Lecture 7: Multiple factors in Statistical Models

An introduction

Terry Neeman

Australian National University

23th August 2022

A familiar scenario

- Gene Knockout experiment;
- Two gene candidates: Gene A and Gene B
- How do these genes (separately or together) impact on drought sensitivity in *Arabidopsis thaliana*?

What is the statistical framework for this biological experiment?

A biologist's thought process

- Three types of mutant plants

  * Gene A knockout
  * Gene B knockout
  * double KO

- Measure leaf water retention (LWR) after 10 days of no watering
- Compare LWR: each mutant genotype with wildtype

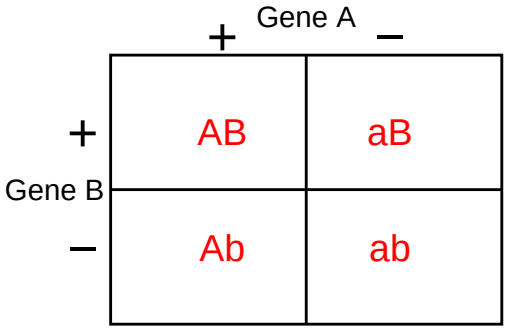
Can we get more information from this biological experiment?

For example, what about the combined effect of A and B? Is it additive, synergistic or antagonistic?

A statistical thought process

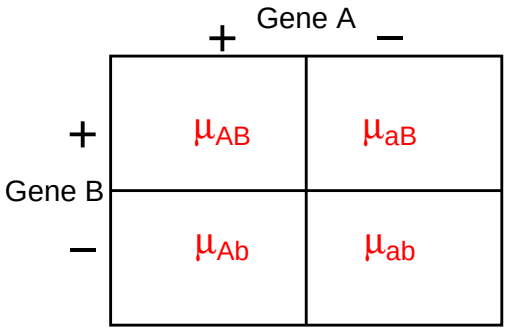
- What is the effect of knocking out A in the presence/absence of B?
- What is the effect of knocking out B in the presence/absence of A?
- What is the statistical evidence that A & B act in synergy?
- What is the "main effect" of knocking out A (B)?

Thinking statistically: as a 2x2 matrix



Experiment with 2 factors. Each factor has 2 levels = 2x2 matrix

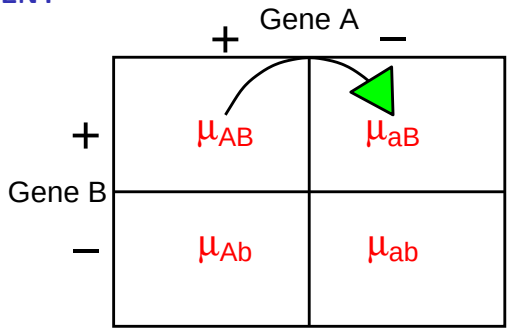
Thinking statistically: a model with (up to) 4 parameters



We are interested in estimating differences between parameters. How can we create a model to test these differences?

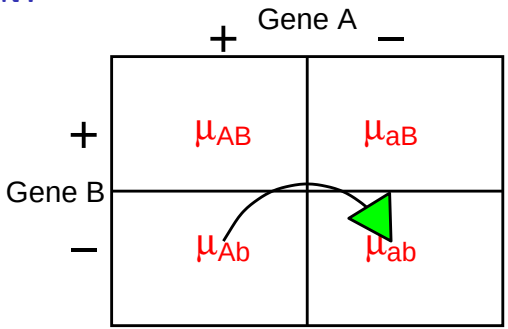
The effect of Gene A knockout on LWR

when Gene B is PRESENT



The effect of Gene A knockout on LWR

when Gene B is ABSENT



The Gene A KO effect is measured under two background conditions: in the presence/absence of B

MAIN EFFECT: the average of the Gene A KO effect across the two backgrounds

Two Scenarios

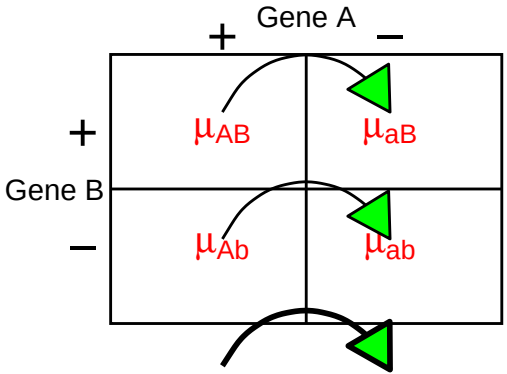
(1) The Gene A KO effect IS INDEPENDENT of the background condition

(2) The Gene A KO effect DEPENDS UPON the background condition

SCENARIO 1

The Gene A KO effect is independent of the background condition

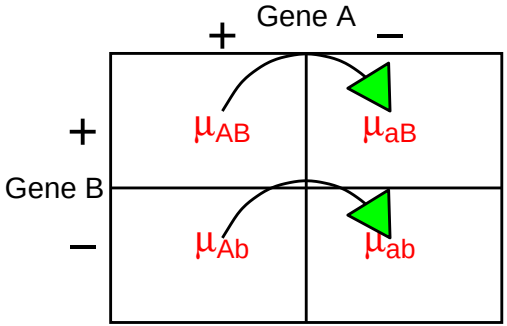
- Combine effect estimates. MORE INFORMATION ABOUT EFFECTS!



SCENARIO 2

The Gene A KO effect depends upon the background condition

- Report on how B impacts on the Gene A KO effect



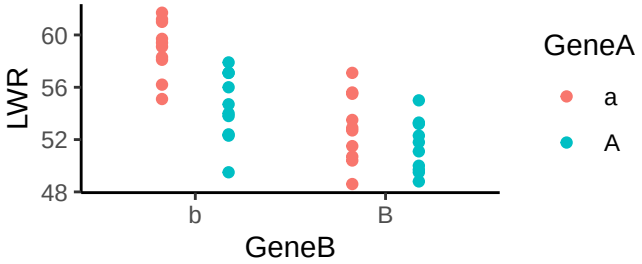
Worked example

```
library(tidyverse)
library(emmeans)
lwr<- read_csv("../Data/mock LWR.csv")
str(lwr)
```

```
## spec_tbl_df [40 x 4] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ PlantID: num [1:40] 1 2 3 4 5 6 7 8 9 10 ...
##  $ GeneA  : chr [1:40] "A" "A" "A" "A" ...
##  $ GeneB  : chr [1:40] "B" "B" "B" "B" ...
##  $ LWR    : num [1:40] 55 51.1 51.8 53.2 49.5 49.7 50 52.3 53.3 48.8 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   PlantID = col_double(),
##   ..   GeneA = col_character(),
##   ..   GeneB = col_character(),
##   ..   LWR = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Data exploration

```
ggplot(lwr, aes(x=GeneB, LWR, colour=GeneA))+
   geom_point(position=position_dodge(width=0.7))+
   theme_classic()
```



What patterns do you notice?

Fit a model to the data

```
lwr$GeneA<- factor(lwr$GeneA, levels=c("A","a")); lwr$GeneB<- factor(lwr$GeneB, le
lm.lwr<-lm(LWR~GeneA*GeneB, data=lwr)
anova(lm.lwr)
```

```
## 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
```

Strong evidence of a Gene A KO effect (averaged over Gene B status) Strong evidence of a Gene B KO effect (averaged over Gene A status) Some evidence of a Gene A: Gene B

Interpreting the model: Model parameter estimates

```
sum_lm.lwr<-summary(lm.lwr)
sum_lm.lwr$coefficients
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)       51.47  0.7511806 68.518813 9.703333e-40
## GeneAa             1.38  1.0623297  1.299032 2.021918e-01
## GeneBb             3.01  1.0623297  2.833395 7.501029e-03
## GeneAa:GeneBb      3.12  1.5023611  2.076731 4.502320e-02
```

Interpreting the model: mean and effect estimates

```
emmeans(lm.lwr, ~GeneA|GeneB)
```

```
## GeneB = B:
##  GeneA emmean    SE df lower.CL upper.CL
##  A       51.5 0.751 36     49.9     53.0
##  a       52.9 0.751 36     51.3     54.4
##
## GeneB = b:
##  GeneA emmean    SE df lower.CL upper.CL
##  A       54.5 0.751 36     53.0     56.0
##  a       59.0 0.751 36     57.5     60.5
##
## Confidence level used: 0.95
```

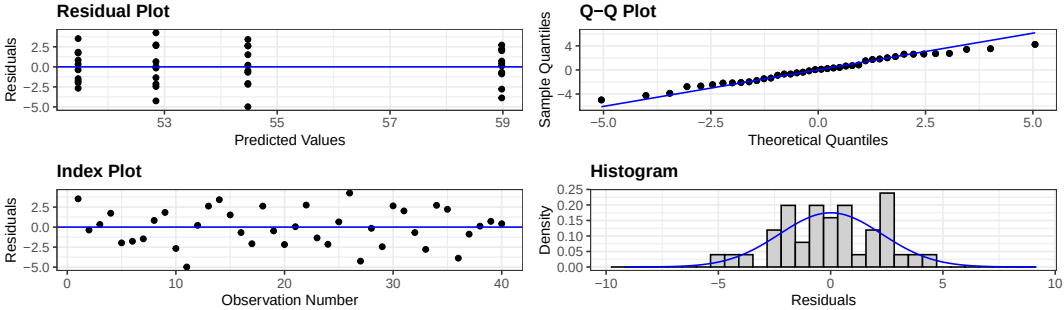
Interpreting the model: mean and effect estimates

```
pairs(emmeans(lm.lwr, ~GeneA|GeneB))
```

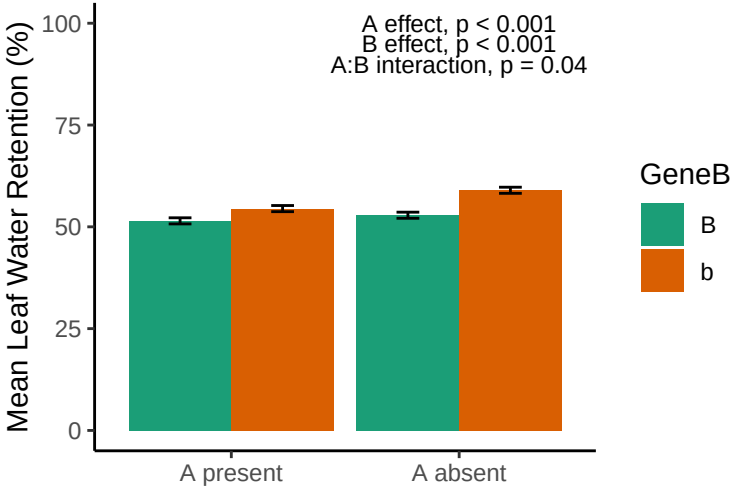
```
## GeneB = B:
##  contrast estimate   SE df t.ratio p.value
##  A - a       -1.38 1.06 36  -1.299  0.2022
##
## GeneB = b:
##  contrast estimate   SE df t.ratio p.value
##  A - a       -4.50 1.06 36  -4.236  0.0002
```

Model assessment

```
library(ggResidpanel)
resid_panel(lm.lwr)
```



Model summary



Model summary code

```
ggplot(results1, aes(x=GeneA, emmean, fill=GeneB))+

geom_bar(stat = "identity", position = "dodge")+

geom_errorbar(aes(ymin=emmean-SE, ymax=emmean+SE), width=.2,
position=position_dodge(width=0.9))+

annotate("text", x=2, y=100, label= "A effect, p < 0.001", size = 3)+

annotate("text", x=2, y=95, label= "B effect, p < 0.001", size = 3)+

annotate("text", x=2, y=90, label = "A:B interaction, p = 0.04", size = 3)+

ylab("Mean Leaf Water Retention (%)")+ xlab(" ")+

scale_fill_brewer(palette="Dark2")+

scale_x_discrete(labels = c("A present", "A absent"))+

theme_classic()
```

Summary: multifactorial experiments

- Multiple factors: increasing complexity of analysis
- But experimenting 1 factor at a time:

  + too much time!
  + inefficient

- Can we generalise to more than 2 factors?