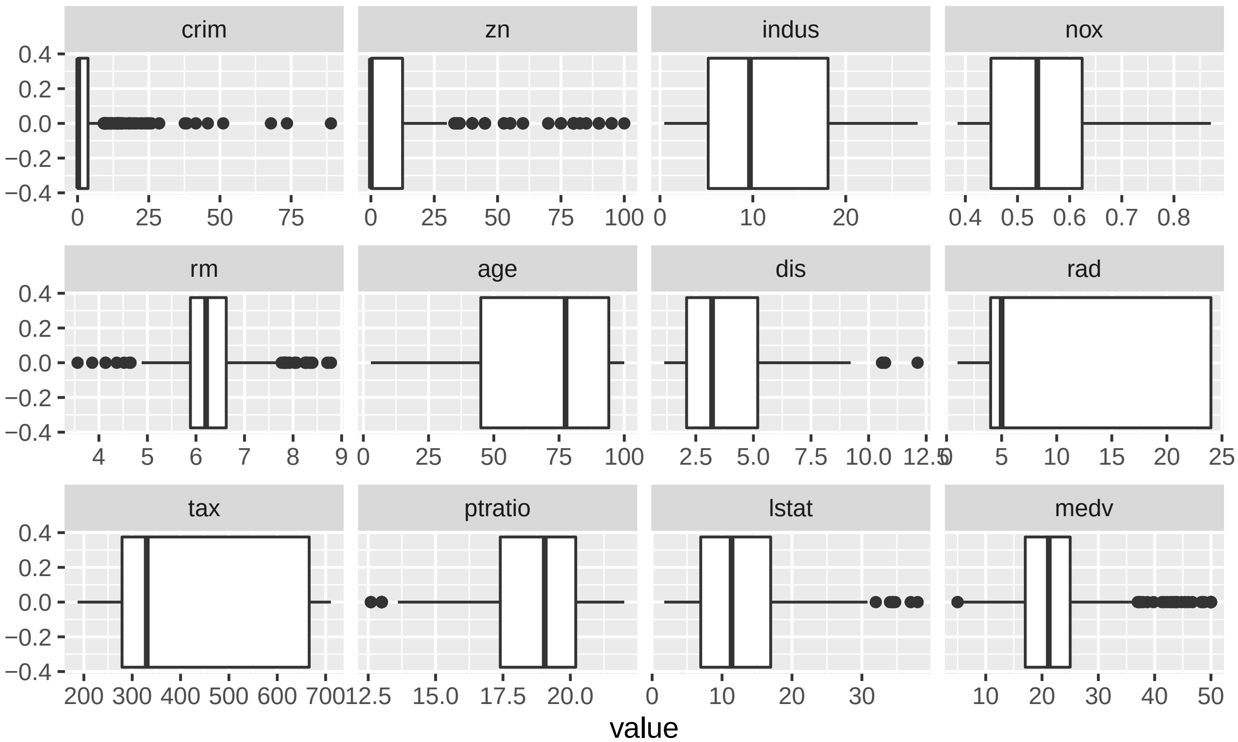
```
summary(dfm)
```

```
##       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
##        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
##       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
```

Box plots:

```r
library(reshape2)
library(ggplot2)
ggplot(melt(dfm[,!(names(dfm) == "chas")], id.vars = c()),aes(x = value)) +
    facet_wrap(~variable,scales = "free_x") +
    geom_boxplot()
```




```r
#par(mfrow=c(3,5))
#for (i in colnames(dfm)){
#  if(class(dfm[[i]])=='numeric'){
#    hist(dfm[,i])
#  }
#}
#apply(dfm[,!(names(dfm) == "chas")],2,hist)
```

Observations

```r
range(dfm[['medv']])
```

```
## [1]  5 50
```

```r
sum(dfm[['medv']] == max(dfm[['medv']]))
```

```
## [1] 16
```

```r
sum(dfm[['medv']] == min(dfm[['medv']]))
```

```
## [1] 2
```

- 16 suburbs attain the maximum value of $50,000, suggesting this variable may have been right-censored.

- 2 suburbs attain the maximum value of $5,000, suggesting this variable may also have been (left) censored.

6 Marks

---

b) Let $n = 506$, the number of suburbs, and $p = 5$. Then let $\mathbf{y} = (y_1, \ldots, y_n) \in \mathbb{R}^n$ denote the nitric oxide concentration (NOX) levels for the n suburbs. Next, let X be an $n \times p$ matrix with ijth entry

$$x_{ij} = \begin{cases} 1 & j = 1 \\ \mathrm{INDUS}_i & j = 2 \\ \mathrm{RAD}_i & j = 3 \\ \mathrm{TAX}_i & j = 4 \\ \mathrm{PTRATIO}_i & j = 5 \end{cases}$$

for $i = 1, \ldots, n$ and NAME_i corresponds to the value of variable NAME for suburb i.
Then under the normal linear model,

$$\mathbf{y} \sim \mathrm{N}_n(X\beta, \sigma^2 I_n),$$

an n-dimensional multivariate normal distribution, with $\beta \in \mathbb{R}^p$ and $\sigma > 0$.

3 Marks

---

c)

Fitting linear model:

```
lmn = lm(nox ~ indus + rad + tax + age, data=dfm)
summary(lmn)
```

```
##
## 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
```

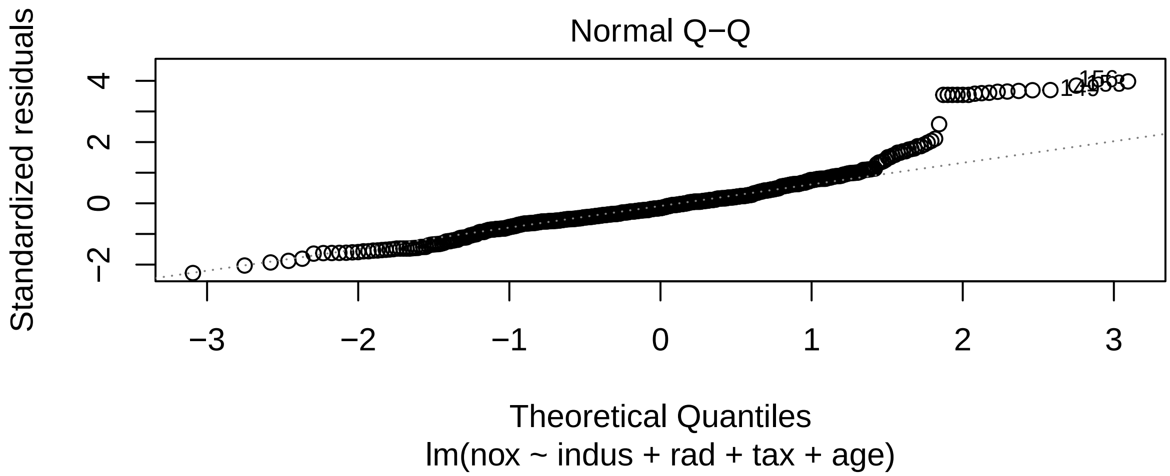
From this summary, the intercept and the the variables INDUS and AGE appear to have near-certain, significant non-zero regression coefficients. The variable RAD has a p-value of about 0.6%, and also appears to be significant but with less certainty. There seems little evidence to support TAX having a non-zero coefficient as part of this regression model.

4 Marks

---

d)

```
plot(lmn,2) #Q-Q plot
```



The jump on the right hand side of the plot (at ≈ 1.5) is suggestive of a bimodally distributed residuals, with a small clump of high, positive residuals. The remainder of the fit looks quite good.

3 Marks

---

e) Performing an analysis of variance twice, once with TAX included last and then with TAX included first:

```
anova(lm(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
```

```
anova(lm(nox ~ tax + indus + rad + age, data=dfm))
```

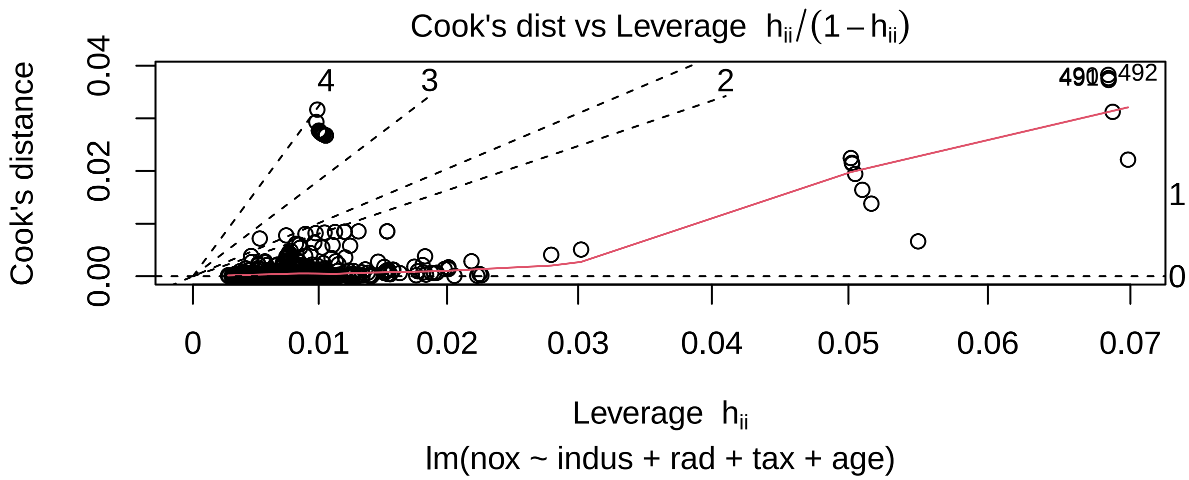
```
## Analysis of Variance Table
##
## Response: nox
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## tax         1 3.02604 3.02604  762.29 < 2.2e-16 ***
## indus       1 1.12358 1.12358  283.04 < 2.2e-16 ***
## rad         1 0.06352 0.06352   16.00 7.288e-05 ***
## age         1 0.57903 0.57903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

When positioned first in the formula, TAX becomes statistically significant in the analysis of variance. This suggests the information in the TAX predictor is well captured by the other predictors INDUS, RAD, TAX, AGE.

3 Marks

f)

```r
plot(lmn,6) #Cook's distances
```



Removing those observations with leverage exceeding 0.05:

```r
idx = hatvalues(lmn)>0.05 | cooks.distance(lmn)>0.02
sum(idx) #Number of observations to be discarded
```

```
## [1] 28
```

```r
dfm2 = dfm[!idx,]
lmn2 = lm(nox ~ indus + rad + tax + age, data=dfm2)
```

This filtering leads to 28 observations being discarded.

Comparing the coefficients side-by-side, they can be seen to be fairly similar despite the deletion of some potentially influential observations.

```r
lmn$coefficients
```

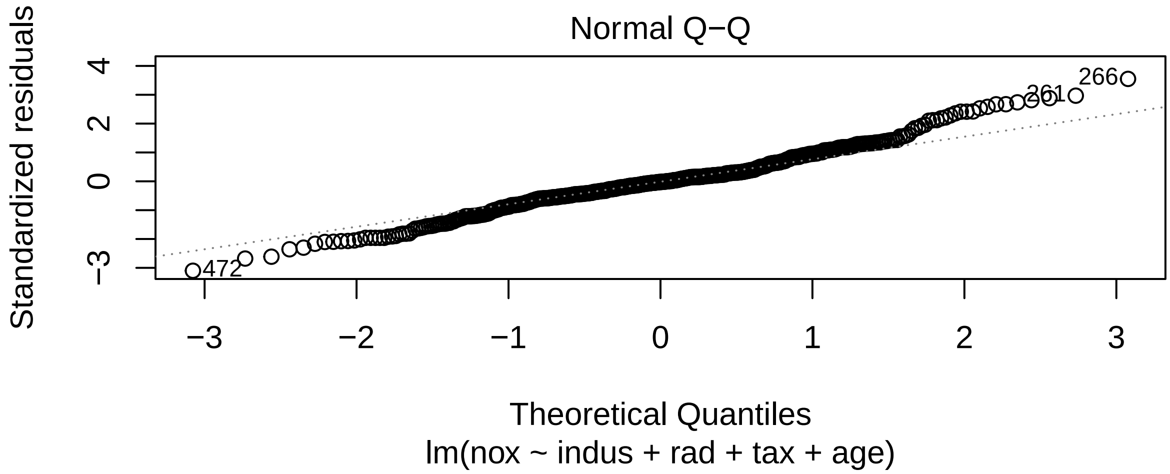
```
##  (Intercept)        indus          rad          tax          age
## 3.401280e-01 6.487624e-03 2.227503e-03 3.001623e-05 1.586452e-03
```

```r
lmn2$coefficients
```

```
##  (Intercept)        indus          rad          tax          age
## 3.526765e-01 4.392310e-03 3.992686e-03 2.629883e-05 1.406443e-03
```

Model fit:

```r
plot(lmn2,2) #Q-Q plot
```



This shows a much closer model fit, although the residuals still display heavier tails than the theoretical normal model.

8 Marks

---

g)

```
dfp=data.frame(indus=median(dfm[['indus']]),rad=median(dfm[['rad']]),
            tax=median(dfm[['tax']]),age=median(dfm[['age']]))
predict.lm(lmn2,newdata=dfp)[[1]]
```

```
## [1] 0.5328794
```

2 Marks

---

h)

```
dfp=data.frame(indus=median(dfm[['indus']]),rad=median(dfm[['rad']]),
            tax=median(dfm[['tax']]),age=median(dfm[['age']]))
predict.lm(lmn2,newdata=dfp,interval="predict",level=0.99)[2:3]
```

```
## [1] 0.4145148 0.6512439
```

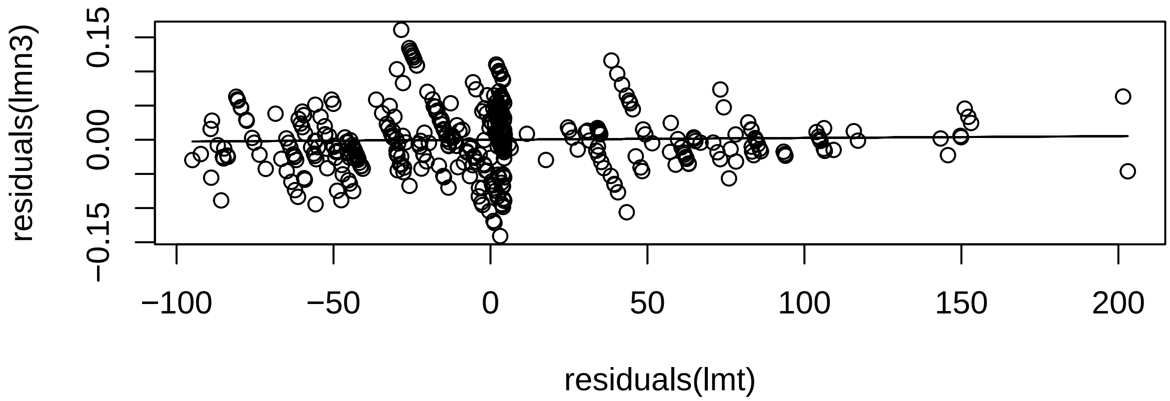
An approximate 99% confidence interval for the nitric oxide concentration level for a suburb with median levels of non-retail business acres, highway accessibility, housing age and tax rate is (0.415,0.651).

2 Marks

---

i)

```
lmn3 = lm(nox ~ indus + rad + age, data=dfm2)
lmt = lm(tax ~ indus + rad + age, data=dfm2)
lmr = lm(residuals(lmn3) ~ residuals(lmt))
plot(residuals(lmt),residuals(lmn3))
lines(residuals(lmt),predict(lmr))
```



The slope of the line is close to zero, further supporting the earlier finding, before filtering, that the TAX variable is well explained by the other predictors in this model.

4 Marks

---

j) Looping through each of the remaining unused variables, performing analysis of variance against the existing linear model:

```
vbls = names(dfm[,!(names(dfm) %in% c("nox","indus","rad","tax","age"))])
invisible(lapply(vbls, function(x) {
  print(anova(lmn,lm(substitute(nox ~  indus + rad + tax + age + i,
                               list(i = as.name(x))), data = dfm)))}))
```

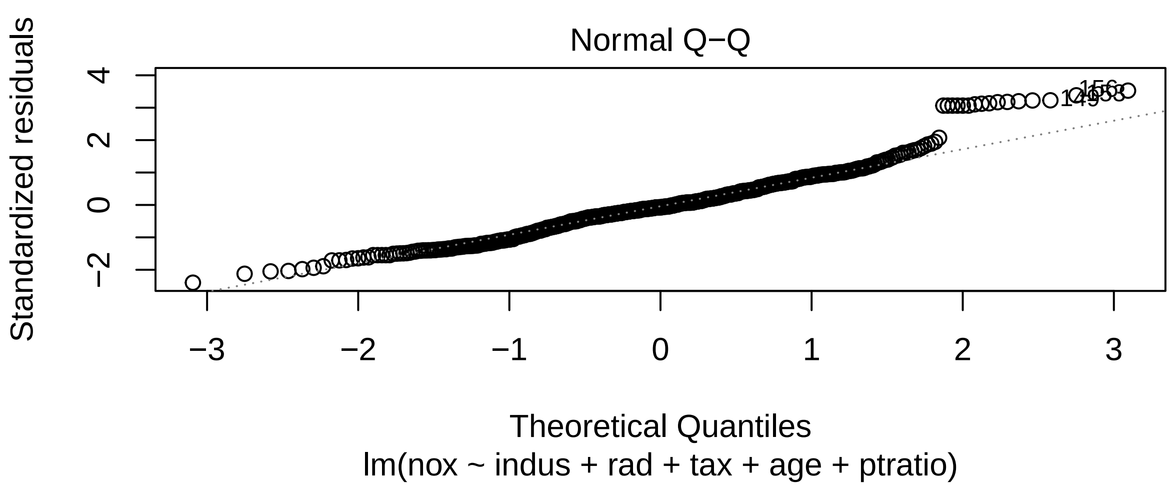
```
## Analysis of Variance Table
##
## Model 1: nox ~ indus + rad + tax + age
## Model 2: nox ~ indus + rad + tax + age + crim
##   Res.Df    RSS Df  Sum of Sq      F Pr(>F)
## 1    501 1.9888
## 2    500 1.9888  1 1.8054e-05 0.0045 0.9463
## Analysis of Variance Table
##
## Model 1: nox ~ indus + rad + tax + age
## Model 2: nox ~ indus + rad + tax + age + zn
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    501 1.9888
## 2    500 1.9810  1 0.0077748 1.9623 0.1619
## Analysis of Variance Table
##
## Model 1: nox ~ indus + rad + tax + age
## Model 2: nox ~ indus + rad + tax + age + chas
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    501 1.9888
## 2    500 1.9795  1 0.0092936 2.3475 0.1261
## Analysis of Variance Table
##
## Model 1: nox ~ indus + rad + tax + age
## Model 2: nox ~ indus + rad + tax + age + rm
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    501 1.9888
## 2    500 1.9878  1 0.0010281 0.2586 0.6113
## Analysis of Variance Table
##
```

```
## Model 1: nox ~ indus + rad + tax + age
## Model 2: nox ~ indus + rad + tax + age + dis
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    501 1.9888
## 2    500 1.7812  1   0.20754 58.257 1.174e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Analysis of Variance Table
##
## Model 1: nox ~ indus + rad + tax + age
## Model 2: nox ~ indus + rad + tax + age + ptratio
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    501 1.9888
## 2    500 1.7706  1    0.2182 61.617 2.558e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Analysis of Variance Table
##
## Model 1: nox ~ indus + rad + tax + age
## Model 2: nox ~ indus + rad + tax + age + lstat
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    501 1.9888
## 2    500 1.9831  1 0.0057063 1.4388 0.2309
## Analysis of Variance Table
##
## Model 1: nox ~ indus + rad + tax + age
## Model 2: nox ~ indus + rad + tax + age + medv
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    501 1.9888
## 2    500 1.9875  1  0.001329 0.3343 0.5634
```

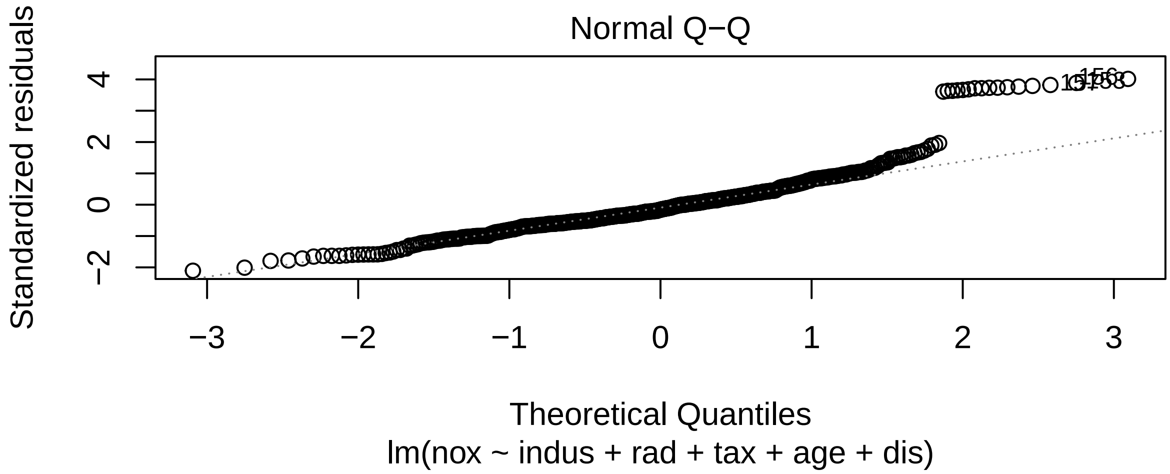
Two strong candidate variables are PTRATIO and DIS, with very low p-values.

Comparing their model fits with Q-Q plots:

```
plot(lm(nox ~ indus + rad + tax + age + ptratio, data=dfm),2) #Q-Q plot
```




```
plot(lm(nox ~ indus + rad + tax + age + dis, data=dfm),2) #Q-Q plot
```



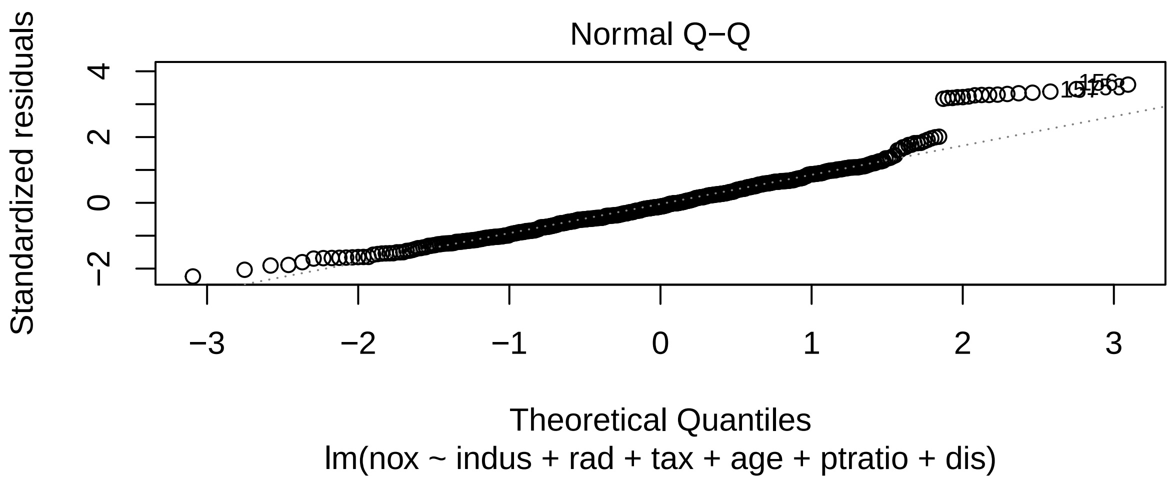
Both extended models still show problematic behaviour at the right upper tail, but this is slightly less pronounced for PTRATIO than DIS. Consequently, PTRATIO would seem the best variable to add into the model.

A subsequent analysis can be performed to check whether DIS should also be added:

```
lmp = lm(nox ~ indus + rad + tax + age + ptratio, data=dfm)
lmpd = lm(nox ~ indus + rad + tax + age + ptratio + dis, data=dfm)
anova(lmp,lmpd)
```

```
## Analysis of Variance Table
##
## Model 1: nox ~ indus + rad + tax + age + ptratio
## Model 2: nox ~ indus + rad + tax + age + ptratio + dis
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    500 1.7706
## 2    499 1.6062  1   0.16435 51.057 3.199e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
plot(lmpd,2)
```



These results suggest that adding both variables would be beneficial.

5 Marks