Exercises

The following question offers you the opportunity to test your understanding of the materials presented to you in Lectures 1-4.

1. In a study of infant feeding, 50 infants aged approximately 2 months were weighed immediately before and after each breast feeding over a period of 24 hours in order to determine their intake of breast milk. This amount, together with five potential explanatory variables, were entered into an R data frame `milkdata`. The variables in the data frame were as follows:

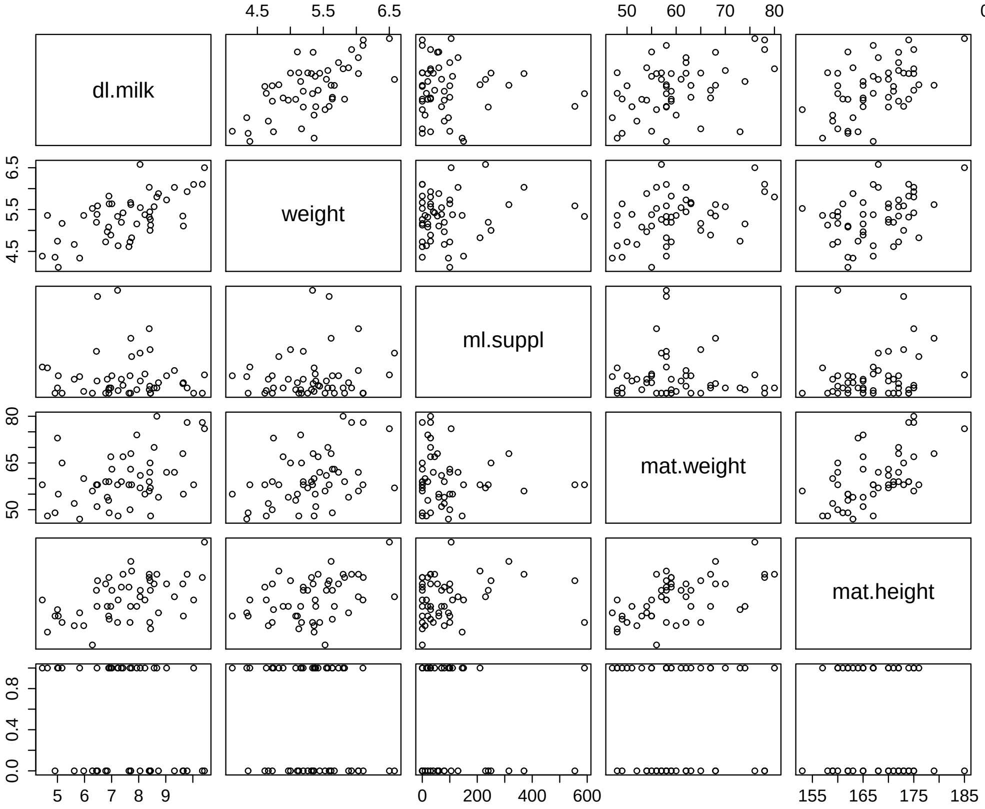
> **dl.milk**: breast milk intake (dl/24 hr)
> **sex**: a factor giving the sex of the infant (boy or girl);
> **weight**: weight of infant (kg);
> **ml.suppl**: amount of milk substitute given to infant (in a period before the breast milk intake measurement) (ml/24 hr);
> **mat.weight**: weight of mother (kg);
> **mat.height**: height of mother (cm).

An extract from the data frame showing the observations relating to the first five boy and first five girl infants (of twenty-five) is shown below.

```
milkdata[c(1:5, 26:30), ]
   no dl.milk  sex weight ml.suppl mat.weight mat.height
1   1    8.42  boy  5.002      250         65        173
2   4    8.44  boy  5.128        0         48        158
3   5    8.41  boy  5.445       40         62        160
4  10    9.65  boy  5.106       60         55        162
5  12    6.44  boy  5.196      240         58        170
26  6   10.03 girl  6.100        0         58        167
27 14    7.42 girl  5.421       45         67        175
28 25    5.00 girl  4.744       30         73        164
29 26    8.67 girl  5.800       30         80        175
30 27    6.90 girl  5.822        0         59        174
```

(a) Overleaf is a scatterplot matrix of the data described above (where the bottom row of plots shows the relationship between the response variable `dl.milk` and each of the potential explanatory variables). Additionally, a variate `sexvar` was calculated, taking the value 0 for a boy and 1 for a girl, so that a correlation matrix between the five explanatory variables was produced, as shown below. Edited output from simple linear regressions on each of the explanatory variables in turn is presented on page 3.

```
milkdata$sexvar <- ifelse(milkdata$sex=="boy",0,1)
pairs(milkdata[,c(-1,-3)])
cor(milkdata[,c(-1,-3)])
               dl.milk     weight    ml.suppl  mat.weight mat.height     sexvar
dl.milk     1.00000000  0.6360448 -0.06351955  0.43427002  0.5050420 -0.29940126
weight      0.63604482  1.0000000  0.12838120  0.40817476  0.3867571 -0.22001058
ml.suppl   -0.06351955  0.1283812  1.00000000 -0.07887363  0.1823026 -0.07136717
mat.weight  0.43427002  0.4081748 -0.07887363  1.00000000  0.5647330 -0.05303191
mat.height  0.50504203  0.3867571  0.18230263  0.56473304  1.0000000 -0.11776734
sexvar     -0.29940126 -0.2200106 -0.07136717 -0.05303191 -0.1177673  1.00000000
```


dl.milk
weight
ml.suppl
mat.weight
mat.height
sexvar
4.5 5.5 6.5
50 60 70 80
0.0 0.4 0.8
9 7 5
6.5 5.5 4.5
500 200 0
80 65 50
185 170 155
0.8 0.4 0.0
5 6 7 8 9
0 200 400 600
155 165 175 185

Individual Simple linear regressions on the explanatory variables

sex

```
summary(lm(dl.milk ~ sexvar, data = milkdata))

Call:
lm(formula = dl.milk ~ sexvar, data = milkdata)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0424 -1.1574  0.1736  0.8496  2.9736 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   7.9524     0.2914  27.288   <2e-16 ***
sexvar       -0.8960     0.4121  -2.174   0.0347 *  
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 1.457 on 48 degrees of freedom
Multiple R-squared:  0.08964,Adjusted R-squared:  0.07068 
F-statistic: 4.726 on 1 and 48 DF,  p-value: 0.03466 
```

weight

```
summary(lm(dl.milk ~ weight, data = milkdata))

Call:
lm(formula = dl.milk ~ weight, data = milkdata)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.9467 -0.8973  0.1148  0.8757  2.5186 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   -1.821      1.641  -1.110    0.273    
weight         1.753      0.307   5.711 6.92e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 1.178 on 48 degrees of freedom
Multiple R-squared:  0.4046,Adjusted R-squared:  0.3921 
F-statistic: 32.61 on 1 and 48 DF,  p-value: 6.915e-07 
```

ml.suppl

```
summary(lm(dl.milk ~ ml.suppl, data = milkdata))

Call:
lm(formula = dl.milk ~ ml.suppl, data = milkdata)
```

```
Residuals:
    Min      1Q  Median      3Q     Max 
-3.0246 -0.7679  0.1023  0.9789  2.9322 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  7.5751805  0.2687400  28.188   <2e-16 ***
ml.suppl    -0.0007373  0.0016720  -0.441    0.661    
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.524 on 48 degrees of freedom
Multiple R-squared:  0.004035,Adjusted R-squared:  -0.01671 
F-statistic: 0.1945 on 1 and 48 DF,  p-value: 0.6612
```

mat.weight

```
summary(lm(dl.milk ~ mat.weight, data = milkdata))

Call:
lm(formula = dl.milk ~ mat.weight, data = milkdata)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.5257 -0.7179 -0.0559  0.9595  2.6791 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  2.80837    1.41935   1.979  0.05361 . 
mat.weight   0.07832    0.02345   3.340  0.00163 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.376 on 48 degrees of freedom
Multiple R-squared:  0.1886,Adjusted R-squared:  0.1717 
F-statistic: 11.16 on 1 and 48 DF,  p-value: 0.001627
```

mat.height

```
summary(lm(dl.milk ~ mat.height, data = milkdata))

Call:
lm(formula = dl.milk ~ mat.height, data = milkdata)

Residuals:
     Min       1Q   Median       3Q      Max 
-3.01288 -0.95128  0.07002  0.76068  2.78263 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -12.10311    4.84011  -2.501 0.015867 *  
mat.height    0.11710    0.02889   4.054 0.000184 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
Residual standard error: 1.318 on 48 degrees of freedom
Multiple R-squared:  0.2551,Adjusted R-squared:  0.2395
F-statistic: 16.44 on 1 and 48 DF,  p-value: 0.0001837
```

On the basis of the information presented to you above, which explanatory variables would you expect to be included in a good multiple regression model which has `dl.milk` as the response variable? Briefly explain your choices.

(b) A multiple regression model (Model A) with `dl.milk` as the response variable and all five explanatory variables was fitted in R, resulting in the following output:

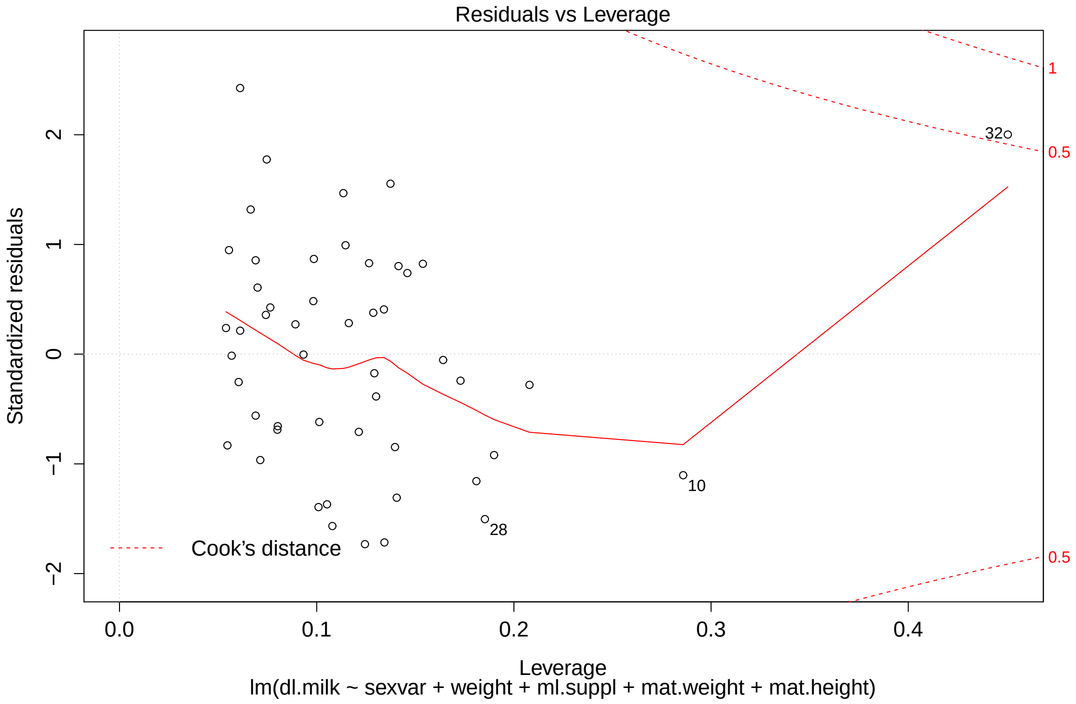
Model A

```
modA <- lm(dl.milk ~ sexvar + weight + ml.suppl + mat.weight + mat.height,
           data = milkdata)
summary(modA)

Call:
lm(formula = dl.milk ~ sexvar + weight + ml.suppl + mat.weight +
    mat.height, data = milkdata)

Residuals:
     Min       1Q   Median       3Q      Max
-1.74201 -0.81173 -0.00926  0.78326  2.52646

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.681839   4.361561  -2.678 0.010363 *
sexvar       -0.499532   0.312672  -1.598 0.117284
weight        1.349124   0.322450   4.184 0.000135 ***
ml.suppl     -0.002233   0.001241  -1.799 0.078829 .
mat.weight    0.006212   0.023708   0.262 0.794535
mat.height    0.072278   0.030169   2.396 0.020906 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.075 on 44 degrees of freedom
Multiple R-squared:  0.5459,Adjusted R-squared:  0.4943
F-statistic: 10.58 on 5 and 44 DF,  p-value: 1.03e-06

dummy.coef(modA)$sexvar
[1] -0.4995322
```

(i) Explain why there are not separate coefficients given for `boy` and `girl` for `sexvar`. What does the given coefficient for `sexvar` represent?

(ii) From this model, what would be the expected breast milk intake in 24 hours for an infant boy weighing 5.5kg, who had no milk substitute in the period before measurement, and whose mother weighed 60kg and was 168cm tall?

(iii) On the residuals vs leverage plot for this model only one point unit 32, stands out. By reference to this plot, explain briefly why this unit has such a high Cook statistic.



(c) The stepwise regression method provided by the `stepAIC` function in the R package MASS was applied to the data. Starting from the null model, Model B was arrived at as follows:

```
library(MASS)
mod0 <- lm(dl.milk ~ 1, data = milkdata)
stepAIC(mod0,  ~ sexvar + weight + ml.suppl + mat.weight + mat.height, data = milkdata)


⋮
output omitted
```

Model B

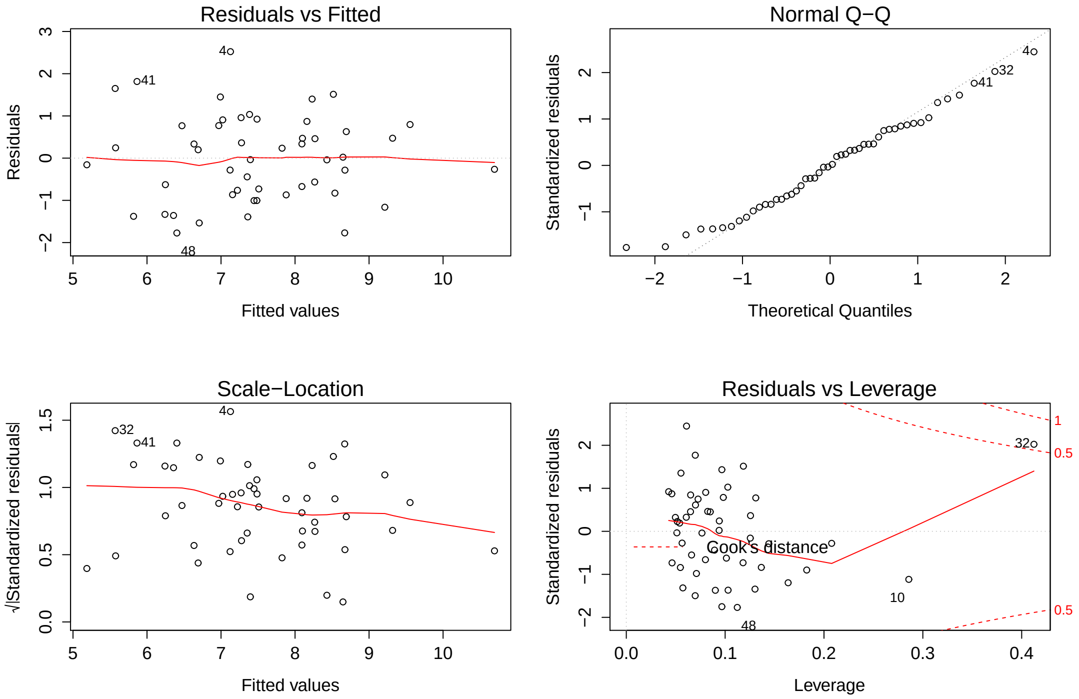
```
modB <- lm(dl.milk ~ sexvar + weight + ml.suppl + mat.height,
           data = milkdata)
summary(modB)

Call:
lm(formula = dl.milk ~ sexvar + weight + ml.suppl + mat.height,
    data = milkdata)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-1.77312 -0.81196 -0.00683  0.76988  2.52240

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -12.112571   3.997860  -3.030  0.00405 **
sexvar       -0.494675   0.308875  -1.602  0.11626
weight        1.372524   0.306612   4.476 5.14e-05 ***
ml.suppl     -0.002313   0.001190  -1.943  0.05824 .
mat.height    0.076363   0.025560   2.988  0.00454 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.064 on 45 degrees of freedom
Multiple R-squared:  0.5452,Adjusted R-squared:  0.5047
F-statistic: 13.48 on 4 and 45 DF,  p-value: 2.658e-07
```

(i) Comment on the selection of variables in Model B.

(ii) Explain briefly why it may be preferable to use Model B than Model A.

(iii) Comment on the residual plots reported below. Do the assumptions underlying the multiple regression model appear to be satisfied in this case?



(d) Further plotting of these data provides some evidence that the pattern of the relationship between breast milk intake and the amount of milk substitute given in the previous period is different for boy and girl infants. Without doing any calculations, briefly outline how you could use a regression model in R to carry out a formal significance test of whether this is in fact true.