

General(ized) linear models

Groups as predictors, then yes/no and count responses



Week 2 recap

Kahoot-style warm-up

12 quick questions on linear regression.

Not graded.

Use a pseudonym if you want.

Fast and correct wins.

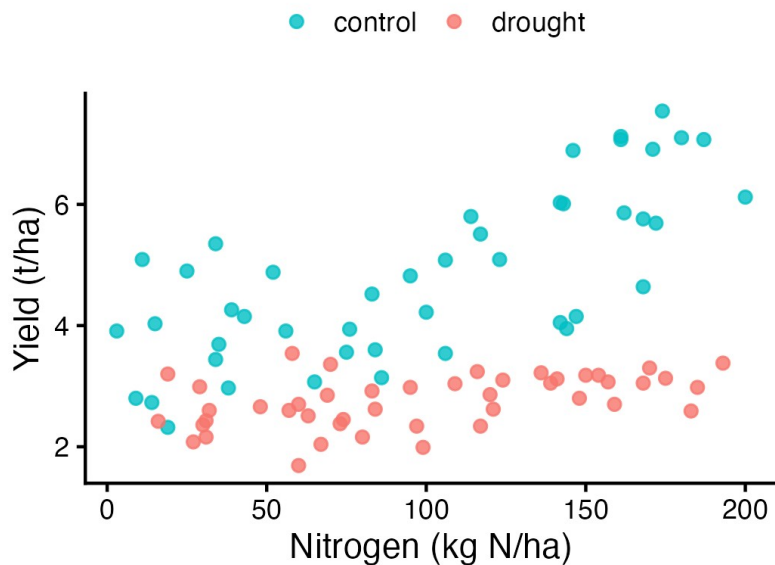
[Join on your phone or laptop: greymonroe.github.io/live-poll-test/quiz-play.html?quiz=pls206-w3-week2-recap](https://greymonroe.github.io/live-poll-test/quiz-play.html?quiz=pls206-w3-week2-recap)



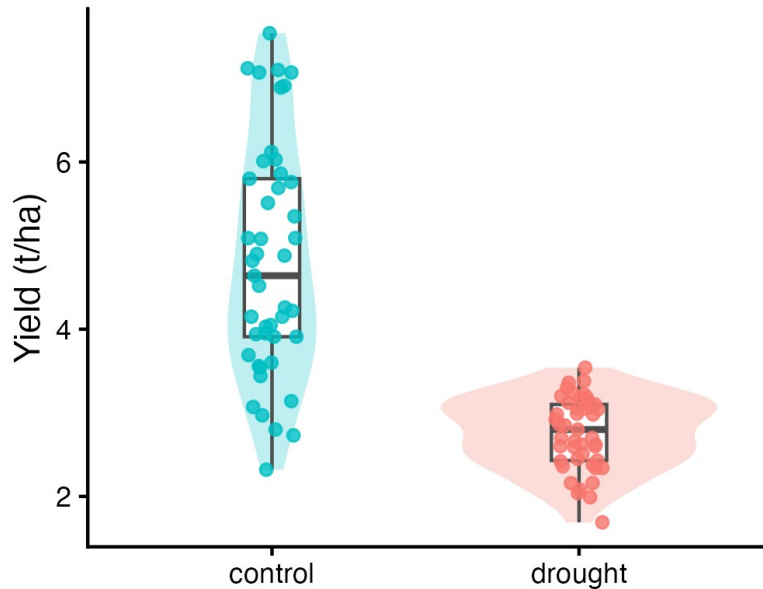
Scan with your phone

Where we are

Last week: x is a number



This week: x is a group



Same machine, **lm()**. Then: what if y is a yes/no, or a count?

Goals

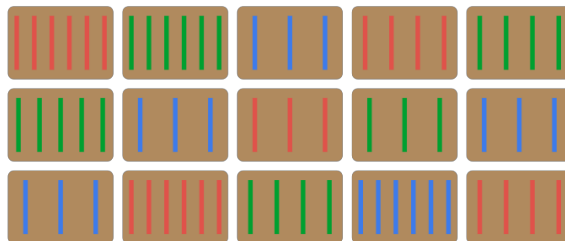
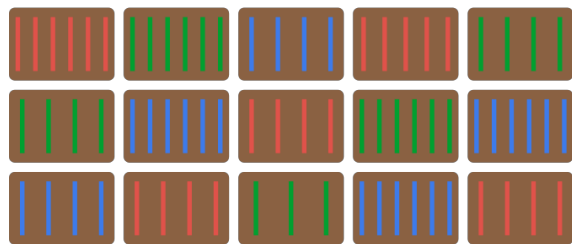
- Groups as predictors: dummy variables
- t-test, ANOVA, ANCOVA: all `lm()`
- Two factors and their interaction
- Yes/no responses: logistic regression
- The GLM family: one recipe, many kinds of y
- Counts: Poisson regression

Part 1: Groups as predictors

Dummy variables



The experiment



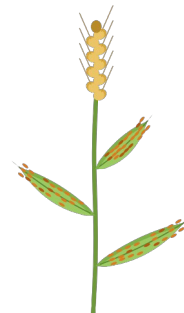
| Amber **| Bronze** **| Copper**



0-200 kg N/ha, varies by plot



yield (t/ha)



rust: yes / no



aphids: a
count

90 plots, three things measured on each. (Hypothetical, simulated data.)

A wheat field trial

```
> trial <- read.csv("https://greymonroe.github.io/PLS_206/data/wheat_trial.csv")
```

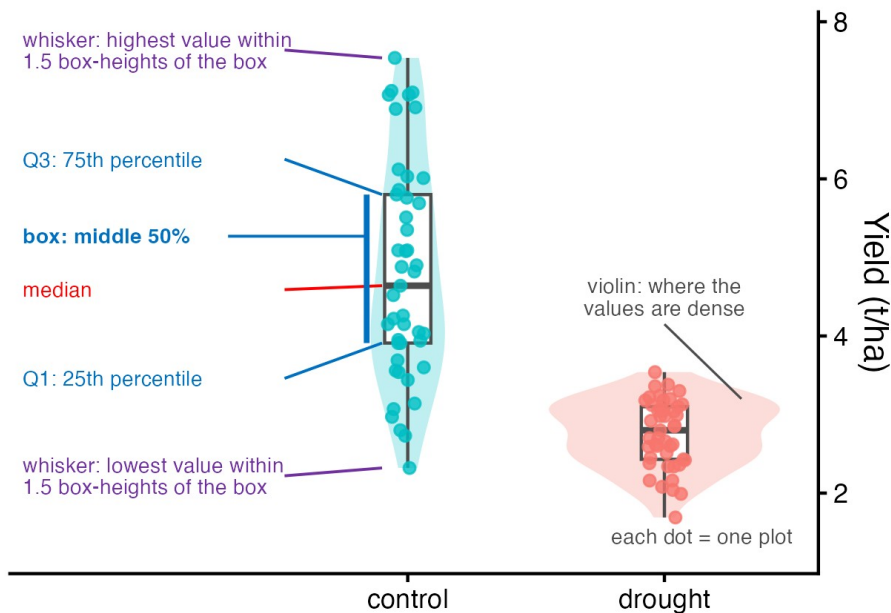
```
> head(trial)
```

	plot	variety	water	nitrogen	yield	rust	aphids
1	1	Copper	control	19	2.32	0	0
2	2	Amber	control	34	3.44	0	8
3	3	Bronze	control	180	7.10	1	12
4	4	Bronze	drought	63	2.51	0	1
5	5	Bronze	control	117	5.51	1	30
6	6	Copper	control	14	2.73	0	4

```
> table(trial$variety, trial$water)
```

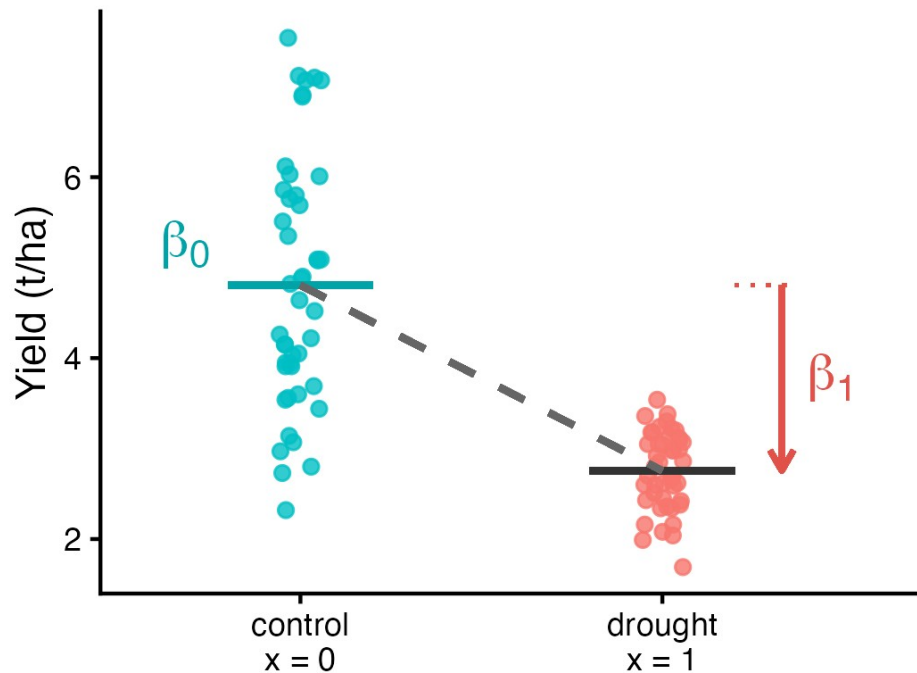
	control	drought
Amber	15	15
Bronze	15	15
Copper	15	15

← 15 plots per cell



90 plots: 3 varieties × 2 water treatments, nitrogen varies by plot

Two groups, one line



$$\text{yield}_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad x_i = \begin{cases} 1 & \text{drought} \\ 0 & \text{control} \end{cases}$$

Code the groups 0 and 1.

Intercept β_0 = mean of control

Slope β_1 = drought - control

lm() with a group

```
> m_water <- lm(yield ~ water, data = trial)
> round(coef(summary(m_water)), 3)
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4.806	0.152	31.523	0	← control mean
waterdrought	-2.051	0.216	-9.513	0	← drought minus control

```
> tapply(trial$yield, trial$water, mean)
control drought
4.806222 2.755111 ← 4.81 - 2.76 = 2.05
```

control: 4.81 drought: $4.81 - 2.05 = 2.76$ t/ha

waterdrought: drought minus control, in t/ha

Live poll

Number: live histogram

Tomatoes, two fertilizers:
 $\text{lm}(\text{fruit_g} \sim \text{fertilizer})$

Intercept = 142

fertilizerorganic = -18

Mean fruit weight (g) with organic
fertilizer?



Scan with your phone

[Join on your phone or laptop: greymonroe.github.io/live-poll-test/submit.html?poll=pls206-w3-tomato](https://greymonroe.github.io/live-poll-test/submit.html?poll=pls206-w3-tomato)

[host screen](#)

It's a t-test

```
> t.test(yield ~ water, data = trial, var.equal = TRUE)
```

Two Sample t-test

data: yield by water

t = 9.5125, df = 88, p-value = 3.613e-15 ← same t as lm()

alternative hypothesis: true difference in means between group
control and group drought is not equal to 0

95 percent confidence interval:

1.622607 2.479615

sample estimates:

mean in group control mean in group drought

4.806222

2.755111

**Same t,
same p,
same difference.**

var.equal = TRUE: lm() assumes
equal variances

Under the hood: a 0/1 column

```
> head(trial$water)
[1] "control" "control" "control" "drought" "control" "control"
> head(model.matrix(m_water))
(Intercept) waterdrought ← 1 = drought, 0 = control
1           1           0
2           1           0
3           1           0
4           1           1
5           1           0
6           1           0
```

R builds the **dummy variable** for you: 1 if drought, 0 if not

Which group is the intercept?

```
> trial2 <- trial
> trial2$water <- factor(trial2$water,
+                         levels = c("drought", "control"))
> coef(lm(yield ~ water, data = trial2))
(Intercept) watercontrol
  2.755111      2.051111  ← now: drought mean
```

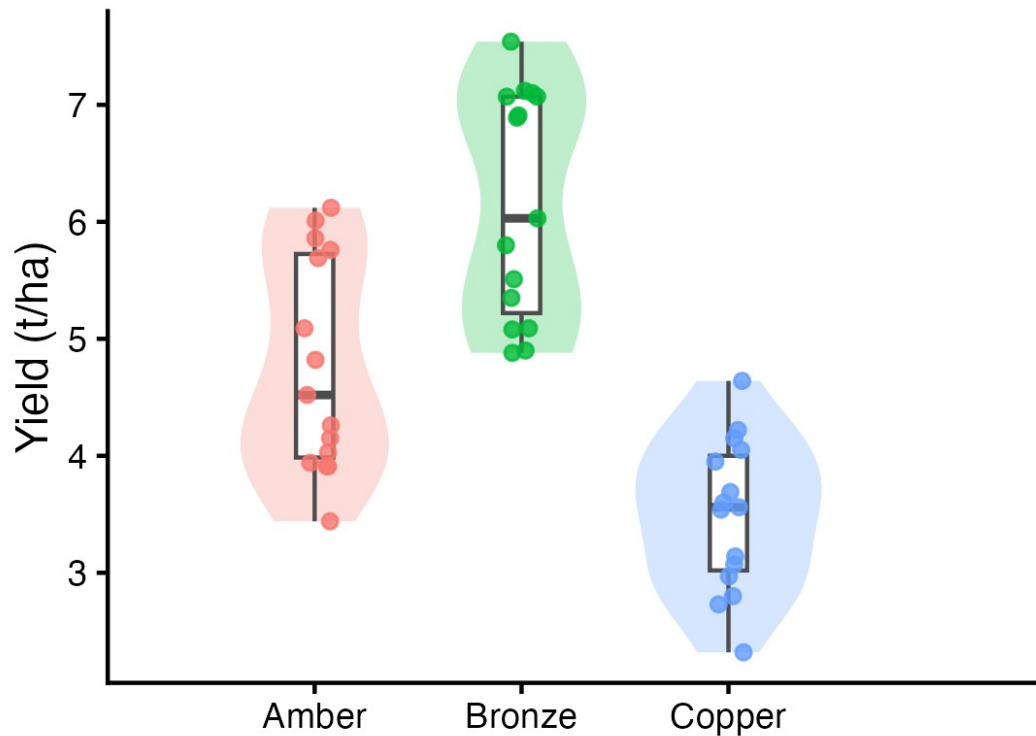
Alphabetical by default. `factor(levels = )` picks the reference.

Part 2: More than two groups

ANOVA is a linear model



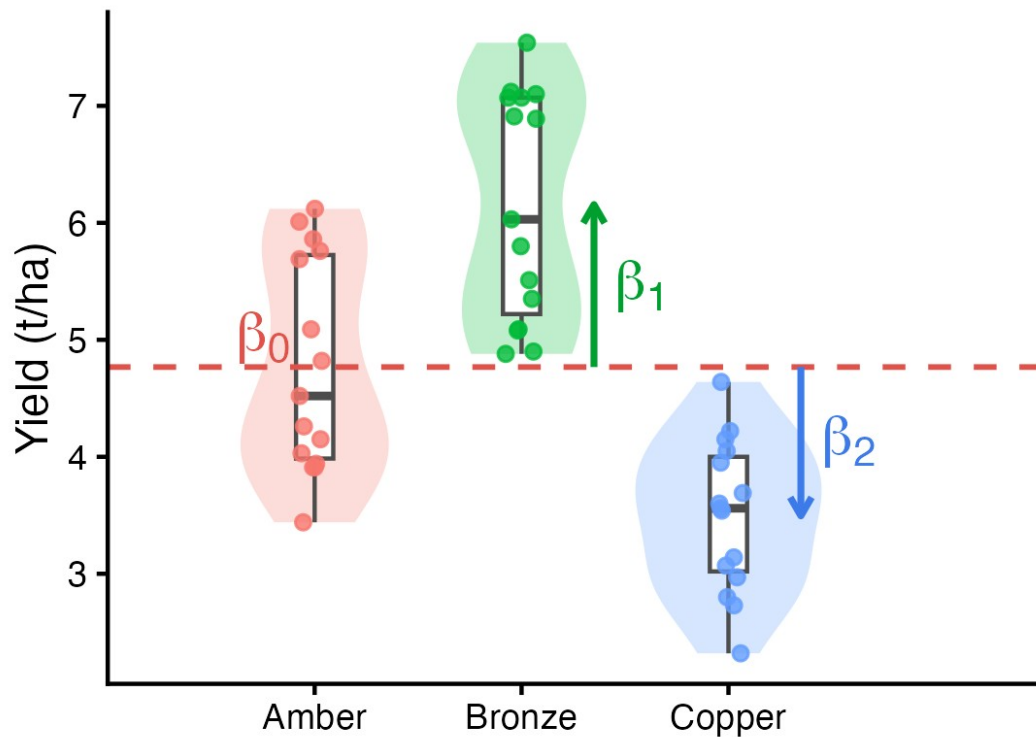
Three varieties



```
> control <- subset(trial, water == "control")  
> tapply(control$yield, control$variety, mean)  
      Amber      Bronze      Copper  
4.767333 6.156000 3.495333 ← three means
```

Do the varieties differ in yield?

Three groups, two dummies



$$\text{yield}_i = \beta_0 + \beta_1 x_{\text{Bronze},i} + \beta_2 x_{\text{Copper},i} + \varepsilon_i$$

β_0 : the Amber mean

β_1 : Bronze - Amber

β_2 : Copper - Amber

The design matrix

```
> m_var <- lm(yield ~ variety, data = control)
> mm <- model.matrix(m_var)
> head(cbind(control["variety"], mm), 7)
```

	variety	(Intercept)	varietyBronze	varietyCopper	
1	Copper	1	0	1	
2	Amber	1	0	0	
3	Bronze	1	1	0	
5	Bronze	1	1	0	
6	Copper	1	0	1	
8	Copper	1	0	1	
12	Bronze	1	1	0	

← two 0/1 columns

Amber plots: all zeros. Each other variety switches on its own column.

lm(yield ~ variety)

```
> round(coef(summary(m_var)), 3)
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4.767	0.221	21.575	0	← Amber mean
varietyBronze	1.389	0.312	4.444	0	← Bronze - Amber
varietyCopper	-1.272	0.312	-4.070	0	← Copper - Amber

Bronze: +1.39 t/ha vs Amber. Copper: -1.27 t/ha vs Amber.

Bronze vs Copper? Not in the table.

Do the varieties differ at all? anova()

```
> anova(m_var)
Analysis of Variance Table

Response: yield
          Df Sum Sq Mean Sq F value    Pr(>F)
variety     2  53.128  26.5638   36.269 7.085e-10 ***
Residuals  42  30.762   0.7324
-----
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*** ← do they differ at all?

One F-test for the whole factor: **F = 36.3, p < 0.001**

Which pairs? TukeyHSD()

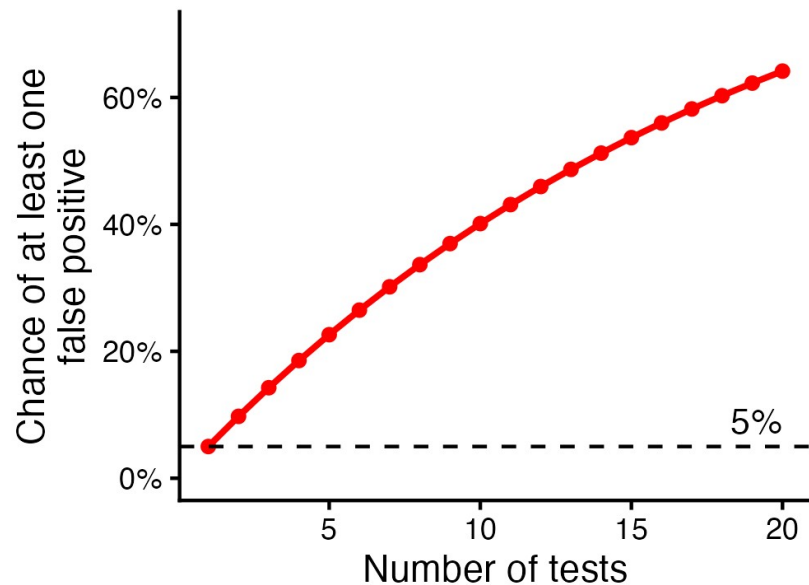
```
> TukeyHSD(aov(m_var))
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = m_var)

$variety
      diff      lwr      upr    p_adj ← adjusted for 3 tests
Bronze-Amber  1.388667  0.629451  2.1478823 0.0001841
Copper-Amber -1.272000 -2.031216 -0.5127843 0.0005836
Copper-Bronze -2.660667 -3.419882 -1.9014510 0.0000000
```

Copper - Bronze = -2.66 t/ha: the biggest gap, missing from the coefficient table

Many tests, many false alarms



$$1 - 0.95^k \quad k = 3 : 14\% \quad k = 20 : 64\%$$

```
> pairwise.t.test(control$yield, control$variety,  
+                 p.adjust.method = "none")
```

Pairwise comparisons using t tests with pooled SD

data: control\$yield and control\$variety

	Amber	Bronze
Bronze	6.3e-05	-
Copper	2e-04	1.1e-10

P value adjustment method: none

```
> pairwise.t.test(control$yield, control$variety,  
+                 p.adjust.method = "holm")
```

Pairwise comparisons using t tests with pooled SD

data: control\$yield and control\$variety

	Amber	Bronze
Bronze	0.00013	-
Copper	0.00020	3.3e-10

P value adjustment method: holm

How big? η^2 (eta-squared)

$$\eta^2 = \frac{SS_{\text{variety}}}{SS_{\text{total}}} = \frac{53.1}{53.1 + 30.8} = 0.63$$

```
> ss <- anova(m_var)$"Sum Sq"
> ss
[1] 53.12763 30.76163
> ss[1] / sum(ss)           # eta-squared
[1] 0.6333067 ← 63% of the variation
> summary(m_var)$r.squared
[1] 0.6333067 ← = R2, one factor
```

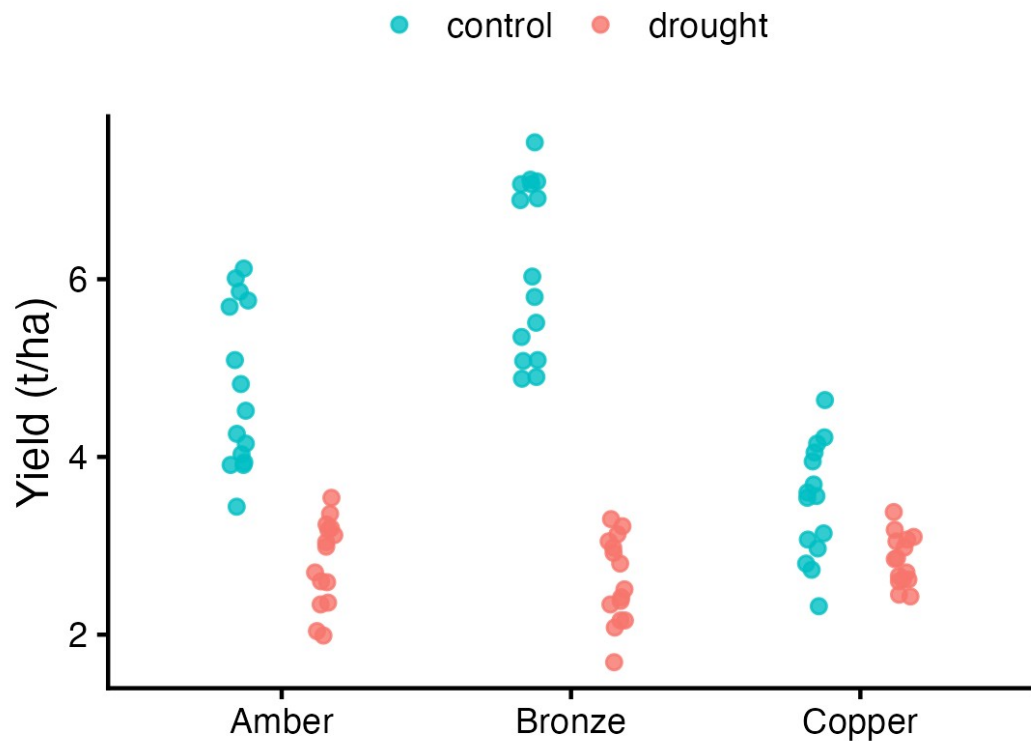
Variety explains **63%**. With one factor $\eta^2 = R^2$; **with several, η^2 splits R^2 by term** (later).

Part 3: Two factors

Variety, water, and their interaction



Variety and water



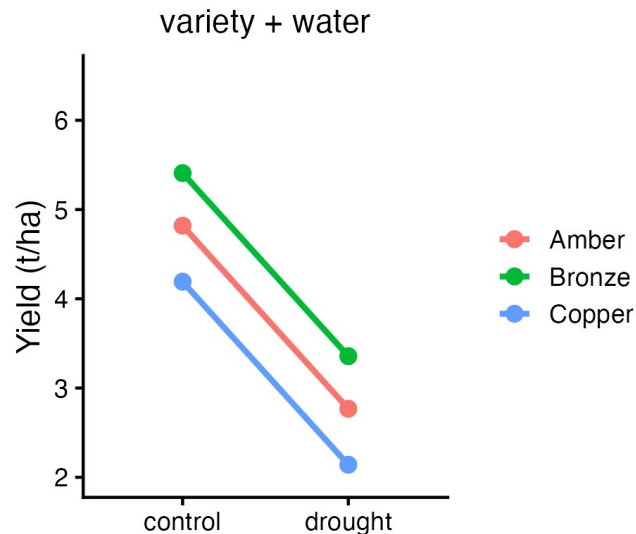
Does drought cost every variety the same?

Adding water: $\text{lm}(\text{yield} \sim \text{variety} + \text{water})$

```
> m_add <- lm(yield ~ variety + water, data = trial)
> round(coef(summary(m_add)), 3)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	4.819	0.190	25.364	0.000
varietyBronze	0.589	0.233	2.533	0.013
varietyCopper	-0.627	0.233	-2.696	0.008
waterdrought	-2.051	0.190	-10.796	0.000

← one drought effect for all

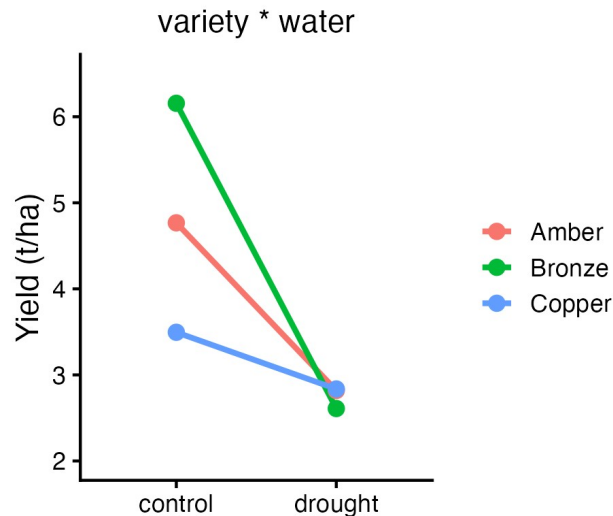


Additive: drought costs the same in every variety. **Parallel lines.**

Interaction: variety * water

```
> m_int <- lm(yield ~ variety * water, data = trial)
> round(coef(summary(m_int)), 3)
```

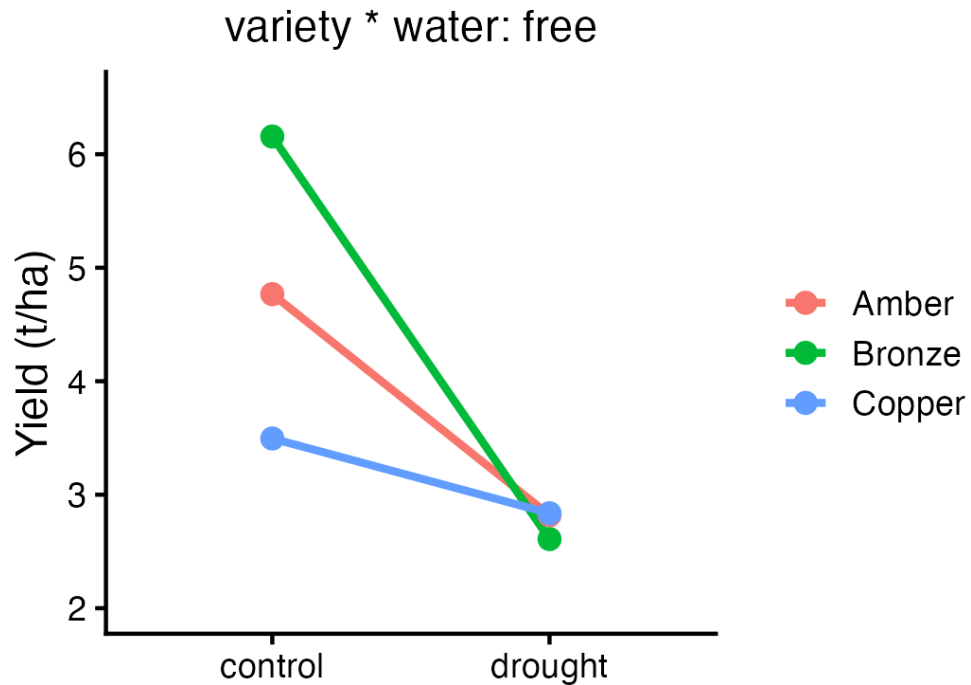
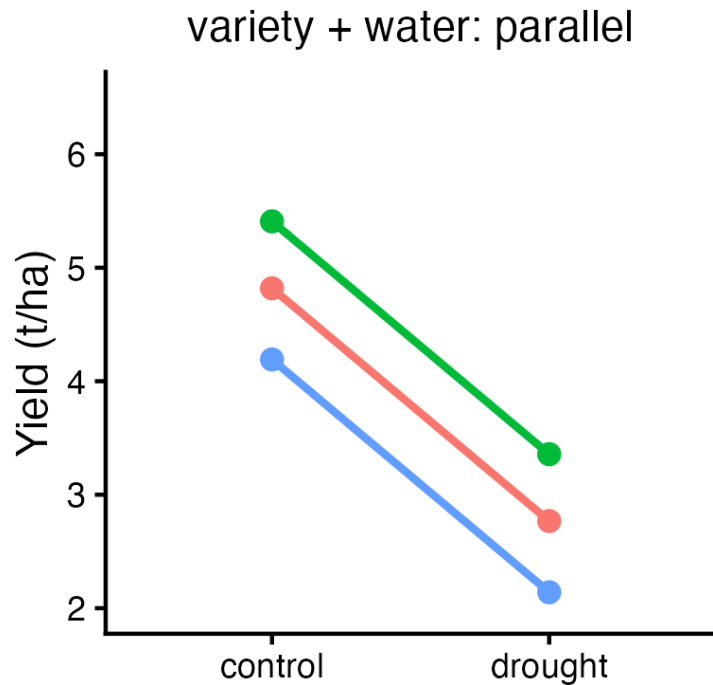
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4.767	0.175	27.291	0	
varietyBronze	1.389	0.247	5.621	0	
varietyCopper	-1.272	0.247	-5.149	0	
waterdrought	-1.948	0.247	-7.885	0	
varietyBronze:waterdrought	-1.599	0.349	-4.576	0	← Bronze loses 1.6 more
varietyCopper:waterdrought	1.289	0.349	3.690	0	← Copper loses 1.3 less



$$\hat{y}_{\text{Copper, drought}} = 4.77 - 1.27 - 1.95 + 1.29 = 2.84$$

varietyCopper:waterdrought: Copper loses 1.29 t/ha less to drought than Amber

Parallel or not?



Is the interaction real?

```
> anova(m_int)
Analysis of Variance Table

Response: yield
      Df Sum Sq Mean Sq F value    Pr(>F)
variety  2  22.211   11.106   24.263 4.820e-09 ***
water    1  94.659   94.659 206.807 < 2.2e-16 ***
variety:water  2  31.397   15.698  34.297 1.292e-11 *** ← test this row first
Residuals 84  38.448    0.458
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Test the interaction first. Here it is real ($p < 0.001$): **report effects within each variety.**

Two tables, two questions

```
> round(coef(summary(m_int)), 3)[, c(1, 4)]
              Estimate Pr(>|t|)
(Intercept)      4.767      0
varietyBronze     1.389      0 ← Bronze vs Amber, control only
varietyCopper    -1.272      0
waterdrought     -1.948      0
varietyBronze:waterdrought -1.599      0
varietyCopper:waterdrought  1.289      0
> anova(m_int)[, c(1, 4, 5)]
      Df F value    Pr(>F)
variety    2 24.263 4.820e-09 *** ← variety overall
water      1 206.807 < 2.2e-16 ***
variety:water 2  34.297 1.292e-11 ***
Residuals  84
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

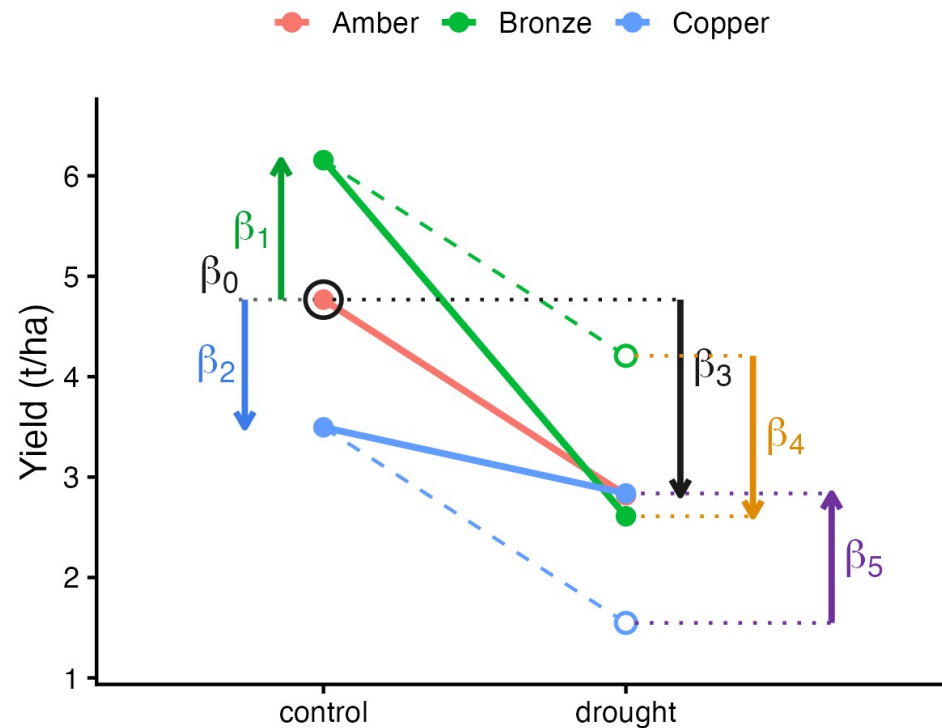
summary()

one row = one comparison
between two cells

anova()

one row = a whole term, all
its levels at once

What each coefficient compares



```
> m_int <- lm(yield ~ variety * water, data = trial)
> round(coef(summary(m_int)), 3)
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4.767	0.175	27.291	0	β_0
varietyBronze	1.389	0.247	5.621	0	β_1
varietyCopper	-1.272	0.247	-5.149	0	β_2
waterdrought	-1.948	0.247	-7.885	0	β_3
varietyBronze:waterdrought	-1.599	0.349	-4.576	0	β_4
varietyCopper:waterdrought	1.289	0.349	3.690	0	β_5

β_0 Amber, control

β_1 Bronze - Amber, under control

β_2 Copper - Amber, under control

β_3 drought - control, in Amber

β_4 Bronze's drought effect - Amber's

β_5 Copper's drought effect - Amber's

Open circles: where parallel lines would put Bronze and Copper. β_4, β_5 = how far off they are.

Change the reference level

```
> trial2$water <- factor(trial$water, levels = c("drought",  
"control"))  
> m_int2 <- lm(yield ~ variety * water, data = trial2)  
> round(coef(summary(m_int))["varietyCopper", ], 3)  
Estimate Std. Error    t value    Pr(>|t|)  
-1.272    0.247    -5.149    0.000 ← Copper – Amber, control  
> round(coef(summary(m_int2))["varietyCopper", ], 3)  
Estimate Std. Error    t value    Pr(>|t|)  
0.017    0.247     0.070    0.944 ← same, under drought  
> anova(m_int)["variety", "Pr(>F)"]  
[1] 4.81955e-09 ← anova(): unchanged  
> anova(m_int2)["variety", "Pr(>F)"]  
[1] 4.81955e-09 ← unchanged
```

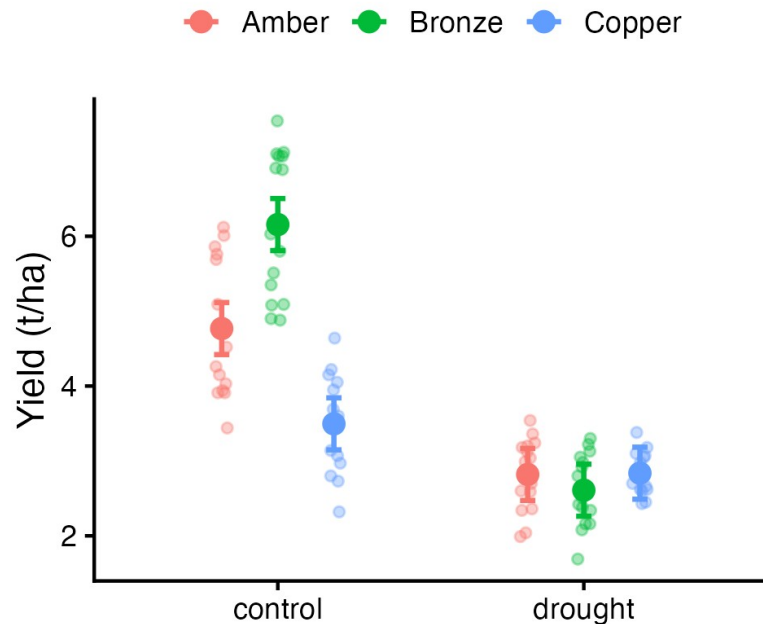
varietyCopper: **-1.27 ($p < 0.001$) vs control, +0.02 ($p = 0.94$) vs drought.** anova(): identical.

emmeans: the model's predicted means

```
> # install.packages("emmeans")
> library(emmeans)
> emmeans(m_int, ~ variety * water)

```

variety	water	emmean	SE	df	lower.CL	upper.CL	
Amber	control	4.77	0.175	84	4.42	5.11	← model's mean per cell
Bronze	control	6.16	0.175	84	5.81	6.50	
Copper	control	3.50	0.175	84	3.15	3.84	
Amber	drought	2.82	0.175	84	2.47	3.17	
Bronze	drought	2.61	0.175	84	2.26	2.96	
Copper	drought	2.84	0.175	84	2.49	3.18	



emmeans package: a mean for every cell (or any average of cells), with a CI, **from the model**.

“Main effects” when there is an interaction

```
> emmeans(m_int, ~ variety)
  variety emmean    SE df lower.CL upper.CL
  Amber    3.79 0.124 84    3.55    4.04
  Bronze    4.38 0.124 84    4.14    4.63
  Copper    3.17 0.124 84    2.92    3.41

Results are averaged over the levels of: water
Confidence level used: 0.95
```

Averaged over water. `emmeans` also prints a NOTE: “may be misleading due to involvement in interactions”.

The drought effect in each variety: emmeans

```
> pairs(emmeans(m_int, ~ water | variety))  
variety = Amber:  
contrast      estimate      SE df t.ratio p.value  
control - drought 1.948 0.247 84  7.885 <0.0001 ← Amber  
  
variety = Bronze:  
contrast      estimate      SE df t.ratio p.value  
control - drought 3.547 0.247 84 14.357 <0.0001 ← Bronze: biggest loss  
  
variety = Copper:  
contrast      estimate      SE df t.ratio p.value  
control - drought 0.659 0.247 84  2.666  0.0092 ← Copper
```

Drought costs Amber 1.95, **Bronze 3.55**, Copper 0.66 t/ha

A second factor removes noise

```
> anova(lm(yield ~ variety, data = trial))
Analysis of Variance Table

Response: yield
      Df Sum Sq Mean Sq F value    Pr(>F)
variety  2  22.211  11.1057   5.8734 0.004049 **
Residuals 87 164.504   1.8908    ← water left in residual
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> anova(m_add)
Analysis of Variance Table

Response: yield
      Df Sum Sq Mean Sq F value    Pr(>F)
variety  2  22.211   11.106  13.675 6.974e-06 ***
water    1  94.659   94.659 116.553 < 2.2e-16 ***
Residuals 86  69.845    0.812    ← water taken out
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

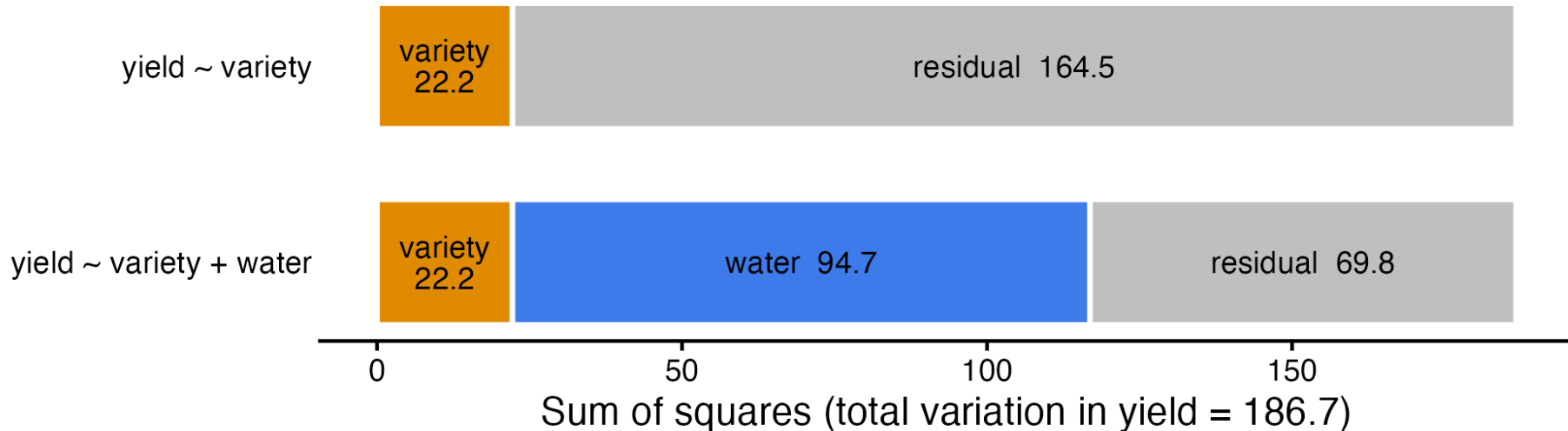
variety alone:

$F = 5.9$

variety + water:

$F = 13.7$

Where the F comes from: slicing the variation



$$F_{\text{variety}} = \frac{MS_{\text{variety}}}{MS_{\text{residual}}} : \quad \frac{11.1}{164.5/87} = \frac{11.1}{1.89} = 5.9 \quad \frac{11.1}{69.8/86} = \frac{11.1}{0.81} = 13.7$$

Same variety slice. **Water takes its slice out of the residual**, so the denominator of F shrinks.

ANOVA types. Balanced: order doesn't matter

	control	drought
Amber	15 plots	15 plots
Bronze	15 plots	15 plots
Copper	15 plots	15 plots

```
> m_vw <- lm(yield ~ variety + water, data = trial)
> m_wv <- lm(yield ~ water + variety, data = trial)
> anova(m_vw)[, c(2, 4)]
              Sum Sq F value
variety      22.211   13.675
water        94.659  116.553
Residuals    69.845
> anova(m_wv)[, c(2, 4)]
              Sum Sq F value
water        94.659  116.553
variety      22.211   13.675
Residuals    69.845
```

← same SS

← either order

15 plots in every cell. **Either order: the same sums of squares, the same F.**

An unbalanced design: the factors overlap

	control	drought
Amber	15 plots	15 plots
Bronze	15 plots	3 plots
Copper	3 plots	15 plots

Same trial, but most **Bronze** plots were irrigated and most **Copper** plots droughted.

Bronze plots yield more. **Is that Bronze, or the water?**

Part of the yield variation can be explained by **either** factor.

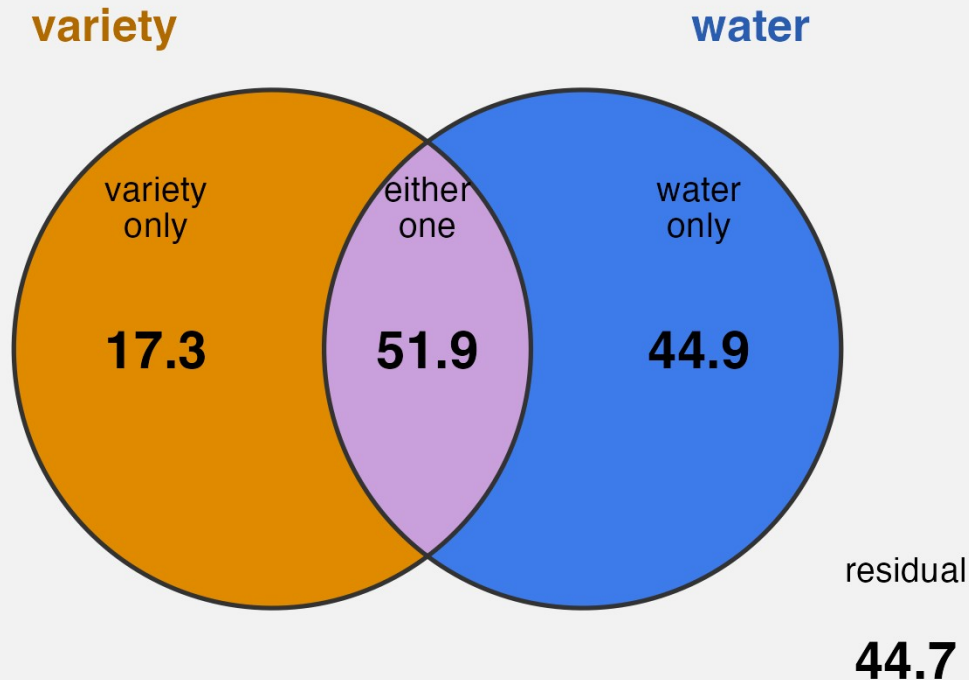
Type I: now the order matters

	control	drought
Amber	15 plots	15 plots
Bronze	15 plots	3 plots
Copper	3 plots	15 plots

```
> m_vw <- lm(yield ~ variety + water, data = lopsided)
> m_wv <- lm(yield ~ water + variety, data = lopsided)
> anova(m_vw)[, c(2, 4)]
              Sum Sq F value
variety      69.277  47.995 ← variety first
water        44.904  62.218
Residuals    44.746
> anova(m_wv)[, c(2, 4)]
              Sum Sq F value
water        96.849 134.193
variety      17.332  12.008 ← variety second
Residuals    44.746
```

Same data, same model. **Variety first: F = 48. Variety second: F = 12.**

The overlap, as a picture



Box: all the variation in yield (158.9).

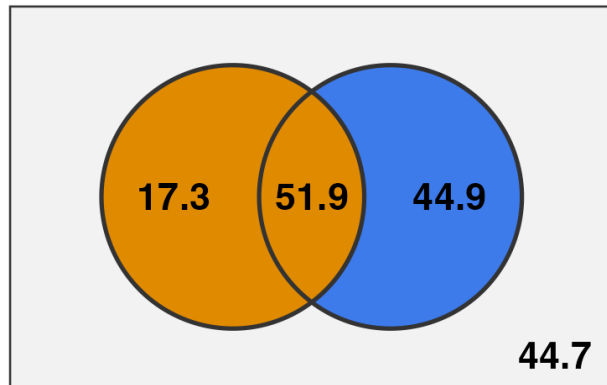
Circles: what **variety** and **water** can each explain.

Overlap: explained by either one. In this design, R can't tell which.

Who gets the overlap?

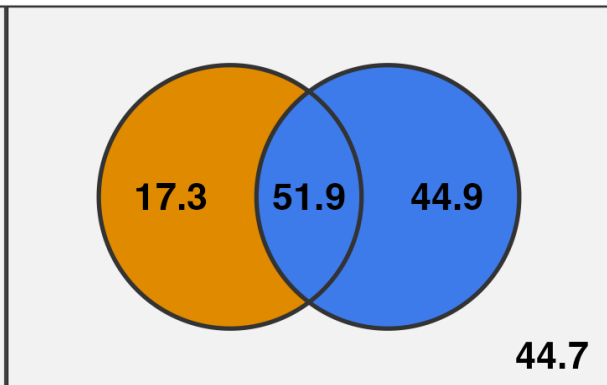
Type I: variety first

variety gets $17.3 + 51.9 = 69.3$
 $F = 48$



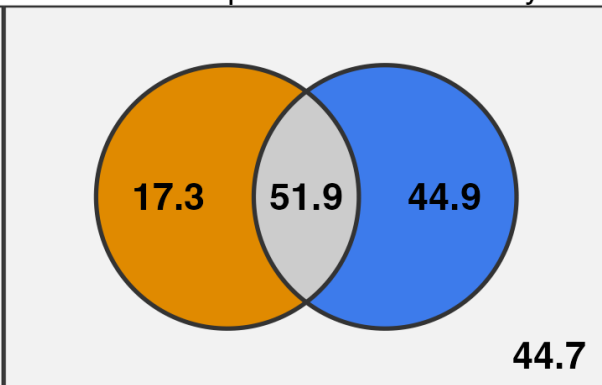
Type I: water first

variety gets only 17.3
 $F = 12$



Type II

variety gets 17.3 ($F = 12$), water 44.9
the overlap counts for nobody



Type I: whoever is first in the formula. **Type II: nobody:** each factor is tested only on what it alone explains.

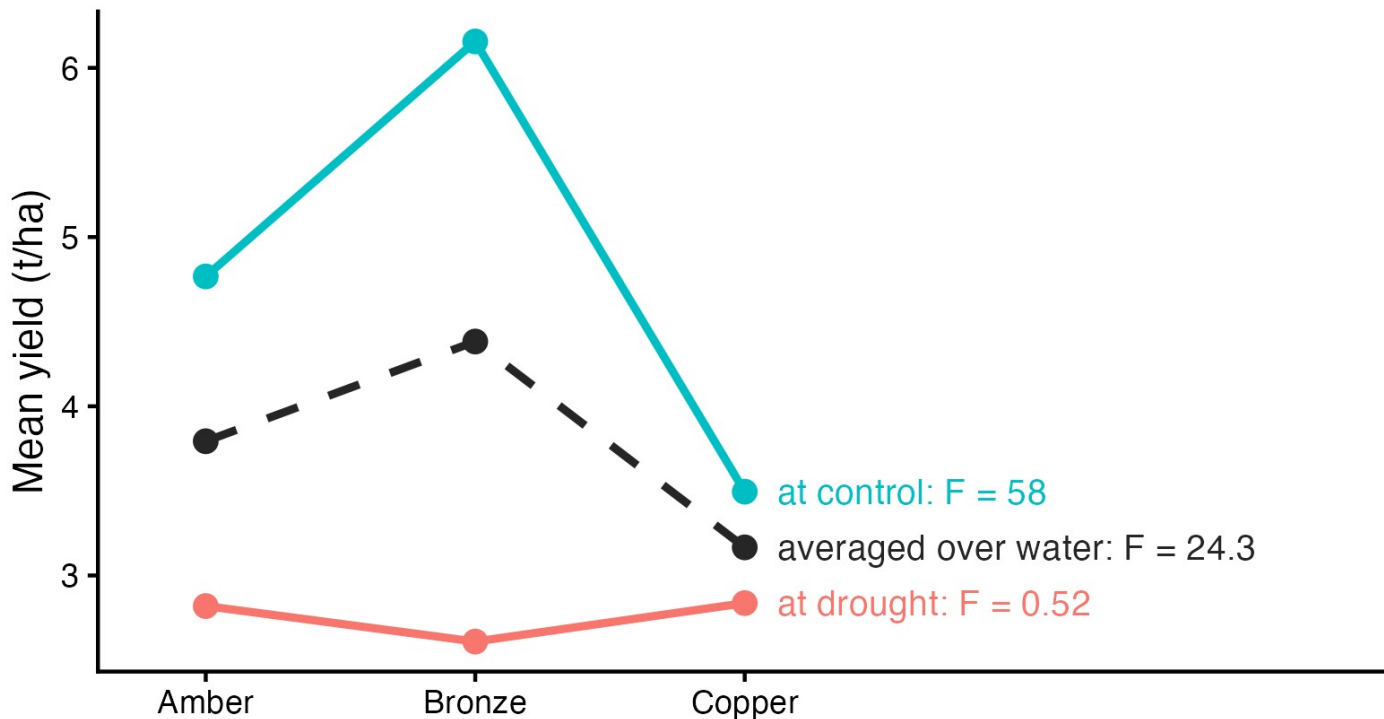
Type II: car::Anova()

```
> library(car)
> Anova(lm(yield ~ variety + water, data = lopsided), type = 2)
Anova Table (Type II tests)

Response: yield
      Sum Sq Df F value    Pr(>F)
variety  17.332  2   12.008 3.911e-05 ***
water    44.904  1   62.218 6.195e-11 ***
Residuals 44.746 62
```

Type II: each main effect adjusted for the other, whatever the order

Type III: which “variety effect”?



With an interaction, variety has **a different effect at each water level.**

Type III must pick one.

Type III, and the contrasts trap

```
> library(car)
> Anova(m_int, type = 3)      # reference: control
```

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	340.91	1	744.811	< 2.2e-16 ***
variety	53.13	2	58.036	< 2.2e-16 ***
water	28.46	1	62.179	1.028e-11 ***
variety:water	31.40	2	34.297	1.292e-11 ***
Residuals	38.45	84		

```
> Anova(m_int2, type = 3)    # reference: drought
```

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	119.230	1	260.4882	< 2.2e-16 ***
variety	0.480	2	0.5248	0.5936
water	28.460	1	62.1789	1.028e-11 ***
variety:water	31.397	2	34.2971	1.292e-11 ***
Residuals	38.448	84		

variety: $F = 58$ vs $F = 0.52$. Same model, opposite answers.

water: the drought effect in Amber, in both. **variety:water**: identical.

R's default coding: each main-effect row is tested at the reference level of the other factor.

Sum-to-zero coding: averaged effects

```
> options(contrasts = c("contr.sum", "contr.poly"))
> m_sum <- lm(yield ~ variety * water, data = trial)
> Anova(m_sum, type = 3)
```

	Sum Sq	Df	F value	Pr(>F)
(Intercept)	1286.41	1	2810.497	< 2.2e-16 ***
variety	22.21	2	24.263	4.820e-09 ***
water	94.66	1	206.807	< 2.2e-16 ***
variety:water	31.40	2	34.297	1.292e-11 ***
Residuals	38.45	84		

```
> options(contrasts = c("contr.treatment", "contr.poly"))
```

```
> joint_tests(m_int)
```

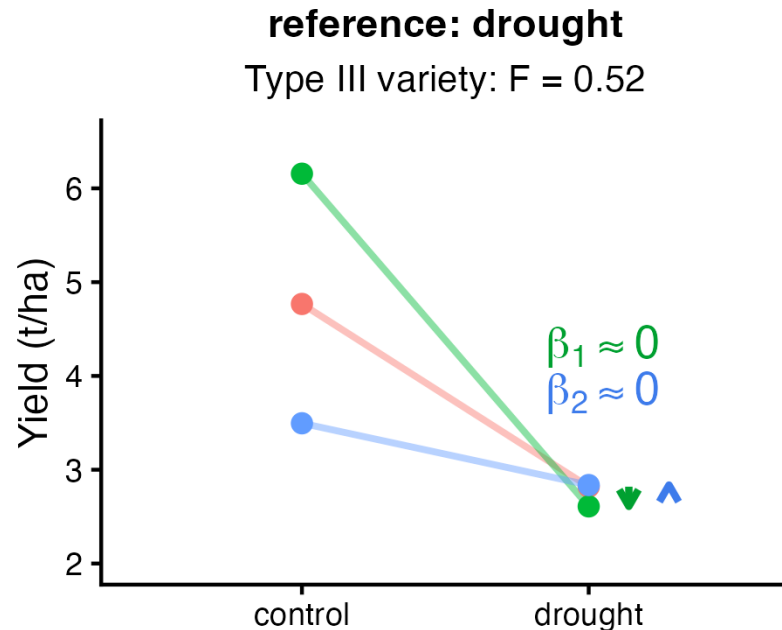
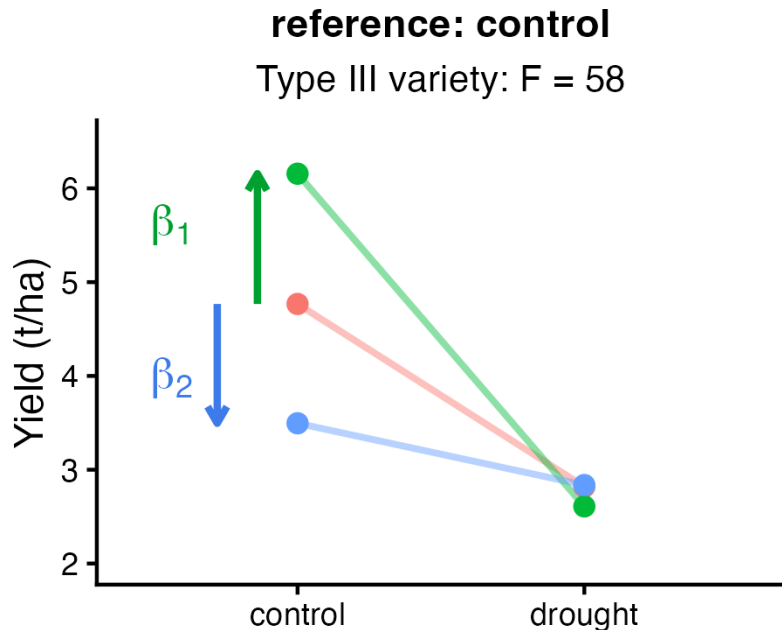
model term	df1	df2	F.ratio	p.value
variety	2	84	24.263	<0.0001
water	1	84	206.807	<0.0001
variety:water	2	84	34.297	<0.0001

← same F as anova() here

variety: $F = 24.3$, averaged over water. **water**: averaged over varieties.

joint_tests() gives the same tests with no contrast settings to remember.

Why: the variety row tests β_1 and β_2



Type III's variety row asks: are β_1 and β_2 both zero? R's default measures them **at the reference level**.

Safe recipes

one F for 'variety, averaged over water'

same thing with car

the interaction is real

no interaction

never

`emmeans::joint_tests(model)`

`contr.sum` first, then `Anova(type = 3)`

report variety within each water level

drop it; `Anova(type = 2)`

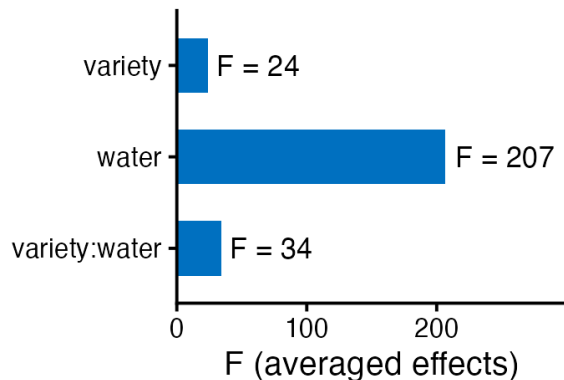
`Anova(type = 3)` with default contrasts

Which tool for which question?

Does this factor matter?

factorial design, interaction in the model

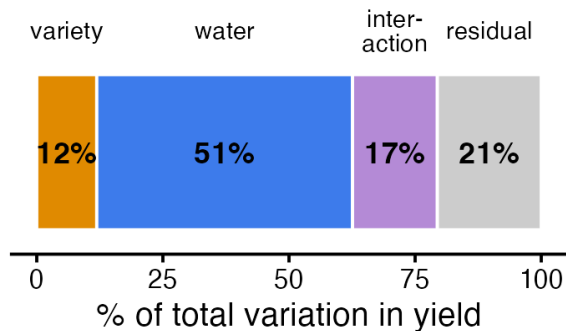
```
> joint_tests(m_int)
model term    df1 df2 F.ratio p.value
variety       2   84  24.263 <0.0001
water         1   84 206.807 <0.0001
variety:water  2   84  34.297 <0.0001
```



How much does it explain?

R^2 for the model; η^2 splits it by term

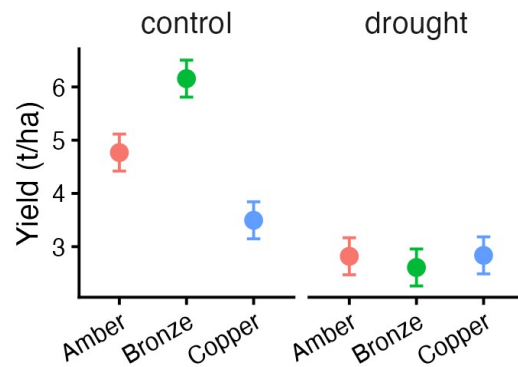
```
> ss <- anova(m_int)$"Sum Sq"
> round(ss / sum(ss), 2) # eta2 per term
[1] 0.12 0.51 0.17 0.21
> summary(m_int)$r.squared # = the first 3
[1] 0.7940811
```



The interaction is real

report effects within each level

```
> pw <- as.data.frame(pairs(emmeans(m_int,
+ ~ variety | water)))
> pw$p <- format.pval(pw$p.value, eps = 1e-4, digits = 2)
> pw$estimate <- round(pw$estimate, 2)
> pw[, c("water", "contrast", "estimate", "p")]
  water contrast estimate p
1 control Amber - Bronze -1.39 <1e-04
2 control Amber - Copper  1.27 <1e-04
3 control Bronze - Copper  2.66 <1e-04
4 drought Amber - Bronze  0.21  0.67
5 drought Amber - Copper -0.02  1.00
6 drought Bronze - Copper -0.23  0.63
```



Which ANOVA type?

balanced design

Type I: `anova()`

Type II: `Anova(type = 2)`

Type III: `Anova(type = 3)`

interaction is real

I, II, and III (`contr.sum`) agree

sequential: order is a choice

no interaction: the default

with interaction; `contr.sum` first

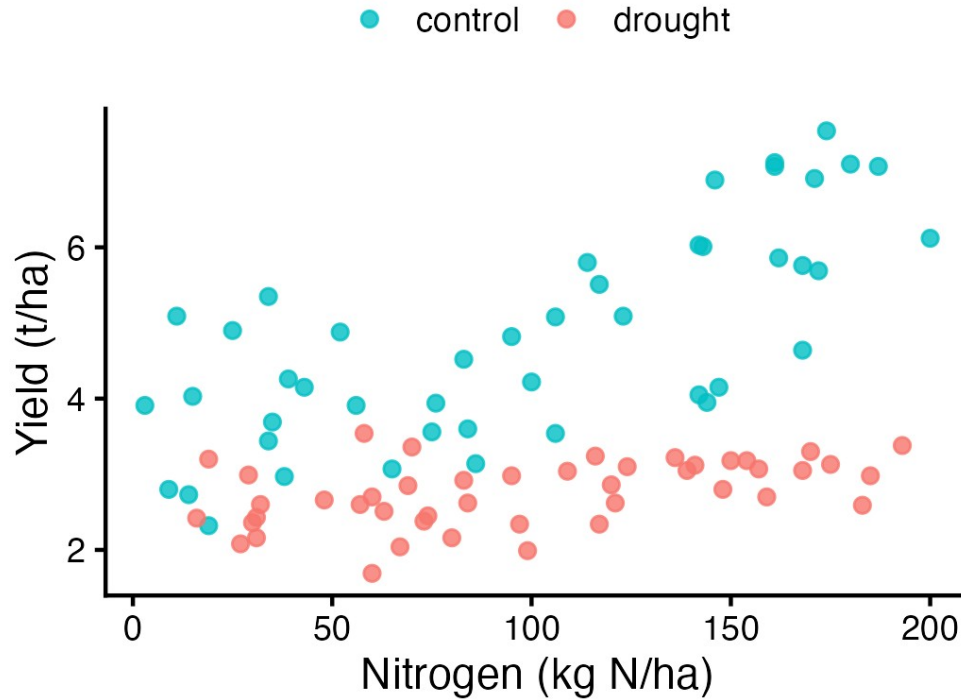
report effects within levels

Part 4: Groups and a slope together

ANCOVA

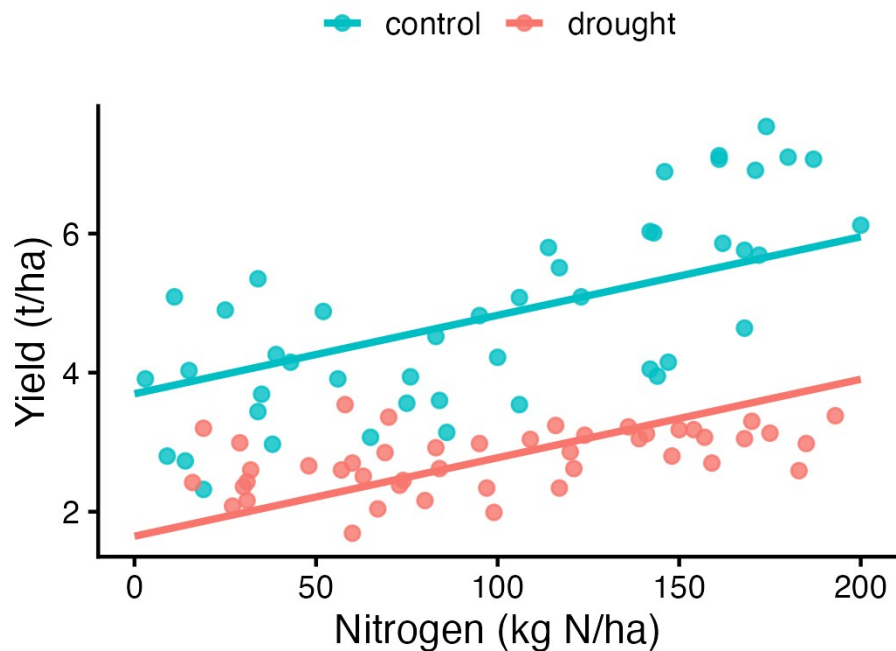


Yield, nitrogen, and drought



A slope and a group in one model?

What is ANCOVA?



ANCOVA = analysis of covariance

= ANOVA (compare groups)
+ a numeric **covariate**

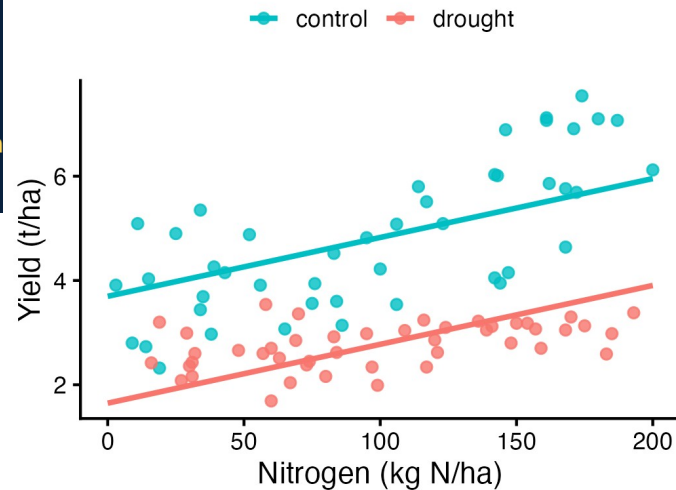
Asks: do control and drought differ **at the same nitrogen rate?**

`lm(yield ~ nitrogen + water)`

Parallel lines: nitrogen + water

```
> m_anc <- lm(yield ~ nitrogen + water, data = trial)
> round(coef(summary(m_anc)), 4)
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	3.6962	0.1942	19.0350	0	
nitrogen	0.0113	0.0015	7.2995	0	← one slope for both
waterdrought	-2.0491	0.1708	-11.9986	0	← drought, same N



Intercept = control yield at 0 kg N, not the control mean

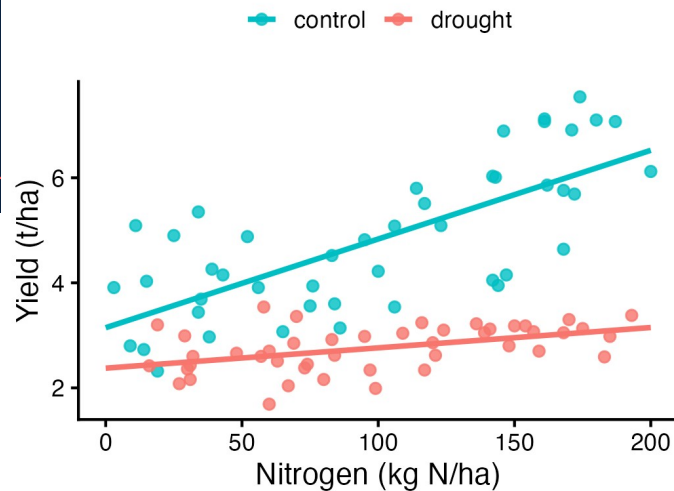
Different slopes: nitrogen * water

```
> m_anc2 <- lm(yield ~ nitrogen * water, data = trial)
> round(coef(summary(m_anc2)), 4)
```

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.1458	0.2113	14.8864	0.0000
nitrogen	0.0169	0.0018	9.1614	0.0000
waterdrought	-0.7704	0.3160	-2.4382	0.0168
nitrogen:waterdrought	-0.0130	0.0028	-4.6312	0.0000

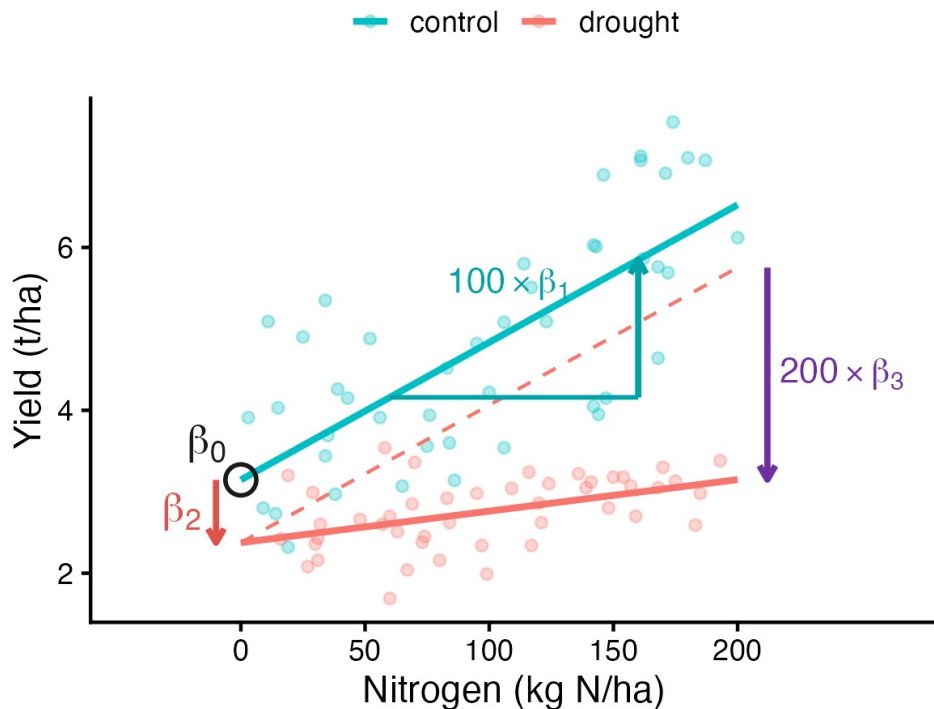
← drought slope flatter

slope under drought = $\beta_1 + \beta_3 = 0.0169 - 0.0130 = 0.0039$ t/ha per kg N



Drought cuts the return on nitrogen by three quarters.

What each coefficient compares: nitrogen * water



```
> m_anc2 <- lm(yield ~ nitrogen * water, data = trial)
> round(coef(summary(m_anc2)), 4)
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	3.1458	0.2113	14.8864	0.0000	β_0
nitrogen	0.0169	0.0018	9.1614	0.0000	β_1
waterdrought	-0.7704	0.3160	-2.4382	0.0168	β_2
nitrogen:waterdrought	-0.0130	0.0028	-4.6312	0.0000	β_3

β_0 control yield at 0 kg N

β_1 nitrogen slope, under control

β_2 drought – control, at 0 kg N

β_3 how much drought changes the slope

Dashed: where the drought line would go if parallel. Drought slope = $\beta_1 + \beta_3$.

Two numeric predictors: a vanishing main effect

```
> m_nw <- lm(yield ~ nitrogen * water, data = irrig)
> round(coef(summary(m_nw)), 4)

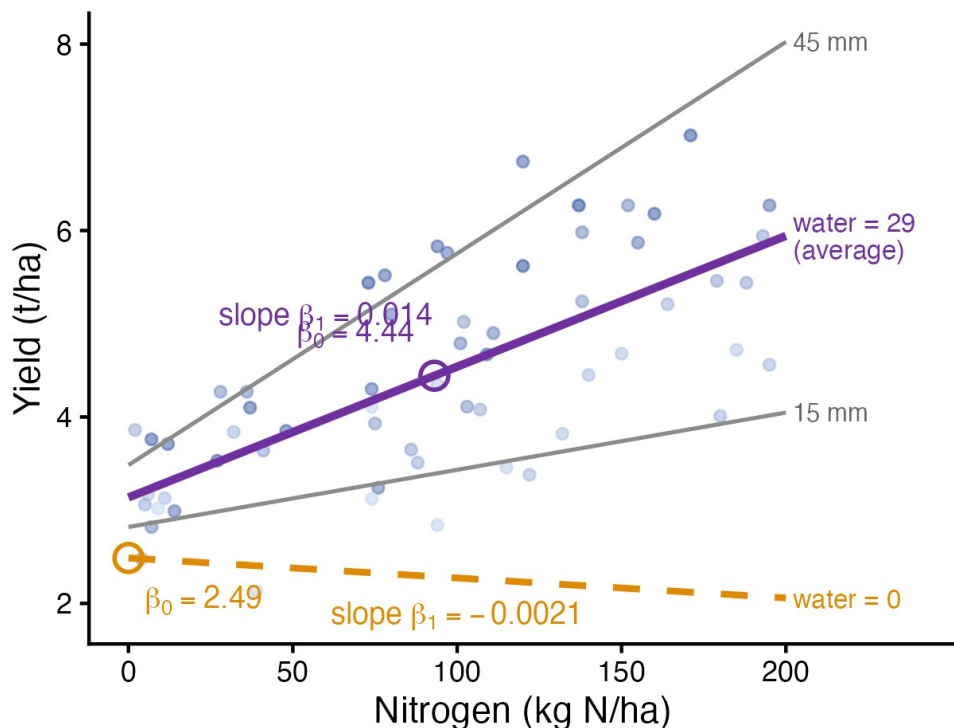
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.4875     0.4687   5.3075   0.0000
nitrogen       -0.0021     0.0044  -0.4826   0.6313 ← slope at water = 0!
water           0.0222     0.0158   1.4020   0.1664
nitrogen:water   0.0006     0.0002   3.6356   0.0006
> anova(m_nw)[, 4:5]
              F value    Pr(>F)
nitrogen    124.773 7.122e-16 *** ← nitrogen matters
water       74.621 6.966e-12 ***
nitrogen:water 13.217 0.0006043 ***
Residuals
```

summary(): **nitrogen p = 0.63**

anova(): **nitrogen p < 0.001**

Same model.

β_1 is the slope of one particular line



as measured: everything at water = 0

```
> m_nw <- lm(yield ~ nitrogen * water, data = irrig)
> round(coef(summary(m_nw)), 4)
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	2.4875	0.4687	5.3075	0.0000	β_0
nitrogen	-0.0021	0.0044	-0.4826	0.6313	β_1
water	0.0222	0.0158	1.4020	0.1664	β_2
nitrogen:water	0.0006	0.0002	3.6356	0.0006	β_3

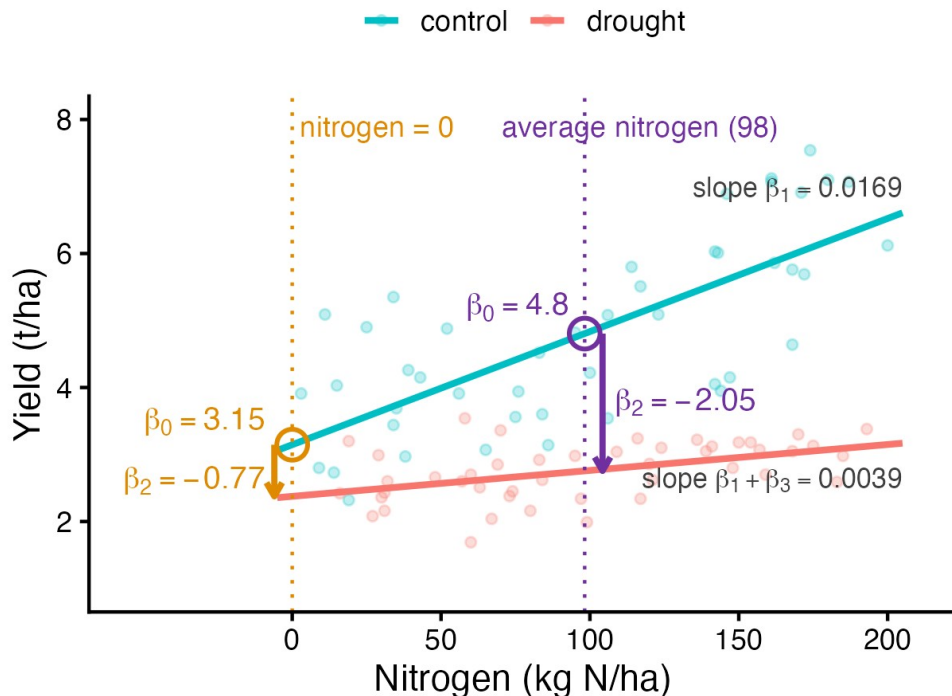
centred: everything at the averages

```
> irrig$nitrogen_c <- irrig$nitrogen - mean(irrig$nitrogen)
> irrig$water_c <- irrig$water - mean(irrig$water)
> m_nwc <- lm(yield ~ nitrogen_c * water_c, data = irrig)
> round(coef(summary(m_nwc)), 4)
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4.4423	0.0700	63.5072	0e+00	β_0
nitrogen_c	0.0140	0.0012	11.5990	0e+00	β_1
water_c	0.0735	0.0083	8.8964	0e+00	β_2
nitrogen_c:water_c	0.0006	0.0002	3.6356	6e-04	β_3 same

β_1 = the nitrogen slope **at water = 0** ($p = 0.63$) or **at average water** ($p < 0.001$). Same model.

Centring: same lines, a new zero point



nitrogen as measured

```
> m_anc2 <- lm(yield ~ nitrogen * water, data = trial)
> round(coef(summary(m_anc2)), 4)
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	3.1458	0.2113	14.8864	0.0000	β_0
nitrogen	0.0169	0.0018	9.1614	0.0000	β_1
waterdrought	-0.7704	0.3160	-2.4382	0.0168	β_2
nitrogen:waterdrought	-0.0130	0.0028	-4.6312	0.0000	β_3

nitrogen centred (minus its mean)

```
> trial$nitrogen_c <- trial$nitrogen - mean(trial$nitrogen)
> m_anc3 <- lm(yield ~ nitrogen_c * water, data = trial)
> round(coef(summary(m_anc3)), 4)
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	4.8047	0.1087	44.2171	0	β_0
nitrogen_c	0.0169	0.0018	9.1614	0	β_1 same
waterdrought	-2.0493	0.1537	-13.3354	0	β_2
nitrogen_c:waterdrought	-0.0130	0.0028	-4.6312	0	β_3 same

β_0 = control yield and β_2 = the drought gap, at 0 kg N or at the average (98 kg N). Slopes: unchanged.

Should you centre?

interaction, reading main-effect rows

no interaction

0 is a real value (0 kg N)

centre (else the row is about $x = 0$)

no need: nothing changes

uncentred rows are fine

```
> library(emmeans)
> emtrends(m_nw, ~ water, var = "nitrogen", at = list(water =
c(15, 30, 45)))
water nitrogen.trend SE df lower.CL upper.CL
15      0.00614 0.00233 56 0.00147 0.0108 ← slope at 15 mm
30      0.01442 0.00123 56 0.01195 0.0169 ← at 30 mm (average)
45      0.02270 0.00282 56 0.01704 0.0283 ← at 45 mm

Confidence level used: 0.95
```

Often better: ask for the slope where you care, **emtrends()**

It's all lm()

two-sample t-test

one-way ANOVA

two-way ANOVA

regression

ANCOVA

different slopes

$\text{lm}(y \sim \text{group})$

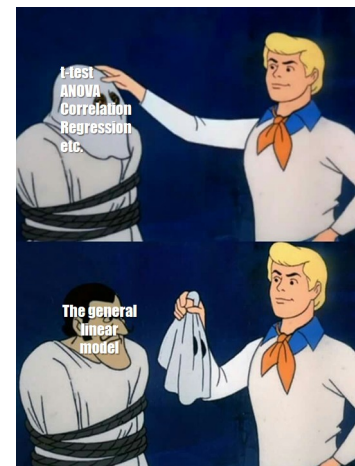
$\text{lm}(y \sim \text{factor})$

$\text{lm}(y \sim f1 * f2)$

$\text{lm}()$

$\text{lm}(y \sim x + \text{group})$

$\text{lm}(y \sim x * \text{group})$



Recap: groups in $\text{Im}()$

Kahoot-style warm-up

12 quick questions on groups in $\text{Im}()$.

Not graded.

Use a pseudonym if you want.

Fast and correct wins.



Scan with your phone

[Join on your phone or laptop: greymonroe.github.io/live-poll-test/quiz-play.html?quiz=pls206-w3-monday-recap](https://greymonroe.github.io/live-poll-test/quiz-play.html?quiz=pls206-w3-monday-recap)

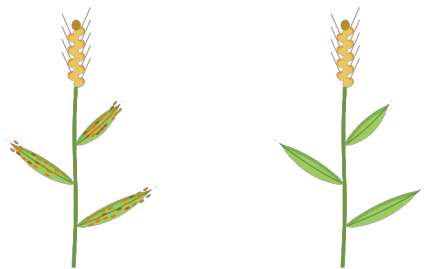
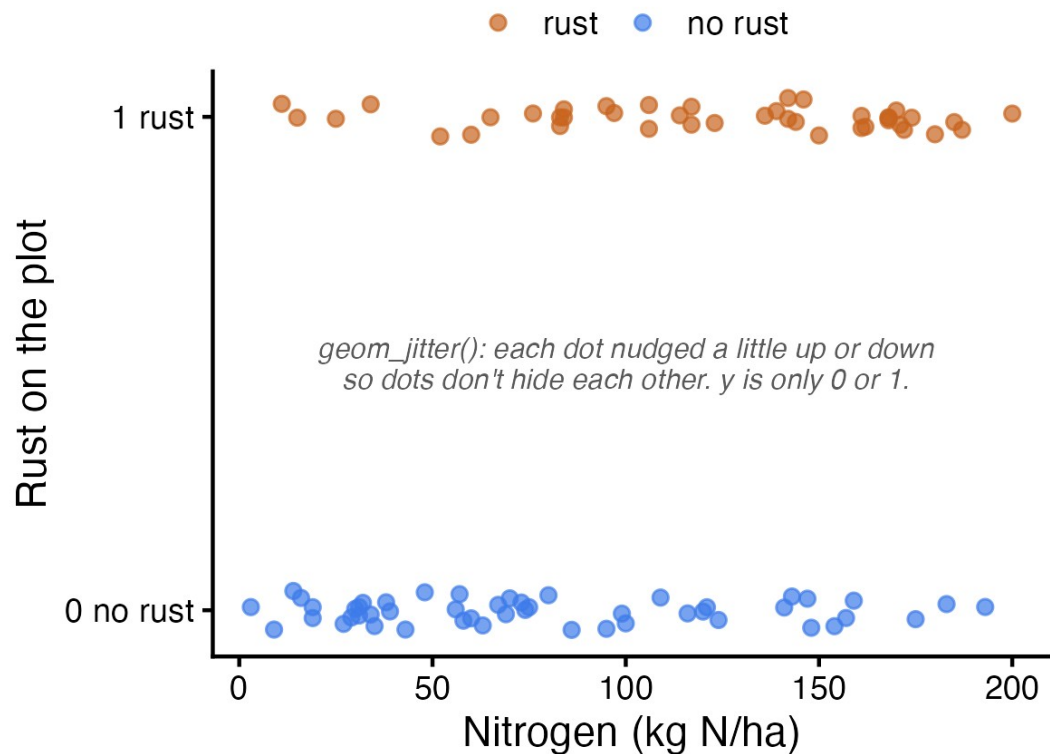
[host screen](#)

Part 5: Yes or no

Logistic regression



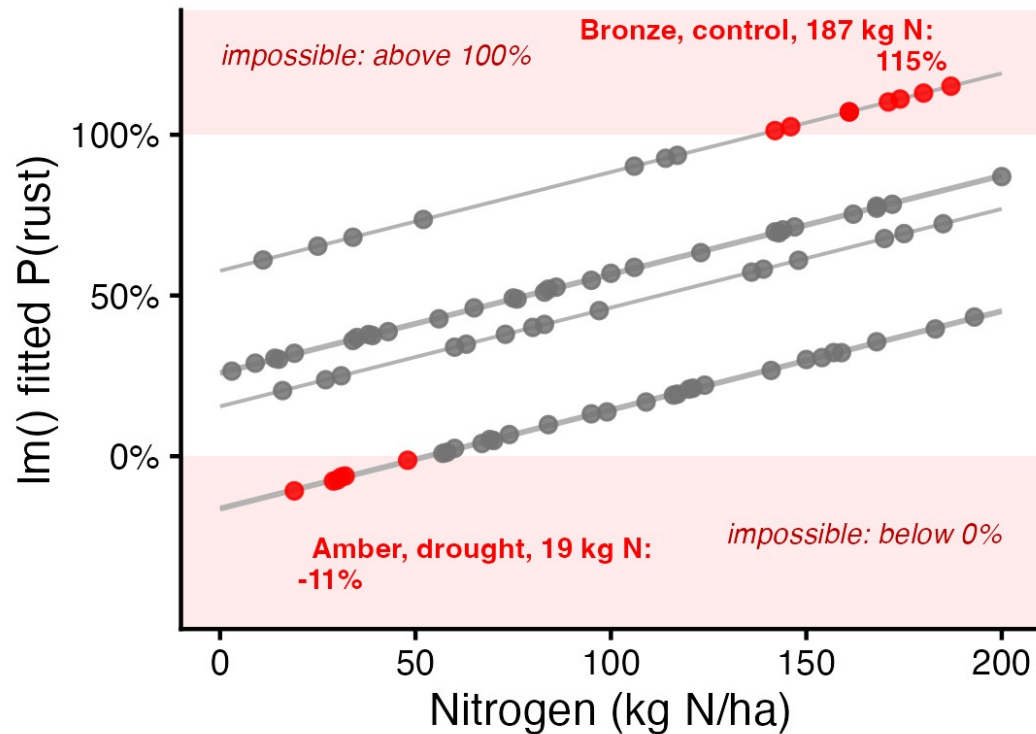
Rust: yes or no



rust no rust

y is **0 or 1**. Can `lm()` handle it?

A straight line fails



```
> m_lin <- lm(rust ~ nitrogen + variety + water, data = trial)
> range(fitted(m_lin))
[1] -0.107848 1.151009 ← below 0, above 1
> sum(fitted(m_lin) < 0 | fitted(m_lin) > 1) # impossible
[1] 14 ← of 90 plots
```

14 of the 90 real plots get an impossible probability

Probability, odds, log-odds

y is 0 / 1	example rust on the plot	family Bernoulli (binomial)	link: $\eta =$ $\log \frac{p}{1-p}$	mean: $\mu =$ $\frac{1}{1+e^{-\eta}}$	in R <code>glm(family = binomial)</code>
---------------	-----------------------------	--------------------------------	--	--	---

probability

odds

log-odds

0.2

1 : 4 = 0.25

-1.39

0.5

1 : 1 = 1

0

0.8

4 : 1 = 4

+1.39

0.99

99 : 1

+4.6

$$\text{odds} = \frac{p}{1-p} \quad \text{log-odds} = \log\left(\frac{p}{1-p}\right)$$

The logit stretches 0-1 into a line

y is

0 / 1

example

rust on the plot

family

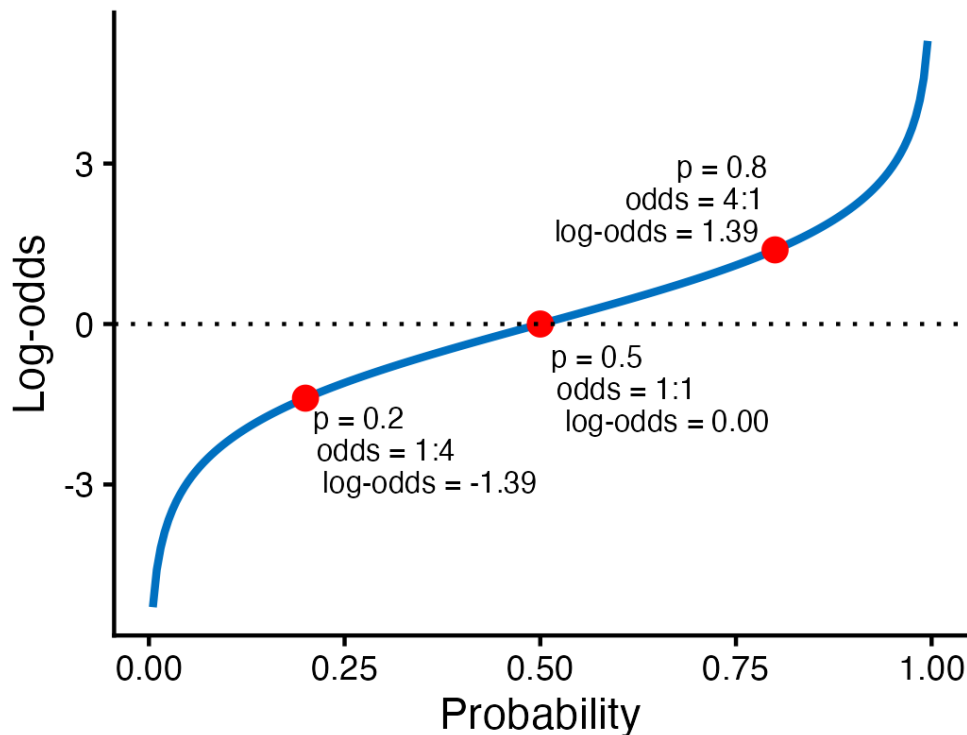
Bernoulli (binomial)

link: $\eta =$

$\log \frac{p}{1-p}$

mean: $\mu =$ in R

$\frac{1}{1+e^{-\eta}}$ `glm(family = binomial)`



$$p = 0.8 : \quad \text{odds} = \frac{0.8}{0.2} = 4 \quad \log(4) = 1.39$$

Near 0 and 1, small changes in p are **big** changes in log-odds.

Undo the logit: log-odds $\rightarrow$ probability

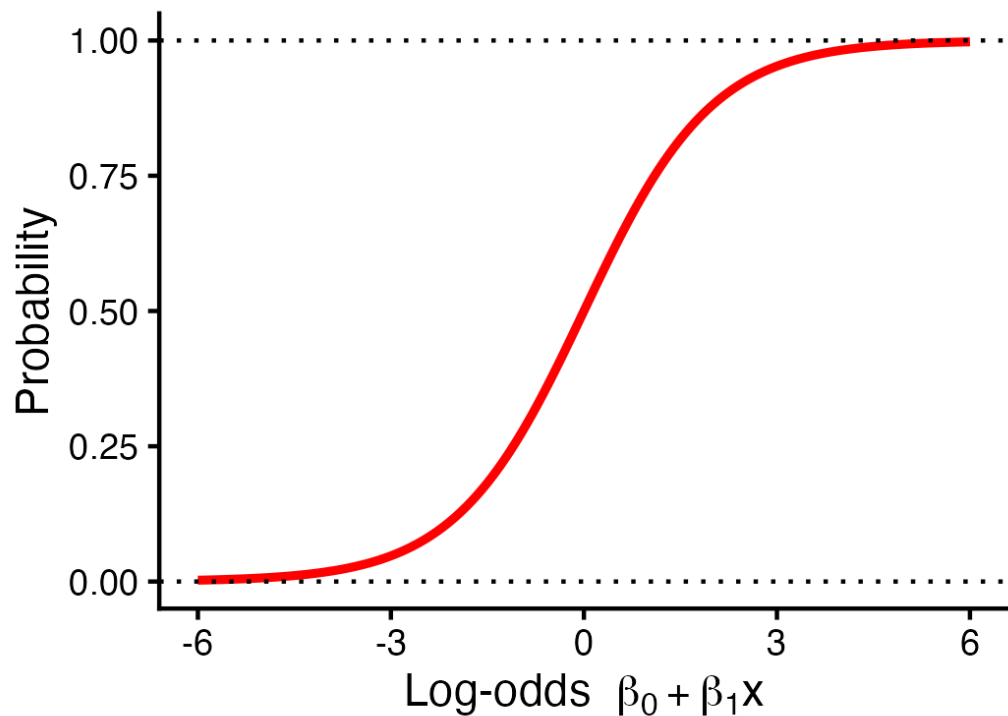
y is
0 / 1

example
rust on the plot

family
Bernoulli (binomial)

link: $\eta =$
 $\log \frac{p}{1-p}$

mean: $\mu =$ in R
 $\frac{1}{1+e^{-\eta}}$ `glm(family = binomial)`



$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \quad \text{plogis()}$$

Any line on the log-odds scale is an **S-curve** between 0 and 1.

glm(rust ~ nitrogen, family = binomial)

y is	example	family	link: $\eta =$	mean: $\mu =$	in R
0 / 1	rust on the plot	Bernoulli (binomial)	$\log \frac{p}{1-p}$	$\frac{1}{1+e^{-\eta}}$	glm(family = binomial)

```
> m_rn <- glm(rust ~ nitrogen, family = binomial, data = trial)
> round(coef(summary(m_rn)), 4)
```

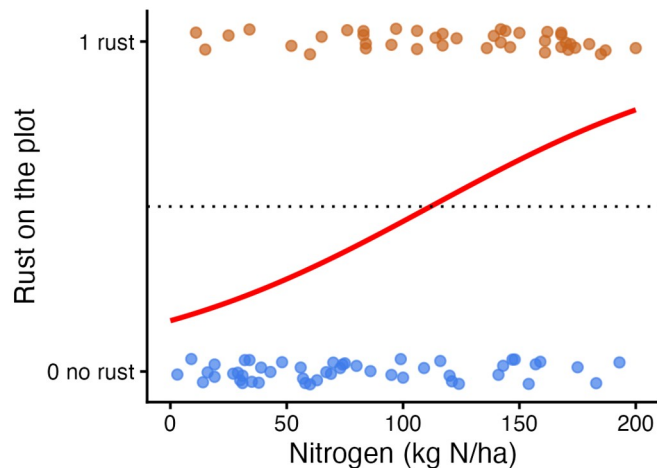
	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-1.7047	0.5046	-3.3784	7e-04	← log-odds at 0 kg N
nitrogen	0.0152	0.0044	3.4533	6e-04	← log-odds per kg N

Coefficients are on the **log-odds** scale:

-1.70 = log-odds of rust at 0 kg N

+0.0152 = change in log-odds per kg N

The S-curve crosses 50:50 at 112 kg N.



predict(): log-odds or probability?

y is	example	family	link: $\eta =$	mean: $\mu =$	in R
0 / 1	rust on the plot	Bernoulli (binomial)	$\log \frac{p}{1-p}$	$\frac{1}{1+e^{-\eta}}$	glm(family = binomial)

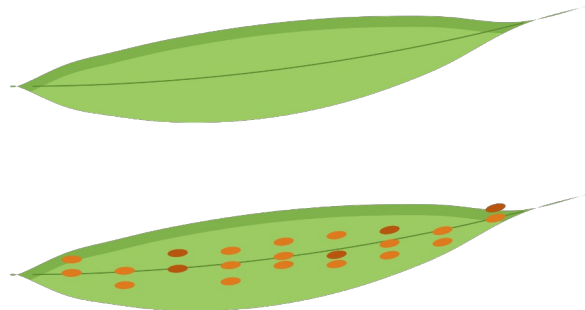
```
> new <- data.frame(nitrogen = c(50, 100, 150))
> predict(m_rn, newdata = new) # log-odds
      1      2      3
-0.9426418 -0.1805881  0.5814655 ← log-odds
> predict(m_rn, newdata = new, type = "response") # probability
      1      2      3
0.2803670 0.4549753 0.6414046 ← P(rust)
```

type = "response" for probabilities. The default is log-odds.

Odds ratio: $\exp(\text{coef})$

y is 0 / 1	example rust on the plot	family Bernoulli (binomial)	link: $\eta =$ $\log \frac{p}{1-p}$	mean: $\mu =$ $\frac{1}{1+e^{-\eta}}$	in R <code>glm(family = binomial)</code>
---------------	-----------------------------	--------------------------------	--	--	---

```
> exp(10 * coef(m_rn)["nitrogen"])          # per 10 kg N
nitrogen
1.164638
> exp(10 * confint(m_rn)["nitrogen", ])
      2.5 %      97.5 %
1.072029 1.276117
```



$$e^{10\beta_1} = e^{10 \times 0.0152} = 1.16 \text{ per 10 kg N}$$

Every 10 kg N multiplies the odds of rust by **1.16 (95% CI 1.07 to 1.28)**

How often is it right?

y is 0 / 1	example rust on the plot	family Bernoulli (binomial)	link: $\eta =$ $\log \frac{p}{1-p}$	mean: $\mu =$ $\frac{1}{1+e^{-\eta}}$	in R glm(family = binomial)
---------------	-----------------------------	--------------------------------	--	--	--------------------------------

```
> pred <- as.numeric(fitted(m_rn) > 0.5)
> table(predicted = pred, observed = trial$rust)
      observed
predicted  0   1
      0  35  16
      1  14  25
> mean(pred == trial$rust)      # accuracy
[1] 0.6666667 ← model: 67% right
> mean(trial$rust == 0)        # always guess "no rust"
[1] 0.5444444 ← always 'no': 54%
```

67% correct, but always guessing "no rust" gets 54%.

A group predictor: glm(rust ~ water)

y is	example	family	link: $\eta =$	mean: $\mu =$	in R
0 / 1	rust on the plot	Bernoulli (binomial)	$\log \frac{p}{1-p}$	$\frac{1}{1+e^{-\eta}}$	glm(family = binomial)

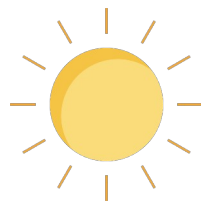
```
> table(trial$rust, trial$water)

  control drought
0       15      34
1       30      11
> m_rw <- glm(rust ~ water, family = binomial, data = trial)
> round(coef(summary(m_rw)), 3)

              Estimate Std. Error z value Pr(>|z|)
(Intercept)    0.693      0.316    2.192   0.028
waterdrought  -1.822      0.469   -3.881   0.000 ← log-odds: drought lower
```



control
30 of 45 rust



drought
11 of 45 rust

Same reading as Part 1, on the **log-odds** scale: intercept = control, waterdrought = difference.

Back to probabilities: plogis()

y is	example	family	link: $\eta =$	mean: $\mu =$	in R
0 / 1	rust on the plot	Bernoulli (binomial)	$\log \frac{p}{1-p}$	$\frac{1}{1+e^{-\eta}}$	glm(family = binomial)

```
> plogis(coef(m_rw)[1])           # control
(Intercept)
0.6666667 ← P(rust), control
> plogis(sum(coef(m_rw)))         # drought
[1] 0.2444444 ← P(rust), drought
> prop.table(table(trial$rust, trial$water), 2)

      control  drought
0 0.3333333 0.7555556
1 0.6666667 0.2444444 ← = the table
```

control: $p = \text{plogis}(0.69) = 0.67$ drought: $p = \text{plogis}(0.69 - 1.82) = 0.24$

The model reproduces the table: **67% vs 24%**

Odds ratio for a group

y is

0 / 1

example

rust on the plot

family

Bernoulli (binomial)

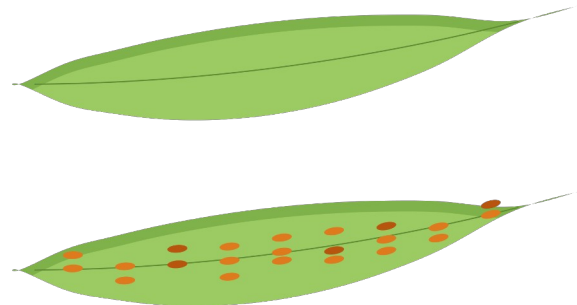
link: $\eta =$

$\log \frac{p}{1-p}$

mean: $\mu =$ in R

$\frac{1}{1+e^{-\eta}}$ `glm(family = binomial)`

```
> exp(coef(m_rw))
(Intercept) waterdrought
2.0000000 0.1617647 ← odds ratio
> exp(confint(m_rw))
                2.5 %    97.5 %
(Intercept) 1.09389806 3.8177252
waterdrought 0.06218335 0.3951711 ← its 95% CI
```



$$\text{odds ratio} = e^{\beta_1} = e^{-1.82} = 0.16$$

Drought multiplies the odds of rust by **0.16 (95% CI 0.06 to 0.40)**

Several predictors

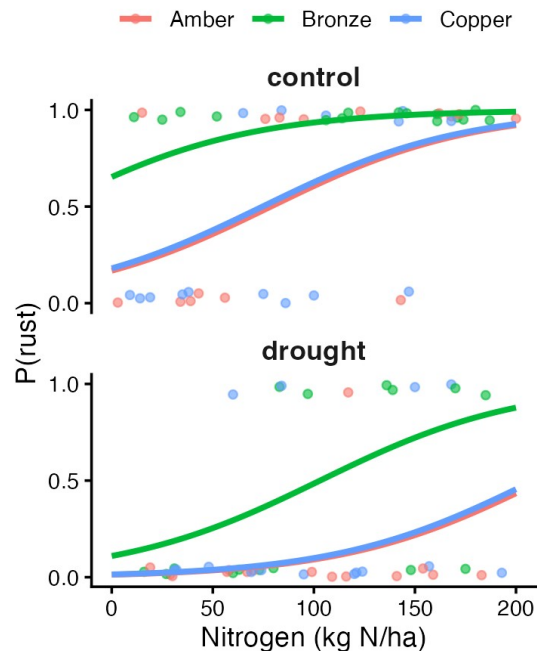
y is	example	family	link: $\eta =$	mean: $\mu =$	in R
0 / 1	rust on the plot	Bernoulli (binomial)	$\log \frac{p}{1-p}$	$\frac{1}{1+e^{-\eta}}$	<code>glm(family = binomial)</code>

```
> m_rall <- glm(rust ~ nitrogen + variety + water,
+               family = binomial, data = trial)
> round(coef(summary(m_rall)), 3)
```

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-1.598	0.718	-2.227	0.026	
nitrogen	0.020	0.006	3.489	0.000	
varietyBronze	2.231	0.768	2.904	0.004	← log-odds
varietyCopper	0.084	0.679	0.123	0.902	
waterdrought	-2.725	0.668	-4.078	0.000	

```
> round(exp(coef(m_rall)), 2)
```

	nitrogen	varietyBronze	varietyCopper	
(Intercept)	0.20	1.02	9.31	← odds ratio x9.3
waterdrought	0.07			



Bronze vs Amber, same nitrogen, same water: **×9.3 the odds of rust.**

Deviance: the GLM's sum of squares

y is 0 / 1	example rust on the plot	family Bernoulli (binomial)	link: $\eta =$ $\log \frac{p}{1-p}$	mean: $\mu =$ $\frac{1}{1+e^{-\eta}}$	in R <code>glm(family = binomial)</code>
---------------	-----------------------------	--------------------------------	--	--	---

```
> m_rall$null.deviance          # intercept only
[1] 124.0544 ← no predictors
> deviance(m_rall)              # with the predictors
[1] 77.89212 ← with predictors
> anova(m_rn, m_rall, test = "Chisq")
Analysis of Deviance Table

Model 1: rust ~ nitrogen
Model 2: rust ~ nitrogen + variety + water
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      88    110.327
2      85     77.892  3   32.435 4.238e-07 *** ← they help
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Compare nested GLMs with
`anova(test = "Chisq")`

$$\text{pseudo-}R^2 = 1 - \frac{D_{\text{residual}}}{D_{\text{null}}} = 1 - \frac{77.9}{124.1} = 0.37$$

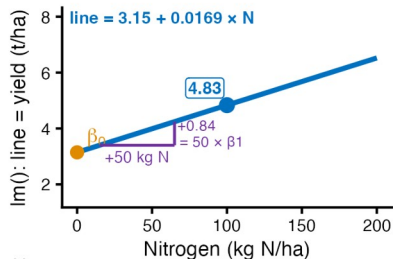
Part 6: One of a family

Logistic regression is a generalized linear model

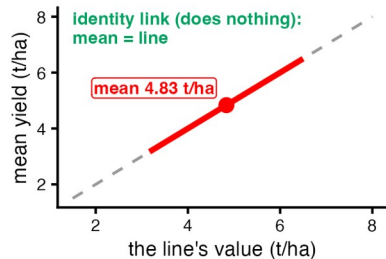


What “generalized” means

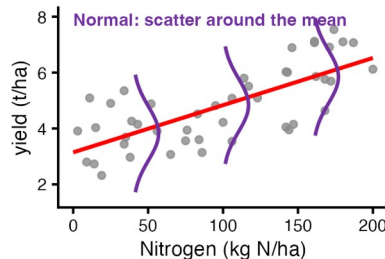
1. the straight line



2. link: line into mean



3. distribution: mean into data



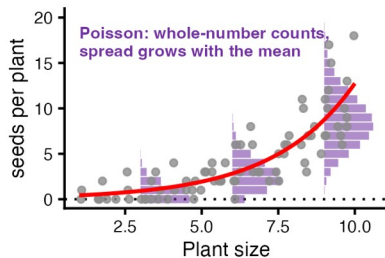
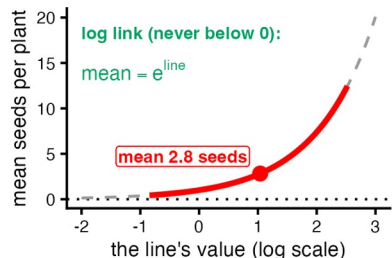
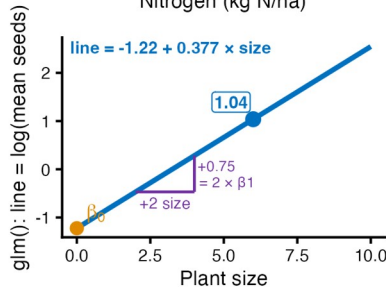
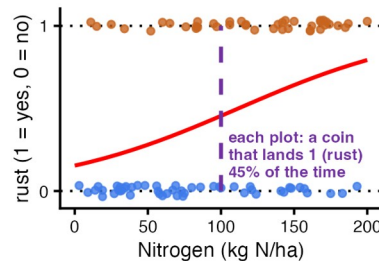
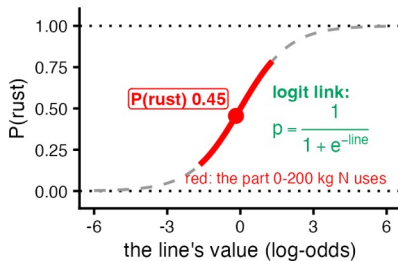
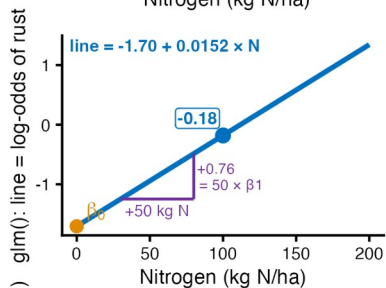
Same straight line.

Generalized = swap the link and the distribution:

yield identity + Normal = lm()

rust logit + binomial

seeds log + Poisson



A GLM has three parts

1. **distribution** $y \sim \text{Binomial}(p)$ or $y \sim \text{Poisson}(\lambda)$

2. **linear predictor** $\eta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$

3. **link** $\log\left(\frac{p}{1-p}\right) = \eta$ or $\log(\lambda) = \eta$

y the response: rust (0 or 1), aphids (a count)

$\sim$ "is drawn from" (in R formulas: "modelled by")

p probability that $y = 1$ (a plot has rust)

λ lambda: the mean count (aphids per plot)

η eta: the straight line, any number

β beta: the coefficients, as in `lm()`

x the predictors (nitrogen, water, ...)

log natural log ($\ln$)

The line lives on the **link scale**, where it can't break the rules.

Choosing the family: numbers

y is	example	family	link: $\eta =$	mean: $\mu =$	in R
any number	yield (t/ha)	Normal	μ	η	<code>lm()</code>
positive, skewed	weed biomass	Gamma (exponential)	$\log \mu$	e^η	<code>glm(family = Gamma("log"))</code>
positive, very skewed	time to germinate	inverse Gaussian	$\log \mu$	e^η	<code>glm(family = inverse.gaussian("log"))</code>
counts 0, 1, 2, ...	seeds / aphids	Poisson	$\log \mu$	e^η	<code>glm(family = poisson)</code>
clumped counts	aphids on 10 tillers	negative binomial	$\log \mu$	e^η	<code>MASS::glm.nb()</code>
counts per effort	lesions on n leaves	Poisson / NB + offset	$\log \mu - \log(\text{effort})$	$\text{effort} \cdot e^\eta$	<code>+ offset(log(effort))</code>
counts, extra zeros	many empty traps	zero-inflated Poisson / NB	$\log \mu$	e^η	<code>pscl::zeroinfl()</code>

η = the straight line, $\beta_0 + \beta_1 x$. μ = the mean of y . **link**: turns μ into η .

Choosing the family: yes/no and categories

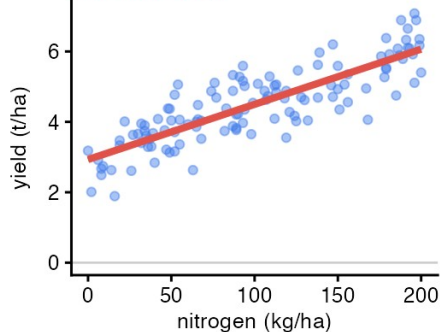
y is	example	family	link: $\eta =$	mean: $\mu =$	in R
0 / 1	rust on the plot	Bernoulli (binomial)	$\log \frac{p}{1-p}$	$\frac{1}{1+e^{-\eta}}$	<code>glm(family = binomial)</code>
k of n	seedlings of seeds sown	binomial	$\log \frac{p}{1-p}$	$\frac{1}{1+e^{-\eta}}$	<code>cbind(k, n - k) ~ x, binomial</code>
proportion, not k of n	% ground cover	beta	$\log \frac{\mu}{1-\mu}$	$\frac{1}{1+e^{-\eta}}$	<code>betareg::betareg()</code>
one of K categories	iris species	multinomial (categorical)	$\log \frac{p_k}{p_1}$	$p_k \propto e^{\eta_k}$	<code>nnet::multinom()</code>
ordered categories	disease score 0–4	ordinal (cumulative logit)	$\log \frac{P(y \leq k)}{P(y > k)}$	per cut-point k	<code>MASS::polr()</code>

yes/no → **binomial** counts → **Poisson** categories → **multinomial**

One recipe, seven kinds of y

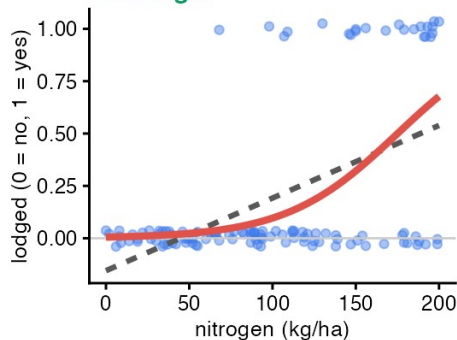
Normal

link: identity



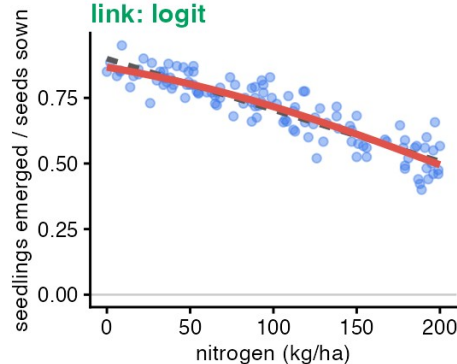
Bernoulli (binomial)

link: logit



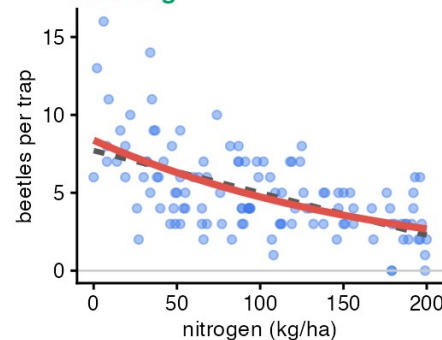
binomial (k of n)

link: logit



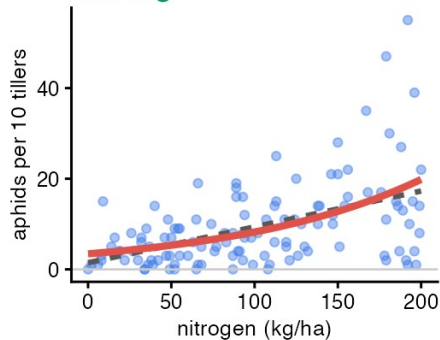
Poisson

link: log



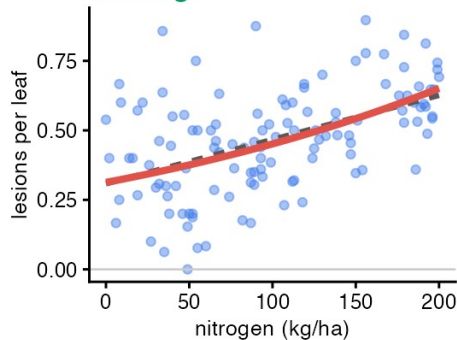
negative binomial

link: log



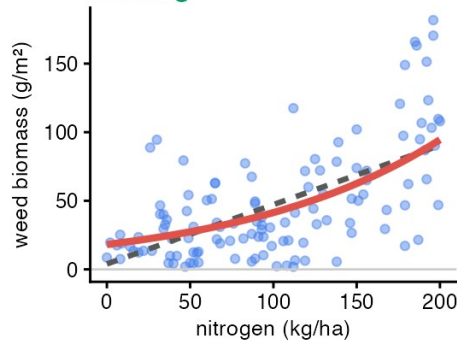
Poisson + offset

link: log



Gamma

link: log



--- $\text{lm}()$: a straight line

— GLM: the straight line, bent by the link

identity link: the two are the same line

Counts out of a total: cbind()

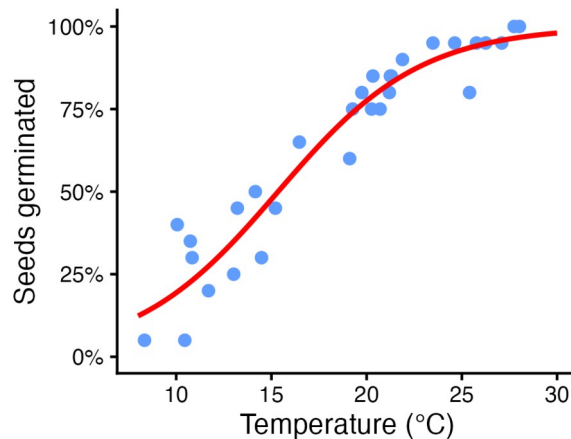
y is
k of n

example
seedlings of seeds
sown binomial

link: $\eta =$
 $\log \frac{p}{1-p}$

mean: $\mu =$ in R
 $\frac{1}{1+e^{-\eta}}$ `cbind(k, n - k) ~ x, binomial`

```
> set.seed(3)
> trays <- data.frame(temp = runif(30, 8, 30)) # 20 seeds per tray
> trays$germ <- rbinom(30, 20, plogis(-4 + 0.25 * trays$temp))
> head(trays, 3)
   temp germ
1 11.69691    4
2 25.76536   19
3 16.46873   13
> m_g <- glm(cbind(germ, 20 - germ) ~ temp, family = binomial,
+            data = trays)
> round(coef(summary(m_g)), 3)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -4.090      0.379  -10.784      0
temp           0.266      0.022   12.281      0 ← log-odds per C
```



cbind(successes, failures)



More than two categories

y is	example	family	link: $\eta =$	mean: $\mu =$	in R
one of K categories	iris species	multinomial (categorical)	$\log \frac{p_k}{p_1}$	$p_k \propto e^{\eta_k}$	<code>nnet::multinom()</code>

```
> library(nnet)
> m_sp <- multinom(Species ~ Petal.Length + Petal.Width,
+                  data = iris, trace = FALSE)
> table(predicted = predict(m_sp), observed = iris$Species)
```

	observed		
predicted	setosa	versicolor	virginica
setosa	50	0	0
versicolor	0	47	3
virginica	0	3	47

← 6 of 150 wrong



setosa, versicolor, virginica

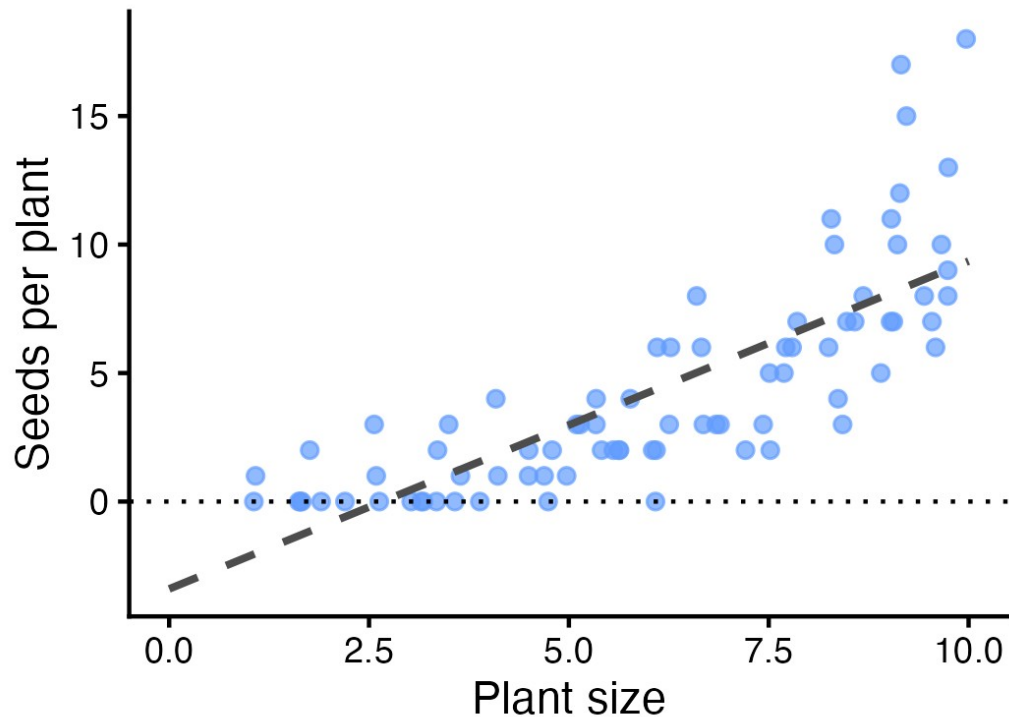
`nnet::multinom()`: one set of log-odds per category vs the reference

Part 7: Counts

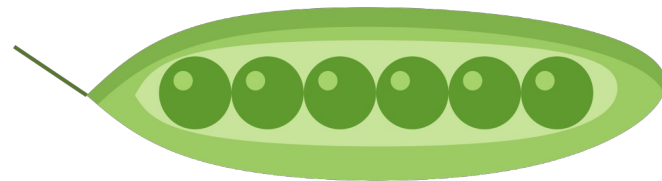
Poisson regression



Last week's Violation 3



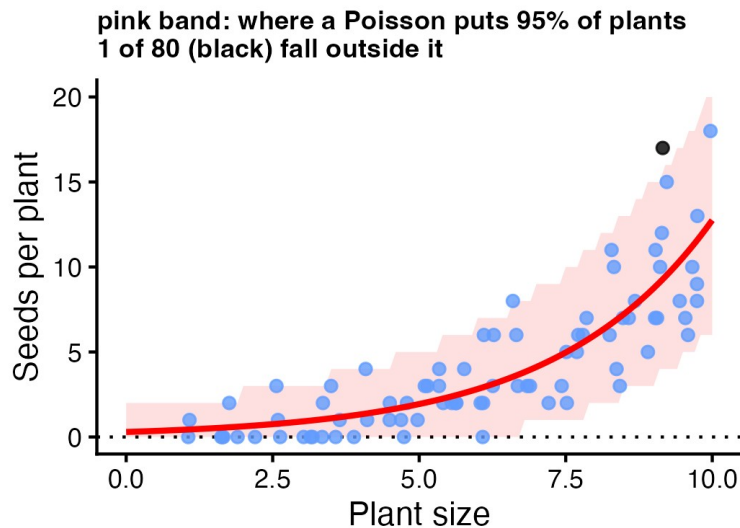
Counts are never negative, and their spread **grows with the mean**.



glm(seeds ~ size, family = poisson)

y is	example	family	link: $\eta =$	mean: $\mu =$	in R
counts 0, 1, 2, ...	seeds / aphids	Poisson	$\log \mu$	e^η	glm(family = poisson)

```
> set.seed(4) # last week's seed counts
> size <- runif(80, 1, 10)
> seeds <- rpois(80, lambda = exp(-1 + 0.35 * size))
> pods <- data.frame(size, seeds)
> m_pod <- glm(seeds ~ size, family = poisson, data = pods)
> round(coef(summary(m_pod)), 3)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.222      0.238   -5.138     0
size           0.377      0.029  13.009     0 ← log scale
> exp(coef(m_pod))
              size
(Intercept)  0.294559 1.457612 ← x1.46 seeds per unit
```



$\text{seeds} \sim \text{Poisson}(\lambda)$ $\log(\lambda) = \beta_0 + \beta_1 \text{ size}$

1 of 80 plants outside the 95% Poisson band: a **Poisson** fits.

Reading a Poisson slope: exp()

y is	example	family	link: $\eta =$	mean: $\mu =$	in R
counts 0, 1, 2, ...	seeds / aphids	Poisson	$\log \mu$	e^η	glm(family = poisson)

```
> new <- data.frame(size = c(5, 6, 7))
> mu <- predict(m_pod, new, type = "response")
> mu
      1      2      3
1.938126 2.825035 4.117805 ← mean seeds
> mu[2] / mu[1]
      2
1.457612 ← size 6 vs 5
> mu[3] / mu[2]
      3
1.457612 ← 7 vs 6: same ratio
> exp(coef(m_pod)["size"])
      size
1.457612 ← = exp(slope)
```

Multiplicative: each unit of size, **+46% seeds**. The same ratio at every size: a rate ratio.

$$e^{\beta_1} = e^{0.377} = 1.46 \Rightarrow \times 1.46 \text{ seeds per unit of size}$$

Aphids: does a Poisson fit?

y is	example	family	link: $\eta =$	mean: $\mu =$	in R
counts 0, 1, 2, ...	seeds / aphids	Poisson	$\log \mu$	e^η	<code>glm(family = poisson)</code>

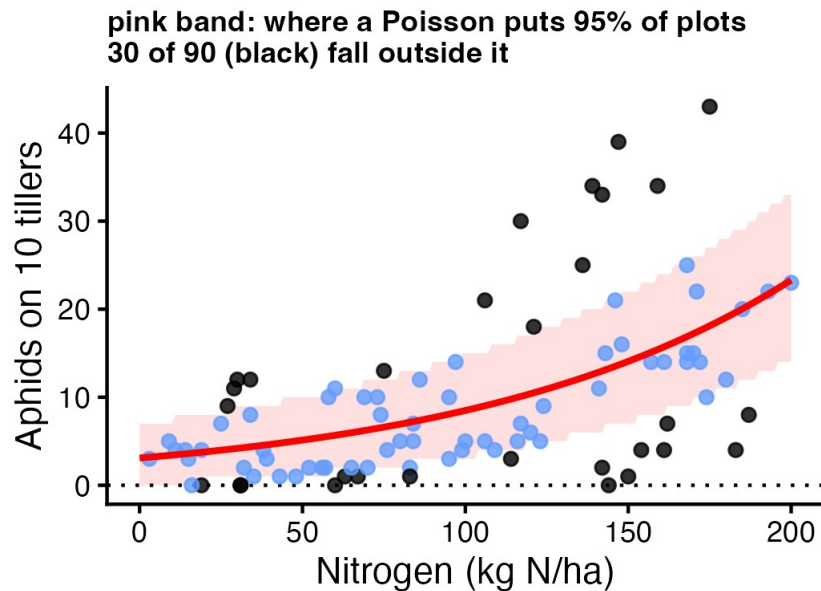
```
> m_aph <- glm(aphids ~ nitrogen, family = poisson, data = trial)
> round(coef(summary(m_aph)), 4)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.1307    0.0914  12.3743     0
nitrogen      0.0101    0.0007  15.1468     0 ← log scale, per kg N
```

A Poisson assumes **variance = mean**

so at each mean it allows only a narrow band (pink).

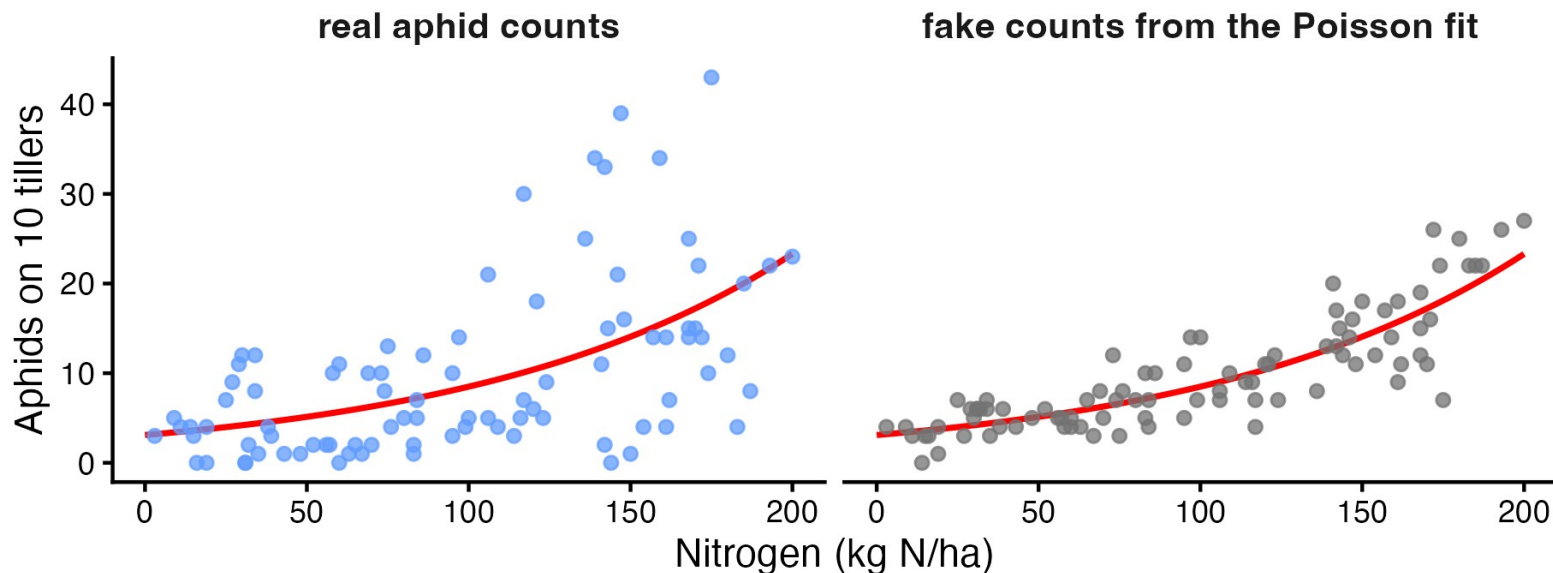
30 of 90 plots fall outside a band that should hold 95%. (Seeds: 1 of 80.)

No: a Poisson doesn't fit.



Could a Poisson have made the aphid data?

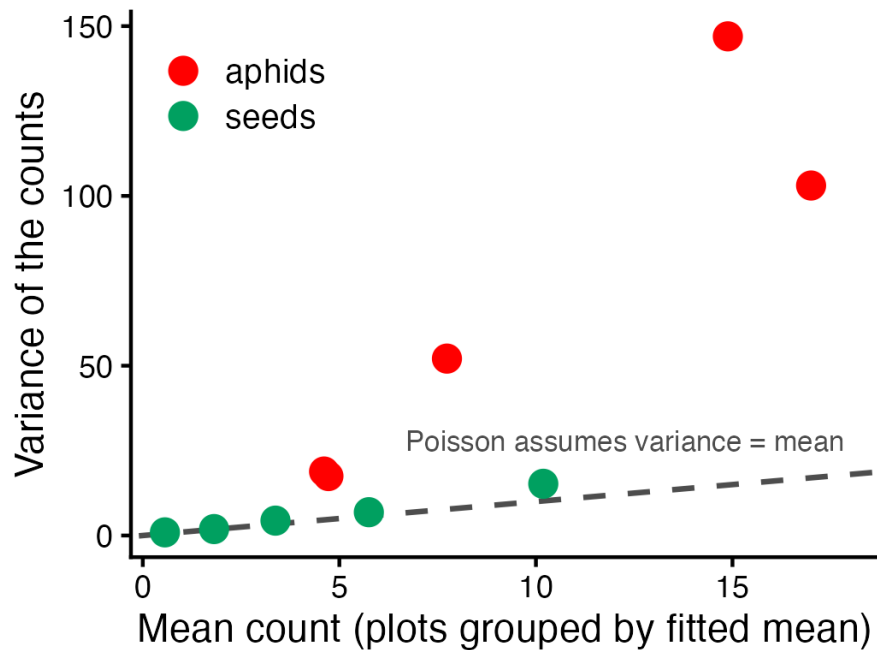
y is	example	family	link: $\eta =$	mean: $\mu =$	in R
counts 0, 1, 2, ...	seeds / aphids	Poisson	$\log \mu$	e^η	<code>glm(family = poisson)</code>



Same mean curve. Real aphids are **far more variable** than Poisson aphids: **overdispersion**

Overdispersion: check it every time

y is	example	family	link: $\eta =$	mean: $\mu =$	in R
counts 0, 1, 2, ...	seeds / aphids	Poisson	$\log \mu$	e^η	<code>glm(family = poisson)</code>



```
> deviance(m_pod) / df.residual(m_pod) # seeds  
[1] 1.187493 ← about 1: fine  
> deviance(m_aph) / df.residual(m_aph) # aphids  
[1] 5.69656 ← 5.7x too spread out
```

$$\frac{\text{residual deviance}}{df} \approx 1 \quad \text{seeds: } 1.19 \quad \text{aphids: } 5.70$$

Aphids: 5.7× more variance than a Poisson allows
so SEs are too small by $\sqrt{5.7} \approx 2.4\times$

Fix: negative binomial

y is	example	family	link: $\eta =$	mean: $\mu =$	in R
clumped counts	aphids on 10 tillers	negative binomial	$\log \mu$	e^η	MASS: :glm.nb()

```
> library(MASS)
> m_nb <- glm.nb(aphids ~ nitrogen, data = trial)
> round(coef(summary(m_apb)), 4) # Poisson
      Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.1307    0.0914 12.3743     0
nitrogen      0.0101    0.0007 15.1468     0 ← Poisson SE: too small
> round(coef(summary(m_nb)), 4) # negative binomial
      Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.1020    0.1872  5.8861     0
nitrogen      0.0103    0.0016  6.4848     0 ← honest SE
```

Same slope, **honest standard error**: 2.4× wider, as $\sqrt{5.7}$ predicted

Reporting a GLM

“The odds of rust increased with nitrogen (**odds ratio 1.16 per 10 kg N, 95% CI 1.07 to 1.28**; $z = 3.5$, $p < 0.001$; logistic regression, $n = 90$).”

“Aphid counts increased with nitrogen (**rate ratio 1.11 per 10 kg N, 95% CI 1.07 to 1.14**; $z = 6.5$, $p < 0.001$; negative binomial GLM, $n = 90$).”

`exp(coef)` and `exp(confint)`, **on a step size that means something.**

Rough summary and guide

y numeric, Normal scatter

x is a group

do groups differ?

y is 0/1

y is k out of n

y is a count

read a GLM

lm()

dummies: intercept = reference

anova(); TukeyHSD(), emmeans

glm(family = binomial)

glm(cbind(k, n - k) ~ ...)

glm(family = poisson); check dispersion

exp(coef): odds or rate ratio

Week 3 showdown

Kahoot-style showdown

12 questions on this week.

Not graded.

Use a pseudonym if you want.

Fast and correct wins.



Scan with your phone

[Join on your phone or laptop: greymonroe.github.io/live-poll-test/quiz-play.html?quiz=pls206-w3-showdown](https://greymonroe.github.io/live-poll-test/quiz-play.html?quiz=pls206-w3-showdown)

[host screen](#)

Bonus: mixed effect models

When plots come in groups



Where this goes next

	independent	grouped (blocks, farms, repeats)
Normal y	<code>lm()</code>	<code>lmer()</code>
yes/no, counts	<code>glm()</code>	<code>glmer()</code>

Same formulas, plus `(1 | farm)`. lme4 package.

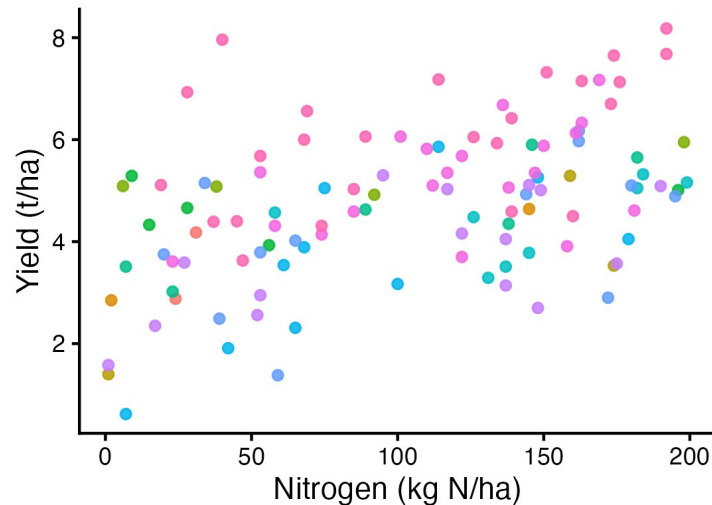
112 plots on 12 farms

```
> head(farms, 4)
  farm cover nitrogen yield
1 farm01  no      24  2.88
2 farm01  no      31  4.18
3 farm02  yes     145  4.64
4 farm02  yes       2  2.85
> table(farms$farm)

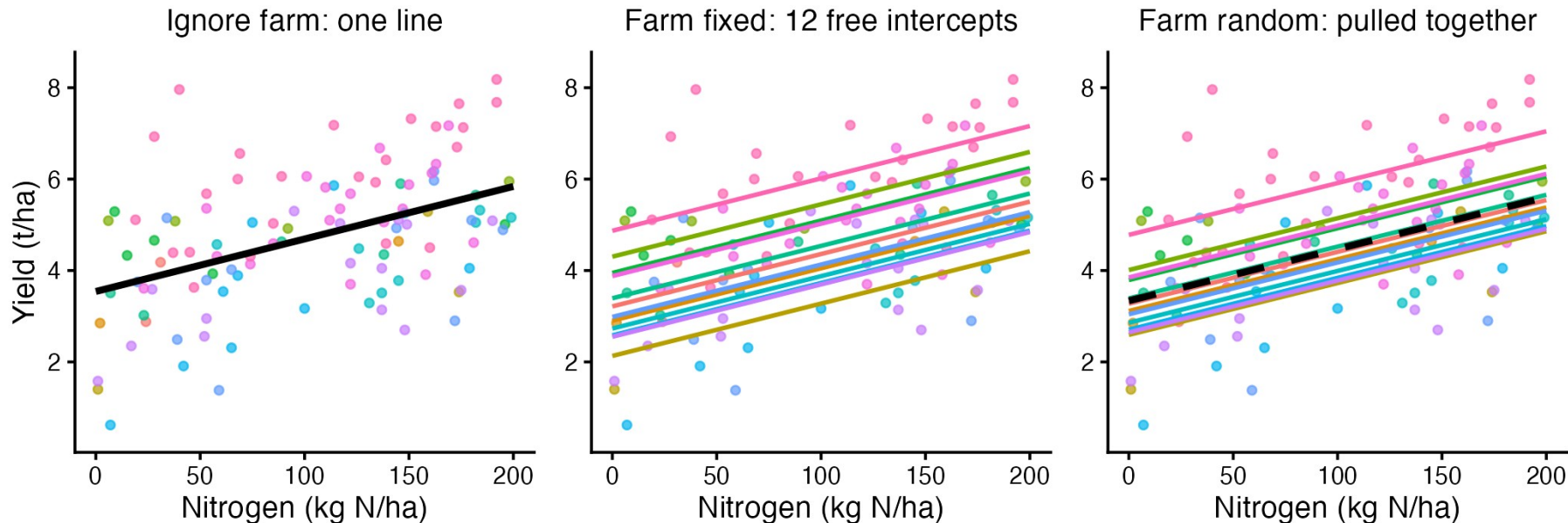
farm01 farm02 farm03 farm04 farm05 farm06 farm07 farm08 farm09
     2      2      3      4      5      6      8     10     12  ← 2 to 25 plots
farm10 farm11 farm12
    15     20     25
```



Each farm has its own baseline. 2 to 25 plots per farm.



Three ways to treat farm



Fixed: every farm on its own. **Random: farms are a sample of all farms.**

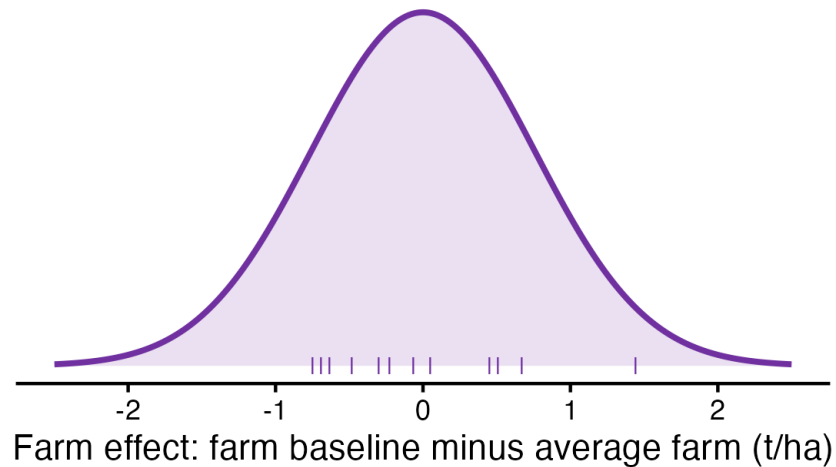
What “random” means

$$\text{yield}_{ij} = \beta_0 + u_j + \beta_1 \text{nitrogen}_{ij} + \varepsilon_{ij}$$

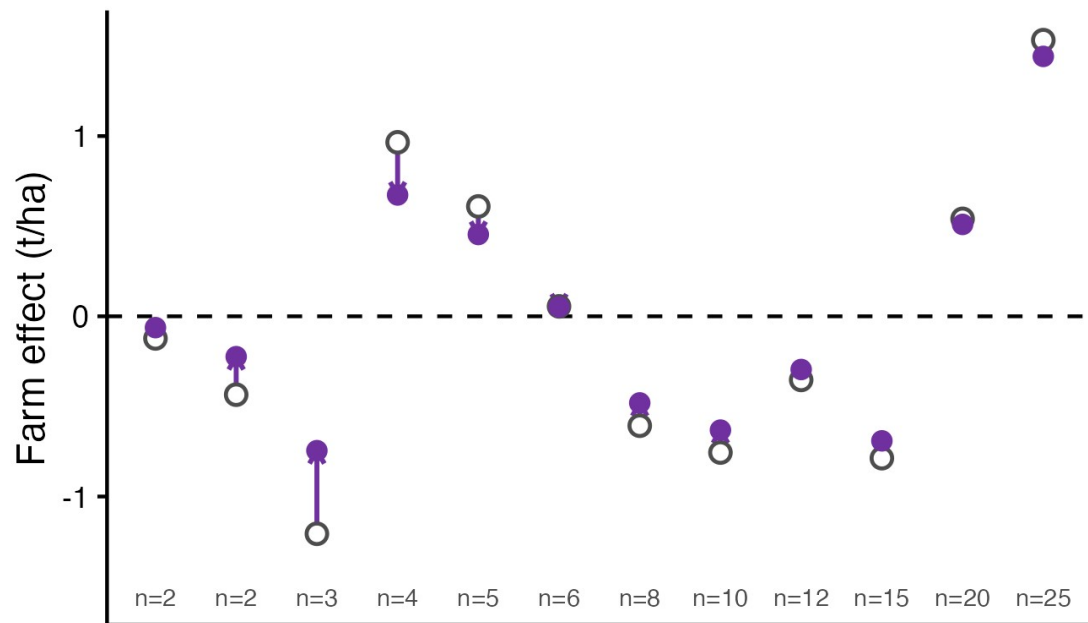
$$u_j \sim N(0, \sigma_{\text{farm}}^2) \quad \varepsilon_{ij} \sim N(0, \sigma^2)$$

Fixed: estimate 11 farm differences.

Random: estimate one spread, $\sigma_{\text{farm}} = 0.76$ t/ha



Shrinkage: small farms borrow strength



$$\hat{u}_j = \frac{\sigma_{\text{farm}}^2}{\sigma_{\text{farm}}^2 + \sigma^2/n_j} \times \text{farm's own estimate}$$

weight: $n = 2$: 0.53 $n = 25$: 0.93

In R: lmer()

```
> library(lme4)
> m_mixed <- lmer(yield ~ nitrogen + (1 | farm), data = farms)
> VarCorr(m_mixed)
Groups   Name             Std.Dev.
farm     (Intercept) 0.75921 ← farm-to-farm SD
Residual                               1.01835
> round(coef(summary(m_mixed)), 4)
              Estimate Std. Error t value
(Intercept)   3.3377    0.3001  11.1219
nitrogen      0.0113    0.0017   6.6802 ← same slope
```

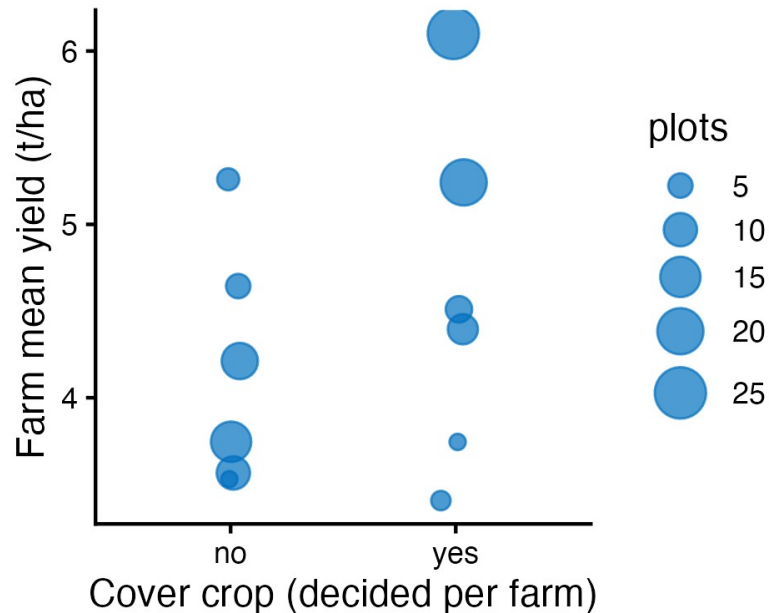
Same nitrogen slope:

0.0113 vs 0.0115

2 variances instead of 12
intercepts

```
> m_fixed <- lm(yield ~ nitrogen + farm, data = farms)
> length(coef(m_fixed)) # coefficients
[1] 13 ← 12 farms + slope + 1
> round(coef(summary(m_fixed))["nitrogen", ], 4)
      Estimate Std. Error    t value    Pr(>|t|)
      0.0115    0.0017     6.6110     0.0000 ← same slope
```

Why it matters: a treatment chosen per farm



```
> m_plots <- lm(yield ~ nitrogen + cover, data = farms)
> round(coef(summary(m_plots))["coveryes", ], 4)
Estimate Std. Error t value Pr(>|t|)
1.0139 0.2398 4.2289 0.0000 ← lm: 'significant'
> m_mixed2 <- lmer(yield ~ nitrogen + cover + (1 | farm), data = farms)
> round(coef(summary(m_mixed2))["coveryes", ], 4)
Estimate Std. Error t value
0.2380 0.5142 0.4627 ← mixed: nothing
```

lm(): $t = 4.2$, $p < 0.001$. lmer(): $t = 0.46$. **The truth: no effect.**

Assumptions

everything from `lm()`

farm effects

groups

enough groups

check

linear, equal variance, Normal residuals

Normal, centred on 0

independent of each other

5 or more to estimate a spread

```
plot(model); qqnorm(ranef(model)$farm[, 1])
```

Fixed or random?

you care about each level

levels sample a population

only 2 to 4 levels

treatment applied to whole groups

y is 0/1 or a count

fixed (these 3 varieties)

random (farms, blocks, years)

fixed

group must be random

glmer(..., family = binomial)

Full session: **Friday December 4**

Next week: Non-linear regression & GAMs

Problem Set 3 due Monday, October 19, 1:10 PM • Friday: data wrangling
(data.table and tidyverse)