Tutorial Problems

Question 1

- A manufacturer of household appliances wants to find the best combination of wash temperature and drying temperature to produce unwrinkled cotton sheets at the end of the laundry session.
- He wants to compare four different wash temperatures (`WTemp`) and three different drying temperatures (`DTemp`).
- He uses eight similar washing machines (`Wash`) and six similar dryers (`Dryer`).
- First, 48 cotton sheets are randomly allocated to the washing machines, six per machine.
- The wash temperatures are randomly allocated to the washing machines so that two machines are run at each temperature.
- After the wash, the six sheets in each machine are randomly allocated to the dryers, one per dryer.
- Then the drying temperatures are randomly allocated to the dryers so that two machines are run at each temperature.
- After the drying, all 48 sheets are scored by experts for how wrinkled they are.

```
[1] 1 4 8 2 6 3 7 5
[1] 2 3 5 1 4 6
```

```
summary(aov(Wrinkle ~ DTemp*WTemp + Wash + Dryer, data=dat))
```

```
             Df Sum Sq Mean Sq F value Pr(>F)
DTemp         2   55.2   27.58   0.267  0.767
WTemp         3  219.4   73.13   0.708  0.555
Wash          4  292.3   73.06   0.707  0.593
Dryer         3  127.7   42.56   0.412  0.746
DTemp:WTemp   6  261.7   43.61   0.422  0.858
Residuals    29 2995.3  103.29
```

(a) Identify the treatment structure and the structural factors within the experiment. What are the experimental and observational units are and how many of each there are?

(b) Complete the ANOVA table below.

Strata	Source	df	SS	MS	F
Dryer	Drying temperature				
	Residual				
Washing machine	Wash temperature				
	Residual				
Sheet	Drying and Wash temperature				
	Residual				

(c) Does the interaction of drying and washing temperature make a difference to the wrinkle of the sheets?

Question 2

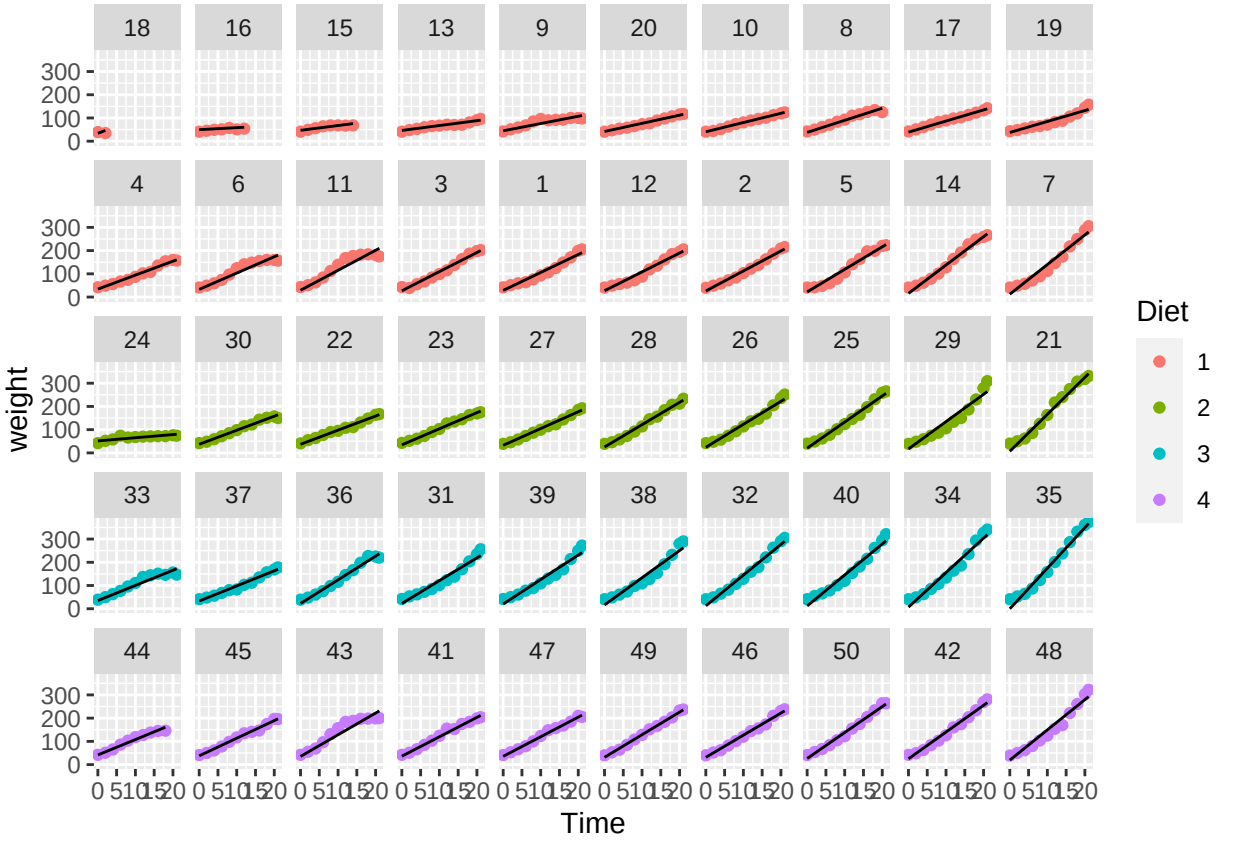
The body weights of the chicks were measured at birth and every second day thereafter until day 20. They were also measured on day 21. There were four groups on chicks on different protein diets. The data frame `ChickWeight` has 578 rows and 4 columns:

- `weight` – a numeric vector giving the body weight of the chick (gm);
- `Time` – a numeric vector giving the number of days since birth;
- `Chick` – unique identifier for the chick; and
- `Diet` – a factor of four levels indicating diet that the chick received.

The data is built-in in `R` and can be viewed by typing `ChickWeight` and pressing enter. The following shows the weight as a function of time for each chick.

```
library(ggplot2)
library(lme4)
```

```
M1 <- lmer(weight ~ Diet*Time + (Time|Chick), data=ChickWeight)
ChickWeight$predict <- predict(M1)
ggplot(ChickWeight, aes(Time, weight)) + geom_point(aes(col=Diet)) +
  geom_line(aes(Time, predict)) + facet_wrap(~Chick, ncol=10)
```



We can see that some chick weights are not observed (specifically Chick 8, 15, 16, 18, and 44).


```
ChickWeight %>%
group_by(Chick, Time) %>%
count() %>%
spread(Time, n) %>%
drop_na() %>%
pull(Chick) %>%
setdiff(levels(ChickWeight$Chick), .)
```

```
[1] "18" "16" "15" "8"  "44"
```

For simplicity, we fitler out those Chick to work with the complete data (not recommended).

```
dat <- ChickWeight %>%
filter(!(Chick %in% c("18","16","15","8","44"))) %>%
droplevels()
nrow(dat)
```

```
[1] 540
```

```
nlevels(dat$Chick)
```

```
[1] 45
```

Suppose that the fitted model is given as

$$\boldsymbol{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{u} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\beta}$ is the vector of fixed effects, $\boldsymbol{u}$ is the vector of random effects and the $\boldsymbol{\epsilon}$ is the vector of error. We assume that

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{\epsilon} \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix} \right).$$

Suppose that $\boldsymbol{Y}$ is ordered by time within chick; $\boldsymbol{t}$ is the 578×1 vector of associated time (in days); $\boldsymbol{t}_u = (0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 21)^\top$ is the vector of days; and $\boldsymbol{\delta}_2, \boldsymbol{\delta}_3, \boldsymbol{\delta}_4$ are the 578×1 dummy vectors with the corresponding entries are one if the associated response is for Diet 2, 3, or 4, respectively, otherwise the entry is zero.

(a) Write what $\boldsymbol{\beta}$, $\boldsymbol{u}$, $\mathbf{X}$, $\mathbf{Z}$, $\mathbf{G}$ and $\mathbf{R}$ are for the model below.

```
M1 <- lmer(weight ~ Diet*Time + (Time|Chick), data=ChickWeight)
```

(b) Write what $\boldsymbol{\beta}$, $\boldsymbol{u}$, $\mathbf{X}$, $\mathbf{Z}$, $\mathbf{G}$ and $\mathbf{R}$ are for the model below.

```
M2 <- lmer(weight ~ Diet*Time + (1|Chick) + (0 + Time|Chick), data=ChickWeight)
```

(c) Write what $\boldsymbol{\beta}$, $\boldsymbol{u}$, $\mathbf{X}$, $\mathbf{Z}$, $\mathbf{G}$ and $\mathbf{R}$ are for the model below.

```
M3 <- lmer(weight ~ Diet*(1 + Time + I(Time^2)) +
             (1 + Time + I(Time^2)|Chick), data=ChickWeight)
```

(d) What is the null hypothesis associated with the following p-value?


```
anova(M1, M2)
```

```
refitting model(s) with ML (instead of REML)

Data: ChickWeight
Models:
M2: weight ~ Diet * Time + (1 | Chick) + (0 + Time | Chick)
M1: weight ~ Diet * Time + (Time | Chick)
   Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
M2 11 4888.3 4936.3 -2433.2   4866.3
M1 12 4824.2 4876.5 -2400.1   4800.2 66.068      1  4.357e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(e) Using `M3`, how would you rank the different `Diet`?

```
VarCorr(M3)
```

```
 Groups   Name        Std.Dev. Corr
 Chick    (Intercept) 5.62318
          Time        3.42026  -0.968
          I(Time^2)   0.21199   0.450 -0.660
 Residual             6.56407
```

```
round(fixef(M3),2)
```

```
      (Intercept)             Diet2             Diet3             Diet4              Time
            37.53              0.15              1.29             -1.71              5.11
       I(Time^2)       Diet2:Time       Diet3:Time       Diet4:Time Diet2:I(Time^2)
             0.05              0.70             -0.05              3.28              0.08
Diet3:I(Time^2) Diet4:I(Time^2)
             0.25              0.00
```

Question 3

(Advanced) Suppose that we are testing t treatments. You employ a randomised complete block design using b blocks and each treatment appearing exactly once in each block. Suppose that model the response as

$$Y_{ij} = \alpha_i + \beta_j + \epsilon_{ij}, \qquad \epsilon_{ij} \sim NID(0, \sigma^2), \quad i = 1, \ldots, t \text{ and } j = 1, \ldots, b. \tag{1}$$

(a) Assume that $\beta_1 = 0$. We rewrite the above model in the matrix notation

$$\boldsymbol{Y} = \mathbf{X}_t \boldsymbol{\tau} + \mathbf{X}_b \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{Y}$ is ordered by treatment within block, i.e. $\boldsymbol{Y} = (Y_{11}, Y_{21}, Y_{31}, \ldots, Y_{tb})^\top$; $\boldsymbol{\tau} = (\alpha_1, \ldots, \alpha_t)^\top$, $\boldsymbol{\beta} = (\beta_2, \ldots, \beta_b)^\top$ and $\boldsymbol{\epsilon}$ is the random error. What is $\mathbf{X}_t$ and $\mathbf{X}_b$?

(b) Show that the estimator $\hat{\boldsymbol{\tau}}_1 = (\mathbf{X}_t^\top \mathbf{S} \mathbf{X}_t)^{-1} \mathbf{X}_t^\top \mathbf{S} \boldsymbol{Y}$ where $\mathbf{S} = \boldsymbol{I}_{bt} - \mathbf{X}_b (\mathbf{X}_b^\top \mathbf{X}_b)^{-1} \mathbf{X}_b^\top$ is the least squares estimator or the maximum likelihood estimator for model (1).

(c) Now suppose that we ignore the block effects and model the response as

$$\boldsymbol{Y} = \mathbf{X}_t\boldsymbol{\tau} + \boldsymbol{\epsilon} \qquad \boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\boldsymbol{I}_{bt}). \tag{2}$$

Show that the estimator $\hat{\boldsymbol{\tau}}_2 = (\mathbf{X}_t^\top\mathbf{X}_t)^{-1}\mathbf{X}_t^\top\boldsymbol{Y}$ is the least square estimator or maximum likelihood estimate of $\boldsymbol{\tau}$ in model (2).

(d) Now suppose that the response is modelled as

$$Y_{ij} = \alpha_i + B_j + \epsilon_{ij}, \qquad B_j \sim NID(0, \sigma_b^2) \quad \epsilon_{ij} \sim NID(0, \sigma^2), \quad i = 1, \ldots, t \text{ and } j = 1, \ldots, b. \tag{3}$$

or equivalently,

$$\boldsymbol{Y} = \mathbf{X}_t\boldsymbol{\tau} + \mathbf{Z}_b\boldsymbol{u}_b + \boldsymbol{\epsilon}; \quad \begin{bmatrix}\boldsymbol{u}_b \\ \boldsymbol{\epsilon}\end{bmatrix} \sim N\left(\begin{bmatrix}\mathbf{0} \\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\sigma_b^2\boldsymbol{I}_b & \mathbf{0} \\ \mathbf{0} & \sigma^2\boldsymbol{I}_{bt}\end{bmatrix}\right)$$

where $\mathbf{Z}_b = \boldsymbol{I}_t \otimes \mathbf{1}_b$ is the design matrix for the random block effects $\boldsymbol{u}_b$. Suppose $\text{var}(\boldsymbol{Y}) = \mathbf{V}$.

(i) Show that $\mathbf{V}^{-1} = \dfrac{1}{\sigma^2}\boldsymbol{I}_b \otimes \left(\boldsymbol{I}_t - \dfrac{\sigma_b^2}{\sigma^2 + t\sigma_b^2}\boldsymbol{J}_t\right)$ using Sherman-Morrison formula $(\mathbf{A} + \boldsymbol{u}\boldsymbol{v}^\top)^{-1} = \mathbf{A}^{-1} - \dfrac{\mathbf{A}^{-1}\boldsymbol{u}\boldsymbol{v}^\top\mathbf{A}^{-1}}{1 + \boldsymbol{v}^\top\mathbf{A}^{-1}\boldsymbol{u}}$ or otherwise.

(ii) Woodbury matrix identity is given as

$$(\mathbf{A} + \mathbf{B})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\boldsymbol{I} + \mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{A}^{-1}$$

where $\mathbf{A}$ and $\mathbf{B}$ are matrices of conformable sizes.

Show that using the result from (a) and the Woodbury matrix identity or otherwise that

$$(\mathbf{X}_t^\top\mathbf{V}^{-1}\mathbf{X}_t)^{-1} = \sigma^2\left((\mathbf{X}_t^\top\mathbf{X}_t)^{-1} + \frac{\sigma_b^2}{b\sigma^2}\boldsymbol{J}_t\right).$$

(iii) What is the maximum likelihood estimate of $\boldsymbol{\tau}$ ($\hat{\boldsymbol{\tau}}_3$) for model (3)?

(iv) How does this estimate compare to the one in (b) and (c)?

Computer Problems

Question 1

In an experiment the yields Y of 3 varieties A, B and C (x_1, x_2 and x_3) of corn are investigated. Besides the variety it is more than possible that the number of plants (x4) per plot that survived until harvest explains variability in the response Y . The data are

Variety	A		B		C	
	x_4	Y	x_4	Y	x_4	Y
	21	165	27	201	24	184
	26	191	27	203	27	186
	29	202	22	145	27	187
	20	134	25	180	28	219
	23	201	29	231	30	262

Consider the model

$$Y_l = \beta_1 x_{l1} + \beta_2 x_{l2} + \beta_3 x_{l3} + \beta_4 x_{l4} + \epsilon_l, \quad \epsilon_l \sim NID(0, \sigma^2), \quad l = 1, 2, \ldots, 15,$$

where β_1, β_2 and β_3 correspond to the effects of varieties A , B and C , respectively. We could alternatively write the model as

$$Y_{ij} = \beta_i + \beta_4 x_{ij,4} + \epsilon_{ij}, \quad \epsilon_{ij} \sim NID(0, \sigma^2), \quad i = 1, 2, 3 \text{ and } j = 1, \ldots, 5.$$

(a) Construct the data frame `dat <- data.frame(Y, x1, x2, x3, x4)` with columns `Y, x4, x1, x2`, and `x3`, where `x1, x2` and `x3` are columns of 0's and 1's corresponding to the coefficients of β_1, β_2 and β_3, respectively. Hint:

```
x1 <- c(1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
x4 <- c(21, 26, ... , 28, 30)
```

(Assumed knowledge and Lecture 8).

(b) Produce a scatterplot of yields (on y-axis) against the number of plants (on x-axis), colour each point according to the variety.

(c) Form the $\mathbf{X}$ matrix corresponding to the above model and use the matrix approach to find the least squares estimates for the parameters.Calculate the vector of residuals $\boldsymbol{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ and find the square root of the diagonal elements of $\hat{\sigma}^2(\mathbf{X}^\top\mathbf{X})^{-1}$.

(d) Use

```
lm(Y ~ x1 + x2 + x3 + x4 - 1, dat)
```

to fit the model using lm. Note the -1 term is included since the model does not have a general constant term. Check the summary values against the answers in (c).

(e) If all the varieties have the same effect then we have the simple linear regression model

$$Y_i = \beta_0 + \beta_4 x_{4i} + \epsilon_i, \quad \epsilon_i \sim NID(0, \sigma^2).$$

Use `lm` to fit this linear model.

(f) Compare the two models considered for this data set in terms of the best fit for the smallest number of parameters.

(g) Select the best model using

1. Forward, starting with the intercept only model (regardless your findings in (f)), using the AIC criterion (in case you are going to use the step function make sure you specify the option `trace=0` to suppress most of the output),
2. Stepwise, starting with the full model, using an F test with $p_{out} = 0.10$ and $p_{in} = 0.05$,
3. Backward, starting with the full model, using the AIC criterion,
4. Stepwise, starting with the full model, using the BIC criterion,
5. For the model of your choice run a robust multiple regression using the `rlm` in `library(MASS)` with `method="MM"`. Are there any differences to the LS solution (with `lm`)?

Question 2

A plant contains a large number of coil winding machines. A production analyst studied a certain characteristic of the wound coils produced by these machines by selecting four machines at random and then choosing 10 coils at random from the day's output of each selected machine. The results in the study were:

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]  205  204  207  202  208  206  209  205  207  206
[2,]  201  204  198  203  209  207  199  206  205  204
[3,]  198  204  196  201  199  203  202  198  202  197
[4,]  210  209  214  215  211  208  210  209  211  210
```

where rows represent the machines and columns the coils.

(a) Enter the data, provide summary measures using `tapply(...,...,summary)` and also provide boxplots for each level of the factor machine.

(b) Run an ANOVA model II and compare its output to the ANOVA model I.

(c) Obtain the residuals and check the normality assumption for the ten observations from each of the four machines.

