ST 517: Data Analytics I

Using Categorical Explanatory Variables

Outline

Example

Categorical Explanatory Variables

Indicator Variables in R

Categorical Variables with Several Levels

Example: Meadowfoam

Meadowfoam is:

- A small flowering plant in the Pacific Northwest
- Useful for its seed oil—non-greasy and highly stable

In a study to determine whether production can be increased to obtain a profitable crop:

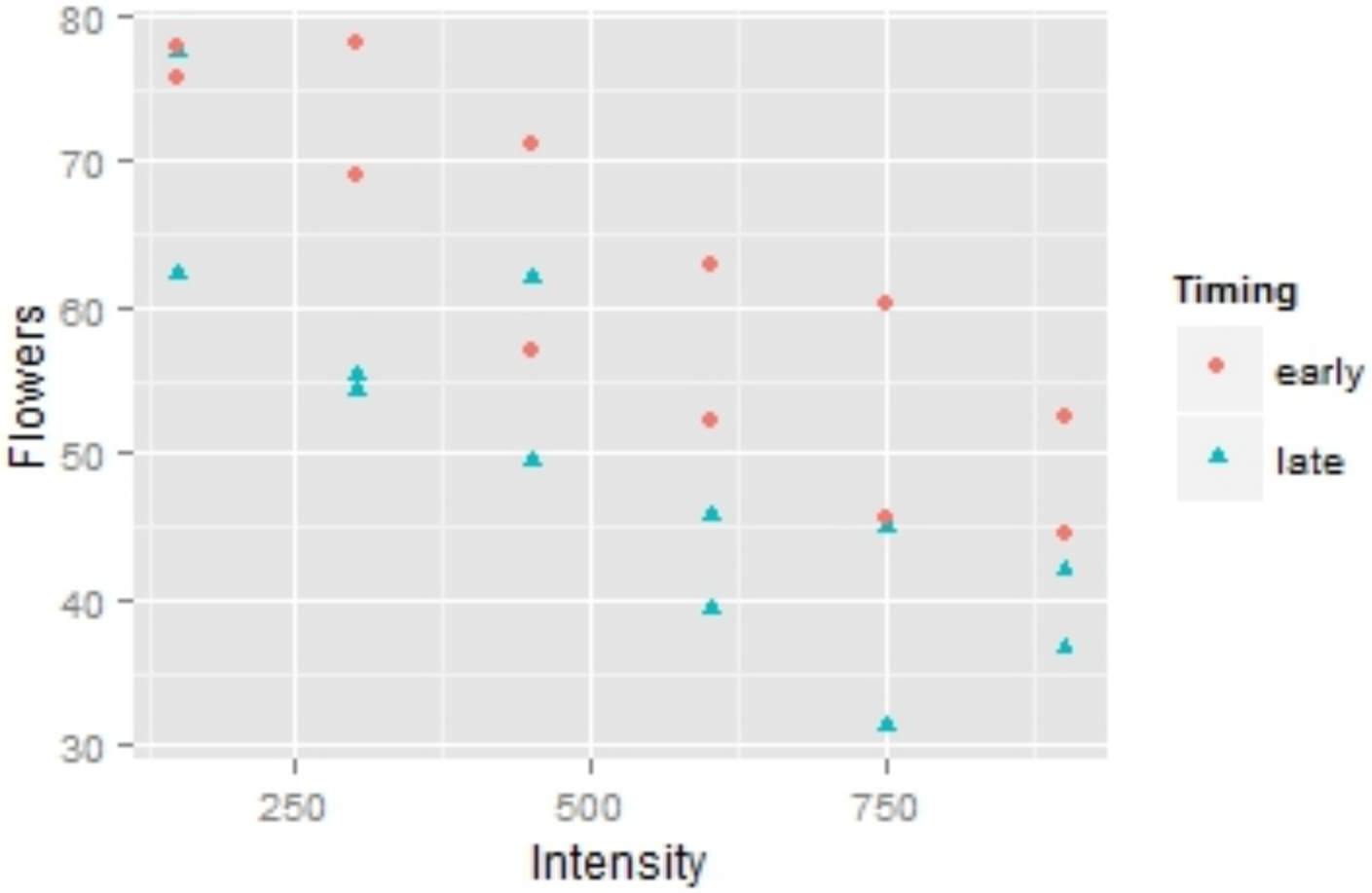
- Six levels of light intensity (150, 300, 450, 600, 750, 900)
- Two levels of starting time for light treatments (early or late)
- Response: average number of flowers in 10 seedlings

Some Questions

For this study, the researchers were interested in the following questions:

- What's the relationship between light intensity and the average number of flowers?
- What's the relationship between timing of light intensity and the average number of flowers?
- Is the relationship between light intensity and the average number of flowers the same for both levels of timing?

The Meadowfoam Data



Some Preliminary Answers

From this *coded* scatterplot:

- It appears that as the light intensity increases, the average number of flowers decreases.
- In general, there appear to be higher numbers of average flowers associated with the early timing light treatment (most of the red circles are above most of the blue triangles).
- Whether the relationship between intensity and average number of flowers is different between the two timing treatments is a question about whether the slope of a line passing through the red circles is the same as the slope of a line passing through the blue triangles.

What to Do?

The Meadowfoam case study provides our first example of a MLR where one of the explanatory variables is continuous (quantitative) and one of the them is categorical.

- One thing to do might be to analyze the early timing light treatment separately from the late timing light treatment.
  - Advantage: It's easy—we know how to run SLR.
  - Disadvantage: We can't make a direct estimate of the effect of timing.
- Another option is to use a MLR, but we have to think about how to include the timing variable.

Dealing with Categorical X's

Here are several rows of the Meadowfoam data:

FLOWERS	TIME	INTENSITY
62.3	1	150
77.4	1	150
55.3	1	300
54.2	1	300
⋮	⋮	⋮
77.8	2	150
75.6	2	150
69.1	2	300
78.0	2	300
⋮	⋮	⋮

Dealing with Categorical X's

Notice that the TIME variable is coded using 1's and 2's, where the 1's correspond to the late timing treatment and the 2's to the early timing treatment.

- But are the 1's and 2's meaningful as numbers?
- That is: is the early timing one unit higher than the late timing; is the early timing twice the late timing?
- The 1's and 2's are just markers denoting which timing treatment is applied.

It wouldn't harm anything to change the coding of the TIME variable so that 1's correspond to the early timing and 0's correspond to the late timing.

Dealing with Categorical X's

Let *late* = 1 when TIME = 1 and let *late* = 0 when TIME = 2. Now, we'll fit the model:

$$\mu(Flowers|Intensity, late) = \beta_0 + \beta_1 Intensity + \beta_2 late$$

Notice that when *late* = 1 (i.e., when timing is late) the right hand side is:

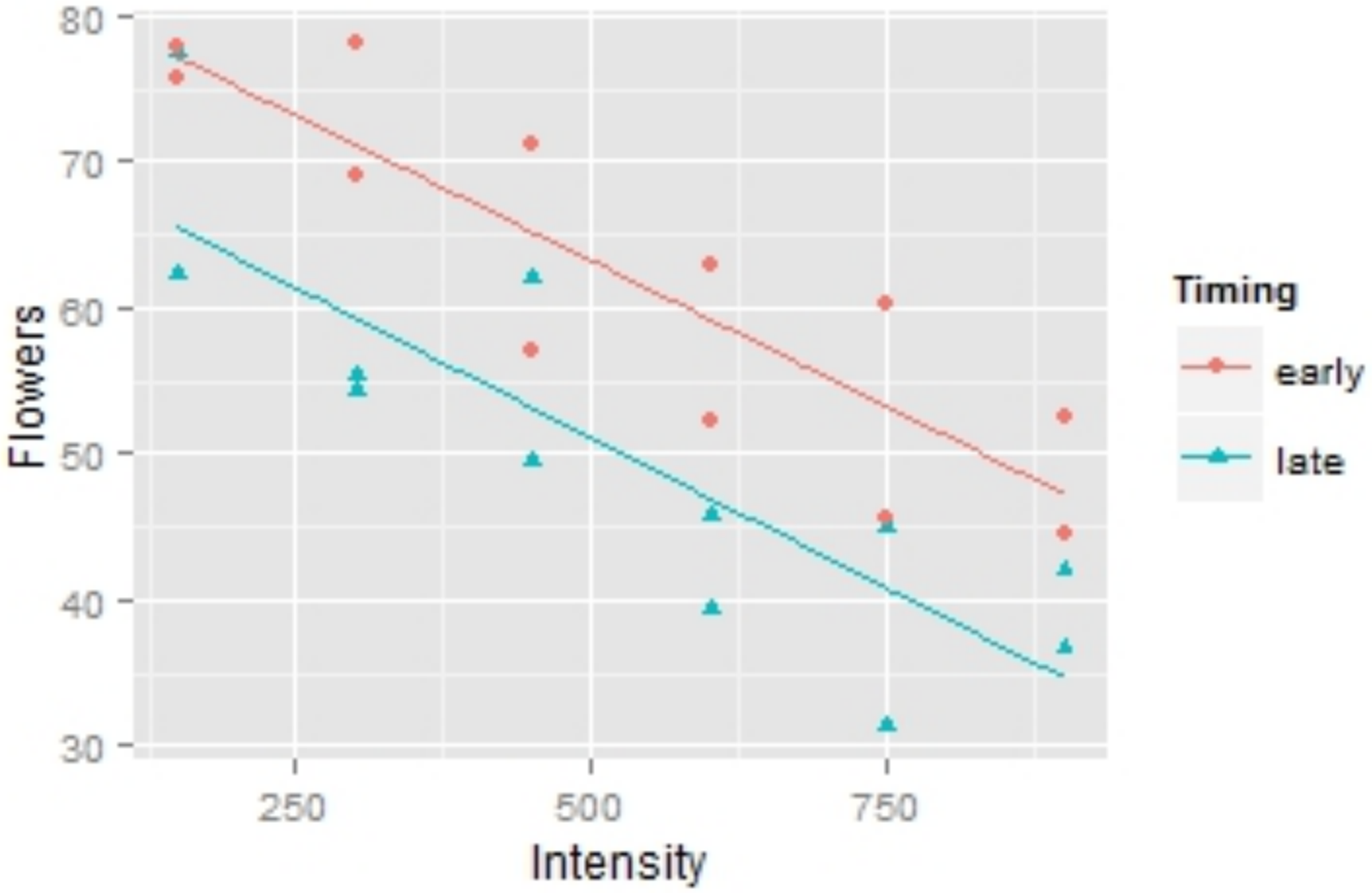
$$(\beta_0+\beta_2) + \beta_1 Intensity$$

and when late = 0, it's:

$$\beta_0 + \beta_1 Intensity$$

So both lines have the same slope, but different *y*-intercepts.

Fitted Model



Indicator Variables in R

The good news is that R will automatically create indicator variables—but you must be sure that the variable you want to have re-coded as an indicator variable is a *factor* variable:

The Meadowfoam data are contained in the `case0901` dataset in the `Sleuth3` library, but the *Time* variable is coded as a numeric variable.

You can use the code:

```
mfoam <- case0901
mfoam$Timing <-
      as.factor(ifelse(mfoam$Time==1,"late","early"))
mod <- lm(Flowers~Intestity+Timing, data=mfoam)
summary(mod)
```

This will produce the output on the next slide

Indicator Variables in R

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  83.464167   3.273772  25.495  < 2e-16 ***
Intensity    -0.040471   0.005132  -7.886 1.04e-07 ***
Timinglate  -12.158333   2.629557  -4.624 0.000146 ***
```

- Notice the "Timinglate" line at the bottom of the left hand column. This is R's way of denoting the indicator variable that corresponds to the "late" level of the Timing variable.
- In this particular analysis, the "early" timing is the **reference level**
- The *y*-intercept for the early timing flowers is estimated to be 83.5. What's the *y*-intercept for the late timing flowers?

More than two Levels

In the Meadowfoam example the categorical explanatory variable, Timing, has two levels—early and late. You will see other examples where a categorical variable has more than two levels.

- Examples:
  - five different varieties of wheat,
  - political party affiliation (Democrat, Independent, Republican).
- In this case, one level of the explanatory variable is the reference level and the multiple linear regression model allows us to estimate the effects of the other levels relative to that reference.

More than Two Levels

Suppose that in the Meadowfoam example, in addition to *early* and *late*, there was another level of Timing, *later*. Then one way possible model is:

$$\mu(Flowers|Intensity, late) = \beta_0 + \beta_1 late + \beta_2 later + \beta_3 Intensity$$

- In this model, β_0 is the *y*-intercept for the *early* Timing level, $\beta_0 + \beta_1$ is the *y*-intercept for the *late* Timing level, and $\beta_0 + \beta_2$ is the *y*-intercept for the *later* Timing level.
- Therefore, β_1 tells us how different the *late* level is from the *early* level, and β_2 tells us how different the *later* level is from the *early* level.