Stat 412/512

Multiple Regression

Jan 12 2015

Charlotte Wickham

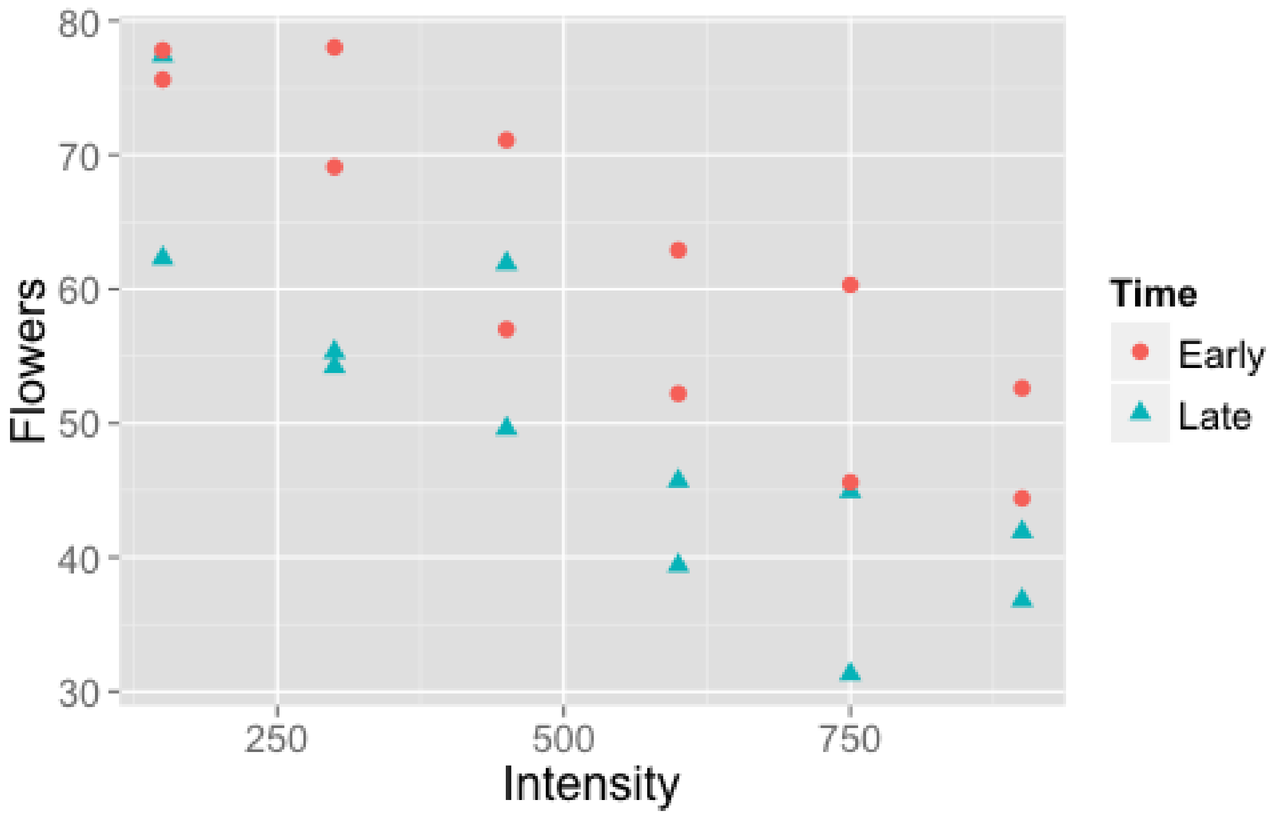
stat512.cwick.co.nz

Case Study 9.1 Effects of light on Meadowfoam

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?



The multiple linear regression model

The mean response, Y, is related to the explanatory variables, X_1 through X_p, through a linear function.

$$\mu\{ Y \mid X_1, X_2, \ldots , X_p\} =$$

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

plus some assumptions

The parameters of the model are $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$

We choose the model so our questions of interest are translated to statements about parameters.

Examples

Multiple linear regression models:

$$\mu\{ Y \mid X_1, X_2, X_3\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

$$\mu\{ Y \mid X_1\} = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$$

$$\mu\{ Y \mid X_1, X_2\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

$$\mu\{ Y \mid X_1, X_2\} = \beta_0 + \beta_1 X_1 + \beta_2 \log(X_2)$$

$$= \beta_0 + \beta_1 \left(\begin{matrix}\text{doesn't} \\ \text{involve } \beta\end{matrix} \right) +$$

A model is linear if it can be written as a sum of terms like: $\beta_i f(X)$ where $f(X)$ doesn't involve any β's.

NOT Multiple linear regression models:

these are example of non-linear regression models

$$\mu\{ Y \mid X_1, X_2\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2^{\beta_3}$$

$$\mu\{ Y \mid X_1, X_2, X_3\} = (\beta_0 + \beta_1 X_1) / (\beta_2 X_2 + \beta_3 X_3)$$

$$\mu\{ Y \mid X_1\} = \beta_0 \exp(\beta_1 X_1)$$

Effect of an explanatory

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

consider E.g. the following model,

$$\mu\{ Y \mid X_1, X_2\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

the effect of X_1 is:

$$\mu\{ Y \mid X_1 = x + 1, X_2\} - \mu\{ Y \mid X_1 = x, X_2\}$$

$$= (\beta_0 + \beta_1(x+1) + \beta_2 X_2) - (\beta_0 + \beta_1 x + \beta_2 X_2)$$

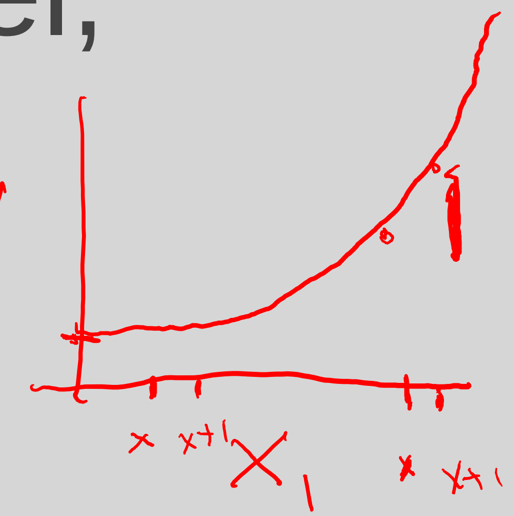
$$= (\cancel{\beta_0} + \cancel{\beta_1 x} + \beta_1 + \cancel{\beta_2 X_2}) - (\cancel{\beta_0} + \cancel{\beta_1 x} + \cancel{\beta_2 X_2})$$

$$= \beta_1$$

Your turn

What's the effect of X_1 in this model,

$\mu\{\, Y \mid X_1\} = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2$? ← quadratic model



$$\mu\{Y \mid X_1 = x+1\} - \mu\{Y \mid X_1 = x\}$$

$$= \left(\beta_0 + \beta_1(x+1) + \beta_2(x+1)^2\right) - \left(\beta_0 + \beta_1 x + \beta_2 x^2\right)$$

$$= \left(\cancel{\beta_0} + \cancel{\beta_1 x} + \beta_1 + \beta_2(\cancel{x^2} + 2x + 1)\right) - \left(\cancel{\beta_0} + \cancel{\beta_1 x} + \cancel{\beta_2 x^2}\right) = \beta_1 + \beta_2(2x+1)$$

Capture

Strategies for understanding a multiple linear regression model

If the model involves **indicator** variables, find the model of the mean for each category. What parameters do the categories share?

For **continuous** variables, find the **effect** of each continuous variable.

Later, we'll see exploring the model through plots of the predictions from the model

Let's practice

Consider this model for the meadowfoam data:

$$\mu\{\, \textit{flowers} \mid \textit{light}, \textit{early} \} = \beta_0 + \beta_1 \textit{light} + \beta_2 \textit{early}$$

indicator

Find the model of the mean for each category. What parameters do the categories share?

For units with late exposure, *early* = *0*:

$\mu\{\, \textit{flowers} \mid \textit{light}, \textit{early} = 0\} =$ $\beta_0 + \beta_1 \textit{light} + \beta_2 (0)$

$= \beta_0 + \beta_1 \textit{light}$

For units with early exposure, *early* = *1*:

$\mu\{\, \textit{flowers} \mid \textit{light}, \textit{early} = 1\} =$ $\beta_0 + \beta_1 \textit{light} + \beta_2 (1)$

$= \beta_0 + \beta_1 \textit{light} + \beta_2 = (\beta_0 + \beta_2) + \beta_1 \textit{light}$

for late group: straight line

$$\mu\{\text{Flowers} \mid \text{light}\} = \beta_0 + \beta_1 \,\text{light}$$

intercept slope

" " early group

$$\mu\{\text{Flowers} \mid \text{light}\} = \underbrace{(\beta_0 + \beta_2)}_{\text{intercept}} + \beta_1 \,\text{light}$$

slope

same slope, different intercepts

Let's practice

Consider this model for the meadowfoam data:

$$\mu\{\,flowers \mid light, early\} = \beta_0 + \beta_1 light + \beta_2 early$$

continuous

Find the effect of each continuous variable.

The effect of light is:

$\mu\{\,flowers \mid \underline{light + 1}, early\} - \mu\{\,flowers \mid \underline{light}, early\}$

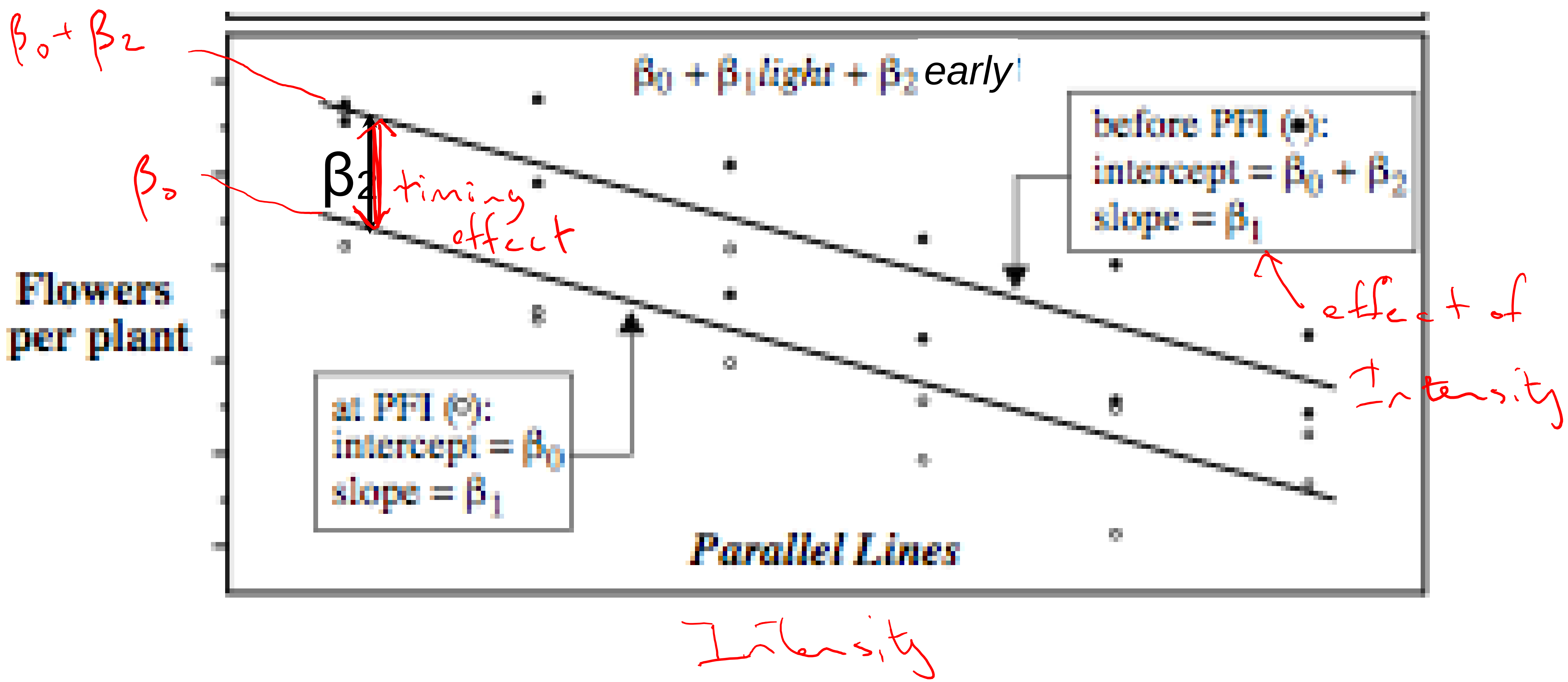
$= (\beta_0 + \beta_1(light+1) + \beta_2\, early) - (\beta_0 + \beta_1\, light + \beta_2\, early)$

$= \beta_1 \underbrace{\qquad\qquad\qquad}_{\text{nothing involving early}}$

This model for the meadowfoam data:

$\mu\{\, flowers \mid light, early\} = \beta_0 + \beta_1 light + \beta_2 early$

is called a Parallel lines model



The effect of light intensity doesn't depend on timing, but timing has an effect.

Your turn

Consider the model:

$\mu\{$ *flowers* | *light*, *early*$\} =$

$\beta_0 + \beta_1 light + \beta_2 early + \beta_3 (light \times early)$

0 0

1 1

late : early = 0

early : early = 1

What is the mean flowers per plant for units in the late treatment group?

$\beta_0 + \beta_1 \text{light}$

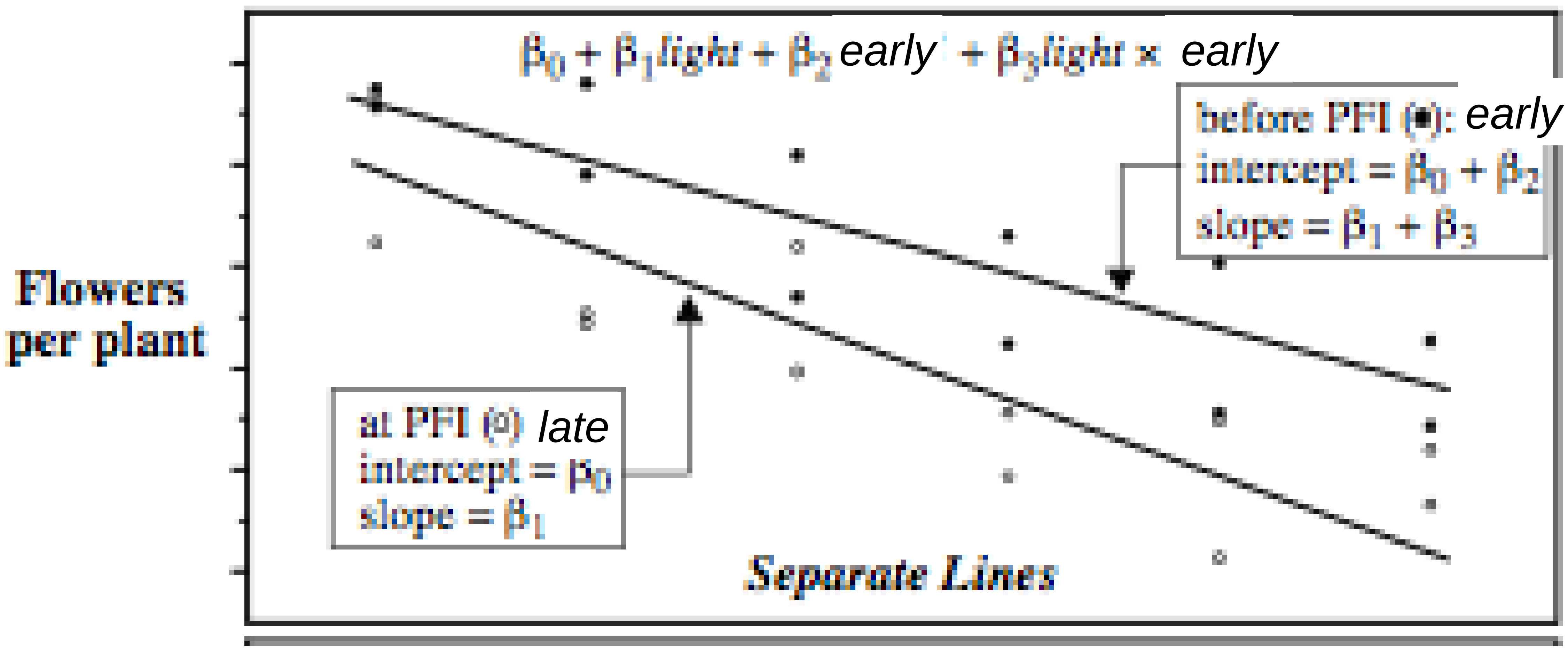
What is the mean flowers per plant for units in the early treatment group?

$\beta_0 + \beta_1 \text{light} + \beta_2 + \beta_3 \text{light} = \beta_0 + \beta_2 + (\beta_1 + \beta_3) \text{light}$

Separate lines model

μ{ *flowers* | *light*, *early*} =

$\beta_0 + \beta_1 light + \beta_2 early + \beta_3 (light \times early)$



The effect of light intensity depends on timing

Interaction terms

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

β_3(*light* × *early)* is called an **interaction** term. In our example it allows the effect of intensity on mean number of flowers to depend on whether the timing was early or late. In this example, it allowed the mean for the *early* units to have a different slope with respect to *light* from the *late* units.

I.e. it allows *light* and *early* to interact.

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

Parallel lines: the effect of light intensity doesn't depend on timing,

$\mu\{ \textit{flowers} \mid \textit{light}, \textit{early} \} = \beta_0 + \beta_1 \textit{light} + \beta_2 \textit{early}$

Separate lines: the effect of light intensity depends on timing

$\mu\{ \textit{flowers} \mid \textit{light}, \textit{early} \} =$

$\beta_0 + \beta_1 \textit{light} + \beta_2 \textit{early} + \beta_3 (\textit{light} \times \textit{early})$

What's the difference?

If $\beta_3 = 0$, the separate lines model reduces to the parallel lines model.

So, to answer our question, we could use the separate lines model and ask is $\beta_3 = 0$?

"...questions of interest are translated to statements about parameters."

separate lines model

```
> fit_sep <- lm(Flowers ~ Intens + early + I(Intens * early), data =
case0901)
> summary(fit_sep)

Call:
lm(formula = Flowers ~ Intens + early + I(Intens * early), data = case09

Residuals:
   Min     1Q Median     3Q    Max 
-9.516 -4.276 -1.422  5.473 11.938 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
β0 (Intercept)       71.623333   4.343305  16.491 4.14e-13 ***
β1 Intens            -0.041076   0.007435  -5.525 2.08e-05 ***
β2 early             11.523333   6.142361   1.876   0.0753 .  
β3 I(Intens * early)  0.001210   0.010515   0.115   0.9096    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.598 on 20 degrees of freedom
Multiple R-squared: 0.7993, Adjusted R-squared: 0.7692 
F-statistic: 26.55 on 3 and 20 DF,  p-value: 3.549e-07
```

There is no evidence that the effect of Intensity depends on timing.

parallel lines model

```
> fit_par <- lm(Flowers ~ Intens + early, data = case0901)
> summary(fit_par)

Call:
lm(formula = Flowers ~ Intens + early, data = case0901)

Residuals:
   Min     1Q Median     3Q    Max 
-9.652 -4.139 -1.558  5.632 12.165 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
β0 (Intercept) 71.305834   3.273772  21.781 6.77e-16 ***
β1 Intens      -0.040471   0.005132  -7.886 1.04e-07 ***
β2 early       12.158333   2.629557   4.624 0.000146 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.441 on 21 degrees of freedom
Multiple R-squared: 0.7992, Adjusted R-squared:  0.78 
F-statistic: 41.78 on 2 and 21 DF,  p-value: 4.786e-08
```

Increasing light intensity decreased the mean number of flowers per plant by 4.0 flowers for every 100 *μmol/m^2/sec.*

Beginning the light treatments 24 days before PFI increased the mean number of flowers per plant by 12.1 compared to beginning light treatments at PFI.