Stat 412/512

Another Multifactor Study

Feb 16 2015

Charlotte Wickham

stat512.cwick.co.nz

Without replication

The estimate of σ depends on the model chosen.

We can't estimate a saturated model.

$df = n - p = n - n = 0$

$\hat{\sigma} = \sqrt{\frac{RSS}{df}}$

Buy back some degrees of freedom by:

- assuming interactions don't exist fri
- some numerical categorical variables can be modelled as continuous today

case1402: Soybeans and ozone

Total of 30 combinations of treatment: Ozone (5 levels), sulfur dioxide (three levels) and water stress (two levels).

One chamber per treatment. I.e. one replicate!

no replicates

Two soybean cultivars in each chamber (we'll analyse them separately)

Response: yield in kg/hectare

Seed yields for soybean cultivars *Forrest* and *Williams* from chambers kept under varying conditions of ozone, sulphur dioxide and water stress

Water Stress	SO_2 (µL/L)	O_3 (µL/L)	Yields (kg/ha) Forrest	Williams
Well-Watered (WW) (-0.05 MPa)	0.0045	0.017	4376	5561
		0.049	4544	5947
		0.067	2806	4273
		0.084	3339	3470
		0.099	3320	3080
	0.0170	0.017	3747	5092
		0.049	4570	4752
		0.067	4635	4232
		0.084	3613	2867
		0.099	3259	3106
	0.0590	0.017	4179	4736
		0.049	5077	3672
		0.067	3401	3386
		0.084	3371	2854
		0.099	2158	2557
Soil Moisture Stress (SMS) (-0.40 MPa)	0.0045	0.017	4977	4520
		0.049	3780	3047
		0.067	3804	3526
		0.084	3941	3357
		0.099	2863	2663
	0.0170	0.017	5573	4869
		0.049	3555	3774
		0.067	3340	2955
		0.084	3243	3513
		0.099	2802	2838
	0.0590	0.017	4589	4056
		0.049	3250	2758
		0.067	3045	3094
		0.084	2827	2398
		0.099	2979	2101

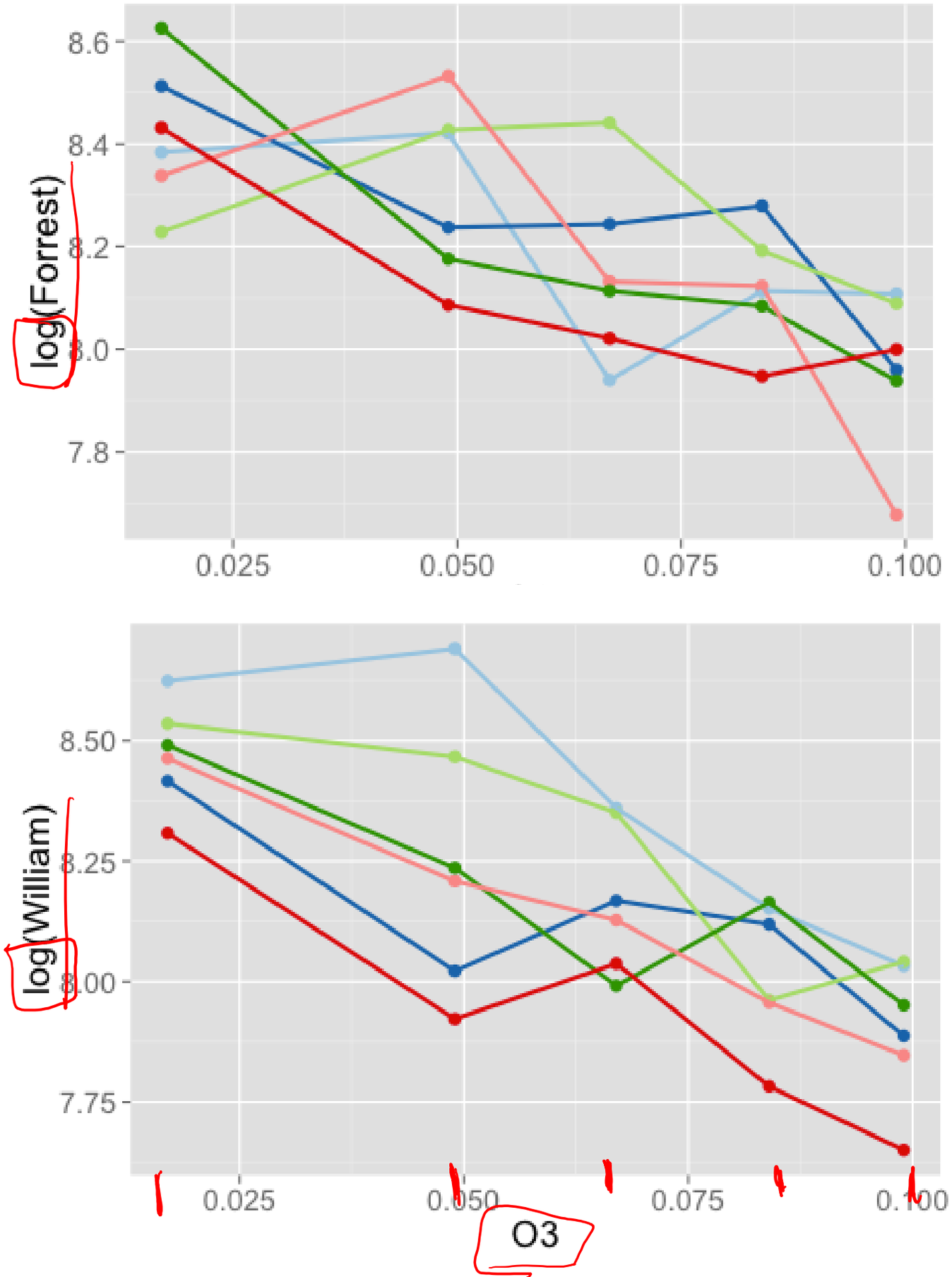
Questions

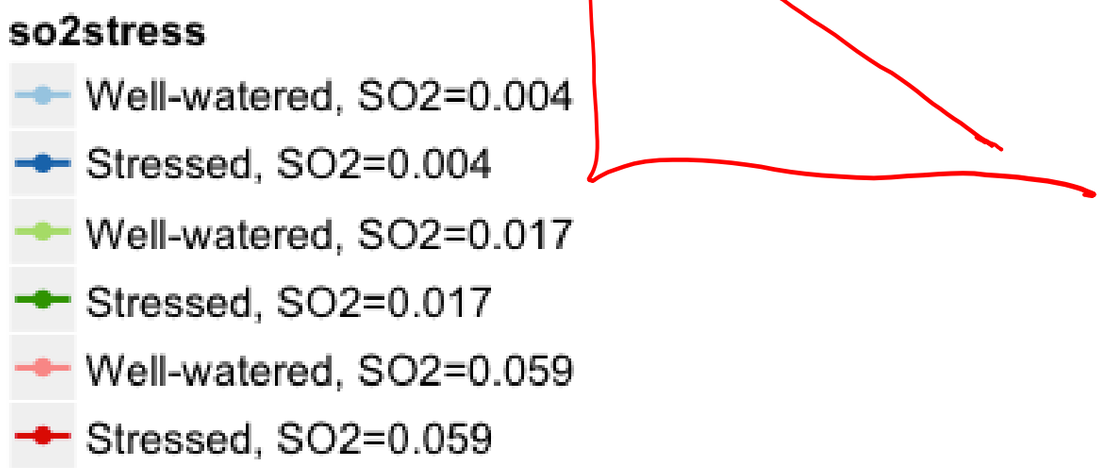
Do the three factors interact?

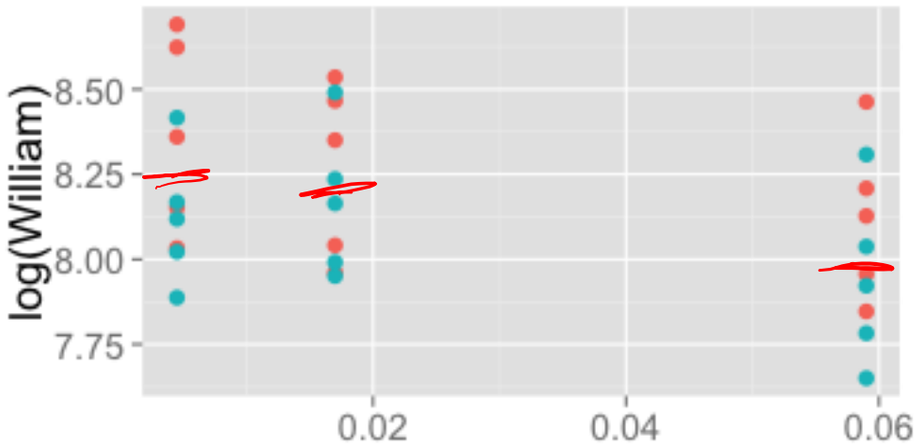
can't assume them away because they are what we are interested in

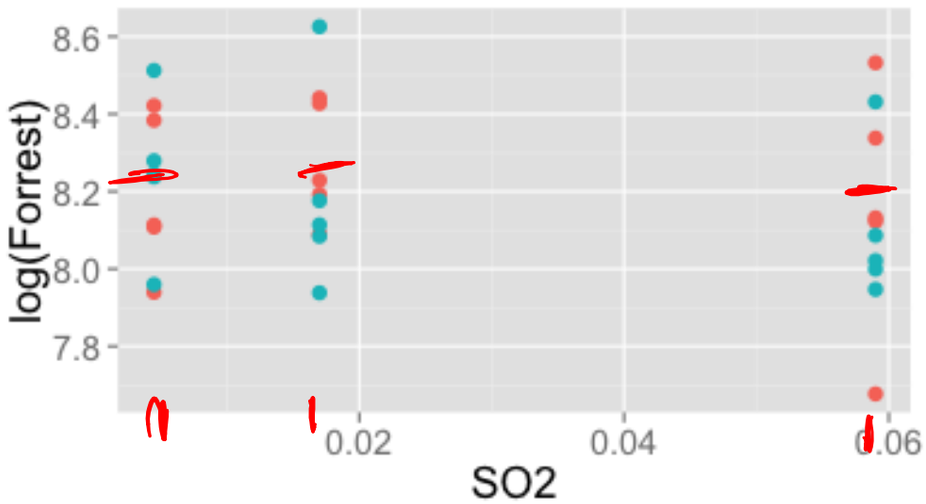
Does water stress affect yield?

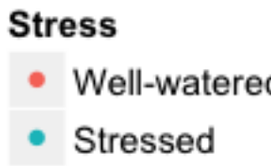
Are there differences between the cultivars?


log(Forrest)
8.6
8.4
8.2
8.0
7.8
0.025
0.050
0.075
0.100
log(William)
8.50
8.25
8.00
7.75
O3


so2stress
Well-watered, SO2=0.004
Stressed, SO2=0.004
Well-watered, SO2=0.017
Stressed, SO2=0.017
Well-watered, SO2=0.059
Stressed, SO2=0.059


log(William)
8.50
8.25
8.00
7.75
0.02
0.04
0.06


log(Forrest)
8.6
8.4
8.2
8.0
7.8
0.02
0.04
0.06
SO2


Stress
Well-watered
Stressed


Stress
Well-watere
Stressed

Your turn

What would the saturated model be:

μ {log Yield | SO2, O3, Water } =

$\beta_0 + \beta_1$ (full model form)

Intl

$SO_2 + O_3 + WATER + SO_2 \times O_3 + SO_2 \times WATER + O_3 \times WATER +$

$SO_2 \times O_3 \times WATER$

(in shorthand)

SO_2 - 3 levels WATER - 2 levels

O_3 - 5 levels

1 + 2 + 4 + 1 + 2 + 4 + 8 + 2 + 4 + 8 = 30 parameters

Tentative model

Looks like we can save some parameters by modeling ozone as continuous

(1 slope parameter, instead of 4 parameters on indicators)

μ {log Yield | SO2, O3, Water } =

SO2 + WATER + ozone + SO2×ozone + WATER×SO2

+ WATER×ozone + WATER×SO2×ozone

SO2 – 3 levels

Water – 2 levels

Your turn

μ {log Yield | SO2, O3, Water } =

SO2 + WATER + ozone + SO2×ozone + WATER×SO2

+ WATER×ozone + WATER×SO2×ozone

1 + 2 + 1 + 1 + 2×1 + 1×2

1×1 1×2×1

How many parameters are in this model?

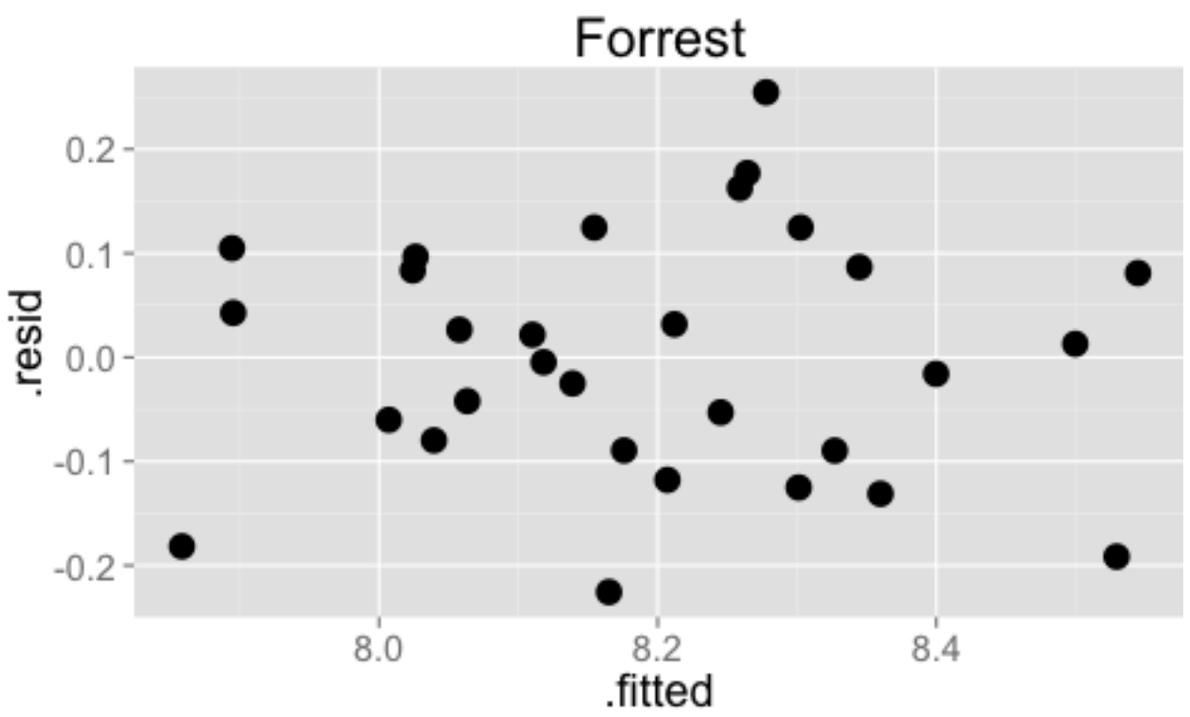
11

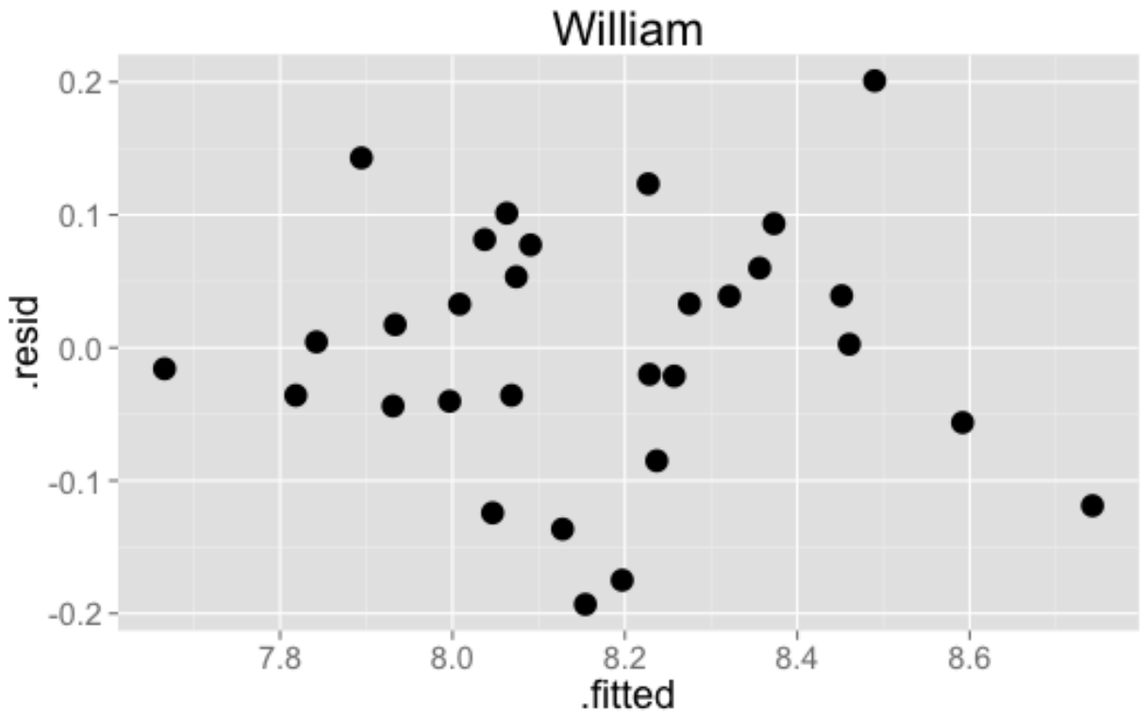
12

ozone

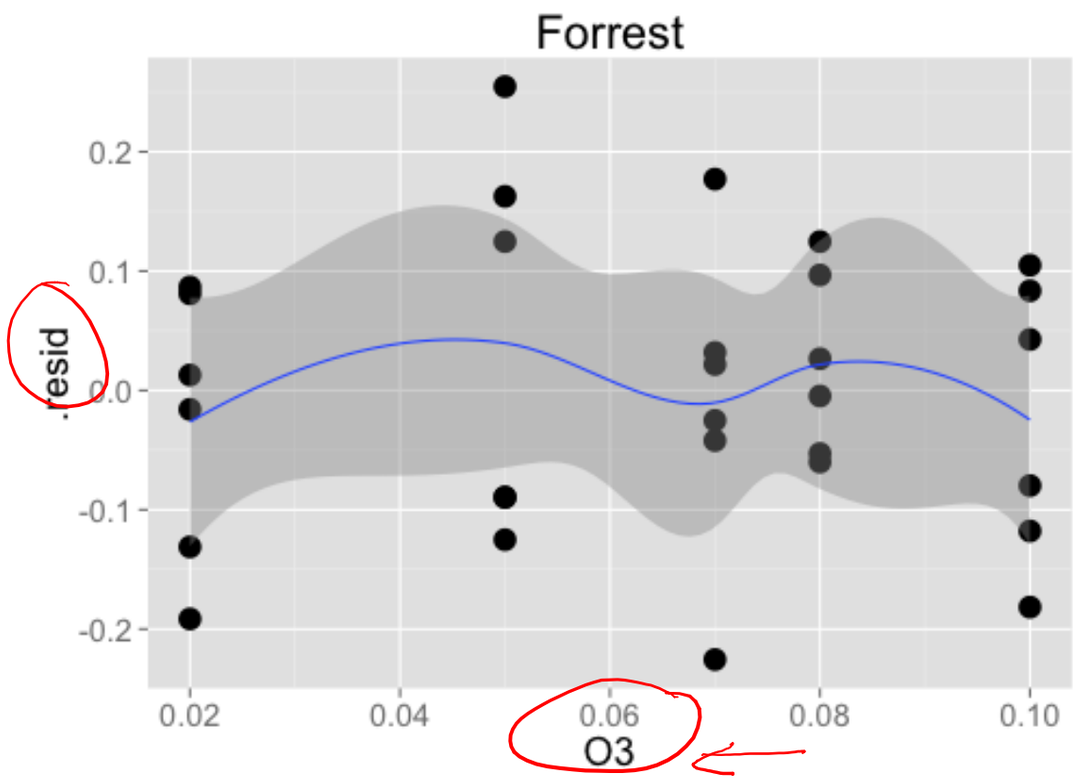
ozone + ozone^2

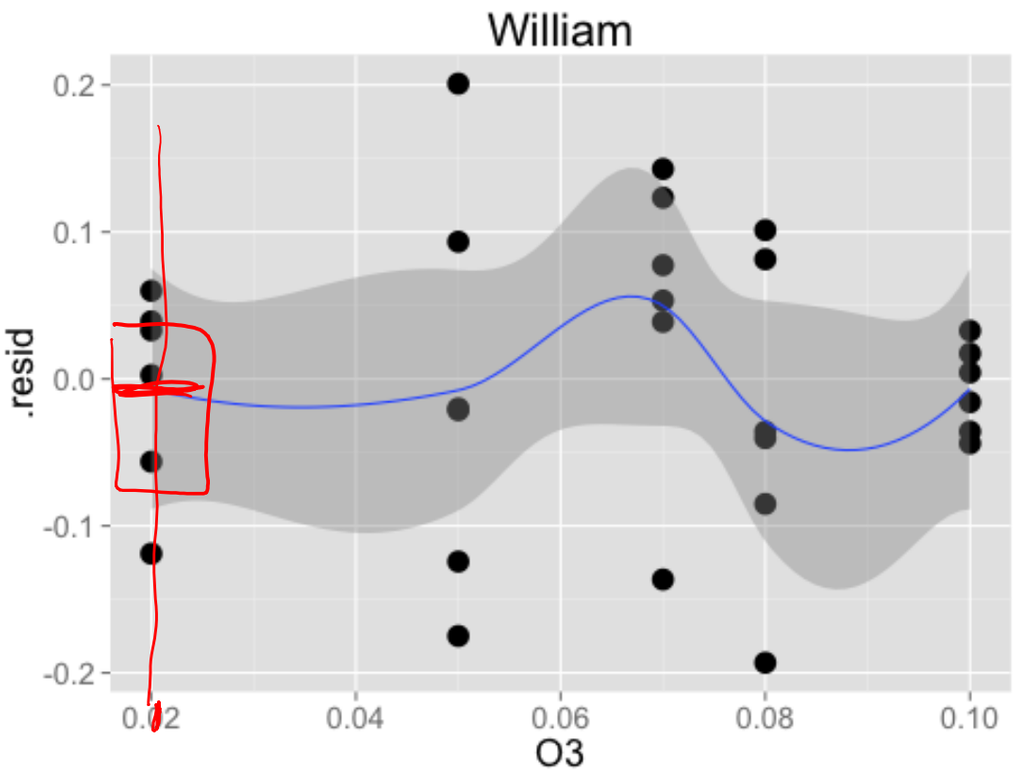
Residuals from tentative model

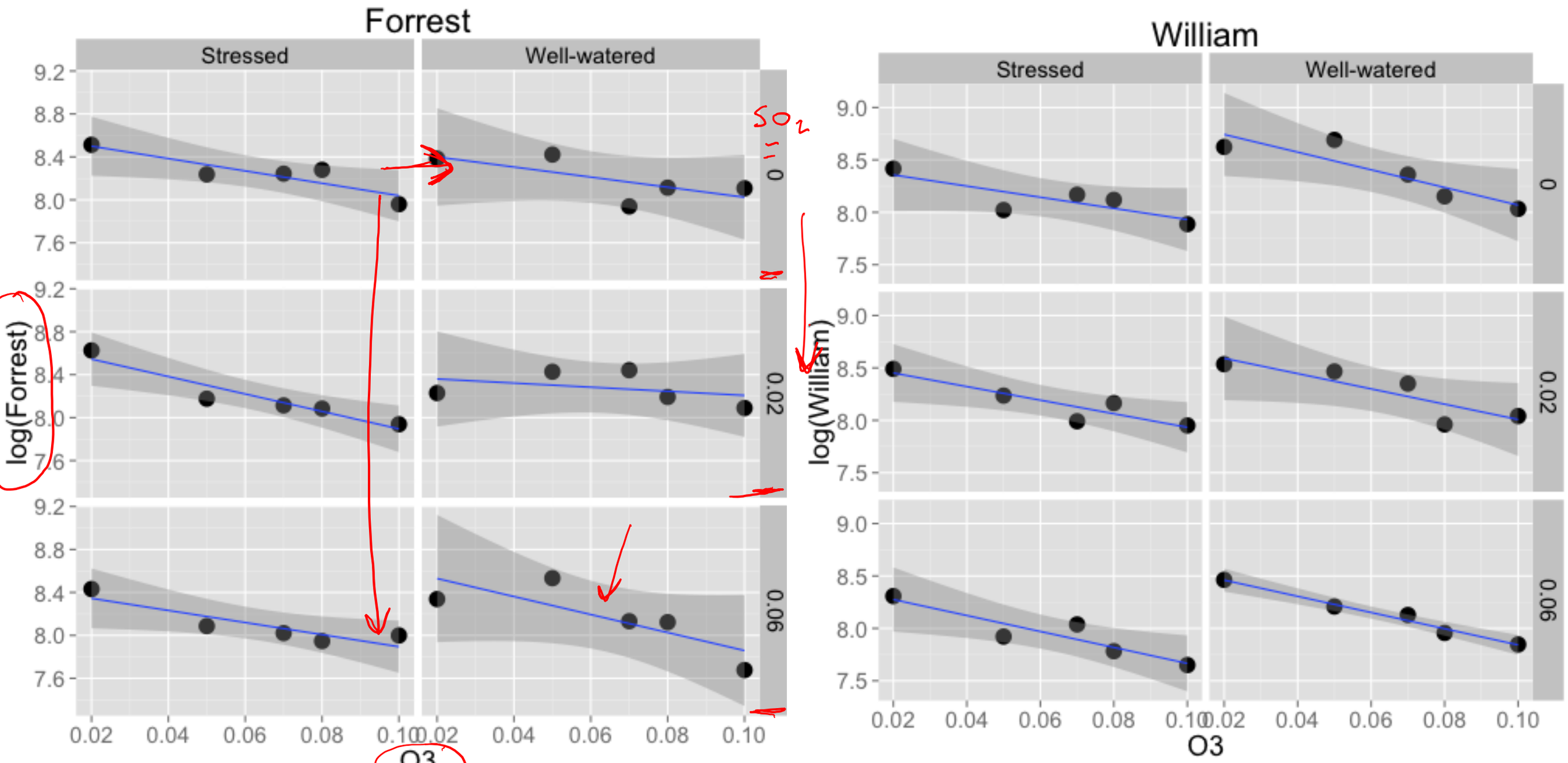




Checking linearity of Ozone







Ozone, SO2 interaction would mean a different slope for ozone at each level of SO2.

Ozone, water stress interaction would mean a different slope for ozone at each level of stress.

A ozone, SO2, water interaction, allows the effect of sulphur on the relationship between yield and ozone to depend on water stress.

Analysis of variance tables for screening effects on log soybean seed yield

Forrest Cultivar

Source	df	Sum of Squares	Mean Square	F-stat	p-value
ozone	1	.7208	.7208	30.050	<.0001
SULPHUR	2	.0635	.0317	1.356	.2827
water	1	.0080	.0080	.343	.5651
ozone×SULPHUR	2	.0173	.0087	.370	.6956
ozone×water	1	.0136	.0136	.583	.4552
SULPHUR×water	2	.0285	.0143	.610	.5543
ozone×SULPHUR×water	2	.0683	.0342	1.461	.2583
residuals	18	.4211	.0234		

Williams Cultivar

Source	df	Sum of Squares	Mean Square	F-stat	p-value
ozone	1	1.150	1.150	86.7999	<.00001
SULPHUR	2	.2780	.1390	10.495	.001
water	1	.2376	.2376	17.943	.0005
ozone×SULPHUR	2	.0037	.0019	.140	.8702
ozone×water	1	.0128	.0128	.964	.3392
SULPHUR×water	2	.0263	.0131	.994	.3896
ozone×SULPHUR×water	2	.0093	.0047	.352	.7079
residuals	18	.2384	.0132		

Full model is: μ {log Yield | SO2, O3, Water } =

SO2 + WATER + ozone + SO2×ozone + WATER×SO2 + WATER×ozone + WATER×SO2×ozone

I think it's easier to think about simplifying the model

three way interaction?

Compare:

SO2 + WATER + ozone + SO2×ozone + WATER×SO2 + WATER×ozone +

WATER×SO2×ozone

big p-value

to

SO2 + WATER + ozone + SO2×ozone + WATER×SO2 + WATER×ozone

two way interactions?

Compare:

SO2 + WATER + ozone + SO2×ozone + WATER×SO2 + WATER×ozone

to

SO2 + WATER + ozone

big p-value

model SO2 as linear too (after checking with a test).

Coefficient estimates and standard errors for the linear soybean models, with Y = log(soybean seed yield)

Variable	Forrest Coefficient	Forrest St. Error	Forrest 2-Sided p-Value	Williams Coefficient	Williams St. Error	Williams 2-Sided p-Value
CONSTANT	8.608	0.080		8.825	0.058	
ozone	-5.397	0.929	<.0001	-6.806	0.679	<.0001
sulphur	-1.566	0.989	.1252	-3.512	0.723	<.0001
water	0.094	0.153	.5453	0.507	0.112	.0001

1page summary in Sleuth.

Includes estimates of possible size of interactions (even though we have no evidence for them).

I.e. The ozone effect when SO2 is 0.0590 is estimated to be only 14.7% of the ozone effect when SO2 is 0.0045 (95% CI, 0.16% and 1365%)

2ndEd

I.e. The ozone effect when SO2 is 0.0590 is estimated to be only 14.7% of the ozone effect when SO2 is 0.0045 (95% CI, 0.16% and 1365%)

Fit: μ {log Yield | SO2, O3, Water }
= so2+ WATER + ozone + so2×ozone
= $\beta_0 + \beta_1 ozone + \beta_2 so2 + \beta_3 water + \beta_4 ozone \times so2$

```
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       8.54408    0.10229  83.528  < 2e-16 ***
SO2               0.34845    2.77868   0.125  0.90121
StressStressed   -0.03274    0.05364  -0.610  0.54718
O3               -4.50785    1.42371  -3.166  0.00404 **
SO2:O3          -34.81736   40.07680  -0.869  0.39324
```

Slope on ozone is : $\beta_1 + \beta_4 so2$
Slope when SO2 is 0.0590: $\beta_1 + \beta_4 0.0590$
Slope when SO2 is 0.0045: $\beta_1 + \beta_4 0.0045$
Difference in slope: $\beta_4 (0.0590 - 0.0045)$

2ndEd

exp(-34.81736*(0.0590 - 0.0045)) = 0.1499361

Difference in slope: β_4 (0.0590 - 0.0045)

```
> confint(fit_for_int)
                          2.5 %       97.5 %
(Intercept)           8.3334107  8.75475270
SO2                  -5.3743554  6.07125724
StressStressed       -0.1432094  0.07773829
O3                   -7.4400394 -1.57565236
SO2:O3             -117.3570749 47.72235133
```

```
exp(c(-117.3570749, 47.7223513) * (0.0590 - 0.0045))
[1]  0.001668283 13.475431598
```

I.e. The ozone effect when SO2 is 0.0590 is estimated to be only 14.7% of the ozone effect when SO2 is 0.0045 (95% CI, 0.16% and 1365%)

3rd ed

I.e. The **detrimental effect of increasing ozone by 0.01** when SO2 is 0.0590 is estimated to be 1.9% greater than when SO2 is 0.0045 (95% CI, 2.6% smaller and 6.7% larger)

A much more sensible comparison!

exp(-34.81736*(0.0590 - 0.0045)***0.01**) = 0.981

To replicate or not?

If interactions are of interest, then replicate!

When experimental units are expensive, you can sometimes gain more by reducing variability, than increasing your replicates.

SE of cell average = σ /√number in cell

reduce this

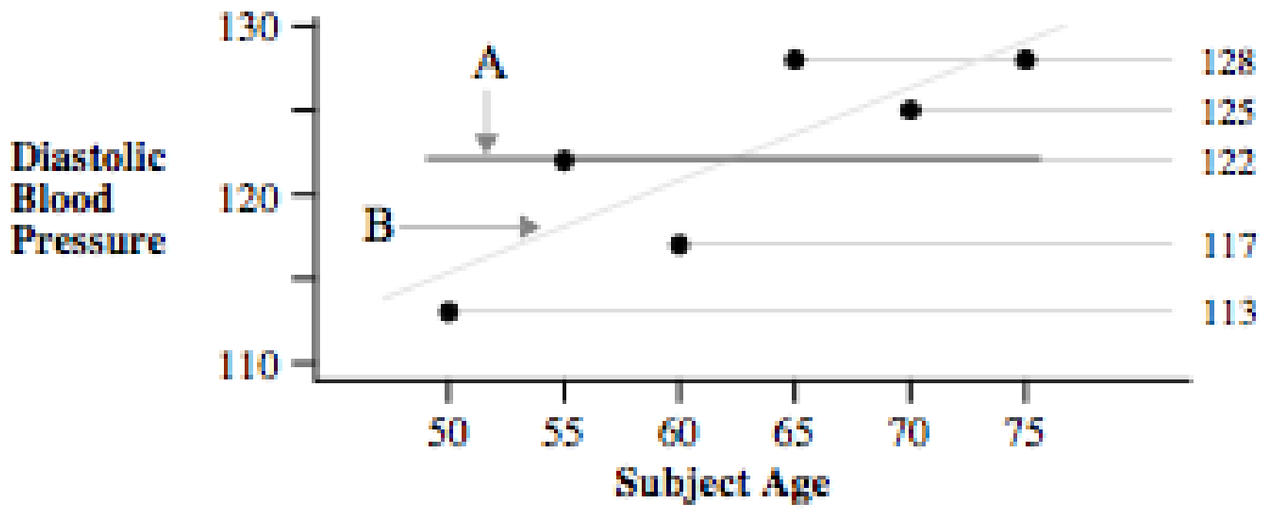
or increase this

Hypothetical example

New method for reducing high blood pressure.

Blood pressure tends to depend on age.

With no treatment the researcher expects something like ->



Option 1:

Ignore age, randomly assign treatment to six people aged 50 to 75. Six replicates, can make causal inferences. Expected variability = 6.1

Option 2:

Pick 6 people of the same age, randomly assign to treatment. Six replicates, can still make causal inferences (to a much reduced population). Expected variability = 3.8

Option 3:

Pick 6 people from 50-75 but pair them by similar ages, within each pair randomly assign to treatment (i.e. block by age). No replicates, can still make causal inferences. Expected variability ~ closer to 3.8

Lesson: Include important sources of variation in the design.

Identifying false replicates

a.k.a pseudo replication

The replication needs to be at the level of experimental unit (the items that are randomly assigned to treatment).

The replicates need to be **independent** applications of the same treatment.

Examples

Pygmalion study: platoon was randomized to treatment. It would be inappropriate to treat individual soldiers scores on the test as replicates.

Soybean study: chambers were randomized to treatment. It would be inappropriate to treat individual soybean plants as replicates.

in both cases we used the average within the experimental unit

our estimate of σ tells us about the variability expected between experimental units

Experimental Design

ST513

read chapters 23 & 24

A good read:

Hurlbert, Stuart H. (1984). "Pseudoreplication and the design of ecological field experiments".*Ecological Monographs* (Ecological Society of America) **54** (2): 187–211. doi:10.2307/1942661