Exercise 1

Repeat the calculations of Subsection 2.3.5, but omitting results from two vines at random. Here is code that will handle the calculation:

\[
\begin{align*}
n_{\text{omit}} & = 2 \\
take & \leftarrow \text{rep}(\text{TRUE}, 48) \\
take[\text{sample}(1:48, 2)] & \leftarrow \text{FALSE} \\
\text{kiwishade.lmer} & \leftarrow \text{lmer}(\text{yield} \sim \text{shade} + (1|\text{block}) + (1|\text{block:plot}), \\
& \quad \text{data} = \text{kiwishade}, \text{subset}=\text{take}) \\
\text{vcov} & \leftarrow \text{show}((\text{VarCorr(\text{kiwishade.lmer})))} \\
\text{gps} & \leftarrow \text{vcov[,]}, "\text{Groups"}] \\
\text{print(vcov[gps=="block:plot", "\text{Variance}"])} \\
\text{print(vcov[gps=="\text{Residual}, "\text{Variance}"])}
\end{align*}
\]

Repeat this calculation five times, for each of \(n_{\text{omit}} = 2, 4, 6, 8, 10, 12\) and 14. Plot (i) the plot component of variance and (ii) the vine component of variance, against number of points omitted. Based on these results, for what value of \(n_{\text{omit}}\) does the loss of vines begin to compromise results? Which of the two components of variance estimates is more damaged by the loss of observations? Comment on why this is to be expected.

For convenience, we place the central part of the calculation in a function. On slow machines, the code may take a minute or two to run.

\[
\begin{align*}
\text{trashvine} & \leftarrow \text{function}(n_{\text{omit}}=2) \\
& + \{ \\
& + \quad k \leftarrow k+1 \\
& + \quad n[k] \leftarrow n_{\text{omit}} \\
& + \quad \text{take} \leftarrow \text{rep}(\text{T}, 48) \\
& + \quad \text{take[\text{sample}(1:48, n_{\text{omit})}] \leftarrow F} \\
& + \quad \text{kiwishade$take < take} \\
& + \quad \text{kiwishade.lmer} \leftarrow \text{lmer}(\text{yield} \sim \text{shade} + (1 | \text{block}) + (1|\text{block:plot}), \\
& \quad \quad \text{data} = \text{kiwishade, subset=take)} \\
& + \quad \text{varv} \leftarrow \text{as.numeric(attr(\text{VarCorr(\text{kiwishade.lmer}), "sc"})^2) \\
& + \quad \text{varp} \leftarrow \text{as.numeric(\text{VarCorr(\text{kiwishade.lmer})$block:plot\')}} \\
& + \quad \text{c(varp, varv)} \\
& + \} \\
\end{align*}
\]

> \text{varp} \leftarrow \text{numeric(35)} \\
> \text{varv} \leftarrow \text{numeric(35)} \\
> \text{n} \leftarrow \text{numeric(35)} \\
> \text{k} \leftarrow 0 \\
> \text{for(n_{\text{omit}} in c(2, 4, 6, 8, 10, 12, 14))} \\
+ \text{for(i in 1:5){} \\
+ \text{k} \leftarrow k+1
We plot the results:

![Graph showing within and between plots variance estimates](image)

Figure 1: Within, and between plots variance estimates, as functions of the number of vines that were omitted at random

As the number of vines that are omitted increases, the variance estimates can be expected to show greater variability. The effect should be most evident on the between plot variance. Inaccuracy in estimates of the between plot variance arise both from inaccuracy in the within plot sums of squares and from loss of information at the between plot level.

At best it is possible only to give an approximate d.f. for the between plot estimate of variance (some plots lose more vines than others), which complicates any evaluation that relies on degree of freedom considerations.

Exercise 2
Repeat the previous exercise, but now omitting 1, 2, 3, 4 complete plots at random.

```r
> trashplot <- function(n.omit=2)
+ {
+   k <- k+1
+   n[k] <- n.omit
+   plotlev <- levels(kiwishade$plot)
+   use.lev <- sample(plotlev, length(plotlev)-n.omit)
+   kiwishade$take <- kiwishade$plot %in% use.lev
+   kiwishade.lmer <- lmer(yield ~ shade + (1 | block) + (1|block:plot),
+     data = kiwishade, subset=take)
+   varv <- as.numeric(attr(VarCorr(kiwishade.lmer), "sc")^2)
+   varp <- as.numeric(VarCorr(kiwishade.lmer)$`block:plot`
+     + c(varp, varv)
+ }>
> varp <- numeric(20)
```
> varv <- numeric(20)
> n <- numeric(20)
> k <- 0
> for(n.omit in 1:4)
+ for(i in 1:5){
+ k <- k+1
+ vec2 <- trashplot(n.omit=n.omit)
+ n[k] <- n.omit
+ varp[k] <- vec2[1]
+ varv[k] <- vec2[2]
+ }

Again, we plot the results:

![within and between plots variance estimates](image)

Figure 2: Within, and between plots variance estimates, as functions of the number of whole plots (each consisting of four vines) that were omitted at random.

Omission of a whole plot loses 3 d.f. out of 36 for estimation of within plot effects, and 1 degree of freedom out of 11 for the estimation of between plot effects, i.e., a slightly greater relative loss. The effect on precision will be most obvious where the d.f. are already smallest, i.e., for the between plot variance. The loss of information on complete plots is inherently for serious, for the estimation of the between plot variance, than the loss of partial information (albeit on a greater number of plots) as will often happen in Exercise 1.

**Exercise 3**

The data set Gun (MEMSS package) reports on the numbers of rounds fired per minute, by each of nine teams of gunners, each tested twice using each of two methods. In the nine teams, three were made of men with slight build, three with average, and three with heavy build. Is there a detectable difference, in number of rounds fired, between build type or between firing methods? For improving the precision of results, which would be better – to double the number of teams, or to double the number of occasions (from 2 to 4) on which each team tests each method?

It probably does not make much sense to look for overall differences in Method; this depends on Physique. We therefore nest Method within Physique.
> library(MEMSS)
> Gun.lmer <- lmer(rounds ~ Physique/Method + (1|Team), data=Gun)
> summary(Gun.lmer)

Linear mixed model fit by REML
Formula: rounds ~ Physique/Method + (1 | Team)
Data: Gun
AIC  BIC logLik deviance REMLdev
143 156 -63.5 134 127
Random effects:
Groups   Name   Variance  Std.Dev.
Team     (Intercept) 1.09   1.04
Residual           2.18   1.48
Number of obs: 36, groups: Team, 9

Fixed effects:         Estimate Std. Error t value
(Intercept)             23.589   0.492  47.9
Physique.L              -0.966   0.853  -1.1
Physique.Q               0.191   0.853   0.2
PhysiqueSlight:MethodM2  -8.450   0.852  -9.9
PhysiqueAverage:MethodM2 -8.100   0.852  -9.5
PhysiqueHeavy:MethodM2   -8.983   0.852 -10.5

Physique.L                0.000
Physique.Q                0.000  0.000
PhysiqueSlight:MM2       -0.289  0.353  -0.204
PhysiqueAverage:MM2      -0.289  0.000   0.408  0.000
PhysiqueHeavy:MM2        -0.289 -0.353  -0.204  0.000   0.000

A good way to proceed is to determine the fitted values, and present these in an interaction plot:

> Gun.hat <- fitted(Gun.lmer)
> interaction.plot(Gun$Physique, Gun$Method, Gun.hat)

Differences between methods, for each of the three physiques, are strongly attested. These can be estimated within teams, allowing 24 degrees of freedom for each of these comparisons.

Clear patterns of change with Physique seem apparent in the plot. There are however too few degrees of freedom for this effect to appear statistically significant. Note however that the parameters that are given are for the lowest level of Method, i.e., for M1. Making M2 the baseline shows the effect as closer to the conventional 5% significance level.

The component of variance at the between teams level is of the same order of magnitude as the within teams component. Its contribution to the variance of team means (1.044^2) is much greater than the contribution of the within team component (1.476^2/4; there are 4 results per team). If comparison between physiques is the concern; it will be much more effective to double the number of teams; compare (1.044^2+1.476^2/4)/2 (=0.82) with 1.044^2+1.476^2/8 (=1.36).
Exercise 4

*The data set ergoStool (MEMSS package) has data on the amount of effort needed to get up from a stool, for each of nine individuals who each tried four different types of stool. Analyse the data both using aov() and using lme(), and reconcile the two sets of output. Was there any clear winner among the types of stool, if the aim is to keep effort to a minimum?

For analysis of variance, specify

```r
> aov(effort~Type+Error(Subject), data=ergoStool)
```

Call:

```r
aov(formula = effort ~ Type + Error(Subject), data = ergoStool)
```

Grand Mean: 10.25

Stratum 1: Subject

Terms:

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>Deg. of Freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Residuals</td>
<td>66.5</td>
<td>8</td>
</tr>
</tbody>
</table>

Residual standard error: 2.883

Stratum 2: Within

Terms:

<table>
<thead>
<tr>
<th></th>
<th>Sum of Squares</th>
<th>Deg. of Freedom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type</td>
<td>81.19</td>
<td>3</td>
</tr>
<tr>
<td>Residuals</td>
<td>29.06</td>
<td>24</td>
</tr>
</tbody>
</table>

Residual standard error: 1.100

Estimated effects may be unbalanced

For testing the Type effect for statistical significance, refer \((81.19/3)/(29.06/24) (=22.35)\) with the \(F_{3,24}\) distribution. The effect is highly significant.

This is about as far as it is possible to go with analysis of variance calculations. When Error() is specified in the aov model, R has no mechanism for extracting estimates. (There are mildly tortuous ways to extract the information, which will not be further discussed here.)

For use of lmer, specify

```r
> summary(lmer(effort~Type + (1|Subject), data=ergoStool))
```

Linear mixed model fit by REML

Formula: effort ~ Type + (1 | Subject)

Data: ergoStool

<table>
<thead>
<tr>
<th>AIC</th>
<th>BIC</th>
<th>logLik</th>
<th>deviance</th>
<th>REMLdev</th>
</tr>
</thead>
<tbody>
<tr>
<td>133</td>
<td>143</td>
<td>-60.6</td>
<td>121</td>
<td>121</td>
</tr>
</tbody>
</table>

Random effects:

<table>
<thead>
<tr>
<th>Groups</th>
<th>Name</th>
<th>Variance</th>
<th>Std.Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>(Intercept)</td>
<td>1.78</td>
<td>1.33</td>
</tr>
<tr>
<td>Residual</td>
<td></td>
<td>1.21</td>
<td>1.10</td>
</tr>
</tbody>
</table>

Number of obs: 36, groups: Subject, 9
Fixed effects:

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>8.556</td>
<td>0.576</td>
<td>14.85</td>
</tr>
<tr>
<td>TypeT2</td>
<td>3.889</td>
<td>0.519</td>
<td>7.50</td>
</tr>
<tr>
<td>TypeT3</td>
<td>2.222</td>
<td>0.519</td>
<td>4.28</td>
</tr>
<tr>
<td>TypeT4</td>
<td>0.667</td>
<td>0.519</td>
<td>1.29</td>
</tr>
</tbody>
</table>

Correlation of Fixed Effects:

<table>
<thead>
<tr>
<th></th>
<th>(Intr)</th>
<th>TypeT2</th>
<th>TypeT3</th>
</tr>
</thead>
<tbody>
<tr>
<td>TypeT2</td>
<td>-0.450</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TypeT3</td>
<td>-0.450</td>
<td>0.500</td>
<td></td>
</tr>
<tr>
<td>TypeT4</td>
<td>-0.450</td>
<td>0.500</td>
<td>0.500</td>
</tr>
</tbody>
</table>

Observe that $1.100295^2$ (Residual StdDev) is very nearly equal to 29.06/24 obtained from the analysis of variance calculation.

Also the Stratum 1 mean square of 66.5/8 (=8.3125) from the analysis of variance output is very nearly equal to $1.3325^2 + 1.100295^2/4 (=2.078)$ from the lme output.

Exercise 5*

In the data set MathAchieve (MEMSS package), the factors Minority (levels yes and no) and sex, and the variable SES (socio-economic status) are clearly fixed effects. Discuss how the decision whether to treat School as a fixed or as a random effect might depend on the purpose of the study? Carry out an analysis that treats School as a random effect. Are differences between schools greater than can be explained by within school variation?

School should be treated as a random effect if the intention is to generalize results to other comparable schools. If the intention is to apply them to other pupils or classes within those same schools, it should be taken as a fixed effect.

For the analysis of these data, both SES and MEANSES should be included in the model. Then the coefficient of MEANSES will measure between school effects, while the coefficient of SES will measure within school effects.

```r
> library(MEMSS)
> MathAch.lmer <- lmer(MathAch ~ Minority * Sex * (MEANSES + SES) + (1 | School),
+ data=MathAchieve)
> options(width=90)
> MathAch.lmer

Linear mixed model fit by REML
Formula: MathAch ~ Minority * Sex * (MEANSES + SES) + (1 | School)
Data: MathAchieve
AIC  BIC logLik deviance REMLdev
46344 46441 -23158 46308 46316
Random effects:
Groups   Name     Variance Std.Dev.
School   (Intercept) 2.51     1.58     
Residual            35.79    5.98     
Number of obs: 7185, groups: School, 160

Fixed effects:

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>12.799</td>
<td>0.179</td>
<td>71.4</td>
</tr>
<tr>
<td>MinorityYes</td>
<td>-2.605</td>
<td>0.279</td>
<td>-9.3</td>
</tr>
</tbody>
</table>
SexMale 1.277 0.186 6.9
MEANSES 2.237 0.504 4.4
SES 2.508 0.185 13.5
MinorityYes:SexMale -0.462 0.376 -1.2
MinorityYes:MEANSES 1.439 0.684 2.1
MinorityYes:SES -1.101 0.319 -3.5
SexMale:MEANSES 0.574 0.574 1.0
SexMale:SES -0.517 0.264 -2.0
MinorityYes:SexMale:MEANSES -0.713 0.903 -0.8
MinorityYes:SexMale:SES 0.110 0.468 0.2

Correlation of Fixed Effects:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>MinorityYes</td>
<td>-0.346</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SexMale</td>
<td>-0.481</td>
<td>0.268</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MEANSES</td>
<td>-0.095</td>
<td>0.066</td>
<td>0.054</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SES</td>
<td>-0.017</td>
<td>0.031</td>
<td>0.007</td>
<td>-0.355</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MntryY:SexM</td>
<td>0.207</td>
<td>-0.671</td>
<td>-0.433</td>
<td>-0.030</td>
<td>-0.010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MnY:MEANSES</td>
<td>0.091</td>
<td>0.161</td>
<td>-0.043</td>
<td>-0.510</td>
<td>0.271</td>
<td>-0.142</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MntryY:SES</td>
<td>0.008</td>
<td>0.117</td>
<td>-0.012</td>
<td>0.211</td>
<td>-0.584</td>
<td>-0.089</td>
<td>-0.446</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SexMale:MEANSES</td>
<td>0.044</td>
<td>-0.035</td>
<td>-0.141</td>
<td>-0.540</td>
<td>0.315</td>
<td>0.092</td>
<td>0.366</td>
<td>-0.181</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SexMale:SES</td>
<td>0.010</td>
<td>-0.017</td>
<td>-0.081</td>
<td>0.252</td>
<td>-0.703</td>
<td>0.046</td>
<td>-0.194</td>
<td>0.409</td>
<td>-0.430</td>
<td></td>
</tr>
<tr>
<td>MY:SM:MEAN</td>
<td>-0.033</td>
<td>-0.140</td>
<td>0.096</td>
<td>0.316</td>
<td>-0.205</td>
<td>0.120</td>
<td>-0.651</td>
<td>0.332</td>
<td>-0.576</td>
<td>0.280</td>
</tr>
<tr>
<td>MnY:SM:SES</td>
<td>-0.011</td>
<td>-0.076</td>
<td>0.056</td>
<td>-0.140</td>
<td>0.397</td>
<td>0.122</td>
<td>0.300</td>
<td>-0.678</td>
<td>0.241</td>
<td>-0.567</td>
</tr>
</tbody>
</table>

> options(width=68)

The between school component of variance (1.585^2) is 2.51, compared with a within school component that equals 35.79. To get confidence intervals (strictly Bayesian credible intervals) for these variance estimates, specify:

```
> MathAch.mcmc <- mcmcsamp(MathAch.lmer, n=10000)
> HPDinterval(VarCorr(MathAch.mcmc, type="varcov"))
```

```
 lower upper
 [1,] 1.626 2.954
 [2,] 34.698 37.061
```

attr(,"Probability")

[1] 0.95

The 95% confidence interval for the between school component of variance ranged, in my calculation, from 1.64 to 3.0. The confidence interval excludes 0.

The number of results for school varies between 14 and 67. Thus, the relative contribution to class means is 5.51 and a number that is at most 5.982429^2/14 = 2.56.