MAS programmes - Statistical Analysis (Autumn Term)

Exercises 4 - SOLUTIONS

1. You may need to reload the data into R as the data frame `oil` [See solutions to Exercises 3].

(a) Using the suggested command shows the values of observation 32 on the explanatory variables. [Note that we are only concerned with the variables `endpoint` and `distil`, which are in the *chosen* model].

```
> oil[32, ]
   spirit gravity pressure distil endpoint
32   45.7    50.8      8.6    190      407
```

We can also consider the following summary statistics for each of these variables, across all observations.

```
> summary(oil$distil)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  190.0   217.0   231.0   241.5   268.8   316.0
> summary(oil$endpoint)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  205.0   274.5   349.0   332.1   383.0   444.0
```

Observation 32 has a high value for `endpoint` and low for `distil` (actually the minimum value). This goes against the correlation between these variables which is positive (0.41, see Exercises 3, question 1(a)), hence the moderately high leverage for this point.

(c) The output relating to the model with observation 32 removed is shown below.

```
> oil232.lm <- lm(spirit ~ distil + endpoint, data = oil, subset = c(-32))
> summary(oil232.lm)

Call:
lm(formula = spirit ~ distil + endpoint, data = oil, subset = c(-32))

Residuals:
    Min      1Q  Median      3Q     Max
-3.7829 -1.6182 -0.1262  1.3979  4.7575

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.656784   2.953484   5.978 1.94e-06 ***
distil      -0.200940   0.013284 -15.126 5.29e-15 ***
endpoint     0.151735   0.007059  21.496  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2.35 on 28 degrees of freedom
Multiple R-squared:  0.946,Adjusted R-squared:  0.9422
F-statistic: 245.4 on 2 and 28 DF,  p-value: < 2.2e-16
```

```
> x <- data.frame(distil = 200, endpoint = 400)
> predict.result <- predict(oil232.lm, x, se.fit = TRUE, pi.fit = T, interval = "prediction")

> predict.result
$fit
       fit      lwr      upr
1 38.16258 32.91981 43.40534

$se.fit
[1] 1.014804

$df
[1] 28

$residual.scale
[1] 2.349654
```

Comparing this with the results reported previously (Exercises 3, question 1(b)), we find:

<u>including observation 32</u>

Model : $\widehat{\texttt{spirit}}$ = 18.4676 + 0.1558 `endpoint` - 0.2093 `distil`
s=2.426, $R^2 = 0.9521$
Predicted value for (`endpoint`, `distil`)=(400,200)
$\hat{y} = 38.93$, 95% PI (33.61, 44.25)

<u>excluding observation 32</u> (from output above)

Model : $\widehat{\texttt{spirit}}$ = 17.6568 + 0.1517 `endpoint` - 0.2009 `distil`
s=2.35, $R^2 = 0.946$
Predicted value for (`endpoint`, `distil`)=(400,200)
$\hat{y} = 38.16$, 95% PI (32.91, 43.41)

[<u>Note</u>: we are not suggesting in parts (b) and (c) that this observation *should* be removed from the model. The exercise was simply intended to illustrate the concept of influence, which is the greatest for observation 32].

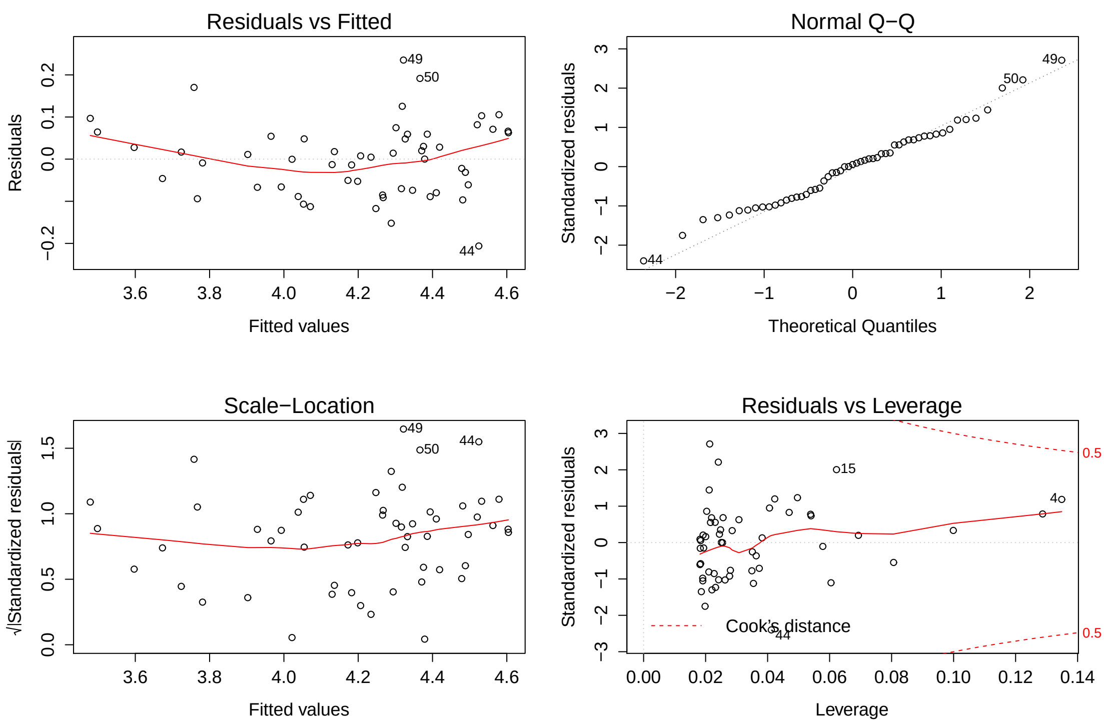
2. You may need to reload the data into R as the data frame `sugar` [See solutions to Exercises 3].

Begin by refitting the model object `sugar.lm`.

```
> sugar.lm <- lm(lconsump~price, data= sugar)
```

Can check the default residual plots

```
> par(mfrow = c(2, 2))
> plot(sugar.lm)
```



In general there is no real cause for concern from these plots (but note that the observation with the largest Cook's distances are 4, 15 and 44).
In particular:

- The normal probability plot indicate a slight skewness, but nothing to be overly concerned about.

- The plot of standardized residuals against fitted values gives the impression that some slight curvature remains, and there are some large positive and negative residuals, but neither of these facts are so strong as to be of great concern. Also the plot does not show any indication of increasing variance with mean, which means that the constant variance assumption holds here.

  Similar conclusions hold for the plot showing the square root of the absolute value of the standardized residuals against fitted value.

- The plot of standardized residuals versus leverages is useful in explaining those points which appear influential. Points appear to be influential based on leverage or standardized residual alone, but none appear to have excessive values of both.

  Recall that Cooks distances over 0.5 are considered borderline problematic, while over 1 is usually considered highly influential, so points to the right of these contours warrant investigation. Although one point has a rather high leverage, its actual influence on the fit is not unduly large.

Overall there is no cause for concern with this model, or any reason to doubt the modelling assumptions (although there are one or two points that are not well fitted by the model which is not ideal).

3. We begin by loading the data into R. [This assumes the file `pollution.txt` is in your *working directory*].

```
> pollution <- read.csv("pollution1.csv", header=T)
```

(a) We fit the *full* model

```
> pollution.lm6 <- lm(SO2 ~ temp + manufact + popul + wind + precip + pdays, data = pollution)
> summary(pollution.lm6)


Call:
lm(formula = SO2 ~ temp + manufact + popul + wind + precip +
    pdays, data = pollution)

Residuals:
    Min      1Q  Median      3Q     Max
-23.004  -8.542  -0.991   5.758  48.758

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 111.72848   47.31810   2.361 0.024087 *
temp         -1.26794    0.62118  -2.041 0.049056 *
manufact      0.06492    0.01575   4.122 0.000228 ***
popul        -0.03928    0.01513  -2.595 0.013846 *
wind         -3.18137    1.81502  -1.753 0.088650 .
precip        0.51236    0.36276   1.412 0.166918
pdays        -0.05205    0.16201  -0.321 0.749972
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 14.64 on 34 degrees of freedom
Multiple R-squared:  0.6695,Adjusted R-squared:  0.6112
F-statistic: 11.48 on 6 and 34 DF,  p-value: 5.419e-07
```

Any of the regressor variables `wind`, `precip`, or `pdays` may be removed (individually) from the regression without significant loss of fit. Especially the variable `pdays` appears to contribute very little to the fit of the model in the presence of the other regressor variables.

(b)

```
library(MASS)
> pollution.lm0 <- lm(SO2 ~ 1, data = pollution)
> stepAIC(pollution.lm0,  ~ temp + manufact + popul + wind + precip + pdays, data = pollution)
Start:  AIC=259.76
SO2 ~ 1

           Df Sum of Sq   RSS    AIC
+ manufact  1    9161.7 12876 239.73
+ popul     1    5373.2 16665 250.31
+ temp      1    4143.3 17895 253.23
+ pdays     1    3009.9 19028 255.74
<none>                  22038 259.76
+ wind      1     197.6 21840 261.40
+ precip    1      65.0 21973 261.64

Step:  AIC=239.73
```

```
SO2 ~ manufact

           Df Sum of Sq     RSS    AIC
+ popul     1    3759.5  9116.6 227.58
+ temp      1    2212.3 10663.8 234.00
+ pdays     1    1816.1 11060.0 235.50
<none>                  12876.2 239.73
+ precip    1     124.7 12751.4 241.33
+ wind      1      80.6 12795.6 241.47
- manufact  1    9161.7 22037.9 259.76

Step:  AIC=227.58
SO2 ~ manufact + popul

           Df Sum of Sq     RSS    AIC
+ pdays     1     685.0  8431.7 226.37
+ temp      1     578.0  8538.7 226.89
<none>                   9116.6 227.58
+ precip    1     148.3  8968.4 228.90
+ wind      1     146.9  8969.7 228.91
- popul     1    3759.5 12876.2 239.73
- manufact  1    7548.0 16664.7 250.31

Step:  AIC=226.37
SO2 ~ manufact + popul + pdays

           Df Sum of Sq     RSS    AIC
<none>                   8431.7 226.37
+ temp      1     257.7  8174.0 227.10
+ wind      1     244.1  8187.6 227.17
- pdays     1     685.0  9116.6 227.58
+ precip    1       4.3  8427.4 228.35
- popul     1    2628.4 11060.0 235.50
- manufact  1    5547.3 13979.0 245.10

Call:
lm(formula = SO2 ~ manufact + popul + pdays, data = pollution)

Coefficients:
(Intercept)     manufact        popul        pdays
    6.96585      0.07433     -0.04939      0.16436


> stepAIC(pollution.lm6,  ~ temp + manufact + popul + wind + precip + pdays, data = pollution)
Start:  AIC=226.37
SO2 ~ temp + manufact + popul + wind + precip + pdays

           Df Sum of Sq     RSS    AIC
- pdays     1      22.1  7305.4 224.50
<none>                   7283.3 226.37
- precip    1     427.3  7710.6 226.71
- wind      1     658.1  7941.4 227.92
- temp      1     892.5  8175.8 229.11
- popul     1    1443.1  8726.3 231.78
- manufact  1    3640.1 10923.4 240.99

Step:  AIC=224.49
SO2 ~ temp + manufact + popul + wind + precip
```

```
           Df Sum of Sq     RSS    AIC
<none>                   7305.4 224.50
- wind      1     636.1  7941.5 225.92
+ pdays     1      22.1  7283.3 226.37
- precip    1     785.4  8090.8 226.68
- popul     1    1447.5  8752.9 229.91
- temp      1    1517.4  8822.8 230.23
- manufact  1    3636.8 10942.1 239.06

Call:
lm(formula = SO2 ~ temp + manufact + popul + wind + precip, data = pollution)

Coefficients:
(Intercept)         temp     manufact        popul         wind       precip
  100.15245     -1.12129      0.06489     -0.03933     -3.08240      0.41947
```

Model A, suggested by `stepAIC` starting with no regressor variables is

$$\widehat{\texttt{SO2}} = 6.96585 + 0.07433\,\texttt{manufact} - 0.04939\,\texttt{popul} + 0.16436\,\texttt{pdays}$$

Model B, suggested by `stepAIC` starting with all six regressor variables, is

$$\begin{aligned}\widehat{\texttt{SO2}} =& 100.15245 - 1.12129\,\texttt{temp} + 0.06489\,\texttt{manufact}\\ & -0.03933\,\texttt{popul} - 3.08240\,\texttt{wind} + 0.41947\,\texttt{precip}\end{aligned}$$

(c) All the required data, apart from the values of the R^2 may be obtained from the stepwise outputs. The values of the R^2 may be obtained by looking at the results of the suggested regressions (using `lm()` and `summary()`).

Model	s	R^2	AIC
A	15.10	0.6174	226.37
B	14.45	0.6685	224.50
Full	14.64	0.6695	226.37

(d) Model A is the most compact one and Model B is the one with the smallest AIC.