Chapter 3

Mark Farrell

10/10/2018

Part 1 - Can drought tolerance in sorghum be improved through genetic modification?

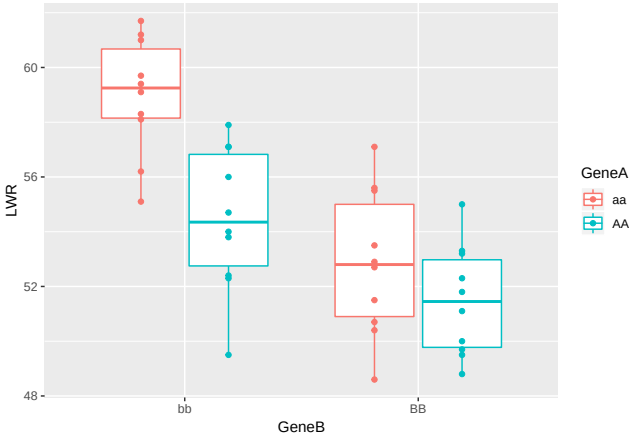
This exercise explores two-factpor ANOVA. First, we need to import the data.

```
setwd("~/DATASCHOOL/r-learning/stats-terry")
data2 <- read_csv("data/working/Prac 3 mock LWR.csv")
```

```
## Parsed with column specification:
## cols(
##   PlantID = col_integer(),
##   GeneA = col_character(),
##   GeneB = col_character(),
##   LWR = col_double()
## )
```

With data imported, a simple plot is produces to both show all the data (dots) and its variance structure (b&w):

```
ggplot(data2, aes(GeneB,LWR,colour=GeneA)) +
  geom_boxplot() +
  geom_point(position=position_dodge(0.75))
```



Now, we can fit a two-factor full factorial model to the data, and test for significance

```
lm1<-lm(LWR~GeneA*GeneB, data = data2)
anova(lm1)
```

```
## Analysis of Variance Table
##
## Response: LWR
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## GeneA        1  86.436  86.436 15.3181 0.0003874 ***
## GeneB        1 208.849 208.849 37.0121 5.372e-07 ***
```

```
## GeneA:GeneB  1  24.336  24.336  4.3128 0.0450232 *
## Residuals   36 203.138   5.643
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As can be seen, both genes AA and BB have a significant impact on leaf water retention (LWR), and there is also a slightly significan interaction effect. To probe this further, we can then output the model summary, and ask emmeans to produce pairwise comparison outputs.

```
summary(lm1)
```

```
##
## Call:
## lm(formula = LWR ~ GeneA * GeneB, data = data2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.980 -1.820  0.085  1.877  4.250
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      58.9800     0.7512  78.516  < 2e-16 ***
## GeneAAA          -4.5000     1.0623  -4.236 0.000151 ***
## GeneBBB          -6.1300     1.0623  -5.770 1.41e-06 ***
## GeneAAA:GeneBBB   3.1200     1.5024   2.077 0.045023 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.375 on 36 degrees of freedom
## Multiple R-squared:  0.6114, Adjusted R-squared:  0.579
## F-statistic: 18.88 on 3 and 36 DF,  p-value: 1.585e-07
```

```
emmeans(lm1, pairwise~GeneA|GeneB)
```

```
## $emmeans
## GeneB = bb:
##  GeneA emmean        SE df lower.CL upper.CL
##  aa     58.98 0.7511806 36 57.45654 60.50346
##  AA     54.48 0.7511806 36 52.95654 56.00346
##
## GeneB = BB:
##  GeneA emmean        SE df lower.CL upper.CL
##  aa     52.85 0.7511806 36 51.32654 54.37346
##  AA     51.47 0.7511806 36 49.94654 52.99346
##
## Confidence level used: 0.95
##
## $contrasts
## GeneB = bb:
##  contrast estimate      SE df t.ratio p.value
##  aa - AA      4.50 1.06233 36   4.236  0.0002
##
## GeneB = BB:
##  contrast estimate      SE df t.ratio p.value
##  aa - AA      1.38 1.06233 36   1.299  0.2022
```

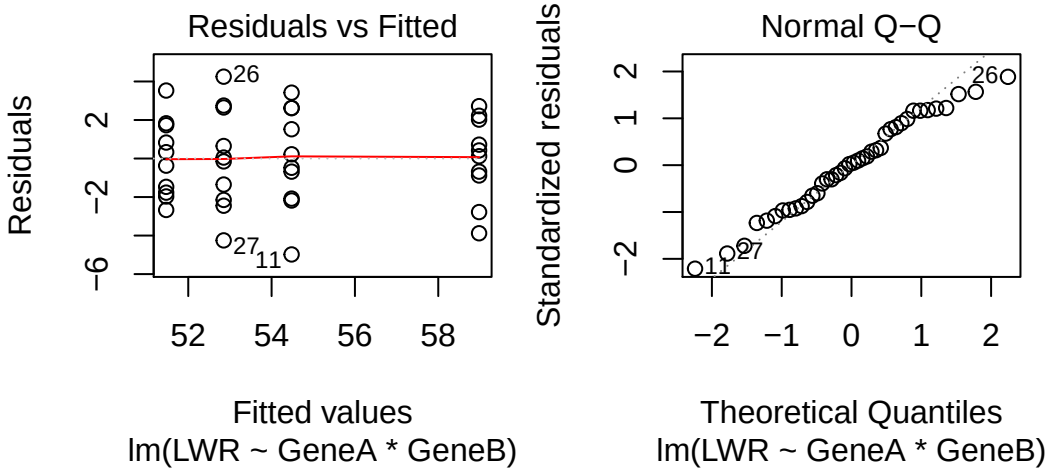
```
emmeans(lm1, pairwise~GeneB|GeneA)
```

```
## $emmeans
## GeneA = aa:
##  GeneB emmean        SE df lower.CL upper.CL
##  bb     58.98 0.7511806 36 57.45654 60.50346
##  BB     52.85 0.7511806 36 51.32654 54.37346
##
## GeneA = AA:
##  GeneB emmean        SE df lower.CL upper.CL
##  bb     54.48 0.7511806 36 52.95654 56.00346
##  BB     51.47 0.7511806 36 49.94654 52.99346
##
## Confidence level used: 0.95
##
## $contrasts
## GeneA = aa:
##  contrast estimate      SE df t.ratio p.value
##  bb - BB      6.13 1.06233 36   5.770  <.0001
##
## GeneA = AA:
##  contrast estimate      SE df t.ratio p.value
##  bb - BB      3.01 1.06233 36   2.833  0.0075
```

This outputs all the multiple comparisons, allowing us to see if bb == or =/= BB when aa or AA are present, and vice versa.

Finally, we need to check the model assumptions using the simple outputs provided by `plot`.

```
plot(lm1, which=1)
plot(lm1, which=2)
```



All looks good :-)

Part 2 Which cabbage cultivar has the higher Vitamin C content on average?

Here, we're exploring model selection. We have two cabbage cultivars (factor 1) planted on three different days (factor 2)

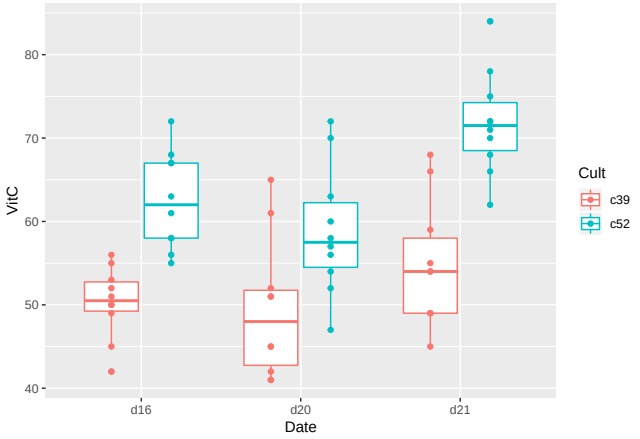
```
setwd("~/DATASCHOOL/r-learning/stats-terry")
cabbage<-read_csv("data/working/Prac 3 cabbage data.csv")
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   X1 = col_integer(),
##   Cult = col_character(),
##   Date = col_character(),
##   HeadWt = col_double(),
##   VitC = col_integer()
## )
```

Plot to see the data:

```
ggplot(cabbage,aes(Date,VitC,colour=Cult))+
  geom_boxplot() +
  geom_point(position=position_dodge(0.75))
```



First, we'll use a full factorial design to confirm no interaction between planting date and cultivar:

```
lm2<-lm(VitC~Cult*Date, data = cabbage)
anova(lm2)
```

```
## Analysis of Variance Table
##
## Response: VitC
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Cult       1 2496.2 2496.15 54.1095 1.089e-09 ***
## Date       2  909.3  454.65  9.8555 0.0002245 ***
## Cult:Date  2  144.3   72.15  1.5640 0.2186275
## Residuals 54 2491.1   46.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As it can be seen that there isn't an interaction, a more appropriate model is an additive model. This is a better test of the question: "Does cultivar affect vitamin C levels?"

```
lm3<-lm(VitC~Cult+Date, data = cabbage)
anova(lm3)
```

```
## Analysis of Variance Table
##
## Response: VitC
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Cult       1 2496.2 2496.15 53.0411 1.179e-09 ***
## Date       2  909.3  454.65  9.6609 0.0002486 ***
## Residuals 56 2635.4   47.06
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
emmeans(lm3, pairwise~Cult)
```

```
## $emmeans
##  Cult emmean       SE df lower.CL upper.CL
##  c39    51.5 1.252474 56 48.99099 54.00901
##  c52    64.4 1.252474 56 61.89099 66.90901
##
## Results are averaged over the levels of: Date
## Confidence level used: 0.95
##
## $contrasts
##  contrast  estimate       SE df t.ratio p.value
##  c39 - c52    -12.9 1.771265 56  -7.283  <.0001
##
## Results are averaged over the levels of: Date
```

It's clear that c39 has lower vitamin C levels than c52, by 12.9 +/- 1.8 units.

```
plot(lm3, which=1)
plot(lm3, which=2)
```