STAT3015/4030/7030 Generalised Linear Modelling

Tutorial 3

1. Reconsider the one-way ANOVA example `coagulation` from class. Refit the one-way ANOVA but this time treat diet effect as random. Write down the structure of your random effects model using mathematical notation. Interpret the output of your model fit. In particular, report the intra-class correlation coefficient, test whether the diet effect is significant, and provide estimates of the random effects for each diet. Also provide 95% confidence interval estimates for the effect of each diet on blood coagulation time.

   Solution: Let Y_{ij} be the blood coagulation time for animal j from diet i, $i \in \{1, 2, 3, 4\}$. The model is

$$
\begin{aligned}
Y_{ij} &= \beta_0 + \alpha_i + \epsilon_{ij}, \\
\alpha_i &\overset{\text{iid}}{\sim} N(0, \sigma_\alpha^2), \\
\epsilon_{ij} &\overset{\text{iid}}{\sim} N(0, \sigma^2),
\end{aligned}
$$

   where α_i is the random effect for diet. σ_α^2 is the variance parameter of the diet averages. σ^2 is the variance parameter for the error terms ϵ_{ij}, which captures the variation of coagulation times within diets. To fit this model, we need the following packages.

```
> library(lme4) # Linear Mixed-Effects Models
> library(arm) # Data Analysis Using Regression and
                Multilevel/Hierarchical Models
> options(digits=5)

> coag <- read.table("coag.csv", sep=",", header=TRUE)
> coag

   X coag diet
1  1   62    A
2  2   60    A
3  3   63    A
4  4   59    A
5  5   63    B
6  6   67    B
```

```
7   7   71    B
8   8   64    B
9   9   65    B
10 10   66    B
11 11   68    C
12 12   66    C
13 13   71    C
14 14   67    C
15 15   68    C
16 16   68    C
17 17   56    D
18 18   62    D
19 19   60    D
20 20   61    D
21 21   63    D
22 22   64    D
23 23   63    D
24 24   59    D
```

```
> model <- lmer(coag~(1|diet), data=coag)
> summary(model)

Linear mixed model fit by REML ['lmerMod']
Formula: coag ~ (1 | diet)
   Data: coag

REML criterion at convergence: 115.8

Scaled residuals:
    Min      1Q  Median      3Q     Max
-2.1849 -0.5992  0.0933  0.5408  2.1751

Random effects:
 Groups   Name        Variance Std.Dev.
 diet     (Intercept) 11.7     3.42
 Residual             5.6     2.37
Number of obs: 24, groups:  diet, 4

Fixed effects:
            Estimate Std. Error t value
(Intercept)    64.01       1.78      36
```

The intraclass correlation coefficient is 11.6915/(11.6915+5.5994)=0.6762. That is, 68% of the variation in coagulation times is explained by the variation between diets.

The estimated coagulation time for diet A is 61.3221424979126. The estimated coagulation time for diet B is 65.8530940333015. The estimated coagulation time for diet C is 67.7052518981989. The estimated coagulation time for diet C is 61.1701693024184. They can be easily found as follows.

```
> ranef(model)

$diet
  (Intercept)
A     -2.6905
B      1.8404
C      3.6926
D     -2.8425

> coef(model)

$diet
  (Intercept)
A      61.322
B      65.853
C      67.705
D      61.170

attr(,"class")
[1] "coef.mer"
```

A 95% confidence interval estimate for the effect of diet A on blood coagulation time is (59.13, 63.51). For diet B the interval is (64.03, 67.68). For diet C the interval is (65.88, 69.53). For diet D the interval is (59.58, 62.76). They can be found using the following commands.

```
> se.ranef(model)
$diet
  (Intercept)
A     1.11811
B     0.92965
C     0.92965
D     0.81265

> coef(model)$diet[, 1] - 1.96*se.ranef(model)$diet[, 1]
     A      B      C      D
59.131 64.031 65.883 59.577
```

```
> coef(model)$diet[, 1] + 1.96*se.ranef(model)$diet[, 1]

     A      B      C      D
63.514 67.675 69.527 62.763
```

Since the design matrix is unbalanced, a likelihood ratio test to compare the two models with and without the random diet effect has $p - value = 6.5 \times 10^{-5}$, and so we conclude that the diet effect is significant.

```
> model2 <- lm(coag~1, data=coag)

> as.numeric(2*(logLik(model)-logLik(model2)))

[1] 15.95

> pchisq(15.95007, 1, lower=FALSE)

[1] 6.5035e-05
```

2. Recall the blood group data from Tutorial 2:

Blood Type	Responses			
$O-$	9	11		
$O+$	20	19	23	19
$A-$	12	10		
$A+$	17	18	21	20
$B-$	16			
$B+$	24	28	25	
$AB-$	15			
$AB+$	25			

(a) Recall that we determined that the A-antigen was not significantly related to the response, so that we could reduce the number of factor levels down to 4. Now, create two categorical variables, one for whether the individual can create the B-antigen and one for whether the individual has the Rh factor, and fit a two-way ANOVA model

$$Y_{ijk} = \mu + \tau_i + \alpha_j + \epsilon_{ijk}$$

to the data. Do both factors appear to be significant?

Solution: First, read in data.

```
> resp <- c(9, 11, 20, 19, 23, 19, 12, 10, 17, 18, 21, 20, 16, 24, 28, 25, 15,
> btyp <- c(rep("O-", 2), rep("O+", 4), rep("A-", 2),
+           rep("A+", 4), "B-", rep("B+", 3), "AB-", "AB+")
```

Second, create the desired indicator variables as follows.

```
> Bfact <- ifelse(btyp=="B-", "B", "notB")
> Bfact <- ifelse(btyp=="B+", "B", Bfact)
> Bfact <- ifelse(btyp=="AB-", "B", Bfact)
> Bfact <- ifelse(btyp=="AB+", "B", Bfact)
> Rhfact <- ifelse(btyp=="O+", "+", "-")
> Rhfact <- ifelse(btyp=="A+", "+", Rhfact)
> Rhfact <- ifelse(btyp=="B+", "+", Rhfact)
> Rhfact <- ifelse(btyp=="AB+", "+", Rhfact)
```

Third, fit two ANOVA models.

```
> btyp.aov <- aov(resp~Bfact+Rhfact)
> summary(btyp.aov)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
Bfact        1    125     125    47.9 4.9e-06 ***
Rhfact       1    355     355   136.2 6.3e-09 ***
Residuals   15     39       3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> btyp.aov <- aov(resp~Rhfact+Bfact)
> summary(btyp.aov)
```

```
            Df Sum Sq Mean Sq F value  Pr(>F)
Rhfact       1    355     355   136.2 6.3e-09 ***
Bfact        1    125     125    47.9 4.9e-06 ***
Residuals   15     39       3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Clearly, both factors are significant, even in the presence of each other. Notice that we fit the model in both orders because the design was unbalanced. Interestingly, the ANOVA tables for both orders of the predictors are the same despite the fact that this is an unbalanced design. This will only rarely be the case, however, as the sequential sums of squares generally do not remain the same when we switch the variable order unless the design is balanced.

(b) Construct appropriate indicator variables for the two categorical predictors and refit the model as a linear regression to verify that the same results are obtained.

Solution: The following commands do the job.

```
> Bind <- ifelse(Bfact=="B", 1, 0)
> Rhind <- ifelse(Rhfact=="+", 1, 0)
> btyp.lm <- lm(resp~Bind+Rhind)
> anova(btyp.lm)

Analysis of Variance Table

Response: resp
          Df Sum Sq Mean Sq F value  Pr(>F)
Bind       1    125     125    47.9 4.9e-06 ***
Rhind      1    355     355   136.2 6.3e-09 ***
Residuals 15     39       3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(c) What are the estimators for τ_2 and α_2 in the two-way ANOVA model (assuming we have used the constraints $\tau_1 = \alpha_1 = 0$). Compare these estimates with the linear contrasts calculated in part (c) of Question 2 on Tutorial 2. What do you notice?

Solution: The following command does the job.

```
> summary(btyp.lm)

Call:
lm(formula = resp ~ Bind + Rhind)

Residuals:
   Min     1Q Median     3Q    Max
-2.722 -0.847 -0.306  0.590  3.278

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   10.306      0.712   14.48  3.2e-10 ***
Bind           5.583      0.807    6.92  4.9e-06 ***
Rhind          9.417      0.807   11.67  6.3e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.61 on 15 degrees of freedom
Multiple R-squared:  0.925,        Adjusted R-squared:  0.915
F-statistic: 92.1 on 2 and 15 DF,  p-value: 3.78e-09
```

These estimates are close to the estimates arrived at in Tutorial 2 (5.4375 for Bind and 9.5626 for Rhind), but they are not exactly the same. This is a consequence of the unbalanced design. If the design had been balanced, then we would have arrived at identical estimates using the two approaches to measuring the effects of the two factors.

(d) Multiply the two indicators together to arrive at the indicator variable for the two-factor interaction, and test for additivity in the model.

Solution: The following commands do the job.

```
> int <- Bind*Rhind
> btyp.lm1 <- lm(resp~Bind+Rhind+int)
> anova(btyp.lm1)
```

```
Analysis of Variance Table

Response: resp
          Df Sum Sq Mean Sq F value  Pr(>F)
Bind       1    125     125   45.49 9.4e-06 ***
Rhind      1    355     355  129.40 1.9e-08 ***
int        1      1       1    0.25    0.63
Residuals 14     38       3
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

additive model: original model with additions only

So, the interaction term is clearly non-significant, indicating that the additive model is appropriate for this data. Note that the lines in the ANOVA table corresponding to the original indicators do not change, which is as it should be, since this is nature of sequential sums of squares, the new interaction term taking its sum of square out of the "unexplained" portion of the response variation as measured by the residual sum of squares in the analysis without the inclusion of the interaction term.

3. This question was adapted from Ramsey and Schafer (2013). A 1989 study investigated the effect of heredity and environment on intelligence. From adoption registers in France, researchers selected samples of adopted children whose biological parents and adoptive parents came from either the very highest or the very lowest socio-economic status (SES) categories (based on years of education and occupation). They attempted to obtain samples of size 10 from each combination: (1) high adoptive SES and high biological SES, (2) high adoptive SES and low biological SES, (3) low adoptive SES and high biological SES, and (4) low SES for both parents. It turned out, however, only eight children belonged to combination three. The 38 children were given IQ tests. The scores are in the data file `SES.csv`.

   (a) Does the difference in mean scores for those with high and low SES biological parents depend on whether the adoptive parents were high or low SES?

   Solution: We fit a two-way ANOVA model with interaction terms to see if there is any evidence of an interaction effect. The output below shows that the interaction can reasonably be omitted from the model (p-value=0.9174):

```
> ses <- read.table("SES.csv", sep=",", header=TRUE)
> ses

    IQ ADOPTIVE BIOLOGIC
1  136     High     High
2   99     High     High
3  121     High     High
4  133     High     High
5  125     High     High
6  131     High     High
```

```
7  103     High     High
8  115     High     High
9  116     High     High
10 117     High     High
11  94     High      Low
12 103     High      Low
13  99     High      Low
14 125     High      Low
15 111     High      Low
16  93     High      Low
17 101     High      Low
18  94     High      Low
19 125     High      Low
20  91     High      Low
21  98      Low     High
22  99      Low     High
23  91      Low     High
24 124      Low     High
25 100      Low     High
26 116      Low     High
27 113      Low     High
28 119      Low     High
29  92      Low      Low
30  91      Low      Low
31  98      Low      Low
32  83      Low      Low
33  99      Low      Low
34  68      Low      Low
35  76      Low      Low
36 115      Low      Low
37  86      Low      Low
38 116      Low      Low
```

```
> ses.lm1 <- lm(IQ~ADOPTIVE*BIOLOGIC, data=ses)
```

```
> summary(ses.lm1)
```

```
Call:
lm(formula = IQ ~ ADOPTIVE * BIOLOGIC, data = ses)

Residuals:
   Min     1Q Median     3Q    Max
-24.40  -9.47  -2.00   8.22  23.60
```

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)               119.60       4.18   28.61   <2e-16 ***
ADOPTIVELow               -12.10       6.27   -1.93    0.062 .
BIOLOGICLow               -16.00       5.91   -2.71    0.011 *
ADOPTIVELow:BIOLOGICLow     0.90       8.62    0.10    0.917
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.2 on 34 degrees of freedom
Multiple R-squared:  0.388,        Adjusted R-squared:  0.334
F-statistic: 7.19 on 3 and 34 DF,  p-value: 0.000726
```

(b) As you should know by now, the answer to part (a) is NO. Now find out how much the mean IQ score is affected by the SES of adoptive parents, and how much it is affected by the SES of the biological parents? Is one of these effects larger than the other?

Solution: First, to see whether there is strong evidence that mean IQ is associated with biological parents' SES after accounting for the association with adoptive parents' SES, OR whether there is strong evidence mean IQ is associated with adoptive parents' SES after accounting for the association with biological parents' SES, we use the `drop1` function in *R* as follows.

```
> ses.lm2 <- lm(IQ~ADOPTIVE+BIOLOGIC, data=ses)
> drop1(ses.lm2, test="F")

Single term deletions

Model:
IQ ~ ADOPTIVE + BIOLOGIC
         Df Sum of Sq  RSS AIC F value  Pr(>F)
<none>                5943 198
ADOPTIVE  1      1276 7219 203    7.51 0.00957 **
BIOLOGIC  1      2291 8235 208   13.49 0.00079 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Second, from the following summary output, it is estimated that the mean IQ for children of high SES biological parents is 15.6 points higher than those of low SES biological parents, regardless of adoptive parents SES. Similarly, it is estimated that the mean IQ of children of high SES adoptive parents is 11.6 points higher than those of low SES adoptive parents, again regardless of biological parents SES.

```
> summary(ses.lm2)

Call:
lm(formula = IQ ~ ADOPTIVE + BIOLOGIC, data = ses)

Residuals:
   Min     1Q Median     3Q    Max
-24.19  -9.62  -1.79   7.97  23.81

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   119.39       3.60   33.13  < 2e-16 ***
ADOPTIVELow   -11.62       4.24   -2.74  0.00957 **
BIOLOGICLow   -15.58       4.24   -3.67  0.00079 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13 on 35 degrees of freedom
Multiple R-squared:  0.388,        Adjusted R-squared:  0.353
F-statistic: 11.1 on 2 and 35 DF,  p-value: 0.000185
```

The above results seem to indicate that the effect of biological parents' SES appears to be stronger than the effect of adoptive parents' SES, although a formal test for equality of the regression coefficients fails to reject the null hypothesis in this case (p-value 0.52).

```
> h <- c(0,-1,1)
> a.low <- with(ses, ifelse(ADOPTIVE=="Low", 1, 0))
> b.low <- with(ses, ifelse(BIOLOGIC=="Low", 1, 0))
> ses.lm3 <- lm(IQ ~ a.low+b.low, ses)
> summary(ses.lm3)

Call:
lm(formula = IQ ~ a.low + b.low, data = ses)

Residuals:
   Min     1Q Median     3Q    Max
-24.19  -9.62  -1.79   7.97  23.81

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   119.39       3.60   33.13  < 2e-16 ***
a.low         -11.62       4.24   -2.74  0.00957 **
b.low         -15.58       4.24   -3.67  0.00079 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13 on 35 degrees of freedom
Multiple R-squared:  0.388,        Adjusted R-squared:  0.353
F-statistic: 11.1 on 2 and 35 DF,  p-value: 0.000185

> sigma <- summary(ses.lm3)$sigma
> Xmat <- cbind(1, a.low,b.low)
> XtXi <- solve(t(Xmat)%*%Xmat)
> est <- t(h)%*%coefficients(ses.lm3)
> sd <- sigma*sqrt(t(h)%*%XtXi%*%h)
> upper <- est + (qt(0.975, df.residual(ses.lm3))*sd)
> lower <- est - (qt(0.975, df.residual(ses.lm3))*sd)
> c(lower, est, upper)

[1] -16.4601  -3.9529   8.5542

> 2*(1-pt(abs(est/sd), ses.lm3$df.residual))

       [,1]
[1,] 0.5253
```

References

F. L. Ramsey and D. W. Schafer. The statistical sleuth: a course in methods of data analysis. Brooks/Cole, 2013.