Stat 412/512

MULTIPLE EXPLANATORIES & INDICATOR VARIABLES Jan 9 2015

Charlotte Wickham

stat512.cwick.co.nz

Today

- An example with two explanatory variables of interest
- Indicator variables, the key to including categorical variables in regression models.

Case Study 9.1 Effects of light on Meadowfoam

Meadowfoam is a seed oil crop.

More flowers = more production = more \$\$\$

Two factors in the experiment:

light intensity: 6 levels

150, 300, 450, 600, 750 & 900 µmol/m²/sec

time of light treatment onset: 2 levels,

at PFI or 24 days before PFI, we call these late and early

120 seedlings, randomly assigned to each combination of treatment (10 seedlings each)

Response: average number of flowers over the 10 seedlings

whole thing repeated

Display 9.1

Time line for light variation experiment on meadowfoam



Questions of interest

What is the effect of **light intensity** on the number of flowers?

What is the effect of the **timing** of the light on the number of flowers?

Does the effect of the **intensity** depend on the **timing** of light treatment?

Display 9.2

Numbers of *flowers* per meadowfoam plant, in twelve treatment groups

		intensity (µmol/m ² /sec)					
		150	300	450	600	750	900
late	at PFI	62.3	55.3	49.6	39.4	31.3	36.8
timing		77.4	54.2	61.9	45.7	44.9	41.9
early	24 days	77.8	69.1	57.0	62.9	60.3	52.6
	before PFI	75.6	78.0	71.1	52.2	45.6	44.4

each number here is the average number of flowers over 10 seedlings

2 times x 6 intensities = 12 treatment groups, 2 replicates for each treatment.

> head(case0901)				
F	lowers	Tin	ne Intensity	
1	62.3	1	150	
2	77.4	1	150	
3	55.3	1	300	
4	54.2	1	300	
5	49.6	1	450	
6	61.9	1	450	

qplot(Intensity, Flowers, data = case0901)



qplot(Time, Flowers, data = case0901)



 \gtrsim



qplot(Intensity, Flowers, data = case0901, shape = Time)





Introducing Indicator variables

An indicator variable is a way to include a categorical variable in a regression model.

An indicator variable for a category takes the value 1 if the observation is in the category and 0 otherwise.

E.g. an *early* indicator variable is:

- 0, if the unit got late light
- 1, if the unit got early light

Flo	owers Time	Intens	early	late
1	62.3 Late	150	0	1
2	77.4 Late	150	0	1
3	55.3 Late	300	0	1
4	54.2 Late	300	0	1
5	49.6 Late	450	0	1
6	61.9 Late	450	0	1
7	39.4 Late	600	0	1
8	45.7 Late	600	0	1
9	31.3 Late	750	0	1
10	44.9 Late	750	0	1
11	36.8 Late	900	0	1
12	41.9 Late	900	0	1
13	77.8 Early	150	1	0
14	75.6 Early	150	1	0
15	69.1 Early	300	1	0
16	78.0 Early	300	1	0
17	57.0 Early	450	1	0
18	71.1 Early	450	1	0
19	62.9 Early	600	1	0
20	52.2 Early	600	1	0
21	60.3 Early	750	1	0
22	45.6 Early	750	1	0
23	52.6 Early	900	1	0
24	44.4 Early	900	1	0

an *early* indicator variable is:

- 0, if the unit got late light
- 1, if the unit got early light

a late indicator variable is:

- 1, if the unit got late light
- 0, if the unit got early light

Flo	wers Time	Intens	early	late	I300
1	62.3 Late	150	0	1	Ø
2	77.4 Late	150	0	1	0
3	55.3 Late	300	0	1	1
4	54.2 Late	300	0	1	1
5	49.6 Late	450	0	1	0
6	61.9 Late	450	0	1	\mathcal{O}
7	39.4 Late	600	0	1	٢
8	45.7 Late	600	0	1	•
9	31.3 Late	750	0	1	-
10	44.9 Late	750	0	1	•
11	36.8 Late	900	0	1	
12	41.9 Late	900	0	1	
13	77.8 Early	150	1	0	
14	75.6 Early	150	1	0	0
15	69.1 Early	300	1	0	1
16	78.0 Early	300	1	0	1
17	57.0 Early	450	1	0	G
18	71.1 Early	450	1	0	
19	62.9 Early	600	1	0	ť
20	52.2 Early	600	1	0	*
21	60.3 Early	750	1	0	•
22	45.6 Early	750	1	0	
23	52.6 Early	900	1	0	
24	44.4 Early	900	1	0	O

Write in the values for an indicator variable for the category: Intensity is 300

Charlotte check capture

Indicator variables in a regression

Consider the following **simple linear** regression model: μ { *flowers* | *early*} = β_0 + β_1 *early*

If the unit is in an early treatment group, *early* = 1, then μ {*flowers* | *early* = 1 } = β_0 + β_1

If the unit is in an late treatment group, early = 0, then $\mu\{ flowers \mid early = 0 \} = \beta_0$ qplot(early, Flowers, data = case0901) +
geom_smooth(method = "Im", se = FALSE)



A two-sample t-test

is the same as a

t-test for zero slope in a simple linear regression on an indicator variable.

> t.test(Flowers ~ Time, data = case0901, var.equal = TRUE)

Two Sample t-test

data: Flowers by Time t = -2.3779, df = 22, p-value = **0.02653** alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -22.762262 -1.554404 sample estimates: mean in group Late mean in group Early 50.05833 62.21667 > summary(Im(Flowers ~ early, data = case0901))

Call: Im(formula = Flowers ~ early, data = case0901)

Residuals:

Min 1Q Median 3Q Max -18.758 -9.717 -1.188 9.623 27.342

Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 50.058 3.616 13.845 2.43e-12 *** early 12.158 5.113 2.378 0.0265 * ---Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.52 on 22 degrees of freedom Multiple R-squared: 0.2045, Adjusted R-squared: 0.1683 F-statistic: 5.654 on 1 and 22 DF, p-value: 0.02653

But this isn't actually what we want to do, it ignores the fact that light intensity was involved.

Your turn

Imagine I **only** had the *Late* time observations.

Write down the simple linear regression model for the mean number of flowers as a function the light intensity.

MEFlowers [Intensity] = for + B. Intensity

Two separate simple linear regression models For the units in the Late treatment groups: M [Flowers [Intensity] = Bo+ B, Intensity (Late) M {Flowers [Intensity, early = 0] = For the units in the *Early* treatment groups: M (Flowers [Intensity] = B2 + B3 Intensity (Early) Next time: capture this idea in a single multiple regression model.