Homework 6

Sarah Adilijiang

Problem 2

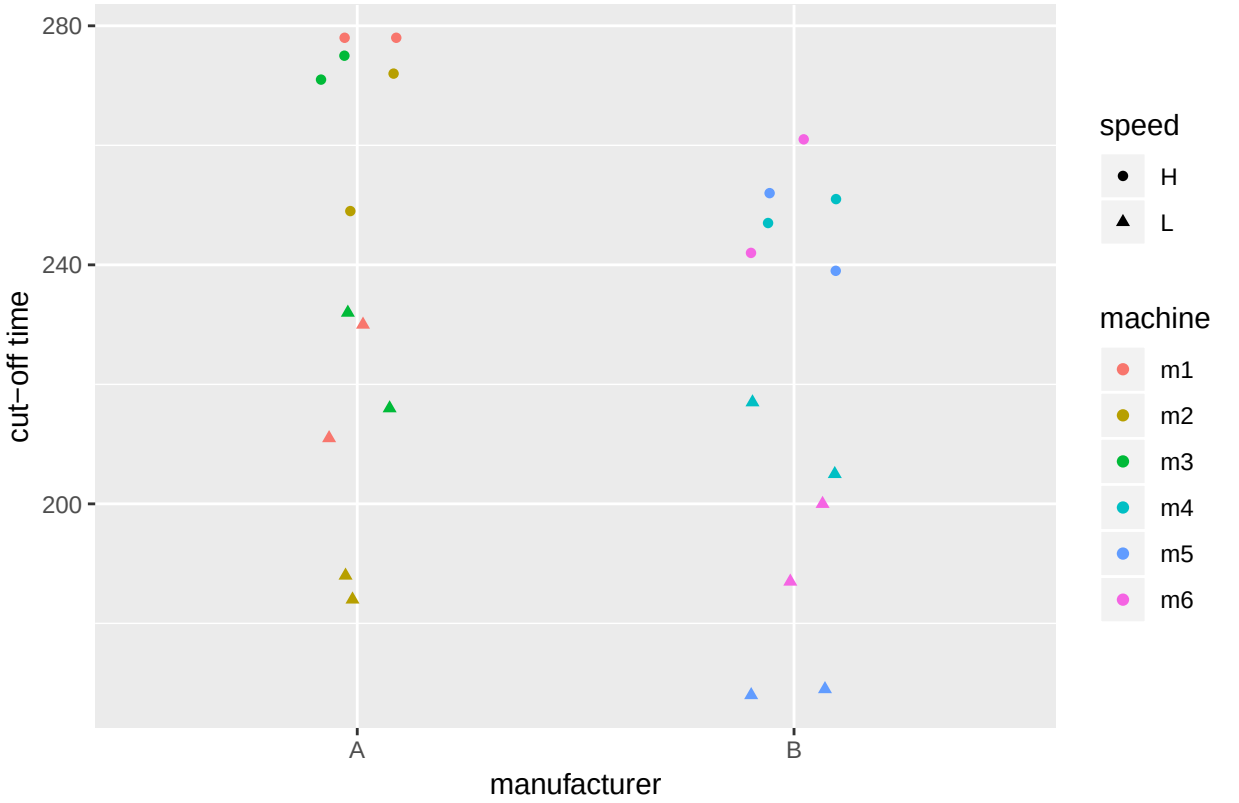
(a)

```
library(faraway)
data(lawn)

# plot the data
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.2
```

```
ggplot(lawn, aes(x=manufact, y=time, color=machine, shape=speed)) +
    geom_point(position=position_jitter(width=0.1, height=0.0)) +
    labs(x="manufacturer", y="cut-off time", caption="Note: Some jittering has been used to make coinci
```



Note: Some jittering has been used to make coincident points apparent.

Comment:

From the data plot, we can see that there is a clearly significant difference in the cut-off times of lawnmowers between different speeds: the high speed ones have much larger cut-off times. So speed seems to be a significant effect for the response cut-off time.

For manufacturers, the cut-off times of manufacturer A is slightly larger than that of manufacturer B, but

it's hard to say whether the difference is significant.

For machines, when using low speed, there is a significant difference in the cut-off times of lawnmowers between different machines, however, when using high speed, the differences between machines become smaller. And the difference in cut-off times between machines is slightly larger for those produced by manufacturer B than those by manufacturer A. To sum up, overall, there is a significant difference in the cut-off times between machines.

(b)

The fixed effects model is:
$time_{ijkl} = \mu + speed_i + manufact_j + machine_k + \epsilon_{ijkl}$, where μ, $speed_i$, $manufact_j$, $machine_k$ are all fixed effects, and error term ϵ_{ijkl} i.i.d. $\sim N(0, \sigma^2)$.

```
# fixed effect model
fmod = lm(time~speed+manufact+machine, lawn)
summary(fmod)
```

```
##
## Call:
## lm(formula = time ~ speed + manufact + machine, data = lawn)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -12.50  -8.50  -3.00   7.50  19.25
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  278.750      6.042  46.138  < 2e-16 ***
## speedL       -59.000      4.567 -12.919 3.23e-10 ***
## manufactB    -26.750      7.910  -3.382  0.00355 **
## machinem2    -26.000      7.910  -3.287  0.00435 **
## machinem3     -0.750      7.910  -0.095  0.92557
## machinem4      7.500      7.910   0.948  0.35635
## machinem5    -15.500      7.910  -1.959  0.06667 .
## machinem6         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.19 on 17 degrees of freedom
## Multiple R-squared:  0.9251, Adjusted R-squared:  0.8986
## F-statistic: 34.97 on 6 and 17 DF,  p-value: 1.183e-08
```

```
# check the data design matrix of the model
X = model.matrix(fmod)
# machinem6 = manufactB - machinem4 - machinem5
rbind(X[,3]-X[,6]-X[,7],  X[,8])
```

```
##      1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24
## [1,] 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  1  1  0  0  1  1
## [2,] 0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  1  1  0  0  1  1
```

Answer:

The model summary shows that the effect of $machinem6$ cannot be estimated. We can see that in the design matrix, the following indicator vectors have a relationship: $machinem6 = manufactB - machinem4 -$

$machinem5$, which makes $machinem6$ a redundant indicator vector that occurs in the last. Thus there is an NA appearing for the $machinem6$ factor level due to the rank deficiency of the design matrix.

(c)

The mixed effects model is:
$time_{ijkl} = \mu + speed_i + manufact_j + machine_k + \epsilon_{ijkl}$, where μ, $speed_i$, $manufact_j$ are fixed effects, and $machine_i$ is random effects.
Random effects $machine_k$ i.i.d. $\sim N(0, \sigma_b^2)$, error term ϵ_{ijkl} i.i.d. $\sim N(0, \sigma^2)$, and $machine_k$ and ϵ_{ijkl} are independent.

```
# mixed effect model
library(lme4)
```

```
## Loading required package: Matrix
```

```
rmod = lmer(time ~ speed * manufact + (1|machine), lawn)

# SD of the same machine with same speed
resid.sd = as.data.frame(VarCorr(rmod))$sdcor[2]
resid.sd
```

```
## [1] 11.50226
```

```
# SD of different machines with same speed
group.sd = as.data.frame(VarCorr(rmod))$sdcor[1]
sqrt(resid.sd^2 + group.sd^2)
```

```
## [1] 16.65919
```

Answer:

If the same machine were tested at the same speed, the variance of response will only come from the variance of residuals, thus $SD(Y) = \sqrt{Var(Y)} = \sqrt{\sigma^2_{residuals}} = 11.50226$.

If different machines were sampled from the same manufacturer and tested at the same speed once only, the variance of response will come from both the variance of randaom effects "machine" and the variance of the residuals, thus $SD(Y) = \sqrt{Var(Y)} = \sqrt{\sigma^2_{residuals} + \sigma^2_{machine}} = 16.65919$.

(d)

```
# test fixed effect: interaction term
nmod = lmer(time ~ speed + manufact + (1|machine), lawn, REML=FALSE)
 mod = lmer(time ~ speed * manufact + (1|machine), lawn, REML=FALSE)
( LRT = as.numeric(-2*(logLik(nmod)-logLik(mod))) )
```

```
## [1] 0.09047376
```

```
# chi-square
1 - pchisq(LRT,1)
```

```
## [1] 0.7635757
```

Answer:

The LRT p-value > 0.05, so we do not reject the null model at the 5% significance level. Therefore, the null model without the interaction term is preferred.

So we drop the interaction term and then continue to test the significance of the two main fixed effects terms.

```
# test fixed effect: "manufact"
nmod = lmer(time~ speed + (1|machine), lawn, REML=FALSE)
 mod = lmer(time~ speed+manufact + (1|machine), lawn, REML=FALSE)
( LRT = as.numeric(-2*(logLik(nmod)-logLik(mod))) )
```

```
## [1] 3.799931
```

```
# chi-square
1 - pchisq(LRT,1)
```

```
## [1] 0.05125469
```

```
# test fixed effect: "speed"
nmod = lmer(time~ manufact + (1|machine), lawn, REML=FALSE)
 mod = lmer(time~ speed+manufact + (1|machine), lawn, REML=FALSE)
( LRT = as.numeric(-2*(logLik(nmod)-logLik(mod))) )
```

```
## [1] 44.69564
```

```
# chi-square
1 - pchisq(LRT,1)
```

```
## [1] 2.301692e-11
```

Answer:

In the LRT test for dropping "manufact", the p-value > 0.05, so we do not reject the null model at the 5% significance level.

In the LRT test for dropping "speed", the p-value < 0.05, so we reject the null model at the 5% significance level.

Therefore, the model without the main fixed effects term "manufact" is preferred at this step. So we drop the "manufact" term and then continue to test the significance of the single fixed effects term "speed".

```
# test the only fixed effect: "speed"
nmod = lmer(time~ 1 + (1|machine), lawn, REML=FALSE)
 mod = lmer(time~ speed + (1|machine), lawn, REML=FALSE)
( LRT = as.numeric(-2*(logLik(nmod)-logLik(mod))) )
```

```
## [1] 43.12815
```

```
# chi-square
1 - pchisq(LRT,1)
```

```
## [1] 5.126932e-11
```

Answer:

The LRT p-value < 0.05, so we reject the null model at the 5% significance level. Therefore, "speed" is a significant fixed effect.

So we get the final model with only "speed" left in the model:

$$time_{ijk} = \mu + speed_i + machine_j + \epsilon_{ijk}$$

where μ, $speed_i$ are fixed effects, and $machine_j$ is random effects.
Random effects $machine_j$ i.i.d. $\sim N(0, \sigma_b^2)$, error term ϵ_{ijk} i.i.d. $\sim N(0, \sigma^2)$, and $machine_j$ and ϵ_{ijk} are independent.

(e)

Here we check whether there is any variation between machines based on the final model in question (d).

```
# test the random effect: "machine"
nmod = lm(time ~ speed, lawn)
rmod = lmer(time ~ speed + (1|machine), lawn)
( LRT = as.numeric(-2*(logLik(nmod, REML=TRUE)-logLik(rmod))) )
```

```
## [1] 11.29984
```

```
set.seed(123)
LRT_stat = numeric(1000)
for (i in 1:1000) {
    y = unlist(simulate(nmod))
    nmod_b = lm(time~ speed*manufact, lawn)
    rmod_b = refit(rmod, y)
    LRT_stat[i] = as.numeric(-2*(logLik(nmod_b, REML=TRUE)-logLik(rmod_b)))
}
mean(LRT_stat > LRT)
```

```
## [1] 0
```

Answer:

The LRT p-value < 0.05, so we reject the null hypothesis that the variance between machines is zero at the 5% significance level. Therefore, there is significant variation between machines.

(f)

The mixed effects model is:
$time_{ijkl} = \mu + speed_i + manufact_j + machine_{jk} + \epsilon_{ijkl}$, where $\mu, speed_i$ are fixed effects, and $manufact_j$, $machine_{jk}$ are random effects, where $machine_{jk}$ is nested within $manufact_j$.
Random effects $manufact_j$ i.i.d. $\sim N(0, \sigma^2_{b1})$, $machine_{jk}$ i.i.d. $\sim N(0, \sigma^2_{b2})$, and error term ϵ_{ijkl} i.i.d. $\sim N(0, \sigma^2)$, and $manufact_j$, $machine_{jk}$ and ϵ_{ijkl} are independent.

```
# mixed effect model with nested random effects
rmod = lmer(time~ speed + (1|manufact) + (1|manufact:machine), lawn)
VarCorr(rmod)
```

```
##  Groups           Name        Std.Dev.
##  manufact:machine (Intercept) 12.125
##  manufact         (Intercept) 12.276
##  Residual                     11.187
```

Answer:

The variability between machines and the variability between manufacturers are similar both in their magnitudes and exact values.

(g)

```
# CI of all effects
set.seed(123)
machine = manufact = residual = intercept = speedL = numeric(1000)
for (i in 1:1000) {
    y = unlist(simulate(rmod))
    bmod = refit(rmod, y)
```

```
    machine[i] = as.data.frame(VarCorr(bmod))$sdcor[1]
    manufact[i] = as.data.frame(VarCorr(bmod))$sdcor[2]
    residual[i] = as.data.frame(VarCorr(bmod))$sdcor[3]
    intercept[i] = summary(bmod)$coefficients[1,1]
    speedL[i] = summary(bmod)$coefficients[2,1]
}
quantile(machine, c(0.025, 0.975))
```

```
##     2.5%    97.5%
##  0.00000 21.32433
```

```
quantile(manufact, c(0.025, 0.975))
```

```
##     2.5%    97.5%
##  0.00000 32.51801
```

```
quantile(residual, c(0.025, 0.975))
```

```
##      2.5%     97.5%
##  7.627496 14.287536
```

```
quantile(intercept, c(0.025, 0.975))
```

```
##     2.5%    97.5%
## 239.8934 279.2409
```

```
quantile(speedL, c(0.025, 0.975))
```

```
##      2.5%     97.5%
## -68.49031 -50.14516
```

```
# check result
#confint(rmod, method="boot")
```

Answer:

Though the 95% confidence interval of the random effects "manufact" is wider than that of the nested random effects "machine", they both cover zero values.

Therefore, we might drop any of these two random effect terms but it is not possible to be sure which is best to go. So it is safest to conclude that there is some variation in the cut-off times of lawnmowers coming from both the two sources: machines and manufacturers.

So the variability of response cannot be ascribed solely to manufacturers or to machines.