ST 517: Data Analytics I

Interactions and squared terms

Outline

Effects

Interactions

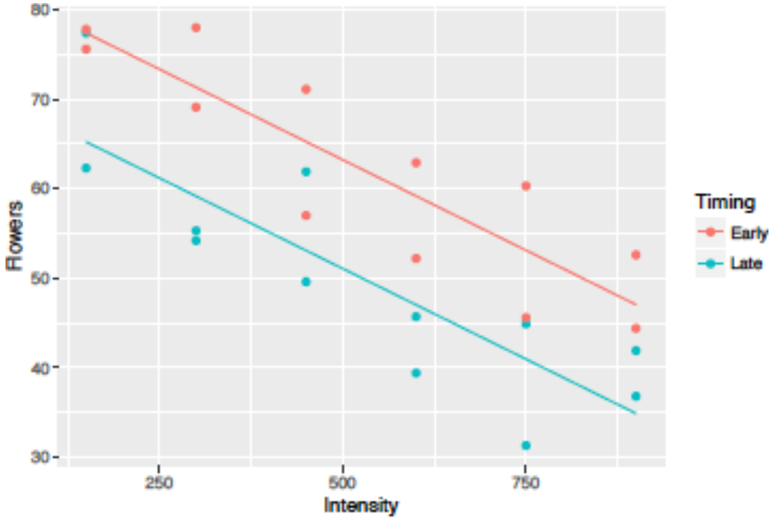
Squared terms

Meadowfoam case study

Recall the Meadowfoam case study from the previous lecture. We fit the model:

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

where *late* was an indicator variable (*late* = 1 when plants received the late treatment, and *late* = 0 when plants received the early treatment).



Effect of a variable

The **effect** of a variable is the change in mean response when the explanatory variable is increased by 1 unit, holding all other variables constant.

The effect of light intensity in the model we fitted is β_1. If the light intensity increases by 1 unit, the associated change in mean number of flowers is β_1.

In this model, the effect of light intensity is the same regardless of the timing.

Or in other words, the slope with respect to light intensity is the same for the early and late treatments.

Interactions

Two variables are said to **interact** if the effect of one variable on the mean response depends on the value of other variable.

Interactions are added to regression models by including a term that consists of the product of the two variables.

For example, consider the model

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

$\beta_3(late \times Intensity)$ is called an **interaction term**.

It allows the effect of intensity on mean number of flowers to depend on whether the timing was early or late (i.e. it allows *Intensity* and *late* to interact).

What is the mean number of flowers for plants with the *early* treatment (*late* = 0) ?

$$\mu(Flowers\ |Intensity, late) = \beta_0 + \beta_1 Intensity + \beta_2 0$$
$$+\beta_3(0 \times Intensity)$$

$$= \beta_0 + \beta_1 Intensity$$

Or in words, for the *early* treatment, the mean response is a straight line function of *Intensity* with intercept β_0 and slope β_1.

What is the mean number of flowers for plants with the *late* treatment (*late* = 1)

$$\mu(Flowers\ |Intensity, late) = \beta_0 + \beta_1 Intensity + \beta_2 1$$
$$+\beta_3(1 \times Intensity)$$

$$= (\beta_0+\beta_2) + (\beta_1+\beta_3)Intensity$$

Or in words, for the late treatment, the mean response is a straight line function of *Intensity* with intercept $(\beta_0+\beta_2)$ and slope $(\beta_1+\beta_3)$.

The coefficient on the interaction term, β_3, can be interpreted as the additional effect of *Intensity* when the plants receive the *late* treatment compared to the *early* treatment.

Or, equivalently, β_3 is the difference in slope with respect to *Intensity* between the *early* and *late* treatment.

Estimating the model

In R, interaction terms are added with “:”.

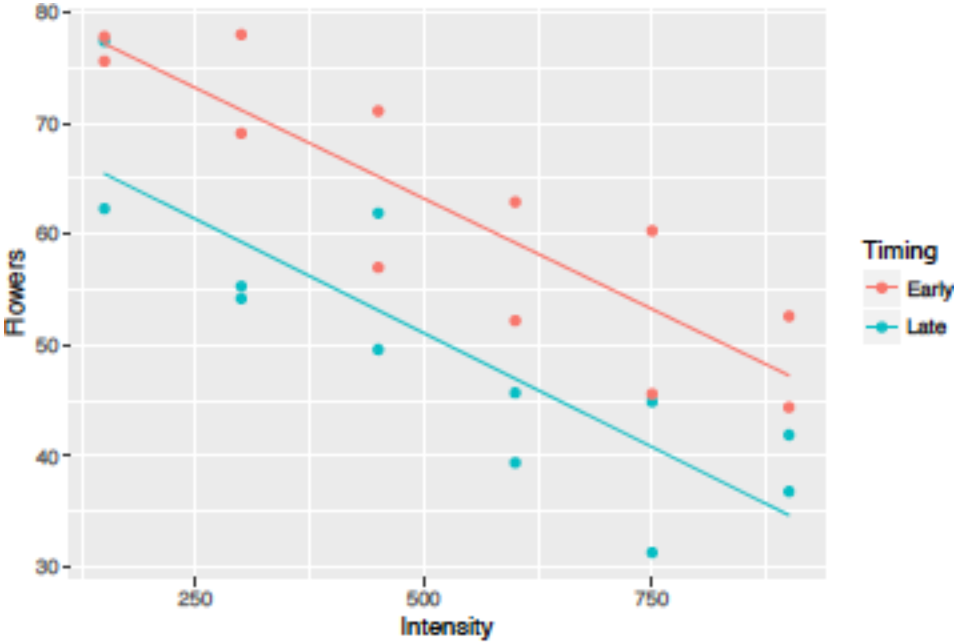
```
fit_sep <- lm(Flowers ~ Intensity + Timing + Intensity:Timing,
         data = case0901)

summary(fit_sep)
```

```
...
Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           83.146667   4.343305  19.144 2.49e-14 ***
Intensity             -0.039867   0.007435  -5.362 3.01e-05 ***
TimingLate           -11.523333   6.142360  -1.876   0.0753 .
Intensity:TimingLate  -0.001210   0.010515  -0.115   0.9096
...
```

There is no evidence the interaction coefficient, β_3 on the line labelled `Intensity:TimingLate`, is not zero. There is no evidence Timing and Intensity interact.

The fitted model with the interaction

These are the fitted lines from the model. The two fitted lines do have different slopes (a difference of 0.0012), but that difference is very small compared to the overall change in mean response, so it’s really hard to see.



Other kinds of interactions

In our example we had a single interaction between a continuous variable and a single indicator variable.

You can have interactions between two continuous variables.

You can have interactions between a continuous variable and a categorical variable, which is represented by a collection of indicator variables. For, example if we had a third timing treatment, say *later*, a model that allows intensity and timing to interact would be:

$$\mu(Flowers\,|Intensity, late, later) = \beta_0 + \beta_1 Intensity + \\ \beta_2 late + \beta_3 later + \\ \beta_4(late \times Intensity) + \beta_5(later \times Intensity)$$

Other kinds of interactions

You can have an interaction between more than two variables by including a product of more than two variables: a higher order interaction. For, example if we had a third treatment variable, say *water*, a model that allows intensity, timing and water to interact would be:

$$\mu(Flowers\,|Intensity, late, water) = \beta_0 + \beta_1 Intensity + \beta_2 late + \beta_3 water + \beta_4(late \times Intensity) + \beta_5(water \times Intensity) + \beta_6(water \times late\ ) + \beta_7(water \times late \times Intensity)$$

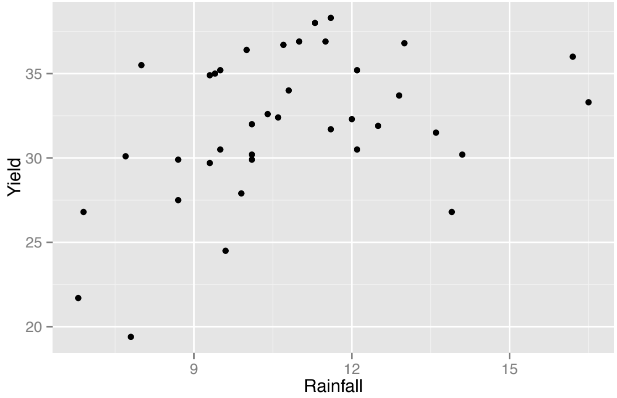
In all cases remember including an interaction means the effect of one variable on the mean response depends on the value of other variable (or variables if it is a higher order interaction).

Squared terms

Multiple linear regression also allows us to model non-linear relationships between the mean response and explanatory variables.

If the relationship appears to have some curvature, one approach is to try adding another term to the model with the explanatory variable squared.

For example, corn yield and rainfall in six U.S. corn–producing states (Iowa, Nebraska, Illinois, Indiana, Missouri and Ohio), are shown below for each year from 1890 to 1927.

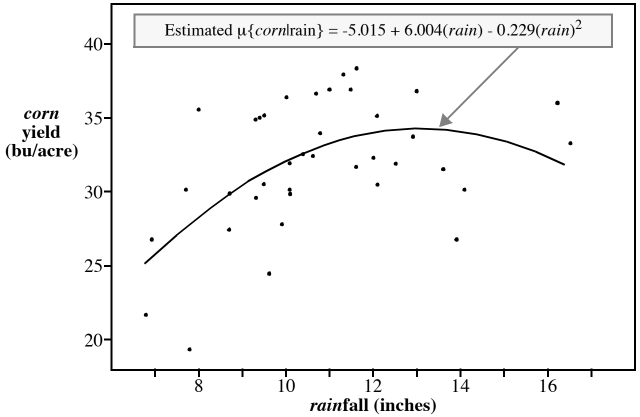


Corn yield

To allow for some curvature we might fit the model,

$$\mu(yield\ |rainfall) = \beta_0 + \beta_1 rainfall + \beta_2 rainfall^2$$

Yearly corn yield versus rainfall (1890-1927) in six U.S. states



The interpretation of the *effect of rainfall* is now complicated, it depends on where on the rainfall scale we are talking about. A plot of the fitted model may be more interpretable than the individual β_1 and β_2 coefficients.