Creating New Explanatory Variables

David Gerard

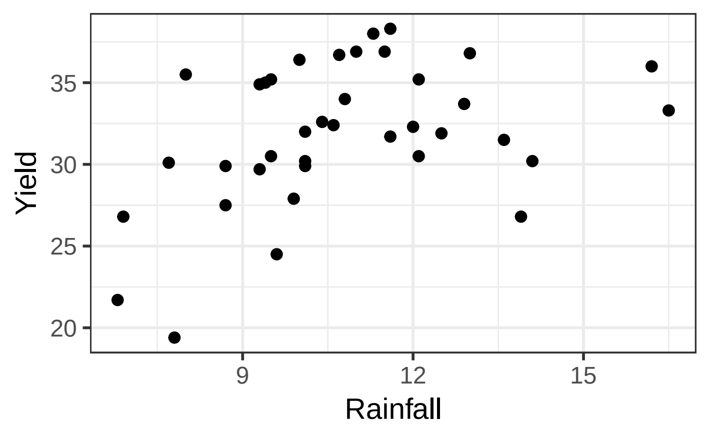
2018-12-07

Objectives

- Create new explanatory variables.
- Chapter 9.

Adding Curvature

Corn and Rain



Quadratic Regression is Multiple Linear Regression

- From last chapter, we said that we should fit

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$

- Relabel $X_{1i} = X_i$
- Relabel $X_{2i} = X_i^2$
- Then this is equivalent to fitting

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

Why is it called linear regression?

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$

- Multiple linear regression represents the mean of the Y's as a linear combination of the β's.
- Even though the mean is a quadratic function of the X_i's, it is still a linear function of the β_j's.

Summary for Curvature

- To fit a polynomial, just create new variables that are powers of existing variables, then include those in the multiple regression model.

```
library(Sleuth3)
data("ex0915")
ex0915$Rainfall2 <- ex0915$Rainfall ^ 2
lmout_quad <- lm(Yield ~ Rainfall + Rainfall2, data = ex0915)
lmout_quad
```

```
##
## Call:
## lm(formula = Yield ~ Rainfall + Rainfall2, data = ex0915)
##
## Coefficients:
## (Intercept)     Rainfall    Rainfall2
##      -5.015        6.004       -0.229
```

Summary for Curvature

```
##
## Call:
## lm(formula = Yield ~ Rainfall + Rainfall2, data = ex0915
##
## Coefficients:
## (Intercept)     Rainfall    Rainfall2
##      -5.015        6.004       -0.229
```

- Interpreting output:

(Intercept)	Rainfall	Rainfall2
$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$

- Estimated Model

$$\mu(Y|Rainfall) = -5.0 + 6.0Rainfall - 0.2Rainfall^2$$

Indicator Variables

Indicator Variables

- If you have a binary explanatory variable, you can include it in your model by representing it as an indicator variable.
- Indicator variable: Only takes on the values of 0 or 1.
- If you include it in your regression model, then you are effectively fitting two lines that have the same slope but a different intercept.

Indicator Variables: Example

- Researchers studied the effect of Time and light intensity on flower yield.
- Response variable: Flower yield (average number of flowers per plant)
- Explanatory variables: Timing of light (early/late), intensity of light (quantitative variable).

```r
data(case0901)
head(case0901)
```

```
##   Flowers Time Intensity
## 1    62.3    1       150
## 2    77.4    1       150
## 3    55.3    1       300
## 4    54.2    1       300
```

Model

- $\mu(Flowers|Time, Intensity) = \beta_0 + \beta_1 Time + \beta_2 Intensity$
- Time can be made into an indicator variable (because it only has two levels).
- The model at Time = 0 is

$$\beta_0 + \beta_2 Intensity$$

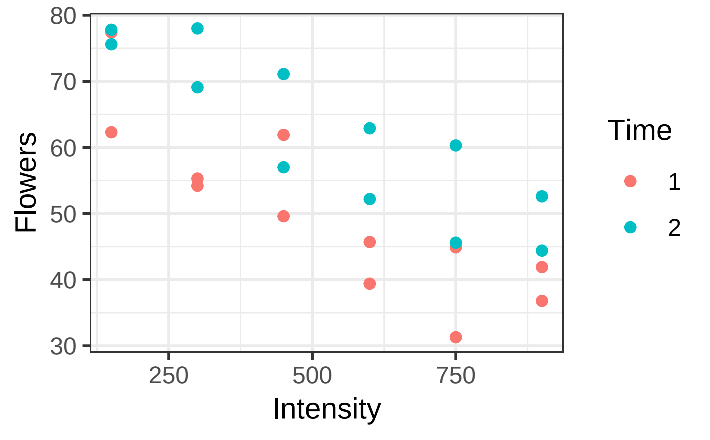
- The model at Time = 1 is

$$\beta_0 + \beta_1 + \beta_2 Intensity$$

- Slope is β_2 both times, but the lines have different intercepts (parallel lines)

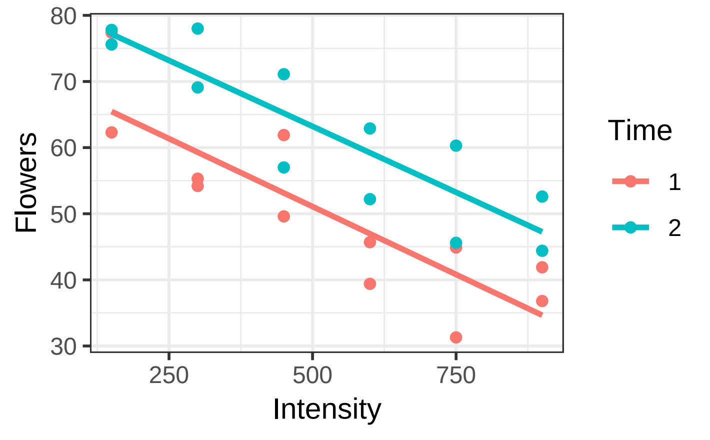
Data

```
case0901$Time <- as.factor(case0901$Time)
qplot(Intensity, Flowers, color = Time, data = case0901)
```



Fit

```r
qplot(Intensity, Flowers, color = Time, data = case0901) +
  geom_smooth(method = "lm", se = FALSE)
```



One-Hot Transformation

One-hot transformation

- You can represent any categorical variable with k levels using $k-1$ indicator variables.
- This representation is called a "one-hot transformation" in the machine learning community.
- Let $X_{\ell i} = 1$ if observational unit i belongs to level ℓ
- Let $X_{\ell i} = 0$ if observational unit i does **not** belong to level ℓ

One-hot transformation: Example

- Let Z be a categorical variable with levels “Bob”, “Cindy”, “Doug”
- $X_{1i} = 1$ if Cindy and 0 otherwise.
- $X_{2i} = 1$ if Doug and 0 otherwise.

One-hot transformation: Example

- Let Z be a categorical variable with levels "Bob", "Cindy", "Doug"
- $X_{1i} = 1$ if Cindy and 0 otherwise.
- $X_{2i} = 1$ if Doug and 0 otherwise.
- Whenever an observational unit is "Bob", it has values $X_{1i} = 0$ and $X_{2i} = 0$
- Whenever an observational unit is "Cindy", it has values $X_{1i} = 1$ and $X_{2i} = 0$
- Whenever an observational unit is "Doug", it has values $X_{1i} = 0$ and $X_{2i} = 1$

One-hot Transformation: Example

- If we have a quantitative response and a categorical explanatory variable, **we can apply a one-hot transformation and use multiple linear regression.**
- $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$
- Mean if “Bob”: $\beta_0 + \beta_1 0 + \beta_2 0 = \beta_0$
- Mean if “Cindy”: $\beta_0 + \beta_1 1 + \beta_2 0 = \beta_0 + \beta_1$
- Mean if “Doug”: $\beta_0 + \beta_1 0 + \beta_2 1 = \beta_0 + \beta_2$

One-hot Transformation: Example

- This is equivalent to One-way ANOVA
- Multiple Regression: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$
- One-way ANOVA: $Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$
- $\mu = \beta_0$, $\alpha_2 = \beta_1$, $\alpha_3 = \beta_2$
- In linear regression, X_{1i} and X_{2i} index the group status of observational unit i.
- In ANOVA, i indexes the group status, and j indexes the observational units in group i.

One-hot transformation: Two-levels

- If a variable only takes on 2 levels, it can be represented by 1 indicator variable.
- `Time` takes on the levels `Late` and `Early`
    - Let $X_{1i} = 0$ if `Late` and $X_{1i} = 0$ if `Early`.

How to include categorical variables in R

- If the variable is a "`factor`", then R will automatically apply a one-hot transformation.
- You can check if a variable is a `factor` using the `class()` function.

```
class(case0901$Time)
```

```
## [1] "factor"
```

- If it is not a `factor`, you can use `as.factor()` to convert it to one.

```
case0901$Time <- as.factor(case0901$Time)
```

Use R

- You can then fit the linear model as before.

```
lmout <- lm(Flowers ~ Time + Intensity, data = case0901)
coef(lmout)
```

```
## (Intercept)       Time2   Intensity
##    71.30583    12.15833    -0.04047
```

Interpreting Output

- Model: $\mu(Y_i | Time, Intensity) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$
- $X_{1i} = 1$ if Time is `Late` and 0 otherwise.
- X_{2i} is the light intensity.

```
## (Intercept)       Time2   Intensity
##    71.30583    12.15833    -0.04047
```

(Intercept)	Time2	Intensity
$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$

Interactions

Interactions

- An interaction between two variables means that the slope with respect to one variable changes with the value of the second variable.
- $\mu(Y_i | Time, Intensity) =$ $\beta_0 + \beta_1 Time + \beta_2 Intensity + \beta_3 Time \times Intensity$
- When $Time = 0$, the model is

$$\mu(Y_i | Time, Intensity) = \beta_0 + \beta_2 Intensity$$

- When $Time = 1$, the model is

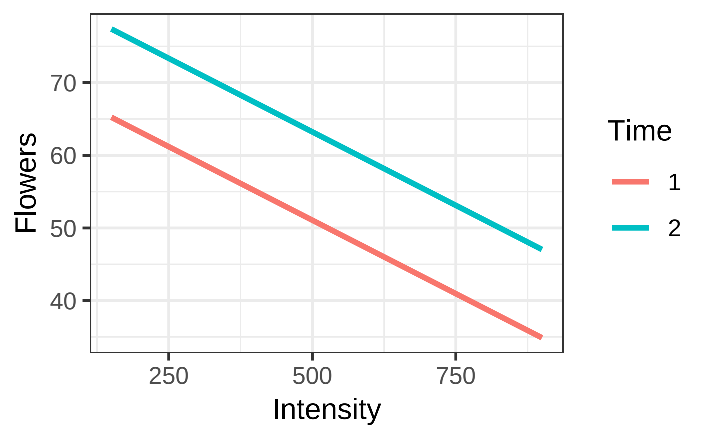
$$\mu(Y_i | Time, Intensity) = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) Intensity$$

Interactions

- Slope when $Time = 0$: β_2
- Slope when $Time = 1$: $\beta_2 + \beta_3$
- Intercept when $Time = 0$: β_0
- Intercept when $Time = 1$: $\beta_0 + \beta_1$
- Each level of the categorical variable ($Time$) has its own line.

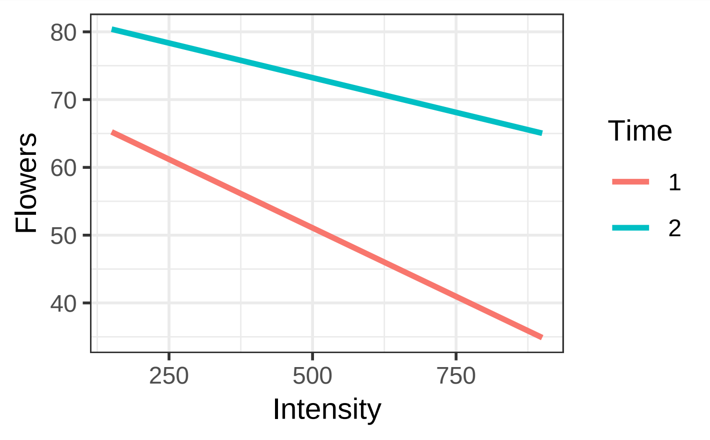
No Interaction

- $\mu(Y_i | Time, Intensity) = \beta_0 + \beta_1 Time + \beta_2 Intensity$



Interaction

- $\mu(Y_i | Time, Intensity) = \beta_0 + \beta_1 Time + \beta_2 Intensity + \beta_3 Time \times Intensity$



Fitting Interactions in R

- $\mu(Y_i | Time, Intensity) =$
  $\beta_0 + \beta_1 Time + \beta_2 Intensity + \beta_3 Time \times Intensity$

```
lmint <- lm(Flowers ~ Time * Intensity, data = case0901)
coef(lmint)
```

```
##     (Intercept)           Time2       Intensity Time2:Intensi
##        71.62333        11.52333        -0.04108          0.001
```

(Intercept)	Time2	Intensity	Time2:Intensity
$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$

Fitting Interactions in R

- Using $*$ fits interactions along with all lower order terms.
- Using $:$ just fits interactions.

Interpreting Interactions

- Reconsider the brain weight data
- $\mu(Brain|Body, Litter) =$
  $\beta_0 + \beta_1 Body + \beta_2 Litter + \beta_3 Body \times Litter$
- What is the slope for `Body` at a given `Litter` size?
- What is the intercept for `Body` at a given `Litter` size?

Intercept Considerations

- Interpreting models with interactions is difficult.
- Include them only if you have to.