Notes on Week 5 Computer Practical

Robert G Aykroyd

2024-02-16

Please note that the following is not meant as a report but instead gives some ideas of the investigation that you can follow. A report would then summaries this, selecting appropriate numerical and graphical output to support and illustrate your conclusions.

Part 1: Linear regression of NOX on HC level

Let's start by reading in the data and looking at some basic information

```
folder = "https://rgaykroyd.github.io/MATH3823/Datasets//"

emissions = read.csv(file=paste(folder,"EngineEmissions-00.csv", sep=""))

head(emissions)
```

```
##     HC   CO  NOX
## 1 0.13 8.30 1.18
## 2 0.38 7.23 1.17
## 3 0.57 8.98 1.21
## 4 0.46 8.83 0.56
## 5 0.21 0.04 2.01
## 6 0.43 4.53 1.45
```

```
summary(emissions)
```

```
##        HC               CO              NOX
##  Min.   :0.0200   Min.   : 0.040   Min.   :0.0500
##  1st Qu.:0.2775   1st Qu.: 4.503   1st Qu.:0.9175
##  Median :0.3950   Median : 7.135   Median :1.1700
##  Mean   :0.3877   Mean   : 7.000   Mean   :1.1924
##  3rd Qu.:0.5025   3rd Qu.: 9.375   3rd Qu.:1.5400
##  Max.   :0.7200   Max.   :17.280   Max.   :2.9800
```

```
cor(emissions)
```

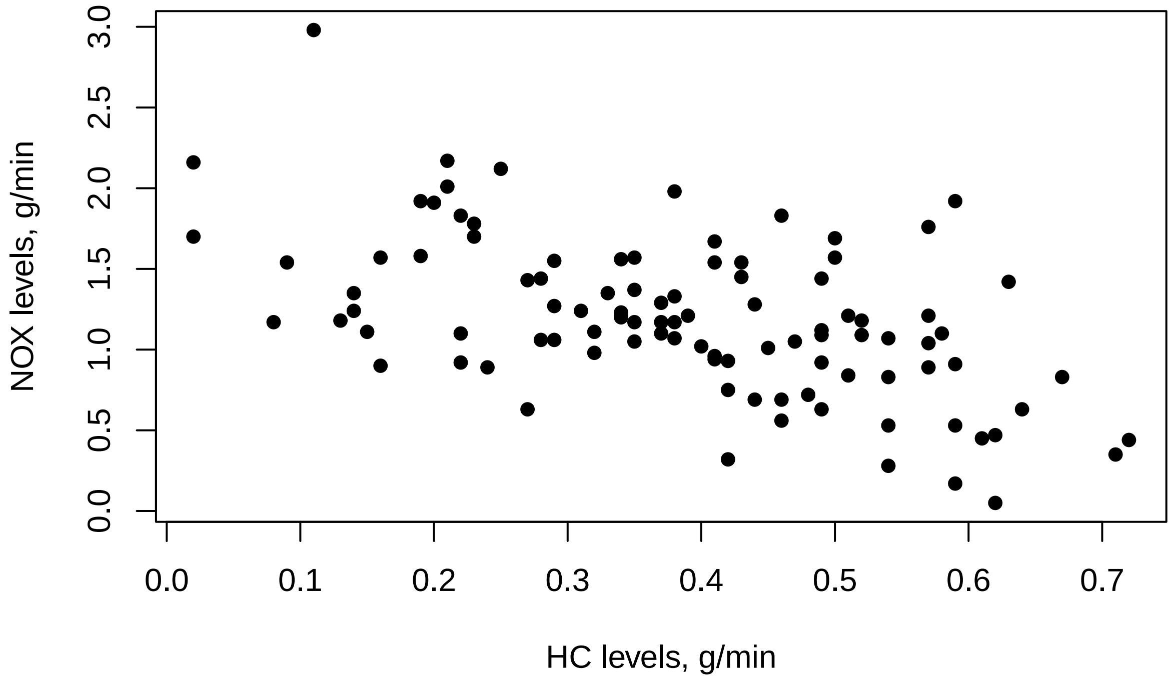
```
##             HC         CO        NOX
## HC   1.0000000  0.6566552 -0.5468423
## CO   0.6566552  1.0000000 -0.4616219
## NOX -0.5468423 -0.4616219  1.0000000
```

Note that the largest (in magnitude) correlation is between CO and HC, but there are negative correlations between NOX and CO and between NOX and HC. This is a contradiction of the statement "There is a general belief that a car engine that is working well will have low readings on all tests and that badly maintained cars will emit high levels of all pollutants."

For later ease, you might like to re-define the variables with simpler variables names.

```
HC = emissions$HC
CO = emissions$CO
NOX = emissions$NOX
```

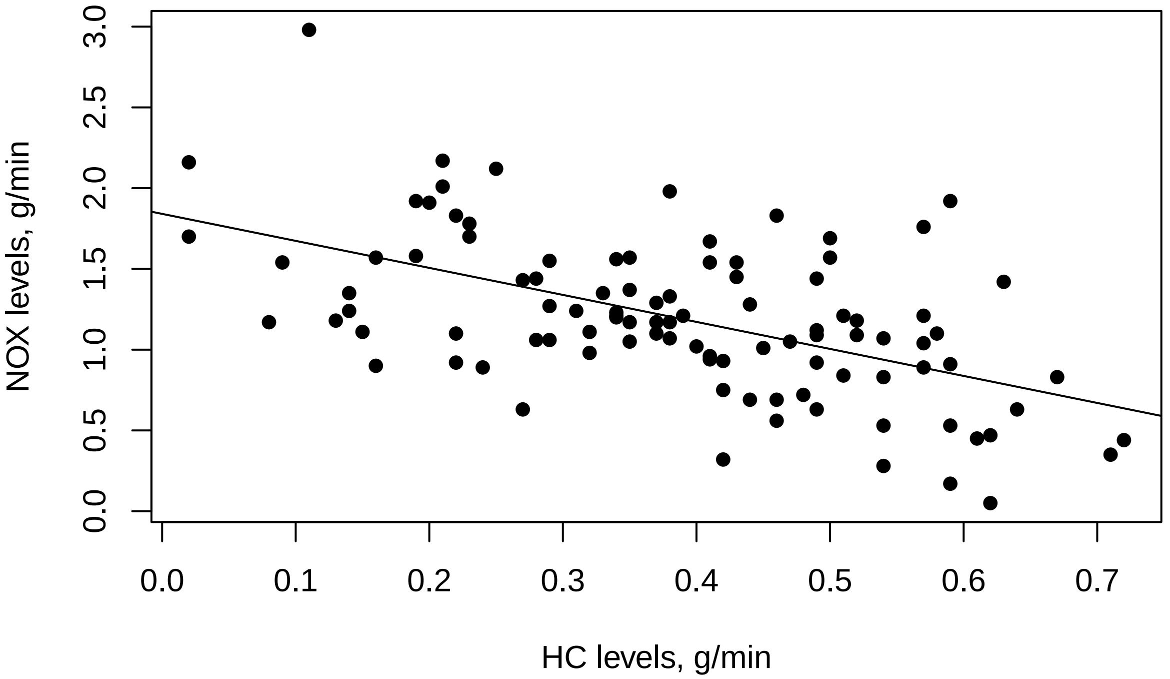
```
plot(HC, NOX, pch=16,
     xlab="HC levels, g/min", ylab="NOX levels, g/min")
```



There is a decreasing relationship between levels of HC and NOX. That is lower levels of HC are associated with higher values of NOX and higher values of HC with lower levels of NOX. Again, this provides evidence against the belief that good cars emit lower levels of all pollutants and that bad cars emit high levels of all. The data, however, suggest that a linear model is suitable.

```
myfitted = lm(NOX ~ HC)

plot(NOX ~ HC, pch=16,
     xlab="HC levels, g/min", ylab="NOX levels, g/min")
abline(myfitted)
```



Although the regression line fits the data reasonably well there is considerable spread around the line.

Parameter values, and other summary output, is given by the `summary` command. This include an intercept parameter of 1.84 and a slope of -1.67, and hence a fitted regression line of $NOX = 1.84 - 1.67HC$.

```
summary(myfitted)
```

```
##
## Call:
## lm(formula = NOX ~ HC)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.81838 -0.30568 -0.04709  0.23286  1.32319
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.8408     0.1084  16.980  < 2e-16 ***
## HC           -1.6723     0.2586  -6.466 3.95e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.412 on 98 degrees of freedom
## Multiple R-squared:  0.299,  Adjusted R-squared:  0.2919
## F-statistic: 41.81 on 1 and 98 DF,  p-value: 3.952e-09
```

The summary output also indicates that the HC is highly significant and hence is very important when modeling NOX. Sometimes the ANOVA table is easier to interpret although it must lead to the same conclusions – note that the test p-values are identical, as they must be in this situation.

```
anova(myfitted)
```

```
## Analysis of Variance Table
##
## Response: NOX
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## HC         1  7.0963  7.0963  41.808 3.952e-09 ***
## Residuals 98 16.6343  0.1697
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A residual analysis should include a scatterplot of residuals against fitted values, and a histogram – here a normal distribution curve has been added to the histogram.

```
fitted.values  = fitted.values(myfitted)
model.residuals = residuals(myfitted)

summary(model.residuals)
```

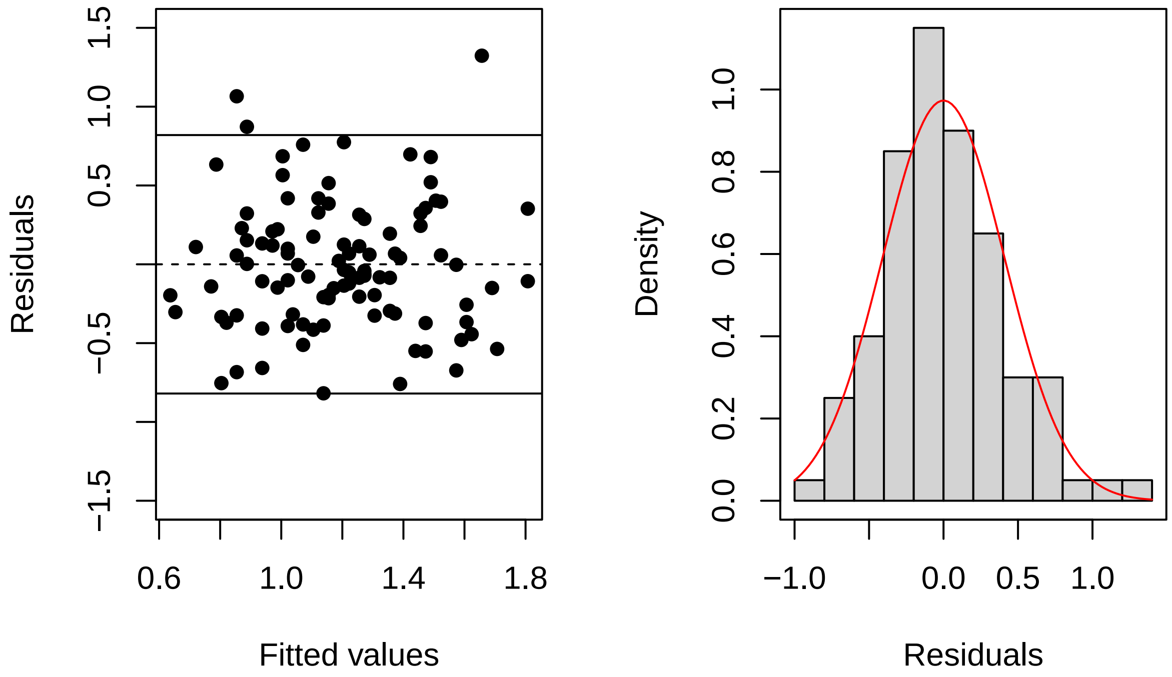
```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -0.81838 -0.30568 -0.04709  0.00000  0.23286  1.32319
```

```
par(mfrow=c(1,2))
plot(fitted.values, model.residuals,
     xlab="Fitted values", ylab="Residuals",
     ylim=1.5*c(-1,1), pch=16)
abline(h=0, lty=2)

r.sd = sd(model.residuals)
abline(h=2*r.sd*c(-1,1))

hist(model.residuals, xlab="Residuals", main="", probability=T)
box()

curve(dnorm(x,0,r.sd), col="red", add=T)
```



These are used to check for any outliers and model assumptions. Here the scatter plot has three points outside a 2-sd interval and there is a very positive skewed shape to the histogram. Together these perhaps suggest that the assumption of normally distributed errors in out linear model may not be correct – it should be investigated further.

The predict command can easily produce predictions.

```
predict(myfitted, newdata=data.frame(HC=c(0.4,0.8)))
```

```
##         1         2
## 1.1718302 0.5028941
```

Given that an HC level of 0.4 is in the main part of the data, but that 0.8 is beyond the right-hand limit of the data values, the former is likely to be a more reliable prediction.

Part 2: Linear regression of NOX on both HC and CO

Now, including the CO level in addition to the HC level – and including the interaction.

```
fit2 = lm(NOX ~ HC*CO)

anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: NOX
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## HC         1  7.0963  7.0963 45.5047 1.146e-09 ***
## CO         1  0.4386  0.4386  2.8126  0.096778 .
## HC:CO      1  1.2248  1.2248  7.8536  0.006136 **
## Residuals 96 14.9709  0.1559
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results indicate that, although the interaction is significant, CO on it's own is not. Hence, consider fitting the reduced model.

```
summary(fit2)
```

```
##
## Call:
## lm(formula = NOX ~ HC * CO)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.93172 -0.28095 -0.05063  0.19453  1.38825
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.27295    0.23666   5.379 5.28e-07 ***
## HC           0.32737    0.67049   0.488  0.62649
## CO           0.05990    0.03308   1.811  0.07333 .
## HC:CO       -0.20223    0.07216  -2.802  0.00614 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3949 on 96 degrees of freedom
## Multiple R-squared:  0.3691, Adjusted R-squared:  0.3494
## F-statistic: 18.72 on 3 and 96 DF,  p-value: 1.215e-09
```

```
fit2 = lm(NOX ~ HC + HC:CO)

anova(fit2)
```

```
## Analysis of Variance Table
##
## Response: NOX
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## HC         1  7.0963  7.0963 44.4604 1.593e-09 ***
## HC:CO      1  1.1521  1.1521  7.2184  0.008491 **
## Residuals 97 15.4822  0.1596
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table shows that both these terms are significant in explaining the NOX level and we can check the estimated model parameters.

```
summary(fit2)
```

```
##
## Call:
## lm(formula = NOX ~ HC + HC:CO)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max 
## -0.89429 -0.26599 -0.04767  0.18194  1.46829 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  1.63193    0.13074  12.482  < 2e-16 ***
## HC          -0.46035    0.51614  -0.892  0.37465    
## HC:CO       -0.08423    0.03135  -2.687  0.00849 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3995 on 97 degrees of freedom
## Multiple R-squared:  0.3476, Adjusted R-squared:  0.3341 
## F-statistic: 25.84 on 2 and 97 DF,  p-value: 1.01e-09
```

This means that the best fit model is $NOX = 1.63 - 0.46HC - 0.08(HC : CO)$.

Performing a residual analysis with the reduced model does not indicate better agreement with model assumptions and hence it may be worthwhile to consider models with other error distributions.

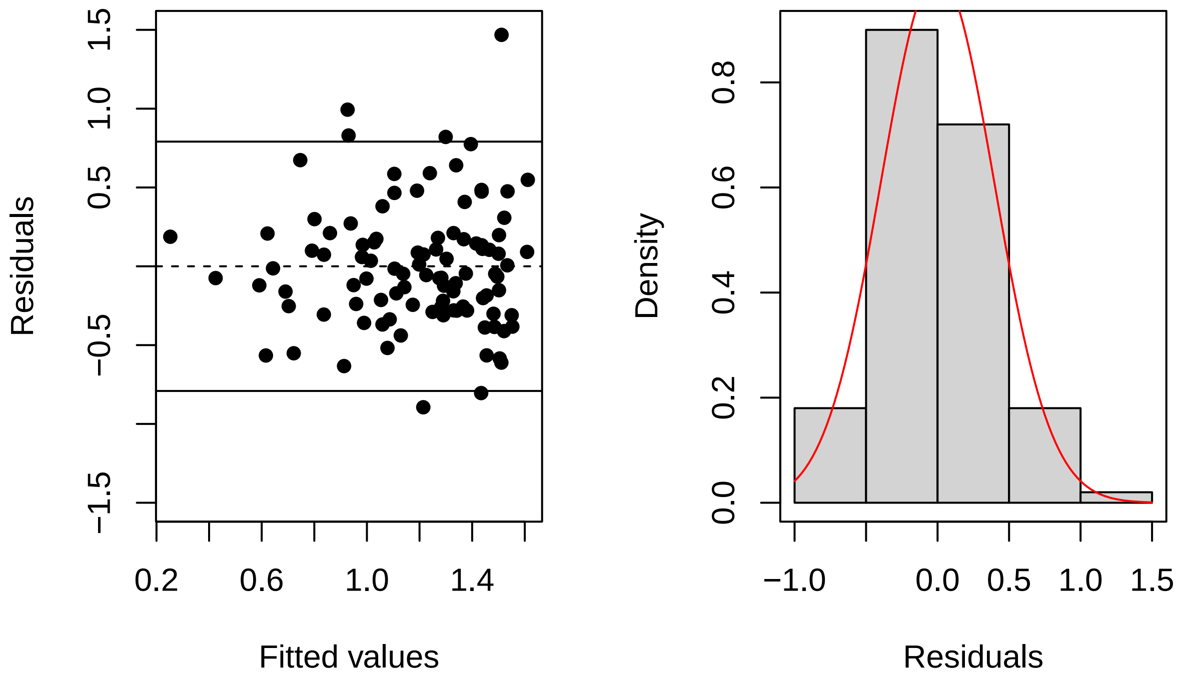
```
fitted.values  = fitted.values(fit2)
model.residuals = residuals(fit2)

par(mfrow=c(1,2))
plot(fitted.values, model.residuals,
     xlab="Fitted values", ylab="Residuals",
     ylim=1.5*c(-1,1), pch=16)
abline(h=0, lty=2)

r.sd = sd(model.residuals)
abline(h=2*r.sd*c(-1,1))

hist(model.residuals, xlab="Residuals", main="", probability=T)
box()

curve(dnorm(x,0,r.sd), col="red", add=T)
```



For completeness, here are the results for the linear model with HC and CO but not their interaction.

```
fit4 = lm(NOX ~ HC + CO)

anova(fit4)
```

```
## Analysis of Variance Table
##
## Response: NOX
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## HC         1  7.0963  7.0963  42.502 3.179e-09 ***
## CO         1  0.4386  0.4386   2.627    0.1083
## Residuals 97 16.1957  0.1670
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The ANOVA table shows that CO is not significant.

```
summary(fit4)
```

```
##
## Call:
## lm(formula = NOX ~ HC + CO)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.84579 -0.24335 -0.04819  0.18996  1.43588
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.86720    0.10875  17.170  < 2e-16 ***
## HC          -1.31034    0.34013  -3.853  0.00021 ***
## CO          -0.02383    0.01470  -1.621  0.10830
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4086 on 97 degrees of freedom
## Multiple R-squared:  0.3175, Adjusted R-squared:  0.3034
## F-statistic: 22.56 on 2 and 97 DF,  p-value: 8.982e-09
```

End of Notes on the Computer Practical