MAS programmes - Statistical Analysis (Autumn Term)

Exercises 5 - SOLUTIONS

1. You will need to load the data into R as the data frame `water`, as seen in Lecture 5. This can be done as follows (assuming the file `water1.csv` is in your *working directory*).

```
> water <- read.csv("water1.csv", header = T)
```

Separating the data frame as suggested, we find

```
> water.s <- water[water$north == 0,  ]
> water.n <- water[water$north == 1,  ]
> water.lmS <- lm(mortality ˜ calcium, data = water.s)
> summary(water.lmS)

Call:
lm(formula = mortality ˜ calcium, data = water.s)

Residuals:
    Min      1Q  Median      3Q     Max 
-218.87  -66.08  -18.93   78.24  218.50 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1522.8150    45.4309  33.519  < 2e-16 ***
calcium       -2.0927     0.5664  -3.695  0.00113 ** 
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 114.3 on 24 degrees of freedom
Multiple R-squared:  0.3626,Adjusted R-squared:  0.336 
F-statistic: 13.65 on 1 and 24 DF,  p-value: 0.001135 

> water.lmN <- lm(mortality ˜ calcium, data = water.n)
> summary(water.lmN)

Call:
lm(formula = mortality ˜ calcium, data = water.n)

Residuals:
    Min      1Q  Median      3Q     Max 
-221.27 -105.57   15.28   90.27  310.14 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1692.3128    33.7846  50.091   <2e-16 ***
calcium       -1.9313     0.8479  -2.278   0.0293 *  
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1 

Residual standard error: 129.2 on 33 degrees of freedom
Multiple R-squared:  0.1359,Adjusted R-squared:  0.1097 
F-statistic: 5.188 on 1 and 33 DF,  p-value: 0.02934 
```

Hence, the individual regressions give the following models

South:

$$\widehat{\texttt{mortality}} = 1522.8150 - 2.0927\ \texttt{calcium} \qquad (s = 114.3)$$

North:

$$\widehat{\texttt{mortality}} = 1692.3128 - 1.9313\ \texttt{calcium} \qquad (s = 129.2)$$

In Model C of the Lecture Notes, we found for towns in the South (`north=0`)

$$\widehat{\texttt{mortality}} = 1522.8151 - 2.0927\ \texttt{calcium}$$

and for towns in the North (`north=1`)

$$\begin{aligned}\widehat{\texttt{mortality}} &= (1522.8151 + 169.4978) + (-2.0927 + 0.1614)\ \texttt{calcium}\\ &= 1692.3129 - 1.9313\ \texttt{calcium}\end{aligned}$$

so that the two approaches lead to the same estimates of the regression parameters.

The advantage of the combined approach using the *dummy* variable (or factor) `north` is that the differences between the two regressions can be formally tested, and we may obtain a more parsimonious model overall. For example in the Lecture we found that there was no significant difference between the two slope parameters so that a parallel lines model (Model B) was adopted, using only three parameters to describe the two lines. This is not possible directly running the regressions separately.

A possible disadvantage of the combined approach however is that we have a single estimate of the error variance, which does not seem unreasonable in the context of these data (the residual standard error given for Model C above was 123.2). The plot of standardized residuals versus fitted values indicated that there was no obvious difference in the spread of the residuals between the two groups (North and South).

Note alternatively that the following code could have been used to fit the individual regressions

```
> lm(mortality~calcium,data=water,subset=(north==0))
> lm(mortality~calcium,data=water,subset=(north==1))
```

2. We begin by reloading the data vectors from Exercises 1.

```
> survived <- c(25.3, 22.6, 25.1, 23.2, 24.4, 25.1, 24.6, 24, 24.2, 24.9,
                24.1, 24, 26, 24.9, 25.5, 23.4, 25.9, 24.2, 24.2, 27.4, 24)
> died <- c(26.3, 25.8, 26, 23.2, 26.5, 24.2, 26.9, 27.7, 23.9, 26.1, 24.6,
            23.6, 26, 25, 24.8, 22.8, 24.8, 24.6, 30.5, 24.8, 23.9, 24.7, 26.9,
            22.6, 26.1, 24.8, 26.2, 26.1)

> weight <- c(survived, died)

> group <- c(rep(0, length(survived)), rep(1, length(died)))
> group.f <- factor(c(rep("surv", length(survived)), rep("died", length(died))))

> sparrows <- data.frame(weight, group, group.f)
> rm(weight, group, group.f)

> sparrows[c(1:5, 22:26), ]
    weight group group.f
1    25.3     0    surv
2    22.6     0    surv
3    25.1     0    surv
4    23.2     0    surv
5    24.4     0    surv
22   26.3     1    died
23   25.8     1    died
24   26.0     1    died
25   23.2     1    died
26   26.5     1    died

> sparrows.lm <- lm(weight ~ group, data = sparrows)
> summary(sparrows.lm)

Call:
lm(formula = weight ~ group, data = sparrows)

Residuals:
    Min      1Q  Median      3Q     Max
-2.7357 -0.6357 -0.3357  0.7643  5.1643

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  24.6190     0.3109  79.189   <2e-16 ***
group         0.7167     0.4113   1.743   0.0879 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.425 on 47 degrees of freedom
Multiple R-squared:  0.06069,Adjusted R-squared:  0.0407
F-statistic: 3.037 on 1 and 47 DF,  p-value: 0.08795
```

The model is

$$\widehat{\texttt{weight}}_i = \alpha + \beta\texttt{group}_i, \quad i = 1, \dots, 49$$

so that β represents the difference in intercept (mean value) for sparrows who died, compared to the baseline mean of sparrows who survived. Hence, if β is significantly different from zero there is a difference between the two mean weights. Looking at the output, we find $p = 0.08795$, so there is no evidence against the null hypothesis that the two groups of sparrows have different mean weights (the *two-sample t*-test we saw in Exercises 1!).

Note also we have that the mean weight of those sparrows who survived is $\alpha = 24.6190$, whilst that of those who died is $\alpha + \beta = 24.6190 + 0.7167 = 25.3357$, matching the values previously found, and the model variance 1.425^2 matches (to within rounding error) the pooled estimate of the sample variance previously found.

```
> t.test(survived, died, var.equal = T)

Two Sample t-test

data:  survived and died
t = -1.7426, df = 47, p-value = 0.08795
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.5440352  0.1107019
sample estimates:
mean of x mean of y
 24.61905  25.33571
```

3. Again, we begin by reloading the data vectors from Exercise 1.

```
> shaded <- c(8.59, 8.59, 8.09, 8.54, 8.09, 8.49, 7.89, 8.59, 8.54, 7.99, 7.89, 8.09, 7.89,
              8.54, 7.84, 7.49, 7.89, 7.79, 7.84, 8.89, 8.54, 8.04, 8.59, 8.19, 8.59)
> exposed <- c(8.49, 8.59, 7.84, 7.89, 8.19, 7.84, 7.89, 7.89, 7.79, 7.84, 7.79, 7.84, 7.89,
               8.07, 7.97, 7.57, 7.92, 7.97, 8.17, 8.67, 8.07, 7.97, 8.62, 7.92, 7.97)

> solids <- c(shaded, exposed)
> exposure <- factor(c(rep(0, 25), rep(1, 25)))
> exposure.f <- factor(c(rep("shaded", 25), rep("exposed", 25)))
> fruit <- factor(rep(1:25, 2))
> grapefruit <- data.frame(fruit, solids, exposure, exposure.f)
> rm(fruit, solids, exposure, exposure.f)
> grapefruit[c(1:5, 26:30),  ]
   fruit solids exposure exposure.f
1      1   8.59        0     shaded
2      2   8.59        0     shaded
3      3   8.09        0     shaded
4      4   8.54        0     shaded
5      5   8.09        0     shaded
26     1   8.49        1    exposed
27     2   8.59        1    exposed
28     3   7.84        1    exposed
29     4   7.89        1    exposed
30     5   8.19        1    exposed
```

The following (general linear model) recovers the *paired t-test* of Exercises 1.

```
> grapefruit.lm <- lm(solids ~ fruit + exposure, data = grapefruit)
> summary(grapefruit.lm)

Call:
lm(formula = solids ~ fruit + exposure, data = grapefruit)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.2782 -0.1118  0.0000  0.1118  0.2782

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.63680    0.15636  55.237  < 2e-16 ***
fruit2       0.05000    0.21683   0.231  0.81959
fruit3      -0.57500    0.21683  -2.652  0.01396 *
fruit4      -0.32500    0.21683  -1.499  0.14695
fruit5      -0.40000    0.21683  -1.845  0.07745 .

⋮   output edited

            Estimate Std. Error t value Pr(>|t|)
fruit20      0.24000    0.21683   1.107  0.27933
fruit21     -0.23500    0.21683  -1.084  0.28923
fruit22     -0.53500    0.21683  -2.467  0.02113 *
fruit23      0.06500    0.21683   0.300  0.76693
fruit24     -0.48500    0.21683  -2.237  0.03486 *
fruit25     -0.26000    0.21683  -1.199  0.24220
exposure1   -0.19360    0.06133  -3.157  0.00426 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.2168 on 24 degrees of freedom
Multiple R-squared:  0.7979,Adjusted R-squared:  0.5873
F-statistic:  3.79 on 25 and 24 DF,  p-value: 0.0008487
```

Note that the model may be written,

$$\widehat{\texttt{solids}}_i = \alpha + \beta_1\texttt{fruit}_i + \texttt{exposure}_i, \quad i = 1, \ldots, 25$$

Hence, α represents the overall average percentage of solids found in a grapefruit, β_1 represents the *fruit* effect (i.e. different fruits may have different underlying levels of solids to begin with which must be accounted for) and β_2 represents the difference in the percentage of solids for the exposed group relative to the shaded group (control), in the presence of the fruit effect. This is a paired t-test, and the significance probability of the term `exposure`, $p = 0.0043$, matches that previously found.

```
> t.test(exposed, shaded, paired = T)

Paired t-test

data:  exposed and shaded
t = -3.1567, df = 24, p-value = 0.004264
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.32017763 -0.06702237
sample estimates:
mean of the differences
                -0.1936
```

Note that the models in questions 2 and 3 contained only qualitative explanatory variables (or *factors*). Models such as these will be the focus of Lectures 6-9.