R07- Multiple (Linear) Regression

HCI/PSYCH 522
Iowa State University

April 5, 2022

Multiple regression

Recall the simple linear regression model is

$$Y_i \stackrel{ind}{\sim} N(\mu_i, \sigma^2), \quad \mu_i = \beta_0 + \beta_1 X_i$$

The multiple regression model has mean

$$\mu_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}$$

where for observation i

- Y_i is the dependent variable and
- $X_{i,p}$ is the p^{th} independent variable.

independent variables

There is a lot of flexibility in the mean

$$\mu_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}$$

as there are many possibilities for the independent variables $X_{i,1}, \ldots, X_{i,p}$:

- Functions ($f(X)$)
- Dummy variables for categorical variables ($X_1 = \mathrm{I}()$)
- Higher order terms (X^2)
- Additional independent variables (X_1, X_2)
- Interactions ($X_1 X_2$)
  - Categorical-categorical
  - Continuous-categorical
  - Continuous-continuous

Parameter interpretation

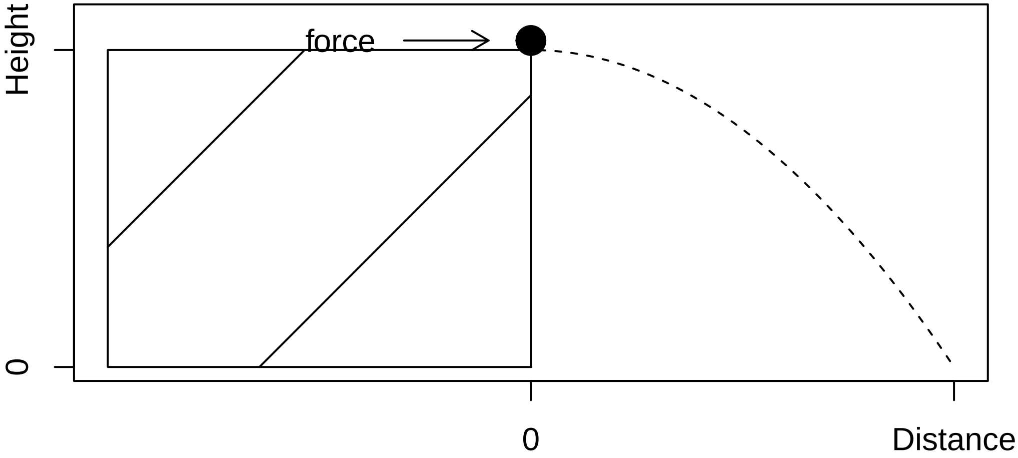
Model:

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}, \sigma^2)$$

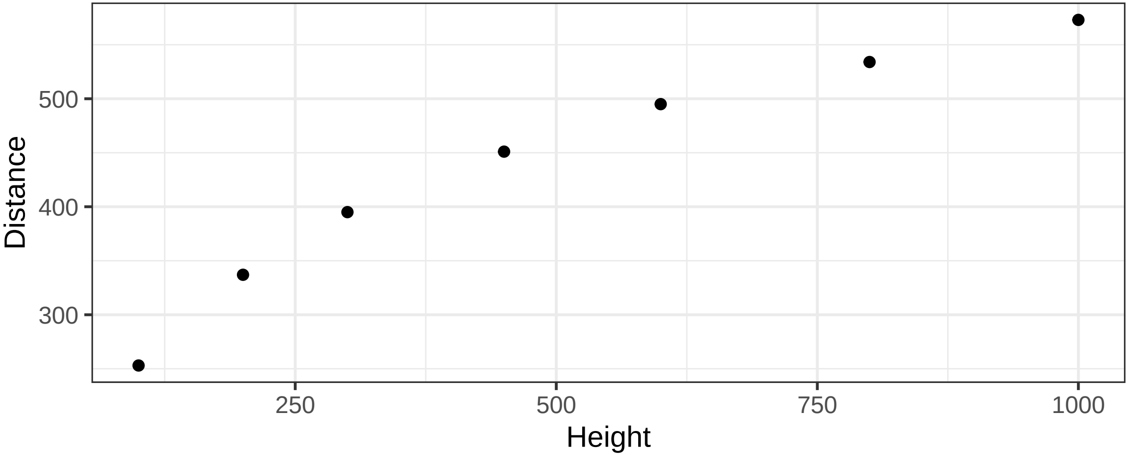
The interpretation is

- β_0 is the mean value of the dependent variable Y_i when all independent variables are zero.
- β_p, $p \neq 0$ is the mean increase in the dependent variable for a one-unit increase in the p^{th} independent variable when all other independent variables are held constant.
- R^2 is the proportion of the variability in the dependent variable explained by the model

Galileo experiment



Galileo data (`Sleuth3::case1001`)



Higher order terms (X^2)

Let

- Y_i be the distance for the i^{th} run of the experiment and
- H_i be the height for the i^{th} run of the experiment.

Simple linear regression assumes

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 H_i \qquad\qquad\qquad\qquad , \sigma^2)$$

The quadratic multiple regression assumes

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 H_i + \beta_2 H_i^2 \qquad\qquad , \sigma^2)$$

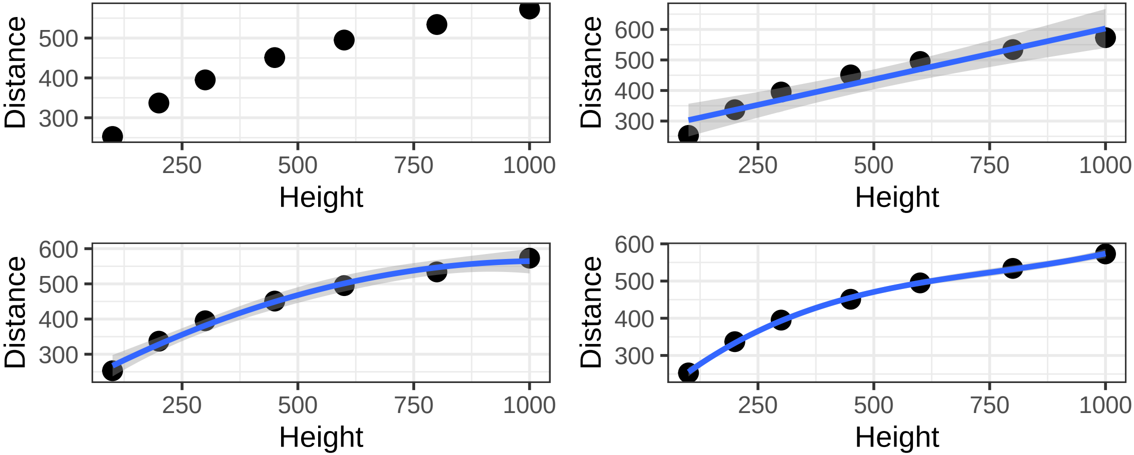
The cubic multiple regression assumes

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 H_i + \beta_2 H_i^2 + \beta_3 H_i^3, \sigma^2)$$

R code and output

```
# Construct the variables by hand
m1 = lm(Distance ~ Height,                              case1001)
m2 = lm(Distance ~ Height + I(Height^2),                case1001)
m3 = lm(Distance ~ Height + I(Height^2) + I(Height^3), case1001)

coefficients(m1)

## (Intercept)      Height
##  269.712458    0.333337

coefficients(m2)

##   (Intercept)        Height   I(Height^2)
##  1.999128e+02  7.083225e-01 -3.436937e-04

coefficients(m3)

##   (Intercept)        Height   I(Height^2)   I(Height^3)
##  1.557755e+02  1.115298e+00 -1.244943e-03  5.477104e-07
```

Galileo experiment (Sleuth3::case1001)



Longnose Dace Abundance

From `http://udel.edu/~mcdonald/statmultreg.html`:

> *I extracted some data from the Maryland Biological Stream Survey. ... The [dependent] variable is the number of Longnose Dace ... per 75-meter section of [a] stream. The [independent] variables are ... the maximum depth (in cm) of the 75-meter segment of stream; nitrate concentration (mg/liter) ....*

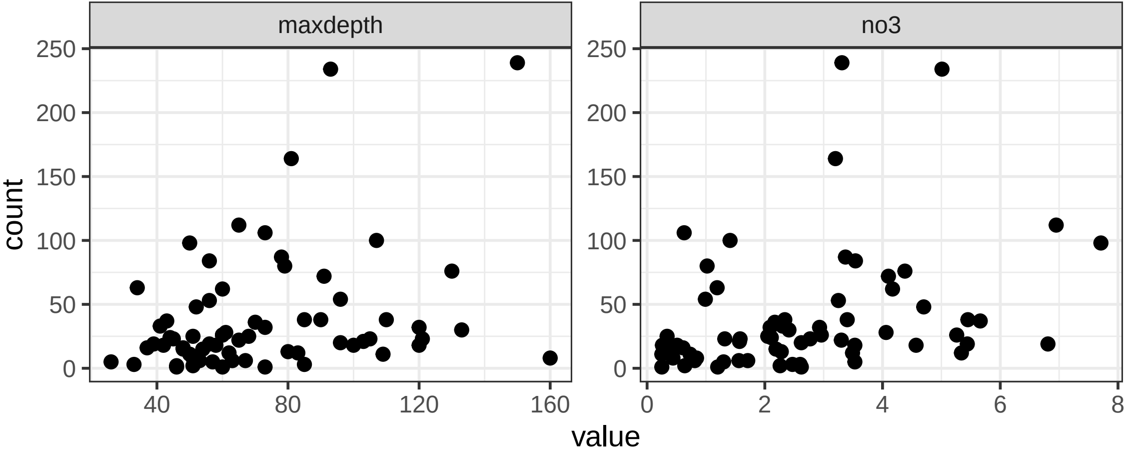
Consider the model

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2}, \sigma^2)$$

where

- Y_i: count of Longnose Dace in stream i
- $X_{i,1}$: maximum depth (in cm) of stream i
- $X_{i,2}$: nitrate concentration (mg/liter) of stream i

Exploratory



R code and output

```
m <- lm(count ~ maxdepth + no3, longnosedace)
summary(m)

##
## Call:
## lm(formula = count ~ maxdepth + no3, data = longnosedace)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.060 -27.704  -8.679  11.794 165.310
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5550    15.9586  -1.100  0.27544
## maxdepth      0.4811     0.1811   2.656  0.00997 **
## no3           8.2847     2.9566   2.802  0.00671 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.39 on 64 degrees of freedom
## Multiple R-squared:  0.1936,Adjusted R-squared:  0.1684
## F-statistic: 7.682 on 2 and 64 DF,  p-value: 0.001022
```

R code and output

```
ci <- confint(m)
ci

##                   2.5 %     97.5 %
## (Intercept) -49.4361015 14.3260349
## maxdepth      0.1192458  0.8428725
## no3           2.3782494 14.1912007
```

Interpretation

- Intercept (β_0): The mean count of Longnose Dace when maximum depth and nitrate concentration are both zero is -18 (-49, 14).
- Coefficient for maxdepth (β_1): Holding nitrate concentration constant, each cm increase in maximum depth is associated with an additional 0.48 (0.12, 0.84) Longnose Dace counted on average.
- Coefficient for no3 (β_2): Holding maximum depth constant, each mg/L increase in nitrate concentration is associated with an addition 8.3 (2.4, 14.2) Longnose Dace counted on average.
- Coefficient of determination (R^2): The model explains 19% of the variability in the count of Longnose Dace.

Interactions

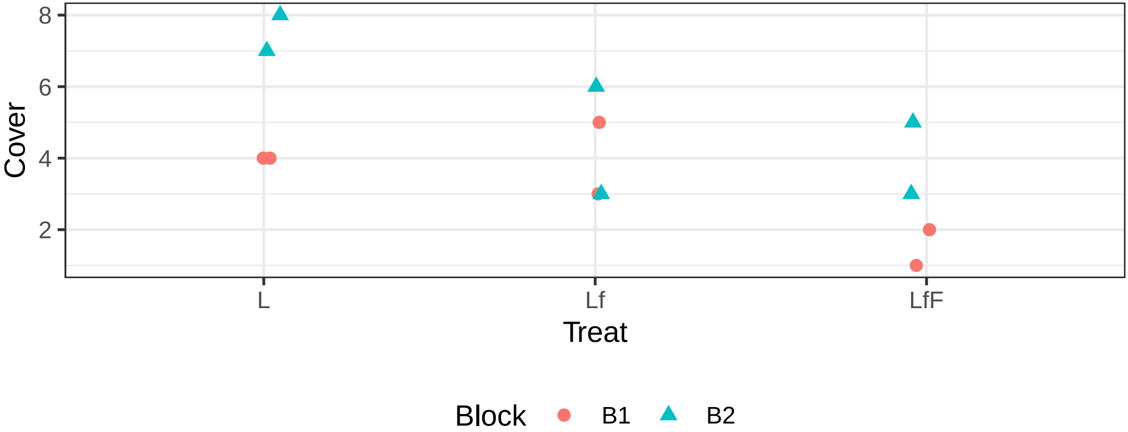
Why an interaction?

> *Two independent variables are said to interact if the effect that one of them has on the mean dependent variable depends on the value of the other.*

For example,

- Crop yield: the effect of tillage method depends on the fertilizer brand (Categorical-categorical)
- Energy expenditure: The effect of mass depends on the species type. (Continuous-categorical)
- Longnose dace count: The effect of nitrate (no3) on longnose dace count depends on the maxdepth. (Continuous-continuous)

Seaweed regeneration (Sleuth3::case1301 subset)



Categorical-categorical

Let category A and type 0 be the reference level. For observation i, let

- Y_i be the dependent variable,
- 1_i be a dummy variable for type 1,
- B_i be a dummy variable for category B, and
- C_i be a dummy variable for category C.

The mean containing only main effects is

$$\mu_i = \beta_0 + \beta_1 1_i + \beta_2 B_i + \beta_3 C_i.$$

The mean with an interaction is

$$\mu_i = \beta_0 + \beta_1 1_i + \beta_2 B_i + \beta_3 C_i + \beta_4 1_i B_i + \beta_5 1_i C_i.$$

Interpretation for the main effects model

The mean containing only main effects is

$$\mu_i = \beta_0 + \beta_1 1_i + \beta_2 B_i + \beta_3 C_i.$$

The means in the main effect model are

	Category		
Type	A	B	C
0	β_0	$\beta_0 \quad + \beta_2$	$\beta_0 \quad + \beta_3$
1	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2$	$\beta_0 + \beta_1 + \beta_3$

Interpretation for the model with an interaction

The mean with an interaction is

$$\mu_i = \beta_0 + \beta_1 1_i + \beta_2 B_i + \beta_3 C_i + \beta_4 1_i B_i + \beta_5 1_i C_i.$$

The means are

	Category				
Type	A	B		C	
0	β_0	β_0	$+\beta_2$	β_0	$+\beta_3$
1	$\beta_0 + \beta_1$	$\beta_0 + \beta_1$	$+\beta_2 + \beta_4$	$\beta_0 + \beta_1$	$+\beta_3 + \beta_5$

This is equivalent to a cell-means model where each combination has its own mean.

R code and output - main effects only

```
##
## Call:
## lm(formula = Cover ~ Block + Treat, data = case1301_subset)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3333 -0.6667  0.0000  0.7917  1.8333
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4.6667     0.7683   6.074 0.000298 ***
## BlockB2       2.1667     0.7683   2.820 0.022491 *
## TreatLf      -1.5000     0.9410  -1.594 0.149578
## TreatLfF     -3.0000     0.9410  -3.188 0.012838 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.331 on 8 degrees of freedom
## Multiple R-squared:  0.6937,Adjusted R-squared:  0.5788
## F-statistic: 6.039 on 3 and 8 DF,  p-value: 0.01881
```

Credible intervals

```
##                  2.5 %     97.5 %
## (Intercept)  2.8949744  6.4383590
## BlockB2      0.3949744  3.9383590
## TreatLf     -3.6698711  0.6698711
## TreatLfF    -5.1698711 -0.8301289
```

Interpretation

- In block B1 and treatment L, the mean cover is 4.7 (2.9, 6.4).
- The mean increase in cover from block B1 to block B2 is 2.2 (0.4, 3.9) while holding treatment constant.
- The mean increase in cover from treatment L to treatment Lf is -1.5 (-3.7, 0.7) while holding block constant.
- The mean increase in cover from treatment L to treatment LfF is -3 (-5.2, -0.8) while holding block constant.
- The model with block and treatment explains 69% of the variability in cover.

R code and output - with an interaction

```
##
## Call:
## lm(formula = Cover ~ Block * Treat, data = case1301_subset)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -1.500 -0.625  0.000  0.625  1.500
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.000e+00  8.898e-01   4.496  0.00412 **
## BlockB2          3.500e+00  1.258e+00   2.782  0.03193 *
## TreatLf         -1.976e-15  1.258e+00   0.000  1.00000
## TreatLfF        -2.500e+00  1.258e+00  -1.987  0.09413 .
## BlockB2:TreatLf  -3.000e+00  1.780e+00  -1.686  0.14280
## BlockB2:TreatLfF -1.000e+00  1.780e+00  -0.562  0.59450
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.258 on 6 degrees of freedom
## Multiple R-squared:  0.7946,Adjusted R-squared:  0.6234
## F-statistic: 4.642 on 5 and 6 DF,  p-value: 0.04429
```

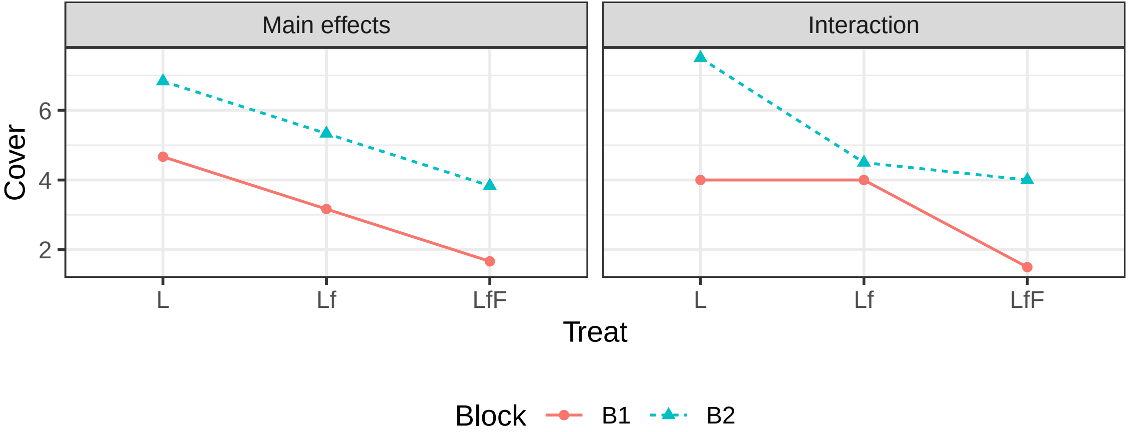
Credible intervals

```
##                       2.5 %    97.5 %
## (Intercept)       1.8228442 6.1771558
## BlockB2           0.4210368 6.5789632
## TreatLf          -3.0789632 3.0789632
## TreatLfF         -5.5789632 0.5789632
## BlockB2:TreatLf  -7.3543116 1.3543116
## BlockB2:TreatLfF -5.3543116 3.3543116
```

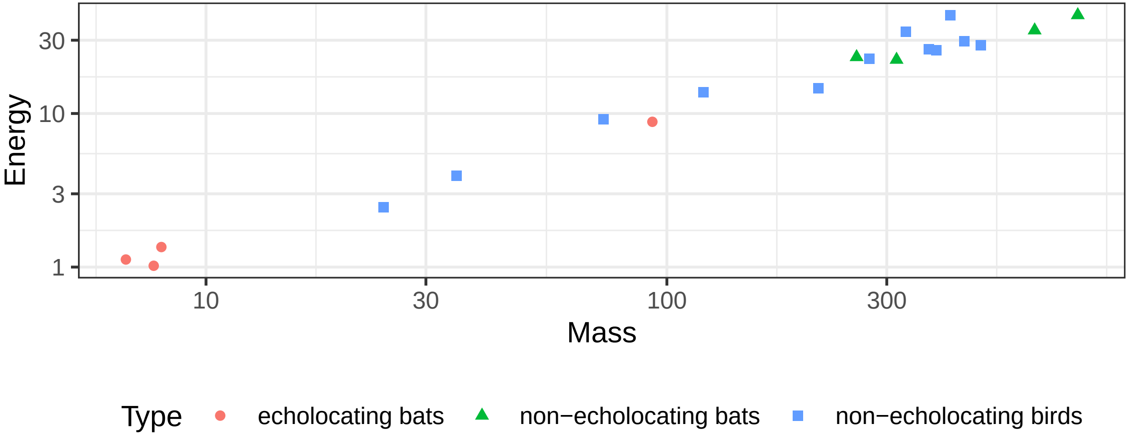
Interpretation

- In block B1 and treatment L, the mean cover is 4 (1.8, 6.2).
- The mean increase in cover from block B1 to block B2 is 3.5 (0.4, 6.6) in treatment L.
- The mean increase in cover from treatment L to treatment Lf is 0 (-3.1, 3.1) in block B1.
- The mean increase in cover from treatment L to treatment LfF is -2.5 (-5.6, 0.6) in block B1.
- The model that includes block, treatment, and their interaction explains 79% of the variability in cover.

Visualizing the models



In-flight energy expenditure (Sleuth3::case1002)



Continuous-categorical interaction

Let category A be the reference level. For observation i, let

- Y_i be the dependent variable
- $X_{i,1}$ be the continuous independent variable,
- B_i be a dummy variable for category B, and
- C_i be a dummy variable for category C.

The mean containing only main effects is

$$\mu_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 B_i + \beta_3 C_i.$$

The mean with the interaction is

$$\mu_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 B_i + \beta_3 C_i + \beta_4 X_{i,1} B_i + \beta_5 X_{i,1} C_i.$$

Interpretation for the main effect model

The mean containing only main effects is

$$\mu_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 B_i + \beta_3 C_i.$$

For each category, the line is

Category	Line (μ)		
A	β_0	$+$	$\beta_1 X$
B	$(\beta_0 + \beta_2)$	$+$	$\beta_1 X$
C	$(\beta_0 + \beta_3)$	$+$	$\beta_1 X$

Each category has a different intercept, but a common slope.

Interpretation for the model with an interaction

The model with an interaction is

$$\mu_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 B_i + \beta_3 C_i + \beta_4 X_{i,1} B_i + \beta_5 X_{i,1} C_i$$

For each category, the line is

Category	Line (μ)		
A	β_0	$+\ \beta_1$	X
B	$(\beta_0 + \beta_2)$	$+(\beta_1 + \beta_4)$	X
C	$(\beta_0 + \beta_3)$	$+(\beta_1 + \beta_5)$	X

Each category has its own intercept and its own slope.

R code and output - main effects only

```
summary(mM <- lm(log(Energy) ~ log(Mass) + Type, case1002))

##
## Call:
## lm(formula = log(Energy) ~ log(Mass) + Type, data = case1002)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23224 -0.12199 -0.03637  0.12574  0.34457
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  -1.49770    0.14987  -9.993 2.77e-08 ***
## log(Mass)                     0.81496    0.04454  18.297 3.76e-12 ***
## Typenon-echolocating bats    -0.07866    0.20268  -0.388    0.703
## Typenon-echolocating birds    0.02360    0.15760   0.150    0.883
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.186 on 16 degrees of freedom
## Multiple R-squared:  0.9815,Adjusted R-squared:  0.9781
## F-statistic: 283.6 on 3 and 16 DF,  p-value: 4.464e-14
```

Credible intervals

```
##                                  2.5 %     97.5 %
## (Intercept)                 -1.8154046 -1.1799884
## log(Mass)                    0.7205339  0.9093811
## Typenon-echolocating bats   -0.5083245  0.3509972
## Typenon-echolocating birds  -0.3104999  0.3576964
```

Interpretation

- For echo-locating bats that weigh 1 gram (log(Mass) is 0), the mean energy expenditure is -1.5 (-1.8, -1.2) watts.
- The mean increase in the logarithm of energy expenditure per unit increase in the logarithm of mass is 0.8 (0.7, 0.9) while holding type of bird or bat constant.
- The mean increase in the logarithm of energy expenditure from echolocating bats to non-echolocating bats is -0.1 (-0.5, 0.4) while mass constant.
- The mean increase in the logarithm of energy expenditure from echolocating bats to non-echolocating birds is 0 (-0.3, 0.4) while mass constant.
- The model the includes main effects for the logarithm of mass and the type of bird or bat explains 98% of the variability in the logarithm of energy expenditure.

R code and output - with an interaction

```
summary(mI <- lm(log(Energy) ~ log(Mass) * Type, case1002))

##
## Call:
## lm(formula = log(Energy) ~ log(Mass) * Type, data = case1002)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25152 -0.12643 -0.00954  0.08124  0.32840
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        -1.47052    0.24767  -5.937 3.63e-05 ***
## log(Mass)                           0.80466    0.08668   9.283 2.33e-07 ***
## Typenon-echolocating bats           1.26807    1.28542   0.987    0.341
## Typenon-echolocating birds         -0.11032    0.38474  -0.287    0.779
## log(Mass):Typenon-echolocating bats -0.21487    0.22362  -0.961    0.353
## log(Mass):Typenon-echolocating birds  0.03071    0.10283   0.299    0.770
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1899 on 14 degrees of freedom
```

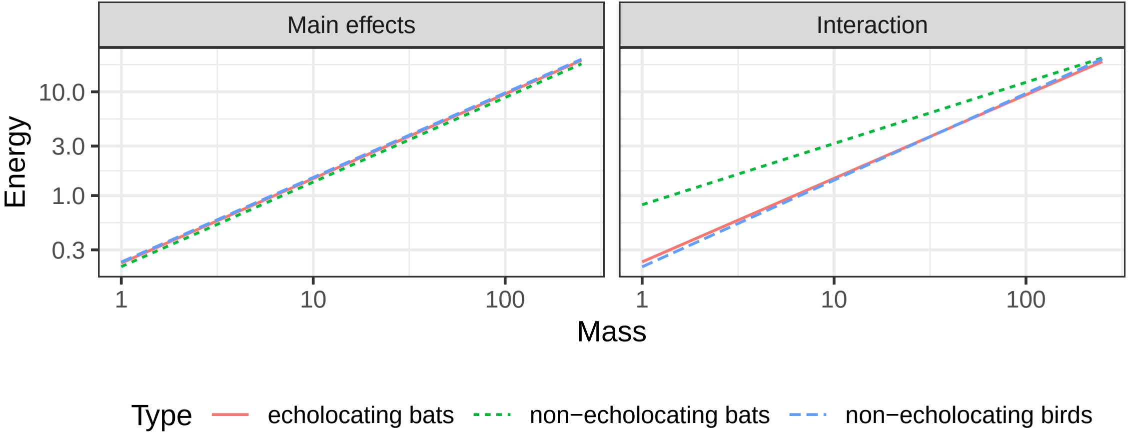
Credible intervals

```
##                                          2.5 %     97.5 %
## (Intercept)                          -2.0017153 -0.9393152
## log(Mass)                             0.6187372  0.9905769
## Typenon-echolocating bats            -1.4888841  4.0250195
## Typenon-echolocating birds           -0.9355124  0.7148674
## log(Mass):Typenon-echolocating bats  -0.6944979  0.2647479
## log(Mass):Typenon-echolocating birds -0.1898417  0.2512682
```

Interpretation

- For echo-locating bats that weigh 1 gram (log(Mass) is 0), the mean logarithm of energy expenditure is -1.5 (-2, -0.9) watts.
- The mean increase in the logarithm of energy expenditure per unit increase in the logarithm of mass is 0.8 (0.6, 1) for echolocating bats.
- The mean increase in the logarithm of energy expenditure from echolocating bats to non-echolocating bats is 1.3 (-1.5, 4) when mass is 1 (log(Mass is 0).
- The mean increase in the logarithm of energy expenditure from echolocating bats to non-echolocating birds is -0.1 (-0.9, 0.7) when mass is 1 (log(Mass is 0).
- The model that includes the logarithm of mass, the type of bird and bat, and their interaction explains 98% of the variability in the logarithm of energy.

Visualizing the models



Continuous-continuous interaction

For observation i, let

- Y_i be the dependent variable
- $X_{i,1}$ be the first continuous independent variable and
- $X_{i,2}$ be the second continuous independent variable.

The mean containing only main effects is

$$\mu_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2}.$$

The mean with the interaction is

$$\mu_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,1} X_{i,2}.$$

Intepretation - main effects only

Let $X_{i,1} = x_1$ and $X_{i,2} = x_2$, then we can rewrite the line (μ) as

$$\mu = (\beta_0 + \beta_2 x_2) + \beta_1 x_1$$

which indicates that the intercept of the line for x_1 depends on the value of x_2.

Similarly,

$$\mu = (\beta_0 + \beta_1 x_1) + \beta_2 x_2$$

which indicates that the intercept of the line for x_2 depends on the value of x_1.

Intepretation - with an interaction

Let $X_{i,1} = x_1$ and $X_{i,2} = x_2$, then we can rewrite the mean (μ) as

$$\mu = (\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_3 x_2) x_1$$

which indicates that both the intercept and slope for x_1 depend on the value of x_2.

Similarly,

$$\mu = (\beta_0 + \beta_1 x_1) + (\beta_2 + \beta_3 x_1) x_2$$

which indicates that both the intercept and slope for x_2 depend on the value of x_1.

R code and output - main effects only

```
##
## Call:
## lm(formula = count ~ no3 + maxdepth, data = longnosedace)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -55.060 -27.704  -8.679  11.794 165.310
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5550    15.9586  -1.100  0.27544
## no3           8.2847     2.9566   2.802  0.00671 **
## maxdepth      0.4811     0.1811   2.656  0.00997 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.39 on 64 degrees of freedom
## Multiple R-squared:  0.1936,Adjusted R-squared:  0.1684
## F-statistic: 7.682 on 2 and 64 DF,  p-value: 0.001022
```

Credible intervals

```
##                  2.5 %     97.5 %
## (Intercept) -49.4361015 14.3260349
## no3           2.3782494 14.1912007
## maxdepth      0.1192458  0.8428725
```

Interpretation

- When nitrate and maximum depth are both zero, the mean longnosedace count is -17.6 (-49.4, 14.3).
- When nitrate increases by 1 mg/L, the mean longnosedace count increases by 8.3 (2.4, 14.2) while holding maximum depth constant.
- When maximum depth increases by 1 cm, the mean longnosedace count increases by 0.5 (0.1, 0.8) while holding nitrate constant.
- The main effects model explains 19% of the variability in longnose dace count.

R code and output - with an interaction

```
##
## Call:
## lm(formula = count ~ no3 * maxdepth, data = longnosedace)
##
## Residuals:
##      Min       1Q  Median       3Q      Max
## -65.111 -21.399  -9.562   5.953 151.071
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.321043  23.455710   0.568   0.5721
## no3          -4.646272   7.856932  -0.591   0.5564
## maxdepth     -0.009338   0.329180  -0.028   0.9775
## no3:maxdepth  0.201219   0.113576   1.772   0.0813 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 42.68 on 63 degrees of freedom
## Multiple R-squared:  0.2319,Adjusted R-squared:  0.1953
## F-statistic: 6.339 on 3 and 63 DF,  p-value: 0.0007966
```

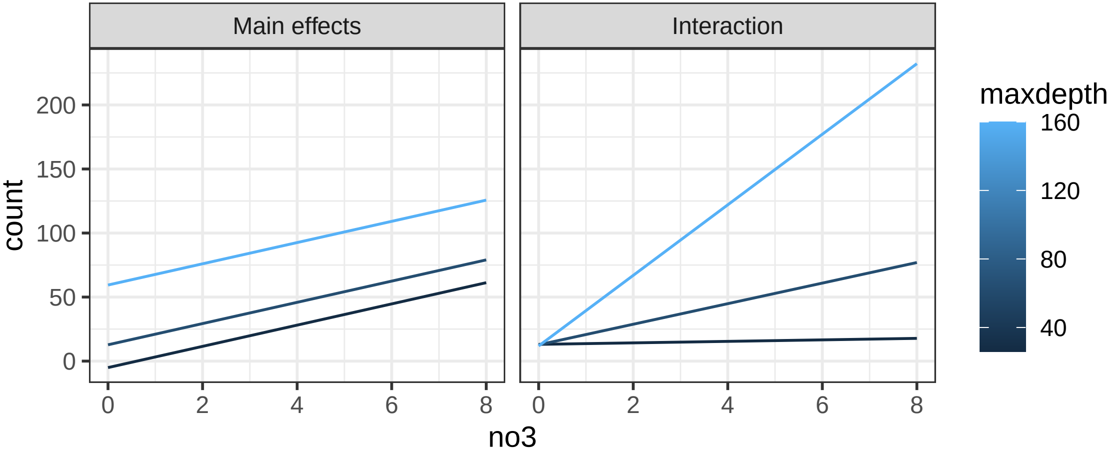
Credible intervals

```
##                    2.5 %     97.5 %
## (Intercept)  -33.55145355 60.1935389
## no3          -20.34709813 11.0545539
## maxdepth      -0.66715251  0.6484768
## no3:maxdepth  -0.02574574  0.4281832
```

Interpretation

- When nitrate and maximum depth are both zero, the mean longnosedace count is 13.3 (-33.6, 60.2).
- When nitrate increases by 1 mg/L, the mean longnosedace count increases by -4.6 (-20.3, 11.1) when maximum depth is zero.
- When maximum depth increases by 1 cm, the mean longnosedace count increases by 0 (-0.7, 0.6) when nitrate is zero.
- The model with maximum depth, nitrate, and their interaction explains 23% of the variability in longnose dace count.

Visualizing the model



When to include interaction terms

From The Statistical Sleuth (3rd ed) page 250:

- when a question of interest pertains to an interaction
- when good reason exists to suspect an interaction or
- when interactions are proposed as a more general model for the purpose of examining the goodness of fit of a model without interaction.

Multiple regression independent variables

The possibilities for independent variables are

- Higher order terms (X^2)
- Additional independent variables (X_1 and X_2)
- Dummy variables for categorical variables ($X_1 = \mathrm{I}()$)
- Interactions (X_1X_2)
  - Categorical-categorical
  - Continuous-categorical
  - Continuous-continuous

We can also combine these independent variables, e.g.

- including higher order terms for continuous variables along with dummy variables for categorical variables and
- including higher order interactions ($X_1X_2X_3$).