S632 HW6

Erik Parker

April 12th, 2018

1. The *denim* dataset in the package *faraway* concerns the amount of waste in material cutting

for a jeans manufacturer due to five suppliers.

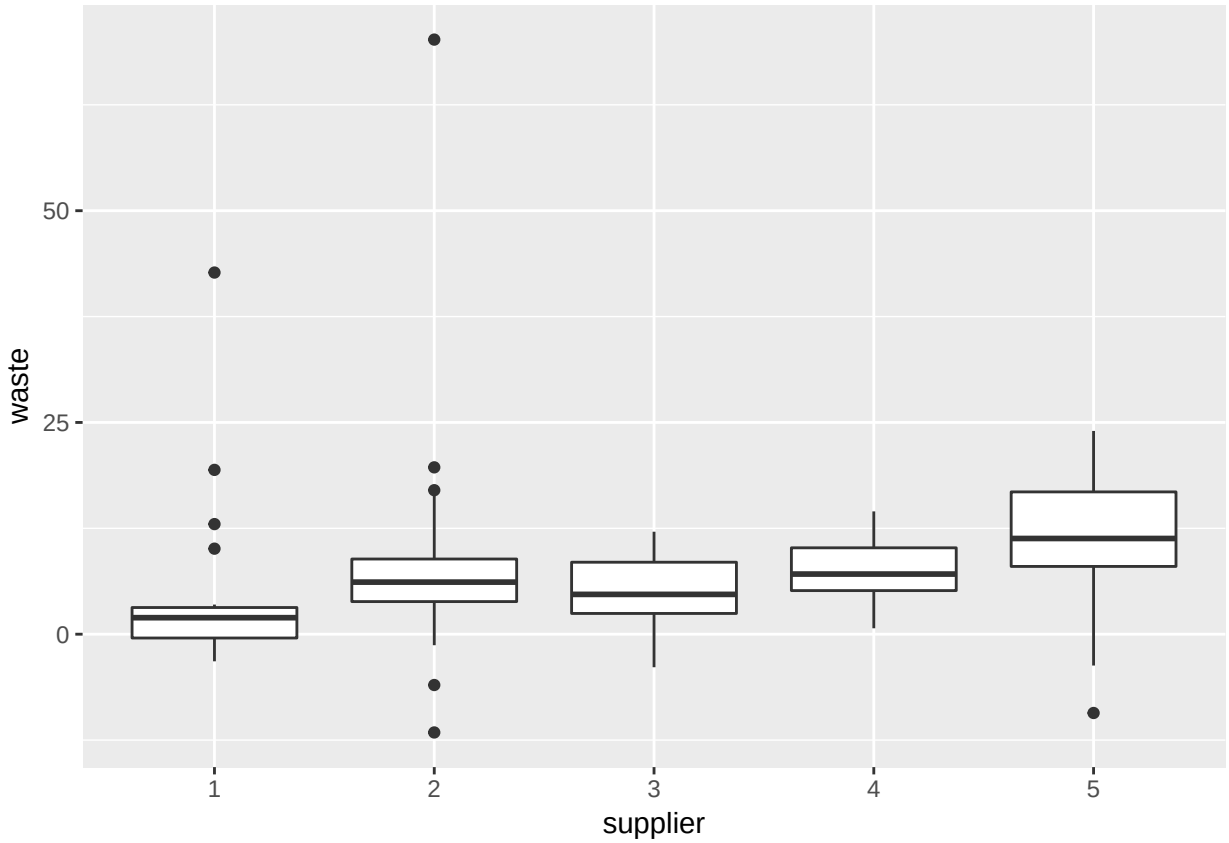
```
rm(list = ls())

library(alr4)
library(ggplot2)
library(faraway)
library(lme4)
library(RLRsim)
library(lattice)


denim <- denim
```

a) Plot the data and comment your results.

```
ggplot(denim, aes(x = supplier, y = waste)) + geom_boxplot()
```



In terms of the median amount of percentage denim waste, we see that there aren't really any huge differences between the suppliers. There are some outlying points, mostly for suppliers 1 and 2, but in general none of the suppliers are particularly wasteful.

b) Fit the linear fixed effects model. Is *supplier* significant?

```
m1 <- lm(waste ~ supplier, data = denim)

Anova(m1)
```

```
## Anova Table (Type II tests)
##
## Response: waste
##           Sum Sq Df F value Pr(>F)
## supplier   450.9  4  1.1596  0.334
## Residuals 8749.1 90
```

Supplier is not significant when we fit a linear fixed effects model.

c) Show the model when supplier is considered a random effect. using the Laird-Ware model (i.e., show what are *X, beta, Z, gamma*, and *epsilon* with the corresponding dimensions).

```
summary(denim)
```

```
##      waste         supplier
##  Min.   :-11.600   1:22
##  1st Qu.:  2.550   2:22
##  Median :  5.200   3:19
##  Mean   :  6.977   4:19
##  3rd Qu.:  9.950   5:13
##  Max.   : 70.200
```

Answer on attached document

d) Using the model with *supplier* as a random effect, is the variance of *supplier* significant? Use two test for this, *LRT* with parametric bootstrapping and any other appropriate test of your choice. In addition, obtain a confidence intervals for the supplier effect standard deviation.

```
mnull <- lm(waste ~ 1, denim)

m2 <- lmer(waste ~ 1 + (1 | supplier), denim, REML = FALSE)

summary(m2)
```

```
## Linear mixed model fit by maximum likelihood  ['lmerMod']
## Formula: waste ~ 1 + (1 | supplier)
##    Data: denim
##
##      AIC      BIC   logLik deviance df.resid
##    710.0    717.7   -352.0    704.0       92
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -1.8877 -0.4498 -0.1806  0.3021  6.4246
##
```

```
## Random effects:
##  Groups   Name        Variance  Std.Dev.
##  supplier (Intercept) 8.263e-17 9.090e-09
##  Residual             9.684e+01 9.841e+00
## Number of obs: 95, groups:  supplier, 5
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)    6.977      1.010    6.91
```

```
lr.1 <- as.numeric(2 * (logLik(m2) - logLik(mnull)))

# Parametric bootstrapping method
y <- simulate(mnull)  #Simulate from the distribution under the null
lrstat <- numeric(1000)
set.seed(142)
for (i in 1:1000) {
    y <- unlist(simulate(mnull))
    bnull <- lm(y ~ 1)
    balt <- lmer(y ~ 1 + (1 | supplier), denim, REML = FALSE)
    lrstat[i] <- as.numeric(2 * (logLik(balt) - logLik(bnull)))
}

phat = mean(lrstat > lr.1)
phat
```

```
## [1] 0.295
```

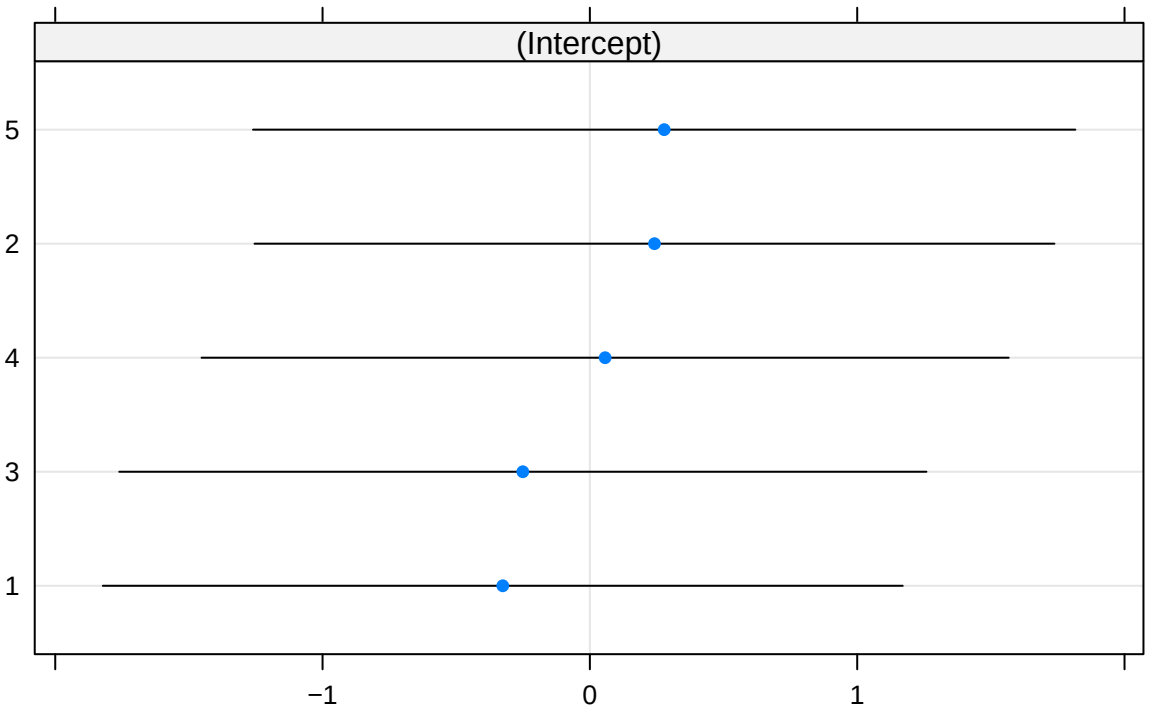
```
m3 <- lmer(waste ~ 1 + (1 | supplier), denim)

exactRLRT(m3)  #use exactRLRT for models obtained with REML
```

```
##
##  simulated finite sample distribution of RLRT.
##
##  (p-value based on 10000 simulated values)
##
## data:
## RLRT = 0.029383, p-value = 0.3444
```

```
dotplot(ranef(m3, condVar = TRUE))
```

```
## $supplier
```



When the parametric bootstrapping method is used, we see a p-value of ~0.313, leading us to the conclusion that we can't reject the null-hypothesis that variance between the different suppliers is zero. So, the variance of *supplier* is not significant according to this test.
And, from the exact test for random effects, we also see a relatively large p-value of ~0.346, leading again to the conclusion that we can't reject our null-hypothesis.

e) Estimate the effect of each supplier. If only one supplier will be used, choose the best.

```
lmod <- aov(waste ~ supplier, denim)

cc <- model.tables(lmod)
cc[[1]]$supplier/ranef(m3)$supplier
```

```
##   (Intercept)
## 1    7.530969
## 2    7.676686
## 3    8.553402
## 4    8.988527
## 5   12.230133
```

```
# Or just this?

ranef(m3)$supplier
```

```
##   (Intercept)
## 1 -0.32586973
## 2  0.24163762
## 3 -0.25080816
## 4  0.05703177
```

```
## 5  0.27800850
```

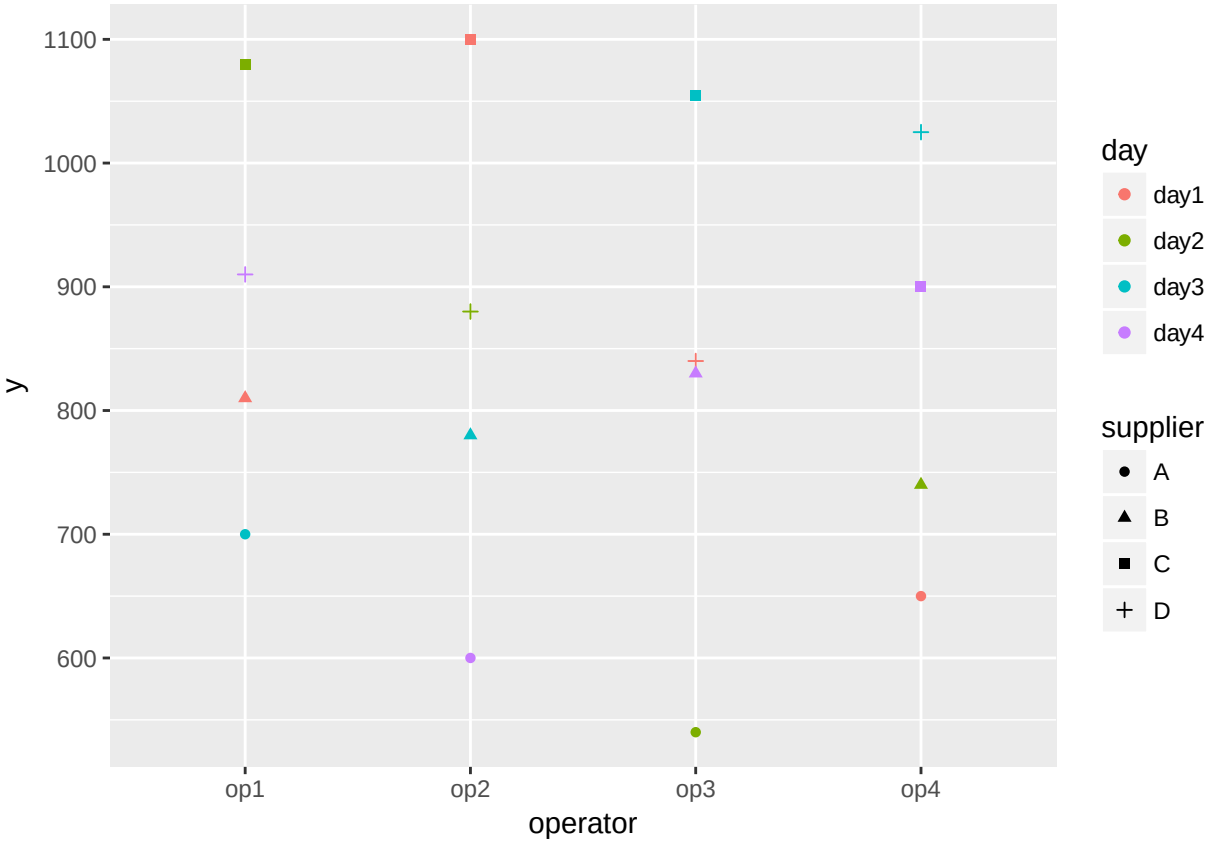
Based on these estimated random effects, the best supplier is number 1 as they have the lowest amount of wastage.

2. An experiment was conducted to select the supplier of raw materials for production of a component. The breaking strength of the component was the objective of interest. Four suppliers were considered. The four operators can only produce one component each per day. A latin square design is used and the data is presented in the *breaking* dataset in the package *faraway*.

a) Run the follow syntax, obtain, and interpret the plot: $ggplot(breaking, aes(y = y, x = operator, color = day, shape = supplier)) + geom_point()$

```
breaking <- breaking

ggplot(breaking, aes(y = y, x = operator, color = day, shape = supplier)) +
    geom_point()
```



This complicated plot shows the difference in breaking strengths of various components, seperated by the operator testing the materials, as well as the day on which the component was produced and the supplier of the raw material. Immediatedly, we can see that supplier A seems to provide the raw material which results in the lowest breaking point of the completed component. Additionally, supplier C seems to generally provide high quality raw materials, with all operators, aside from 4, finding that the components made with raw material C had the highest breaking point.

b) Using the Laird-Ware notation for a mixed effects model with operators and days as random effects but the suppliers as fixed effects.

Answer on attached document

c) Fit a fixed effects model for the main effects. Determine which factors are significant.

```
m1 <- lm(y ~ operator + day + supplier, breaking)

Anova(m1)
```

```
## Anova Table (Type II tests)
##
## Response: y
##           Sum Sq Df F value   Pr(>F)
## operator    7662  3  0.4114 0.750967
## day        17600  3  0.9450 0.475896
## supplier  371138  3 19.9268 0.001602 **
## Residuals  37250  6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From a type II Anova test of the main effects in the *breaking* dataset, we see that *supplier* is the only factor which is significant.

d) Fit a mixed effects model with *operators* and *days* as random effects but *supplier* as a fixed effect. Why is this a natural choice of fixed and random effects? Which supplier results in the highest breaking point?

```
m2 <- lmer(y ~ supplier + (1 | operator) + (1 | day), breaking)

sumary(m2)
```

```
## Fixed Effects:
##             coef.est coef.se
## (Intercept) 622.50    36.09
## supplierB   167.50    49.95
## supplierC   411.25    49.95
## supplierD   291.25    49.95
##
## Random Effects:
##  Groups   Name        Std.Dev.
##  operator (Intercept)  0.00
##  day      (Intercept) 14.80
##  Residual             70.64
## ---
## number of obs: 16, groups: operator, 4; day, 4
## AIC = 156.3, DIC = 213.1
## deviance = 177.7
```

The assignment of *supplier* as the fixed effect, and *operator* and *day* as the random effects is a natural choice as the operator testing the breaking point of some raw material, as well as the day on which that material was tested are quite likely determined at random. The supplier of a sample of raw material is in no way random though, and so should not be treated as such. Furthermore, from the plot in part a, we see that the only real discernable pattern was in terms of the supplier - *day* and *operator* showed no non-random pattern.

From the summary of this model, we see that supplier C makes raw materials with the highest breaking point - as we first noticed from the plot in part a.

e) Test the *operator* and *day* effects

```
m2 <- lmer(y ~ supplier + (1 | operator) + (1 | day), breaking)

m3 <- lmer(y ~ supplier + (1 | operator), breaking)

m4 <- lmer(y ~ supplier + (1 | day), breaking)


exactRLRT(m3, m2, m4)
```

```
##
##  simulated finite sample distribution of RLRT.
##
##  (p-value based on 10000 simulated values)
##
## data:
## RLRT = 0, p-value = 1
```

```
# Testing day

exactRLRT(m4, m2, m3)
```

```
##
##  simulated finite sample distribution of RLRT.
##
##  (p-value based on 10000 simulated values)
##
## data:
## RLRT = 0.030235, p-value = 0.3683
```

```
# Testing operator


mnull <- lm(y ~ supplier, breaking)

m2 <- lmer(y ~ supplier + (1 | operator) + (1 | day), breaking, REML = FALSE)

lr.1 <- as.numeric(2 * (logLik(m2) - logLik(mnull)))

y <- simulate(mnull)  #Simulate from the distribution under the null
lrstat <- numeric(1000)
set.seed(142)
for (i in 1:1000) {
    y <- unlist(simulate(mnull))
    bnull <- lm(y ~ 1)
    balt <- lmer(y ~ supplier + (1 | day) + (1 | operator), breaking, REML = FALSE)
    lrstat[i] <- as.numeric(2 * (logLik(balt) - logLik(bnull)))
}

phat = mean(lrstat > lr.1)
phat
```

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

When we test the two random effects seperately using the exact test, and when we test them together using the parametric bootstrapping method, we see that we get large p-values, meaning that we can't reject the null hypothesis that the variance of both random effects *operator* and *day* is zero.