**R code and model output for egg counter data from egg laying assay**

This document presents the code and output from the first attempted block of the female fecundity assay which ran from Tuesday 29th to 30th of May 2018. This is part of the experiments focussing on the M lines 5, 8, 16, 19, 24, and 95 who have had either male or female benefit fruitless alleles introgressed into a deletion stock background. In this assay, these six lines have been crossed with a deletion/balancer stock to produce offspring with two different genotypes per line, either introgressed/balancer (B) or introgressed/deletion. This assay assessed the fecundity of females by measuring their egg output over an 18hour period from ~3pm Tuesday to 9am on Wednesday. Females had been collected shortly after eclosion and stored in normal food vials with males from the same line. After 1 day three females were placed into an agar vial along with 3 males to ensure that the females were not virgins. These flies were kept in these vials for several days. Agar vials were fed through a capillary tube containing yeast solution and flies were observed feeding on this tube. Flies were collected on two successive days, 24th and 25th of May, meaning that all females were at least 4 or 5 days only by the start of the assay to ensure adequate number of replicate vials and to minimise the differences between flies of different ages. All flies within the same vial emerged on the same day. On Tuesday flies were transferred to a new agar vial. This vial was made with 0.8% agar that morning and was carefully made to avoid bubbles or other features. Females were then free to lay eggs on the agar for 18 hours. These vials were stored at 25C and 76% humidity. After 18 hours, all flies were then removed. The vials were then photographed using the equipment on the third floor. While waiting to be photographed, vials were kept in the cold room to prevent eggs from hatching. Once all vials were photographed, the pictures were transferred to the computer for analysis. The egg counter program uses a machine learning algorithm to count eggs from an array of pictures. After training the program to recognise an egg, this program was used to count the eggs across all the pictures taken. This produced an output of the number of eggs laid per vial by 3 females over 18 hours.

setwd("/Users/michaeljardine/Desktop/DTP/Datasets/Eggs")

eggs <- read.csv("blockone300518.csv")

### convert line to a factor with 6 levels

eggs$line <- factor(eggs$line)

str(eggs)

> str(eggs)

'data.frame': 197 obs. of 9 variables:

$ Frame.number. : int 1 1 1 1 1 1 1 1 1 1 ...

$ Predicted.count.: num 13 27.4 19.2 16.9 13 ...

$ Corrected.count.: num 13 27.3 19.2 16.9 13 ...

$ CI : num 0.43 0.433 0.431 0.431 0.43 ...

$ photo.number : int 5 6 7 8 9 10 11 12 13 14 ...

$ line : Factor w/ 6 levels "5","8","16","19",..: 1 1 1 1 1 1 1 2 2 2 ...

$ genotype : Factor w/ 2 levels "B","D": 1 1 1 1 1 1 1 1 1 1 ...

$ allele : Factor w/ 2 levels "F","M": 2 2 2 2 2 2 2 2 2 2 ...

$ age : Factor w/ 2 levels "five","four": 1 1 1 1 1 1 1 1 1 1 ...

### all good

### we have two numbers produced for the number eggs, predicted and corrected

x = eggs$Predicted.count.

y = eggs$Corrected.count.

cor(x, y)

[1] 1

### since these correlate exactly, it dosen't matter which one we use

### I will use the predicted count for consistency

> ### summary statistics for reference, relating to the plots created in JMP

> summarySE(eggs, measurevar=c("Predicted.count."), groupvars=c("allele"))

allele N Predicted.count. sd se ci

1 F 78 20.14823 9.254343 1.0478480 2.086532

2 M 119 18.08451 8.442731 0.7739439 1.532620

> summarySE(eggs, measurevar=c("Predicted.count."), groupvars=c("line"))

line N Predicted.count. sd se ci

1 5 39 18.70973 9.491612 1.5198743 3.076825

2 8 51 20.52653 7.913421 1.1081006 2.225686

3 16 29 12.94910 5.251748 0.9752251 1.997658

4 19 30 19.98574 11.238318 2.0518267 4.196457

5 24 38 20.14622 8.262545 1.3403618 2.715831

6 95 10 20.64335 6.688180 2.1149884 4.784436

> summarySE(eggs, measurevar=c("Predicted.count."), groupvars=c("genotype"))

genotype N Predicted.count. sd se ci

1 B 97 18.38554 8.216324 0.8342413 1.655956

2 D 100 19.40221 9.360627 0.9360627 1.857352

> summarySE(eggs, measurevar=c("Predicted.count."), groupvars=c("age"))

age N Predicted.count. sd se ci

1 five 107 19.70216 8.636208 0.8348938 1.655258

2 four 90 17.94985 8.963196 0.9448038 1.877305

### Models

### we want to look att eh differenecs in the number of eggs liad between the different lines, genotypes and alleles, while also controlling for variatrion due to age.

**#### Full model ####**

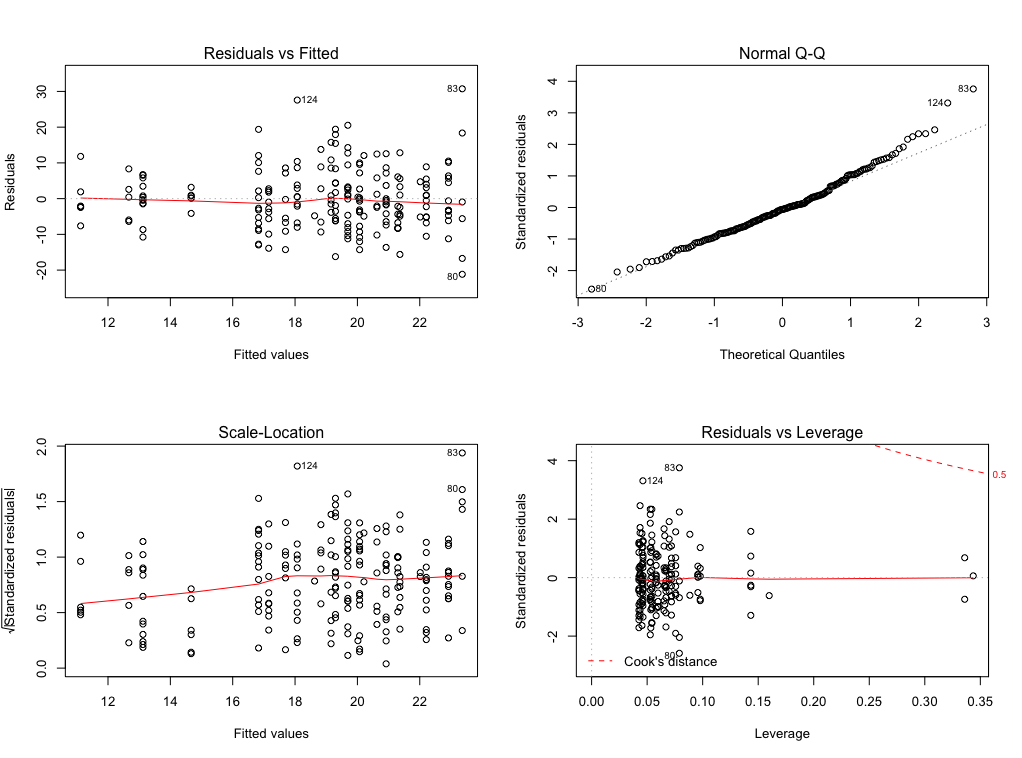
# we'll first do a model including all factors and an interaction between line and genotype

# this should work within a simple linear model

> full <- lm(Predicted.count. ~ allele + line\*genotype + age, data=eggs)

> par(mfrow=c(2,2))

> plot(full)



> ### looks reasonably good, slight devaition within the Q-Q plot at the right-hand end

> summary(full)

Call: lm(formula = Predicted.count. ~ allele + line \* genotype + age, data = eggs)

Residuals:

Min 1Q Median 3Q Max

-21.1698 -5.5588 -0.5009 4.4276 30.7366

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 22.01643 4.94100 4.456 1.45e-05 \*\*\*

alleleM -0.72012 5.29425 -0.136 0.89196

line8 -1.22601 2.55266 -0.480 0.63159

line16 -8.17511 2.80012 -2.920 0.00394 \*\*

line19 -2.86135 5.37188 -0.533 0.59492

line24 0.19286 5.42942 0.036 0.97170

line95 NA NA NA NA

genotypeD -2.46723 2.73346 -0.903 0.36792

agefour -1.99998 1.29841 -1.540 0.12520

line8:genotypeD 5.31792 3.62791 1.466 0.14440

line16:genotypeD 4.01348 4.25510 0.943 0.34681

line19:genotypeD 6.67575 4.14766 1.610 0.10922

line24:genotypeD -0.04931 3.96045 -0.012 0.99008

line95:genotypeD 1.07711 6.49596 0.166 0.86849

Residual standard error: 8.525 on 184 degrees of freedom

Multiple R-squared: 0.1206, Adjusted R-squared: 0.06321

F-statistic: 2.102 on 12 and 184 DF, p-value: 0.01868

> anova(full)

Analysis of Variance Table

Response: Predicted.count.

Df Sum Sq Mean Sq F value Pr(>F)

allele 1 200.7 200.669 2.7610 0.098288 .

line 4 1087.4 271.857 3.7405 0.005957 \*\*

genotype 1 9.5 9.524 0.1310 0.717767

age 1 182.1 182.074 2.5052 0.115188

line:genotype 5 353.7 70.732 0.9732 0.435493

Residuals 184 13372.8 72.678

### so it appears that there is a variation in egg number due to line.

### however in the breakdown of the lines this may only be due to the very low numbers from M16

### no differences due to other factors, although allele is close

**#### remove genotype from the model ####**

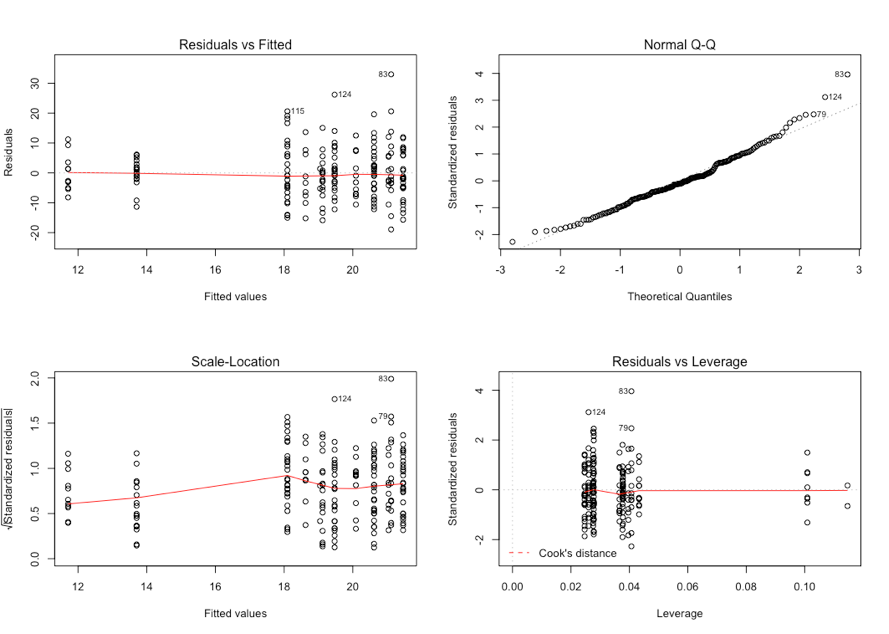
### since genotype explained the least in the full model I'll run the model again from the next one

### this also removes the interaction between line and genotype but this explained the second least amount of variation.

> minusgenotype <- lm(Predicted.count. ~ allele + line + age, data=eggs)

> par(mfrow=c(2,2))

> plot(minusgenotype)



> summary(minusgenotype)

Call: lm(formula = Predicted.count. ~ allele + line + age, data = eggs)

Residuals:

Min 1Q Median 3Q Max

-18.922 -5.161 -0.873 5.567 32.984

Coefficients: (1 not defined because of singularities)

Estimate Std. Error t value Pr(>|t|)

(Intercept) 21.04225 2.70187 7.788 4.29e-13 \*\*\*

alleleM -0.95171 3.08060 -0.309 0.758

line8 1.37458 1.83157 0.750 0.454

line16 -6.38491 2.12404 -3.006 0.003 \*\*

line19 0.07371 3.14131 0.023 0.981

line24 -0.42365 3.02320 -0.140 0.889

line95 NA NA NA NA

agefour -1.99451 1.28695 -1.550 0.123

Residual standard error: 8.505 on 190 degrees of freedom

Multiple R-squared: 0.09613, Adjusted R-squared: 0.06759

F-statistic: 3.368 on 6 and 190 DF, p-value: 0.003533

> anova(minusgenotype)

Analysis of Variance Table

Response: Predicted.count.

Df Sum Sq Mean Sq F value Pr(>F)

allele 1 200.7 200.669 2.7740 0.09745 .

line 4 1087.4 271.857 3.7581 0.00575 \*\*

age 1 173.7 173.748 2.4019 0.12285

Residuals 190 13744.3 72.339

### the minus genotype model has a lower AIC score (not by much) largly due to dropping 6 extra degrees of freedom

#### But really line is not of specific interest as the two triplets of lines are replicates for the fruitless allele intrgressed ####

### nowa new model will be made that will use line as a random factor to look at variation due to the two alleles

### this will also include genotype, both as a factors and also an inclusion in as econd model as part of the random variaance component

**### 1st mixed model - random intercept for line**

> ranline <- lmer(Predicted.count. ~ allele\*genotype + (1 | line), data=eggs)

> summary(ranline)

Linear mixed model fit by REML ['lmerMod']

Formula: Predicted.count. ~ allele \* genotype + (1 | line)

Data: eggs

REML criterion at convergence: 1397.4

Scaled residuals:

Min 1Q Median 3Q Max

-2.0977 -0.5990 -0.0854 0.6413 3.9574

Random effects:

Groups Name Variance Std.Dev.

line (Intercept) 6.076 2.465

Residual 73.485 8.572

Number of obs: 197, groups: line, 6

Fixed effects:

Estimate Std. Error t value

(Intercept) 20.0537 2.1251 9.437

alleleM -2.8150 2.7751 -1.014

genotypeD 0.2558 1.9865 0.129

alleleM:genotypeD 0.4715 2.5390 0.186

Correlation of Fixed Effects:

(Intr) allelM gntypD

alleleM -0.766

genotypeD -0.548 0.420

alllM:gntyD 0.429 -0.490 -0.782

> anova(ranline)

Analysis of Variance Table

Df Sum Sq Mean Sq F value

allele 1 87.526 87.526 1.1911

genotype 1 14.228 14.228 0.1936

allele:genotype 1 2.535 2.535 0.0345

> wald.test(b=fixef(ranline), Sigma=vcov(ranline), Terms = 2, df=1)

Wald test:

Chi-squared test: X2 = 1.0, df = 1, P(> X2) = 0.31

F test: W = 1.0, df1 = 1, df2 = 1, P(> W) = 0.5

> wald.test(b=fixef(ranline), Sigma=vcov(ranline), Terms = 3, df=1)

Wald test:

Chi-squared test: X2 = 0.017, df = 1, P(> X2) = 0.9

F test: W = 0.017, df1 = 1, df2 = 1, P(> W) = 0.92

> wald.test(b=fixef(ranline), Sigma=vcov(ranline), Terms = 4, df=1)

Wald test:

Chi-squared test: X2 = 0.034, df = 1, P(> X2) = 0.85

F test: W = 0.034, df1 = 1, df2 = 1, P(> W) = 0.88

### and now model with random genotype component as well

ranlinegeno <- lmer(Predicted.count. ~ allele\*genotype + (1 + genotype | line), data=eggs)

> summary(ranlinegeno)

Linear mixed model fit by REML ['lmerMod']

Formula: Predicted.count. ~ allele \* genotype + (1 + genotype | line)

Data: eggs

REML criterion at convergence: 1397.1

Scaled residuals:

Min 1Q Median 3Q Max

-2.1741 -0.6280 -0.0815 0.5838 3.9069

Random effects:

Groups Name Variance Std.Dev. Corr

line (Intercept) 7.632 2.763

genotypeD 3.989 1.997 -0.48

Residual 72.861 8.536

Number of obs: 197, groups: line, 6

Fixed effects:

Estimate Std. Error t value

(Intercept) 20.1101 2.2690 8.863

alleleM -2.8462 2.9745 -0.957

genotypeD 0.2803 2.3443 0.120

alleleM:genotypeD 0.4387 3.0583 0.143

Correlation of Fixed Effects:

(Intr) allelM gntypD

alleleM -0.763

genotypeD -0.631 0.481

alllM:gntyD 0.483 -0.591 -0.767

> anova(ranlinegeno)

Analysis of Variance Table

Df Sum Sq Mean Sq F value

allele 1 88.500 88.500 1.2147

genotype 1 9.308 9.308 0.1277

allele:genotype 1 1.499 1.499 0.0206

> wald.test(b=fixef(ranlinegeno), Sigma=vcov(ranlinegeno), Terms = 2, df=1)

Wald test:

Chi-squared test: X2 = 0.92, df = 1, P(> X2) = 0.34

F test: W = 0.92, df1 = 1, df2 = 1, P(> W) = 0.51

> wald.test(b=fixef(ranlinegeno), Sigma=vcov(ranlinegeno), Terms = 3, df=1)

Wald test:

Chi-squared test: X2 = 0.014, df = 1, P(> X2) = 0.9

F test: W = 0.014, df1 = 1, df2 = 1, P(> W) = 0.92

> wald.test(b=fixef(ranlinegeno), Sigma=vcov(ranlinegeno), Terms = 4, df=1)

Wald test:

Chi-squared test: X2 = 0.021, df = 1, P(> X2) = 0.89

F test: W = 0.021, df1 = 1, df2 = 1, P(> W) = 0.91