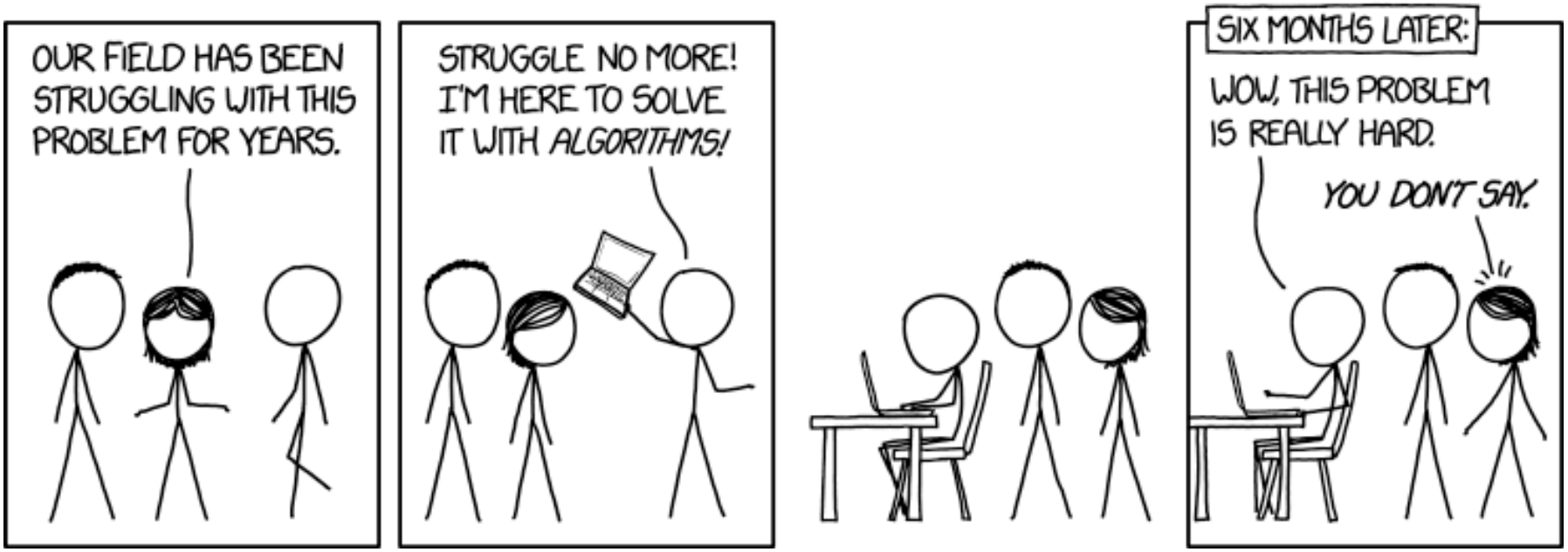
# Multiple Linear Regression

ENTMLGY 6702 Entomological Techniques and Data Analysis



# Learning objectives

1. Compare and contrast simple vs. multiple linear regression
2. Become familiar with additional assumptions when using multiple linear regression and how to check for and deal with them.

# Comparing simple vs. polynomial linear regression

Simple linear regression

$$Y_i = \beta_0 + \beta_1 X_1 + \varepsilon_i$$

$$Height \sim DBH$$

Polynomial regression

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon_i$$

$$Height \sim DBH + +DBH^2$$

# Comparing simple vs. multiple linear regression

Simple linear regression

$$Y_i = \beta_0 + \beta_1 X_1 + \varepsilon_i$$

$$Height \sim DBH$$

Multiple linear regression

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i$$

$$Height \sim DBH + Nitrogen$$

# Comparing polynomial vs. multiple linear regression

Polynomial regression

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon_i$$

$$Height \sim DBH + DBH^2$$

Multiple linear regression

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon_i$$

$$Height \sim DBH + Nitrogen$$

# A word of advice on polynomials

The below R code would result in the exact same line (=predicted *y* values), but different slope coefficients.

Polynomial regression (option 1)

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon_i$$

```
fit1 <- lm(Height ~ DBH + I(DBH^2), data=df)
```

Polynomial regression (option 2)

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \varepsilon_i$$

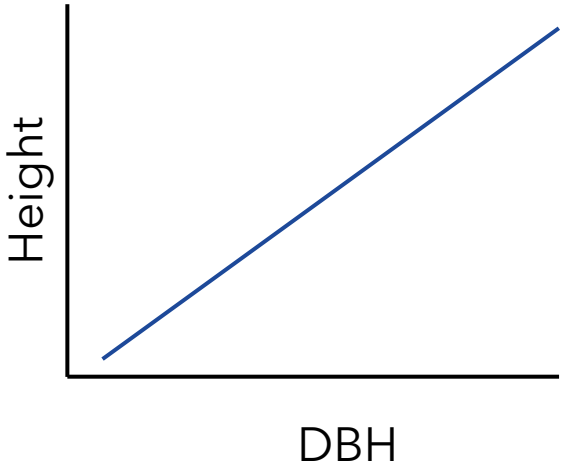
```
fit2 <- lm(Height ~ poly(DBH,2), data=df)
```

# Building large (multiple predictor) models

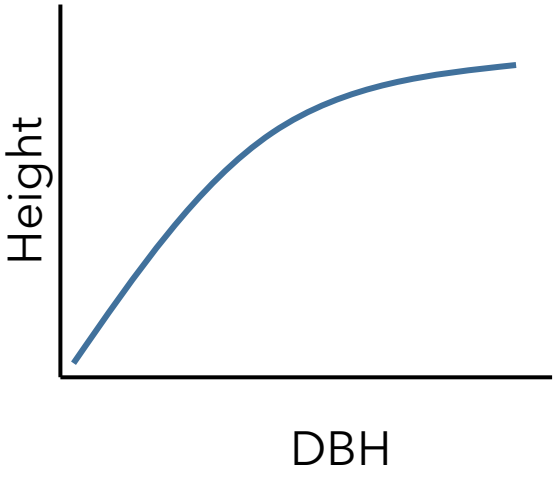
$$Height \sim DBH + Nitrogen + Nitrogen^2$$

```
fit1 <- lm(Height ~ DBH + Nitrogen + I(Nitrogen^2), data=df)
```

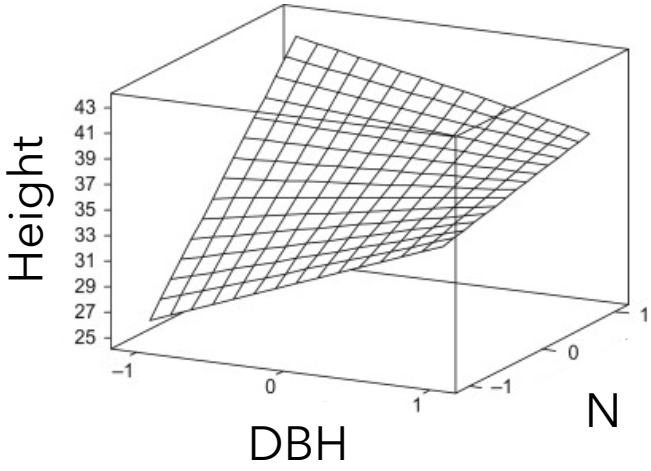
# Single vs. polynomial vs. multiple regression models




$Height \sim DBH$

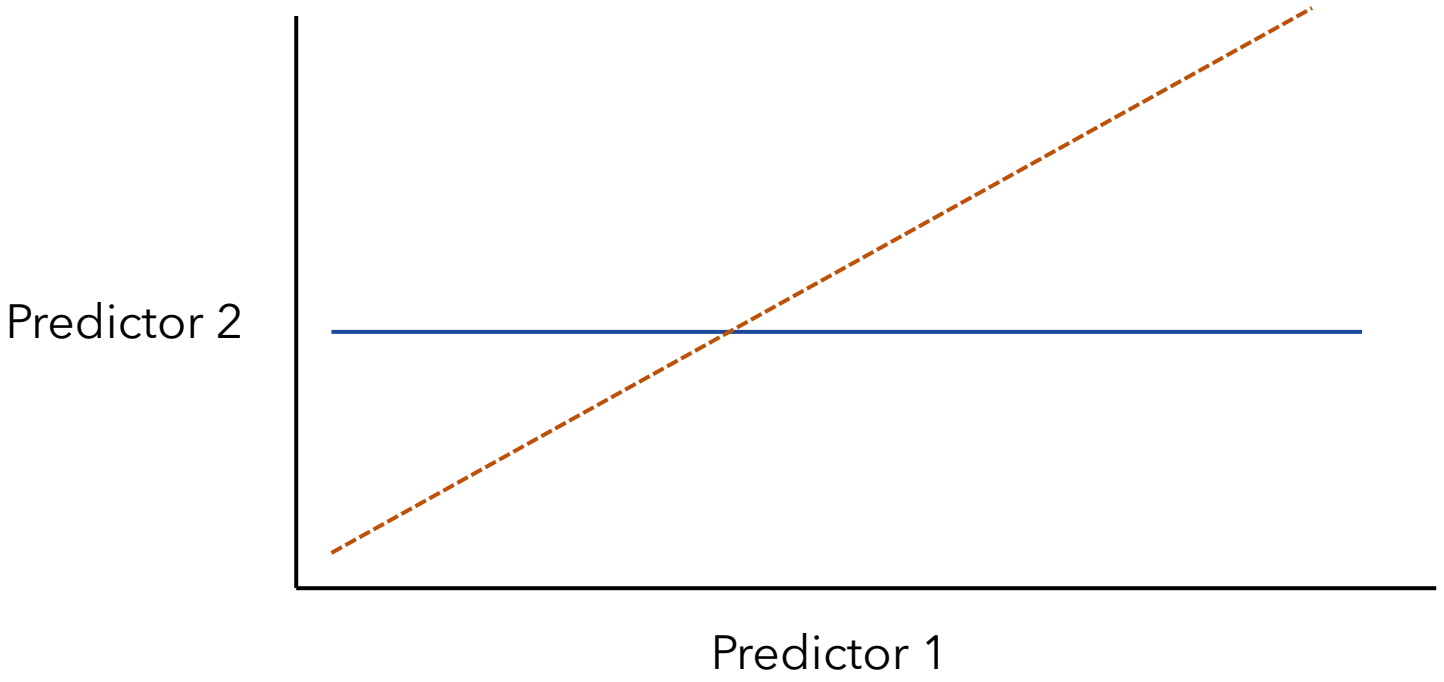



$Height \sim DBH + +DBH^2$




$Height \sim DBH + Nitrogen$

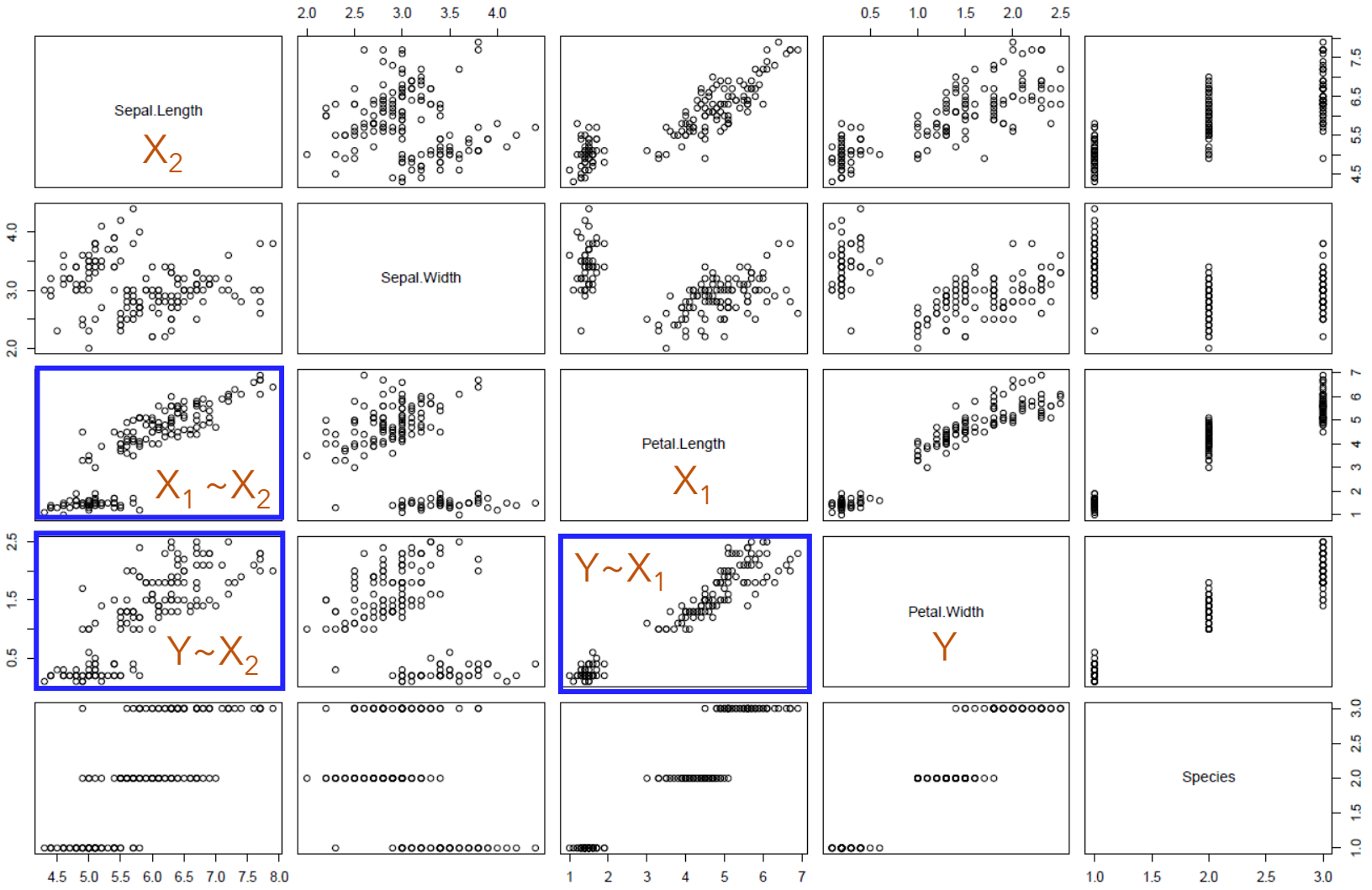
# Multiple regression model: additional assumptions




A correlation – negative or positive – between predictors is called "collinearity" which can cause problems in model fitting. A common sign of collinearity is "large" changes in slope coefficients, including sign flipping (e.g., slope coefficient goes from negative to positive), depending on which predictors are fit in a model.

```
plot(iris)
```

$$Petal.Width \sim Petal.Length + Sepal.Length$$



```
fitA <- lm(Petal.Width~Petal.Length, data=iris)
summary(fitA)


Call:
lm(formula = Petal.Width ~ Petal.Length, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.56515 -0.12358 -0.01898  0.13288  0.64272

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.363076   0.039762  -9.131  4.7e-16 ***
Petal.Length  0.415755   0.009582  43.387  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2065 on 148 degrees of freedom
Multiple R-squared:  0.9271,  Adjusted R-squared:  0.9266
F-statistic:  1882 on 1 and 148 DF,  p-value: < 2.2e-16
```

```
fitB <- lm(Petal.Width~Sepal.Length, data=iris)
summary(fitB)


Call:
lm(formula = Petal.Width ~ Sepal.Length, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.96671 -0.35936 -0.01787  0.28388  1.23329

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.20022    0.25689  -12.46   <2e-16 ***
Sepal.Length  0.75292    0.04353   17.30   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.44 on 148 degrees of freedom
Multiple R-squared:  0.669,   Adjusted R-squared:  0.6668
F-statistic: 299.2 on 1 and 148 DF,  p-value: < 2.2e-16
```

```
fitC <- lm(Petal.Width~Petal.Length+Sepal.Length, data=iris)
summary(fitC)


Call:
lm(formula = Petal.Width ~ Petal.Length + Sepal.Length, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.60598 -0.12560 -0.02049  0.11616  0.59404

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.008996   0.182097  -0.049   0.9607
Petal.Length  0.449376   0.019365  23.205   <2e-16 ***
Sepal.Length -0.082218   0.041283  -1.992   0.0483 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2044 on 147 degrees of freedom
Multiple R-squared:  0.929,   Adjusted R-squared:  0.9281
F-statistic: 962.1 on 2 and 147 DF,  p-value: < 2.2e-16
```

**Coefficients:**

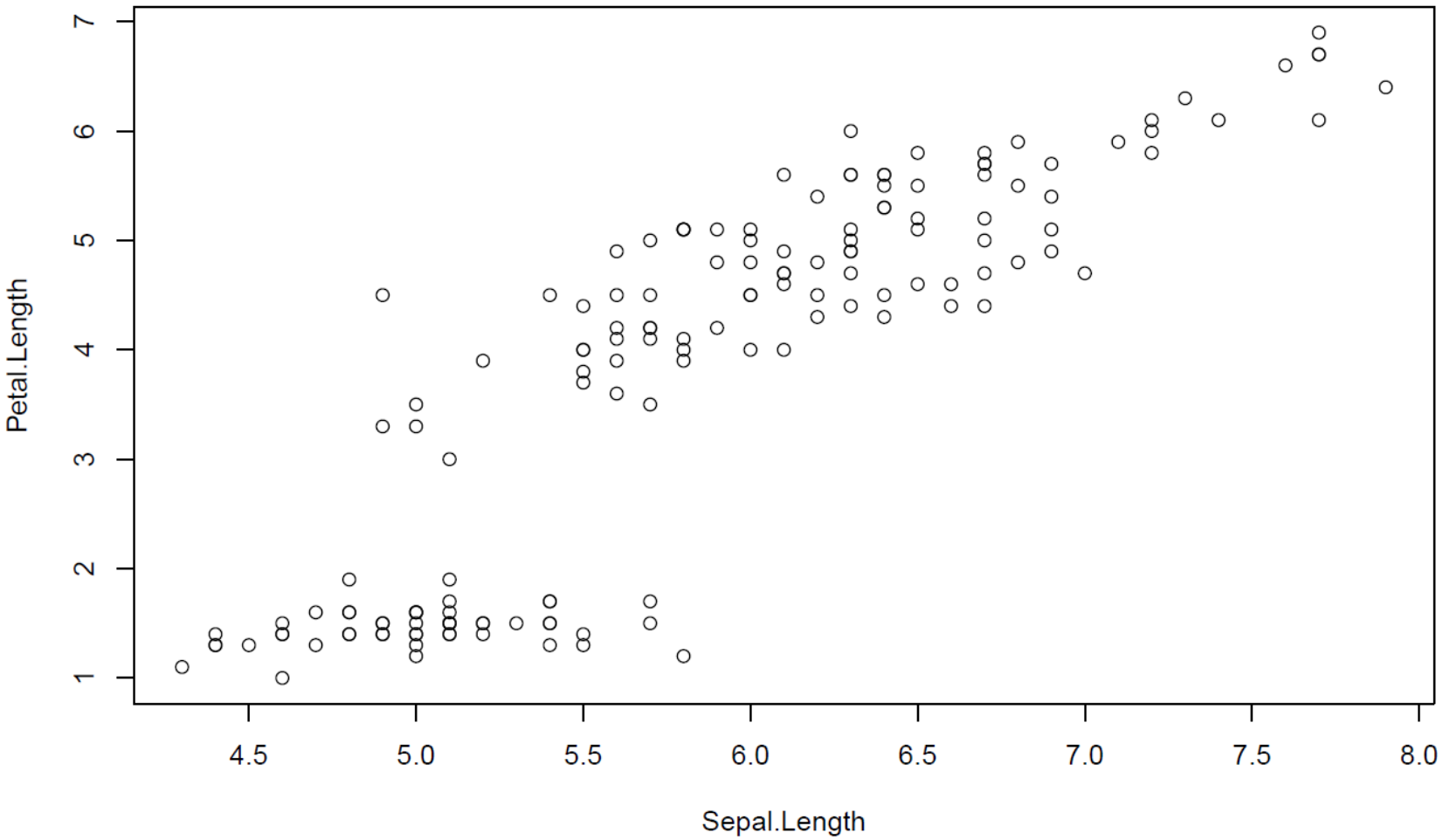
| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | -0.363076 | 0.039762 | -9.131 | 4.7e-16 | *** |
| Petal.Length | 0.415755 | 0.009582 | 43.387 | < 2e-16 | *** |

**Coefficients:**

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | -3.20022 | 0.25689 | -12.46 | <2e-16 | *** |
| Sepal.Length | 0.75292 | 0.04353 | 17.30 | <2e-16 | *** |

**Coefficients:**

| | Estimate | Std. Error | t value | Pr(>\|t\|) | |
|---|---|---|---|---|---|
| (Intercept) | -0.008996 | 0.182097 | -0.049 | 0.9607 | |
| Petal.Length | 0.449376 | 0.019365 | 23.205 | <2e-16 | *** |
| Sepal.Length | -0.082218 | 0.041283 | -1.992 | 0.0483 | * |

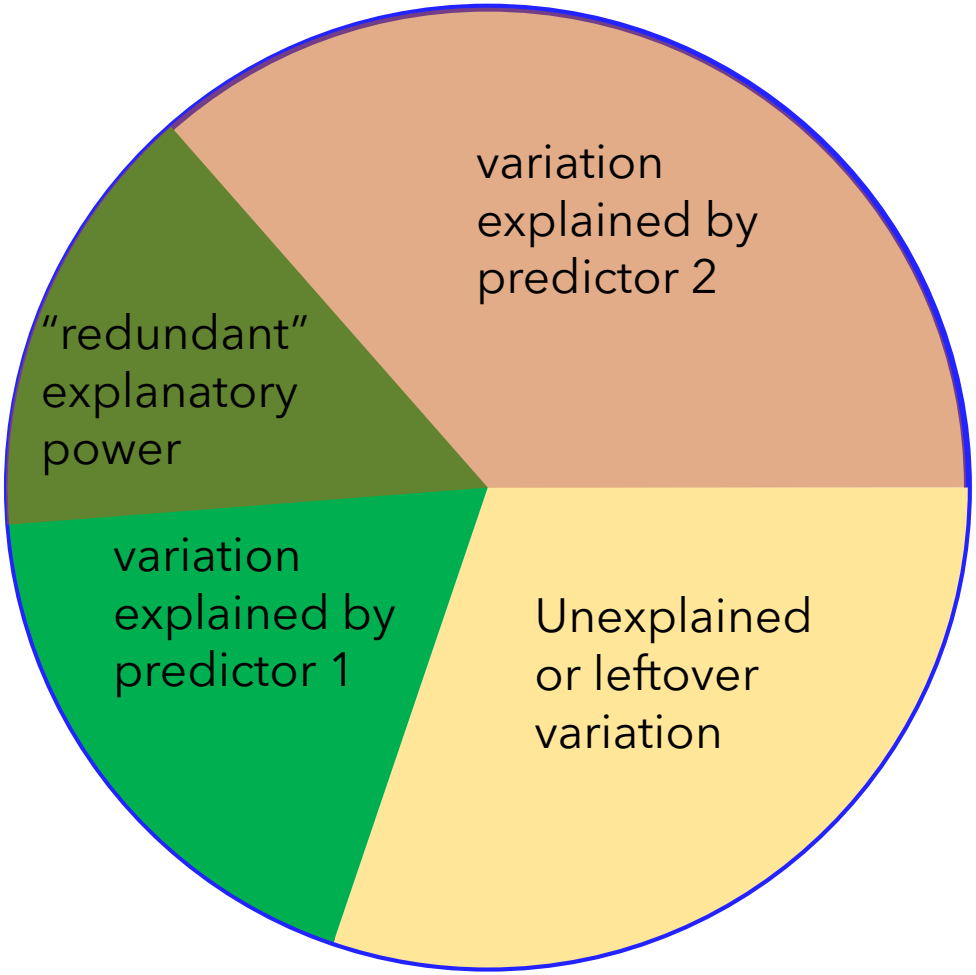

Petal.Length
1
2
3
4
5
6
7
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
Sepal.Length

# Interpreting simple linear regression vs. multiple regression?

In SLR, we interpret the slope coefficient as follows: "a one unit increase in $X_1$ was associated with a $\beta_1$ unit increase in Y ."

In MLR, we interpret coefficients as follows: "holding all else equal, a one unit increase in $X_1$ was associated with a $\beta_1$ unit increase in Y ." That is part of the reason collinearity causes problems. If predictors $X_1$ and $X_2$ are highly correlated, it is difficult to "hold $X_2$ equal or constant" while estimating the effect of $X_1$.

# Sequential vs. marginal fits in ANOVA



# Sequential vs. marginal fitting

```
grouseticks$f_YEAR <- as.factor(grouseticks$YEAR)
fit_ex1 <- lm(TICKS ~ f_YEAR + HEIGHT, data=grouseticks)
anova(fit_ex1)
```

```
## Analysis of Variance Table
##
## Response: TICKS
##             Df Sum Sq Mean Sq F value    Pr(>F)
## f_YEAR       2   7050  3524.9  24.995 5.928e-11 ***
## HEIGHT       1   6092  6092.0  43.199 1.550e-10 ***
## Residuals 399  56268   141.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit_ex2 <- lm(TICKS ~ HEIGHT + f_YEAR, data=grouseticks)
anova(fit_ex2)
```

```
## Analysis of Variance Table
##
## Response: TICKS
##             Df Sum Sq Mean Sq F value    Pr(>F)
## HEIGHT       1   7692  7692.2  54.546 8.948e-13 ***
## f_YEAR       2   5450  2724.8  19.321 9.788e-09 ***
## Residuals 399  56268   141.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Sequential vs. marginal fitting

```
library(car)
```

Marginal fits give you these…

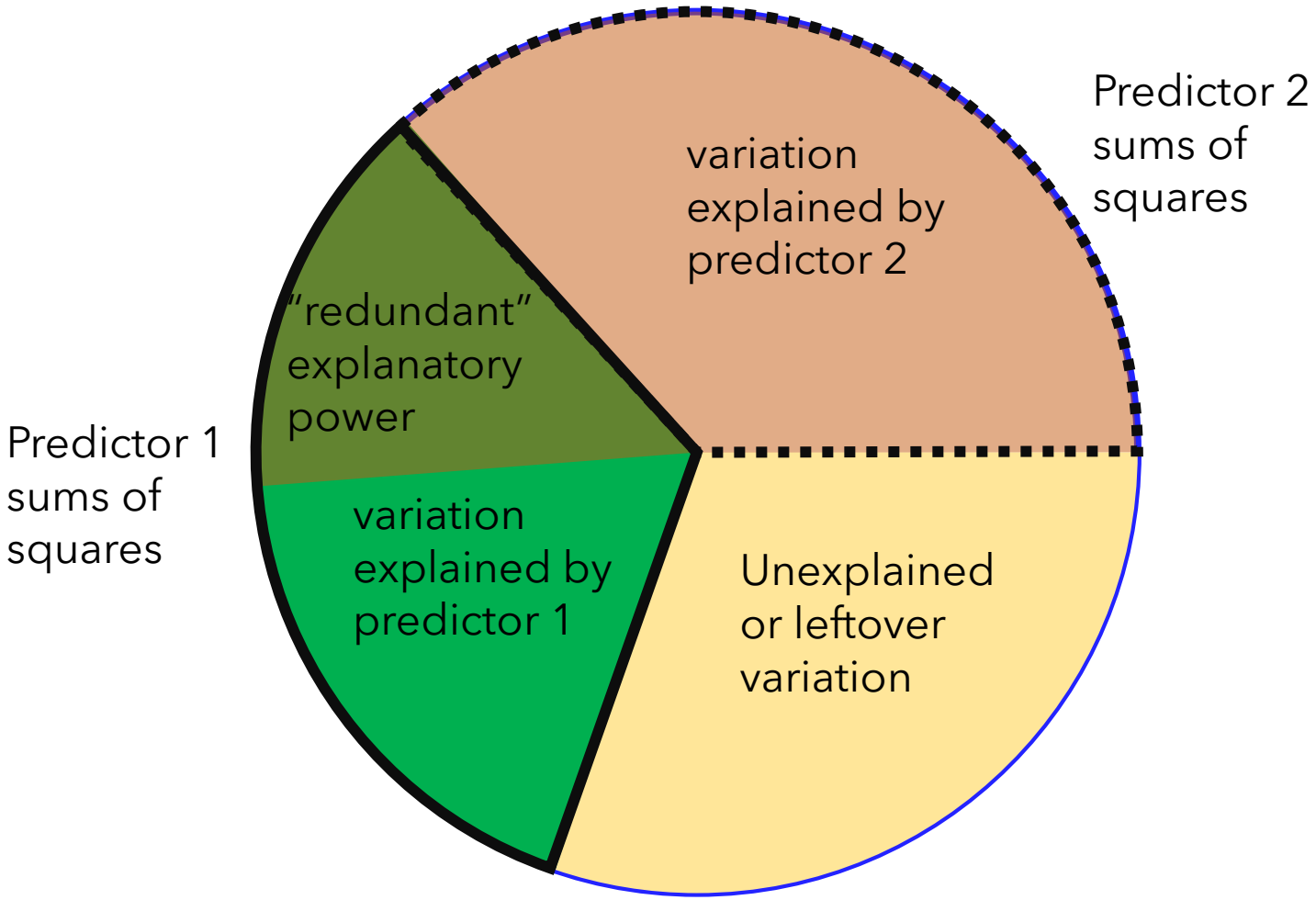
```
Anova(fit_ex1, type="III")
```

```
## Anova Table (Type III tests)
##
## Response: TICKS
##             Sum Sq  Df F value    Pr(>F)
## (Intercept)   7444   1  52.786 1.970e-12 ***
## f_YEAR        5450   2  19.321 9.788e-09 ***
## HEIGHT        6092   1  43.199 1.550e-10 ***
## Residuals    56268 399
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Anova(fit_ex2, type="III")
```

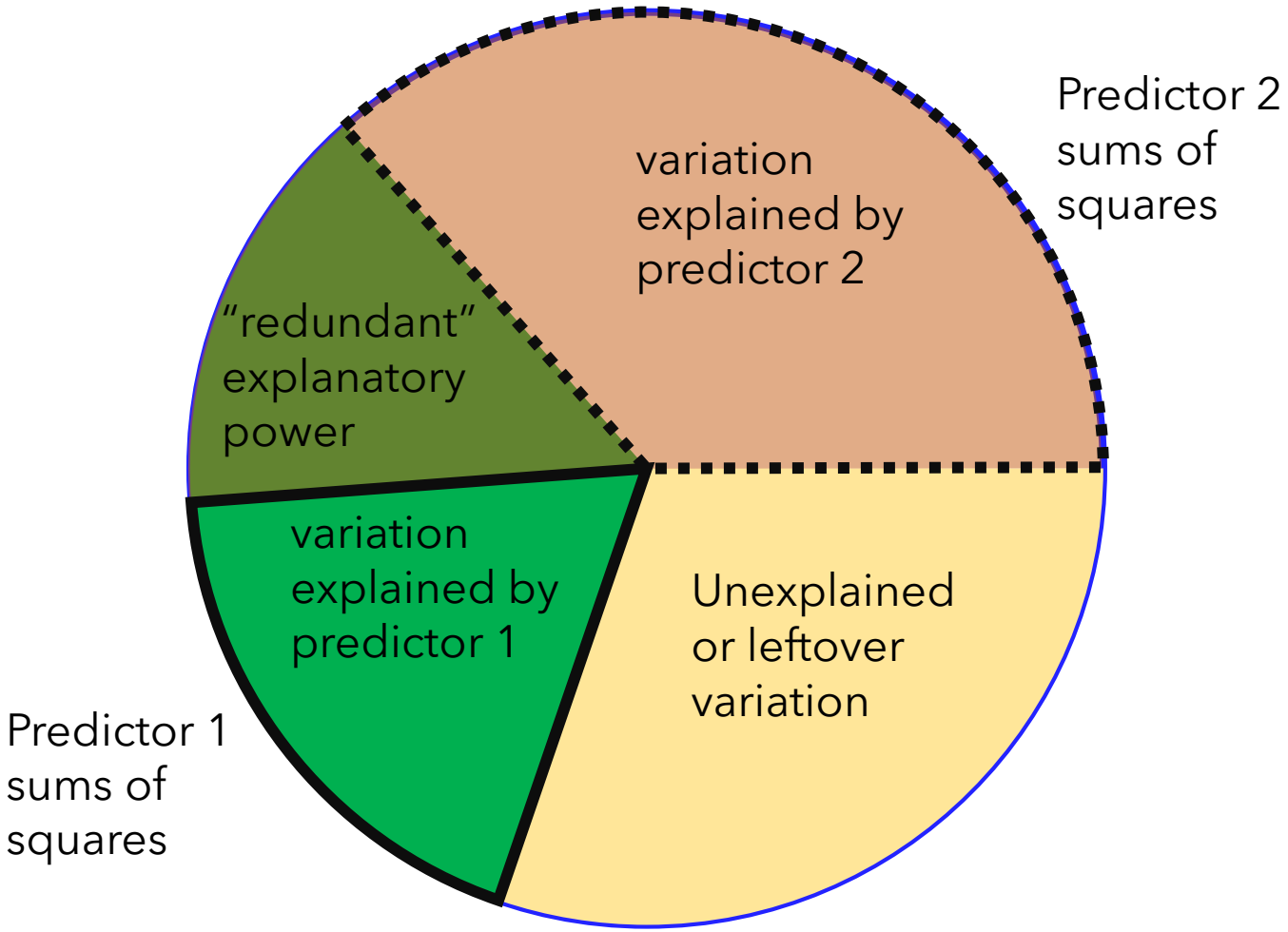
```
## Anova Table (Type III tests)
##
## Response: TICKS
##             Sum Sq  Df F value    Pr(>F)
## (Intercept)   7444   1  52.786 1.970e-12 ***
## HEIGHT        6092   1  43.199 1.550e-10 ***
## f_YEAR        5450   2  19.321 9.788e-09 ***
## Residuals    56268 399
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Sequential fit




Sequential fits change their answer depending on which variable/predictor is fit first.

# Marginal fit



# Sequential vs. marginal fitting

Table 1: ANOVA Table for fit_ex1_df (sequential fit)

| Predictor | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| f_YEAR | 2 | 7049.72 | 3524.86 | 24.99 | <0.0001 |
| HEIGHT\|f_YEAR | 1 | 6091.98 | 6091.98 | 43.2 | <0.01 |
| Residuals | 399 | 56268.21 | 141.02 | | |

Table 2: ANOVA Table for fit_ex2_df (sequential fit)

| Predictor | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| HEIGHT | 1 | 7692.19 | 7692.19 | 54.55 | <0.0001 |
| f_YEAR\|HEIGHT | 2 | 5449.52 | 2724.76 | 19.32 | <0.0001 |
| Residuals | 399 | 56268.21 | 141.02 | | |

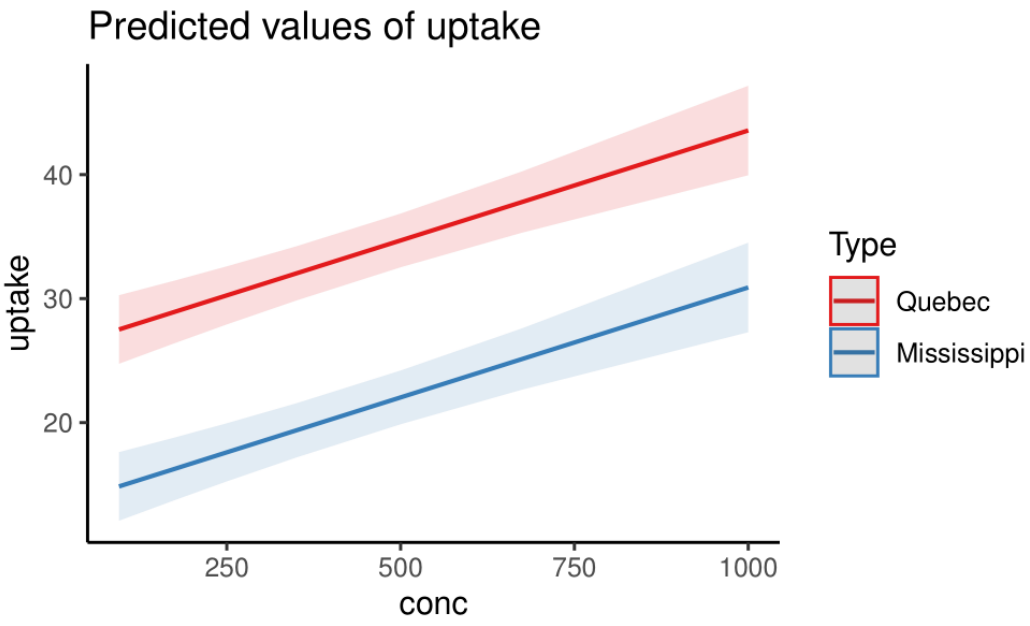
Table 3: ANOVA Table for fit_ex1_df (marginal fit)

| Predictor | DF | SS | MS | F | P |
|---|---|---|---|---|---|
| f_YEAR\|HEIGHT | 2 | 5449.52 | 2724.7600 | 19.32 | <0.0001 |
| HEIGHT\|f_YEAR | 1 | 6091.98 | 6091.9800 | 43.2 | <0.01 |
| Residual | 399 | 56268.21 | 141.0231 | | |

# "Main effects"

```
fit_plants_1_nointeraction <- lm(uptake~conc+Type,data=CO2)
anova(fit_plants_1_nointeraction)
```

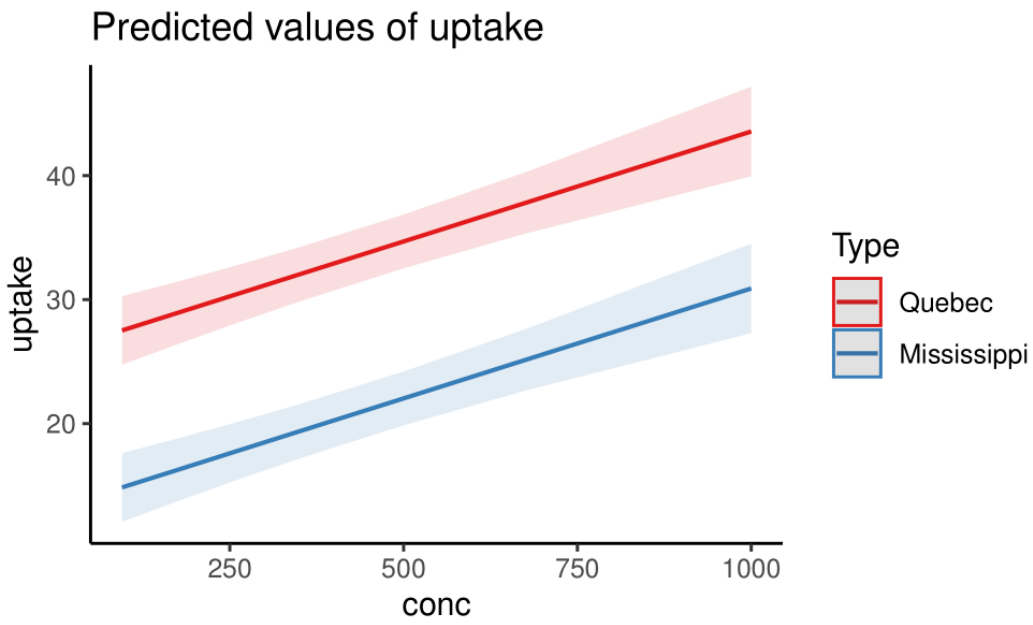
```
## Analysis of Variance Table
##
## Response: uptake
##           Df Sum Sq Mean Sq F value    Pr(>F)
## conc       1 2285.0  2285.0  45.627 1.997e-09 ***
## Type       1 3365.5  3365.5  67.204 3.061e-12 ***
## Residuals 81 4056.4    50.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



# "Main effects"

```
summary(fit_plants_1_nointeraction)
```

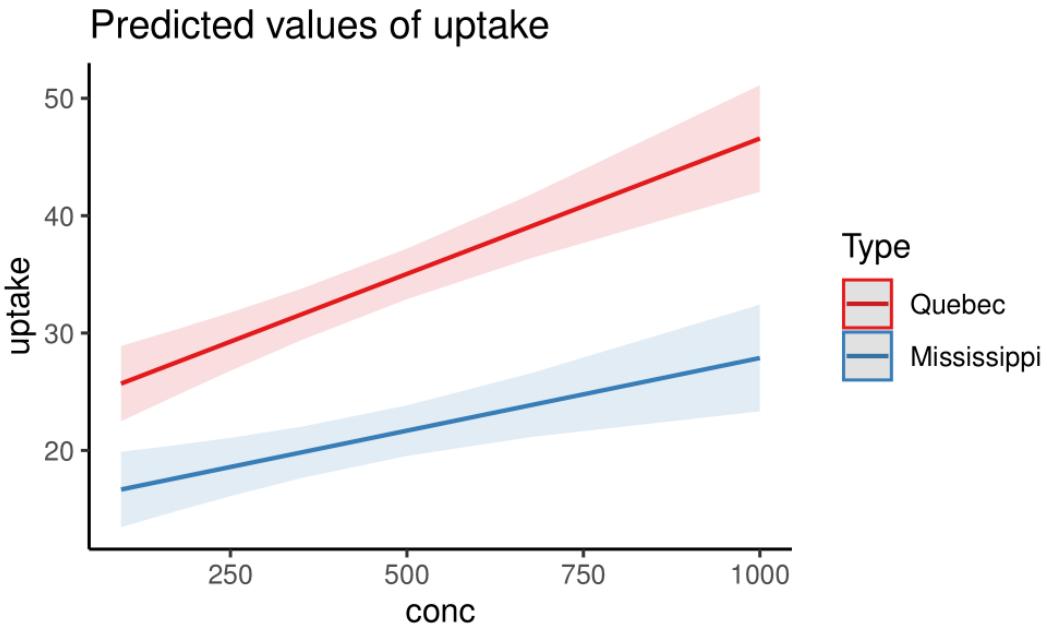
```
##
## Call:
## lm(formula = uptake ~ conc + Type, data = CO2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.2145  -4.2549   0.5479   5.3048  12.9968
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)      25.830052   1.579918  16.349  < 2e-16 ***
## conc              0.017731   0.002625   6.755 2.00e-09 ***
## TypeMississippi -12.659524   1.544261  -8.198 3.06e-12 ***
```



# Interactions

```
fit_plants_1_interaction <- lm(uptake~conc*Type,data=CO2)
anova(fit_plants_1_interaction)
```

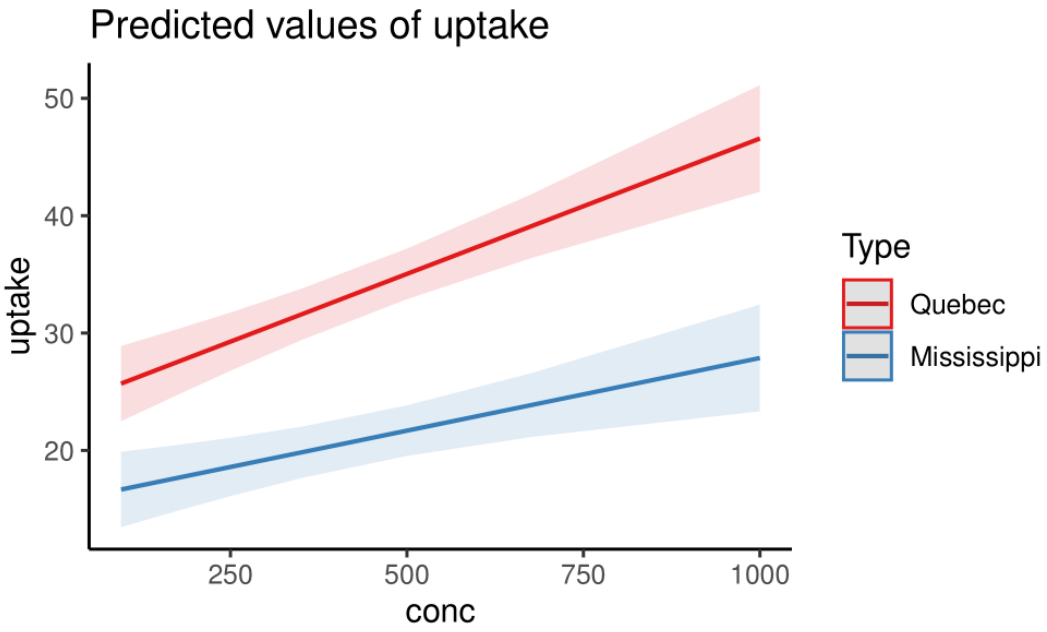
```
## Analysis of Variance Table
##
## Response: uptake
##           Df Sum Sq Mean Sq F value    Pr(>F)
## conc       1 2285.0  2285.0 47.4995 1.143e-09 ***
## Type       1 3365.5  3365.5 69.9614 1.560e-12 ***
## conc:Type  1  208.0   208.0  4.3238   0.04079 *
## Residuals 80 3848.4    48.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



# Interactions

```
summary(fit_plants_1_interaction)
```

```
##
## Call:
## lm(formula = uptake ~ conc * Type, data = CO2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.3956  -5.5250  -0.1604   5.5724  12.0072
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          23.503038   1.910531  12.302  < 2e-16 ***
## conc                  0.023080   0.003638   6.344 1.25e-08 ***
## TypeMississippi      -8.005495   2.701899  -2.963  0.00401 **
## conc:TypeMississippi -0.010699   0.005145  -2.079  0.04079 *
```



# *p*-hacking

There are two widely recognized types of researcher-driven publication bias: selection (also known as the "file drawer effect", where studies with nonsignificant results have lower publication rates [7]) and inflation [12]. Inflation bias, also known as "p-hacking" or "selective reporting," is the misreporting of true effect sizes in published studies (Box 1). It occurs when researchers try out several statistical analyses and/or data eligibility specifications and then selectively report those that produce significant results [12–15]. Common practices that lead to p-hacking include: conducting analyses midway through experiments to decide whether to continue collecting data [15,16]; recording many response variables and deciding which to report postanalysis [16,17], deciding whether to include or drop outliers postanalyses [16], excluding, combining, or splitting treatment groups postanalysis [2], including or excluding covariates postanalysis [14], and stopping data exploration if an analysis yields a significant *p*-value [18,19].

**Citation:** Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD (2015) The Extent and Consequences of P-Hacking in Science. PLoS Biol 13(3): e1002106. doi:10.1371/journal.pbio.1002106