STAT 401A - Statistical Methods for Research Workers

Multiple regression models

Jarad Niemi (Dr. J)

Iowa State University

last updated: November 7, 2014

Multiple regression

Recall the simple linear regression model is

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

The multiple regression model is

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

where

- Y_i is the response for observation i and
- $X_{i,p}$ is the p^{th} explanatory variable for observation i.

We may also write

$$Y_i \stackrel{ind}{\sim} N(\mu_i, \sigma^2) \quad \text{or} \quad Y_i = \mu_i + e_i, e_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

where

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

Explanatory variables

There is a lot of flexibility in the mean

$$\mu_i = E[Y_i|X_{i,1}, \ldots, X_{i,p}] = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}$$

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

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

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

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

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

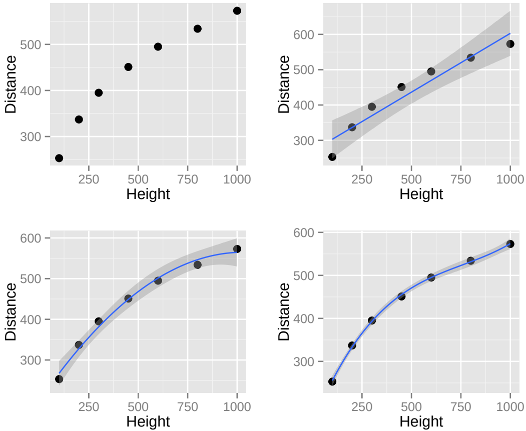
The quadratic multiple regression assumes

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

The cubic multiple regression assumes

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

Case1001



SAS code and output

```
DATA case1001;
  INFILE 'case1001.csv' DSD FIRSTOBS=2;
  INPUT distance height;
  height2 = height*height;
  height3 = height*height2;

# PROC REG allows multiple MODEL statements
PROC REG DATA=case1001;
  MODEL distance = height;
  MODEL distance = height height2;
  MODEL distance = height height2 height3;
  RUN;
```

```
                              Parameter Estimates

              Parameter     Standard
Variable  DF   Estimate        Error   t Value   Pr > |t|

Intercept  1  269.71246     24.31239     11.09     0.0001
height     1    0.33334      0.04203      7.93     0.0005

Intercept  1  199.91282     16.75945     11.93     0.0003
height     1    0.70832      0.07482      9.47     0.0007
height2    1 -0.00034369   0.00006678    -5.15     0.0068

Intercept  1  155.77551      8.32579     18.71     0.0003
height     1    1.11530      0.06567     16.98     0.0004
height2    1   -0.00124   0.00013842     -8.99     0.0029
height3    1 5.477104E-7  8.327329E-8     6.58     0.0072
```

SAS code and output

```
DATA case1001;
  INFILE 'case1001.csv' DSD FIRSTOBS=2;
  INPUT distance height;
  height2 = height ** 2;
  height3 = height ** 3;

PROC GLM DATA=case1001;
  MODEL distance = height height2 height3;
```

```
/* PROC GLM allows the variable construction within the MODEL statement
   and provides nicer output (not shown here) */
DATA case1001;
  INFILE 'case1001.csv' DSD FIRSTOBS=2;
  INPUT distance height;

/* This shorthand puts in H, H^2, and H^3 */
PROC GLM DATA=case1001;
  MODEL distance = height|height|height;

/* This only puts H^3 */
PROC GLM DATA=case1001;
  MODEL distance = height*height*height;
```

R code and output

```
# Construct the variables by hand
case1001$Height2 = case1001$Height^2
case1001$Height3 = case1001$Height^3

m1 = lm(Distance~Height,                 case1001)
m2 = lm(Distance~Height+Height2,         case1001)
m3 = lm(Distance~Height+Height2+Height3, case1001)

coefficients(m1)


(Intercept)      Height
   269.7125      0.3333


coefficients(m2)


(Intercept)       Height     Height2
  1.999e+02   7.083e-01  -3.437e-04


coefficients(m3)


(Intercept)       Height     Height2      Height3
  1.558e+02   1.115e+00  -1.245e-03   5.477e-07
```

R code and output

```
# Let R construct the variables for you
m = lm(Distance~poly(Height, 3, raw=TRUE), case1001)
summary(m)
```

```
Call:
lm(formula = Distance ~ poly(Height, 3, raw = TRUE), data = case1001)

Residuals:
      1       2       3       4       5       6       7
-2.4036  3.5809  1.8917 -4.4688 -0.0804  2.3216 -0.8414

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                   1.56e+02   8.33e+00   18.71  0.00033 ***
poly(Height, 3, raw = TRUE)1  1.12e+00   6.57e-02   16.98  0.00044 ***
poly(Height, 3, raw = TRUE)2 -1.24e-03   1.38e-04   -8.99  0.00290 **
poly(Height, 3, raw = TRUE)3  5.48e-07   8.33e-08    6.58  0.00715 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.01 on 3 degrees of freedom
Multiple R-squared:  0.999, Adjusted R-squared:  0.999
F-statistic: 1.6e+03 on 3 and 3 DF,  p-value: 2.66e-05
```

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 (Rhinichthys cataractae) per 75-meter section of [a] stream. The independent variables are the area (in acres) drained by the stream; the dissolved oxygen (in mg/liter); the maximum depth (in cm) of the 75-meter segment of stream; nitrate concentration (mg/liter); sulfate concentration (mg/liter); and the water temperature on the sampling date (in degrees C).*

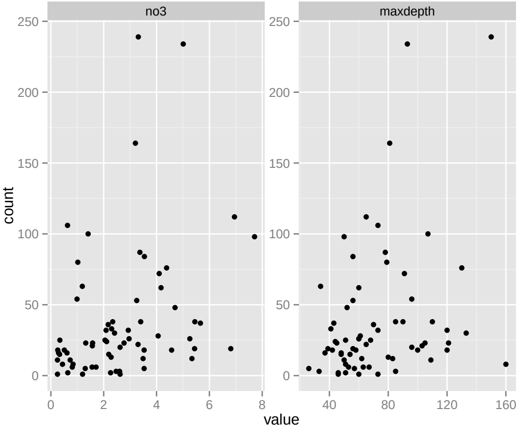
Consider the model

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

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



```
DATA dace;
  INFILE 'Longnose Dace.csv' DSD FIRSTOBS=2;
  INPUT stream $ count acreage do2 maxdepth no3 so4 temp;

PROC REG DATA=dace;
  MODEL count = maxdepth no3;
  RUN;
```

```
                                The REG Procedure
                                 Model: MODEL1
                            Dependent Variable: count

                       Number of Observations Read          67
                       Number of Observations Used          67


                               Analysis of Variance
                                      Sum of           Mean
        Source                DF     Squares         Square    F Value    Pr > F
        Model                  2       28930          14465       7.68    0.0010
        Error                 64      120503     1882.85220
        Corrected Total       66      149432

                    Root MSE             43.39184    R-Square     0.1936
                    Dependent Mean       39.10448    Adj R-Sq     0.1684
                    Coeff Var           110.96388

                               Parameter Estimates
                                 Parameter       Standard
             Variable     DF      Estimate          Error    t Value    Pr > |t|
             Intercept     1     -17.55503       15.95865      -1.10      0.2754
             maxdepth      1       0.48106        0.18111       2.66      0.0100
             no3           1       8.28473        2.95659       2.80      0.0067
```

R code and output

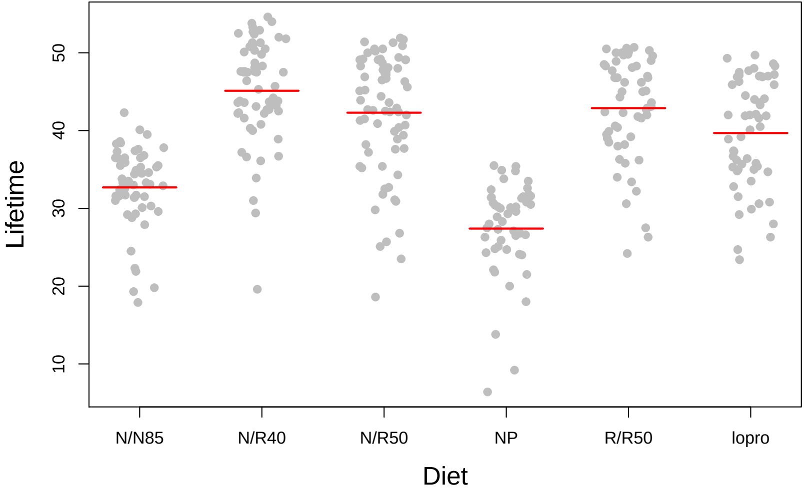
```
d = read.csv("longnosedace.csv")
m = lm(count~no3+maxdepth,d)
summary(m)
```

```
Call:
lm(formula = count ~ no3 + maxdepth, data = d)

Residuals:
   Min     1Q Median     3Q    Max
-55.06 -27.70  -8.68  11.79 165.31

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -17.555     15.959   -1.10   0.2754
no3            8.285      2.957    2.80   0.0067 **
maxdepth       0.481      0.181    2.66   0.0100 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43.4 on 64 degrees of freedom
Multiple R-squared:  0.194, Adjusted R-squared:  0.168
F-statistic: 7.68 on 2 and 64 DF,  p-value: 0.00102
```

Interpretation

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

Using a categorical variable as an explanatory variable.



Regression with a categorical variable

- Choose one of the levels as the reference level, e.g. N/N85
- Construct dummy variables using indicator functions, i.e.

$$\mathrm{I}(A) = \begin{cases} 1 & A \text{ is TRUE} \\ 0 & A \text{ is FALSE} \end{cases}$$

  for the other levels, e.g.

$$\begin{aligned} X_{i,1} &= \mathrm{I}(\text{diet for observation } i \text{ is N/R40}) \\ X_{i,2} &= \mathrm{I}(\text{diet for observation } i \text{ is N/R50}) \\ X_{i,3} &= \mathrm{I}(\text{diet for observation } i \text{ is NP}) \\ X_{i,4} &= \mathrm{I}(\text{diet for observation } i \text{ is R/R50}) \\ X_{i,5} &= \mathrm{I}(\text{diet for observation } i \text{ is lopro}) \end{aligned}$$

- Estimate the parameters of a multiple regression model using these dummy variables.

SAS code and output

```
DATA case0501;
  INFILE 'case0501.csv' DSD FIRSTOBS=2;
  INPUT lifetime diet $;

PROC GLM DATA=case0501;
  CLASS diet(REF='N/N85'); /* by default, SAS uses the alphabetically last group as the reference level */
  MODEL lifetime=diet / SOLUTION;
  RUN;
                                       The GLM Procedure

Dependent Variable: lifetime
                                                Sum of
       Source                      DF          Squares     Mean Square    F Value    Pr > F
       Model                        5      12733.94181      2546.78836      57.10    <.0001
       Error                      343      15297.41532        44.59888
       Corrected Total            348      28031.35713

                      R-Square     Coeff Var      Root MSE    lifetime Mean
                      0.454275      17.21323      6.678239         38.79713

                                                       Standard
           Parameter                 Estimate             Error    t Value    Pr > |t|
           Intercept              32.69122807 B      0.88455439      36.96      <.0001
           diet      N/R40        12.42543860 B      1.23521298      10.06      <.0001
           diet      N/R50         9.60595503 B      1.18768248       8.09      <.0001
           diet      NP           -5.28918725 B      1.30100640      -4.07      <.0001
           diet      R/R50        10.19448622 B      1.25652099       8.11      <.0001
           diet      lopro         6.99448622 B      1.25652099       5.57      <.0001
           diet      N/N85         0.00000000 B       .                .           .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the
```

R code and output

```
# by default, R uses the alphabetically first group as the reference level
case0501$Diet = relevel(case0501$Diet, ref='N/N85')

m = lm(Lifetime~Diet, case0501)
summary(m)



Call:
lm(formula = Lifetime ~ Diet, data = case0501)

Residuals:
    Min      1Q  Median      3Q     Max
-25.517  -3.386   0.814   5.183  10.014

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   32.691      0.885   36.96  < 2e-16 ***
DietN/R40     12.425      1.235   10.06  < 2e-16 ***
DietN/R50      9.606      1.188    8.09  1.1e-14 ***
DietNP        -5.289      1.301   -4.07  5.9e-05 ***
DietR/R50     10.194      1.257    8.11  8.9e-15 ***
Dietlopro      6.994      1.257    5.57  5.2e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.68 on 343 degrees of freedom
Multiple R-squared:  0.454, Adjusted R-squared:  0.446
F-statistic: 57.1 on 5 and 343 DF,  p-value: <2e-16
```

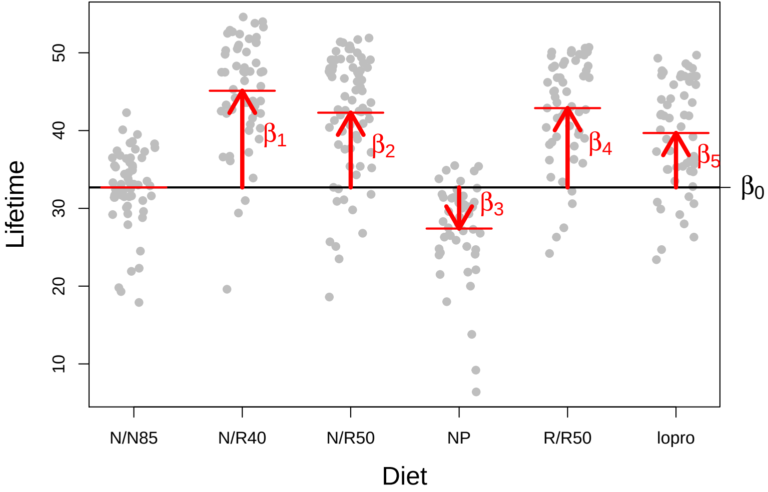
Interpretation

- $\beta_0 = E[Y_i|\text{reference level}]$, i.e. expected response for the reference level
  Note: the only way $X_{i,1} = \cdots = X_{i,p} = 0$ is if all indicators are zero, i.e. at the reference level.
- $\beta_p, p > 0$: expected change in the response moving from the reference level to the level associated with the p^{th} dummy variable
  Note: the only way for $X_{i,p}$ to increase by one and all other indicators to stay constant is if initially $X_{i,1} = \cdots = X_{i,p} = 0$ and now $X_{i,p} = 1$

For example,

- The expected lifetime for mice on the N/N85 diet is 32.7 weeks.
- The expected increase in lifetime for mice on the N/R40 diet compared to the N/N85 diet is 12.4 weeks.
- The model explains 45% of the variability in mice lifetimes.

Using a categorical variable as an explanatory variable.



Interactions

Why an interaction?

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

For example,

- Longnose dace: The effect of nitrate (no3) on longnose dace count depends on the maxdepth. (Continuous-continuous)
- Case1002: The effect of mass on energy depends on the species type. (Continuous-categorical)
- Yield: the effect of tillage method depends on the fertilizer brand (Categorical-categorical)

Continuous-continuous interaction

For observation i, let

- Y_i be the response
- $X_{i,1}$ be the first explanatory variable and
- $X_{i,2}$ be the second explanatory 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

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

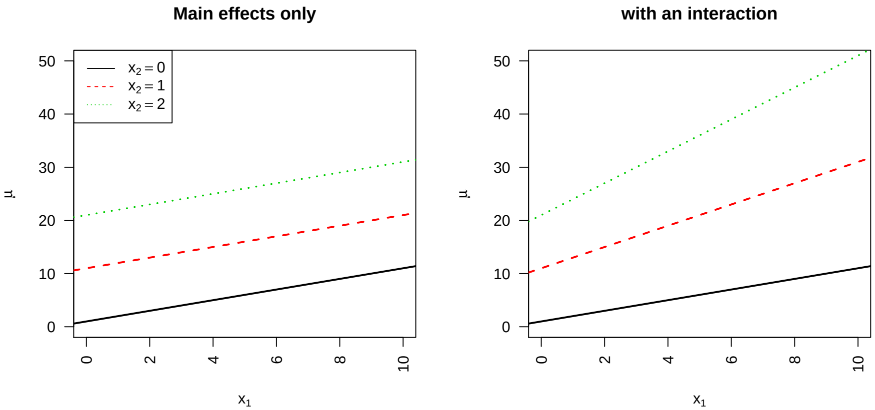
which indicates that both the intercept and slope for x_1 depend on the value of x_2.

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.

Visualizing the models



SAS code and output - main effects only

```
DATA longnosedace;
  INFILE 'longnosedace.csv' DSD FIRSTOBS=2;
  INPUT stream $ count acreage do2 maxdepth no3 so4 temp;

PROC GLM DATA=longnosedace;
  MODEL count = no3 maxdepth;
  RUN;
```

```
                                    The GLM Procedure

Dependent Variable: count

                                             Sum of
       Source                      DF       Squares     Mean Square    F Value    Pr > F
       Model                        2    28929.7279      14464.8639       7.68    0.0010
       Error                       64   120502.5408       1882.8522
       Corrected Total             66   149432.2687

                    R-Square     Coeff Var      Root MSE    count Mean
                    0.193598      110.9639      43.39184      39.10448


                                                   Standard
              Parameter         Estimate              Error    t Value    Pr > |t|
              Intercept     -17.55503330        15.95864994      -1.10      0.2754
              no3             8.28472502         2.95659408       2.80      0.0067
              maxdepth        0.48105914         0.18111227       2.66      0.0100
```

SAS code and output - with an interaction

```
PROC GLM DATA=longnosedace;
  MODEL count = no3|maxdepth;
  RUN;

                                    The GLM Procedure

Dependent Variable: count

                                              Sum of
       Source                      DF        Squares     Mean Square    F Value    Pr > F
       Model                        3     34648.4646      11549.4882       6.34    0.0008
       Error                       63    114783.8040       1821.9651
       Corrected Total             66    149432.2687

                       R-Square     Coeff Var      Root MSE    count Mean
                       0.231867      109.1550      42.68448      39.10448

                                                  Standard
              Parameter          Estimate            Error    t Value    Pr > |t|
              Intercept       13.32104269      23.45570999       0.57      0.5721
              no3             -4.64627211       7.85693213      -0.59      0.5564
              maxdepth        -0.00933787       0.32918045      -0.03      0.9775
              no3*maxdepth     0.20121872       0.11357647       1.77      0.0813
```

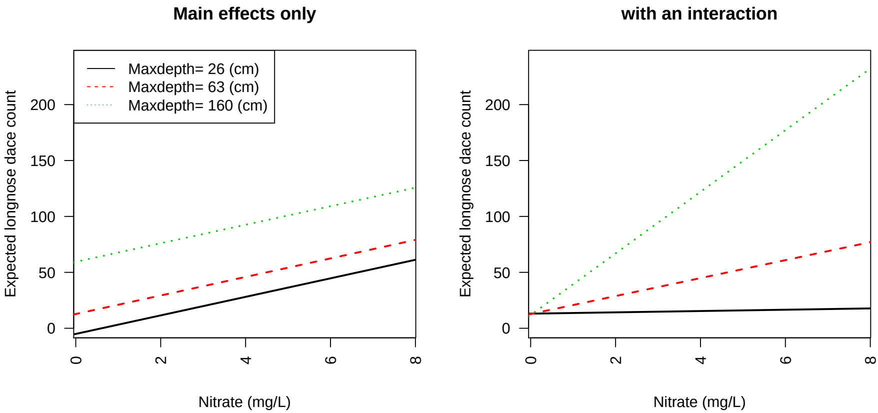
R code and output - main effects only

```
d = read.csv("longnosedace.csv")
mM = lm(count ~ no3+maxdepth, d)
summary(mM)



Call:
lm(formula = count ~ no3 + maxdepth, data = d)

Residuals:
   Min     1Q Median     3Q    Max 
-55.06 -27.70  -8.68  11.79 165.31 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  -17.555     15.959   -1.10   0.2754   
no3            8.285      2.957    2.80   0.0067 **
maxdepth       0.481      0.181    2.66   0.0100 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43.4 on 64 degrees of freedom
Multiple R-squared:  0.194, Adjusted R-squared:  0.168 
F-statistic: 7.68 on 2 and 64 DF,  p-value: 0.00102
```

R code and output - with an interaction

```
mI = lm(count ~ no3*maxdepth, d)
summary(mI)



Call:
lm(formula = count ~ no3 * maxdepth, data = d)

Residuals:
   Min     1Q Median     3Q    Max
-65.11 -21.40  -9.56   5.95 151.07

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  13.32104   23.45571    0.57    0.572
no3          -4.64627    7.85693   -0.59    0.556
maxdepth     -0.00934    0.32918   -0.03    0.977
no3:maxdepth  0.20122    0.11358    1.77    0.081 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 42.7 on 63 degrees of freedom
Multiple R-squared:  0.232, Adjusted R-squared:  0.195
F-statistic: 6.34 on 3 and 63 DF,  p-value: 0.000797
```

Visualizing the model



Continuous-categorical interaction

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

- Y_i be the response
- $X_{i,1}$ be the continuous explanatory 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

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

Think about this model as a different line for each level of the categorical explanatory variable.

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

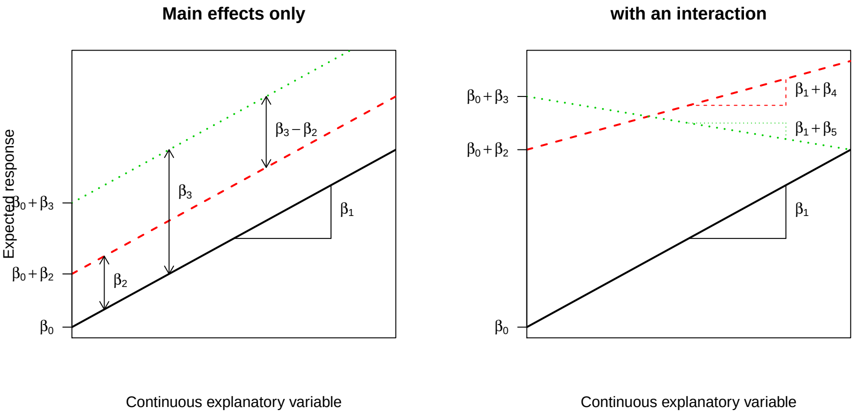
$$\mu_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 B_i + \beta_3 C_i + \beta_4 X_{i,1} B_i + \beta_5 X_{i,1} C_i$$

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.

Visualizing the models



SAS code and output - main effects only

```
DATA case1002;
  INFILE 'case1002.csv' DSD FIRSTOBS=2;
  LENGTH Type $22.;
  INPUT Mass Type $ Energy;
  lMass   = log(Mass);
  lEnergy = log(Energy);

PROC GLM DATA=case1002;
  CLASS Type(REF='non-echolocating bats');
  MODEL lEnergy = Type lMass / SOLUTION;
```

```
                                        Sum of
      Source                      DF       Squares     Mean Square    F Value    Pr > F
      Model                        3   29.42148268      9.80716089     283.59    <.0001
      Error                       16    0.55331753      0.03458235
      Corrected Total             19   29.97480021

                  R-Square     Coeff Var      Root MSE    lEnergy Mean
                  0.981541      7.491872      0.185963        2.482201

                                                             Standard
   Parameter                            Estimate             Error    t Value    Pr > |t|
   Intercept                        -1.576360194 B      0.28723642      -5.49      <.0001
   Type      echolocating bats       0.078663681 B      0.20267926       0.39      0.7030
   Type      non-echolocating birds  0.102261918 B      0.11418264       0.90      0.3837
   Type      non-echolocating bats   0.000000000 B       .               .          .
   lMass                             0.814957494        0.04454143      18.30      <.0001

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve
      the normal equations.  Terms whose estimates are followed by the letter 'B' are not
      uniquely estimable.
```

SAS code and output - with an interaction

```
PROC GLM DATA=case1002;
  CLASS Type(REF='non-echolocating bats');
  MODEL lEnergy = Type|lMass / SOLUTION;

                                               Sum of
      Source                      DF          Squares     Mean Square    F Value    Pr > F
      Model                        5      29.46993221      5.89398644     163.44    <.0001
      Error                       14       0.50486800      0.03606200
      Corrected Total             19      29.97480021

                     R-Square     Coeff Var      Root MSE    lEnergy Mean
                     0.983157      7.650468      0.189900        2.482201

                                                                Standard
   Parameter                                   Estimate            Error    t Value    Pr > |t|
   Intercept                               -0.202447571 B     1.26133425      -0.16      0.8748
   Type       echolocating bats            -1.268067693 B     1.28542004      -0.99      0.3406
   Type       non-echolocating birds       -1.378390198 B     1.29524130      -1.06      0.3053
   Type       non-echolocating bats         0.000000000 B      .                .         .
   lMass                                    0.589782057 B     0.20613801       2.86      0.0126
   lMass*Type echolocating bats             0.214874992 B     0.22362264       0.96      0.3529
   lMass*Type non-echolocating birds        0.245588273 B     0.21343221       1.15      0.2691
   lMass*Type non-echolocating bats         0.000000000 B      .                .         .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve
      the normal equations.  Terms whose estimates are followed by the letter 'B' are not
      uniquely estimable.
```

R code and output - main effects only

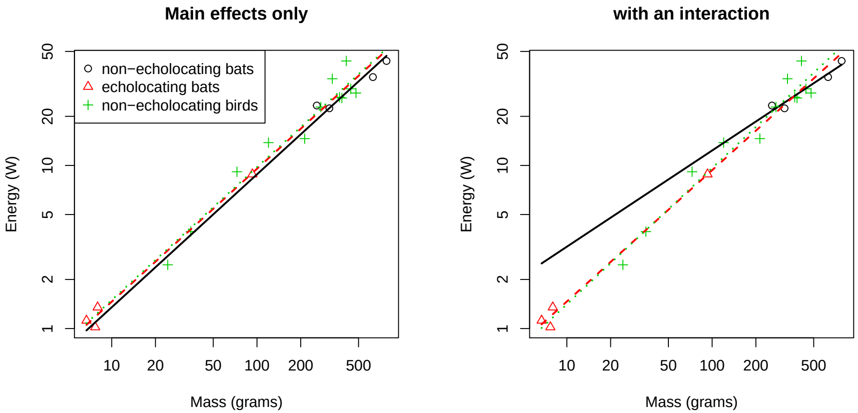
```
case1002$Type = relevel(case1002$Type, ref='non-echolocating bats') # match SAS
summary(mM <- lm(log(Energy)~log(Mass)+Type, case1002))
```

```
Call:
lm(formula = log(Energy) ~ log(Mass) + Type, data = case1002)

Residuals:
    Min      1Q  Median      3Q     Max
-0.2322 -0.1220 -0.0364  0.1257  0.3446

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                 -1.5764     0.2872   -5.49  5.0e-05 ***
log(Mass)                    0.8150     0.0445   18.30  3.8e-12 ***
Typeecholocating bats        0.0787     0.2027    0.39     0.70
Typenon-echolocating birds   0.1023     0.1142    0.90     0.38
---
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:  0.982, Adjusted R-squared:  0.978
F-statistic:  284 on 3 and 16 DF,  p-value: 4.46e-14
```

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
-0.2515 -0.1264 -0.0095  0.0812  0.3284

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)
(Intercept)                            -0.202      1.261   -0.16    0.875
log(Mass)                               0.590      0.206    2.86    0.013 *
Typeecholocating bats                  -1.268      1.285   -0.99    0.341
Typenon-echolocating birds             -1.378      1.295   -1.06    0.305
log(Mass):Typeecholocating bats         0.215      0.224    0.96    0.353
log(Mass):Typenon-echolocating birds    0.246      0.213    1.15    0.269
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.19 on 14 degrees of freedom
Multiple R-squared:  0.983, Adjusted R-squared:  0.977
F-statistic:  163 on 5 and 14 DF,  p-value: 6.7e-12
```

Visualizing the models



Categorical-categorical

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

- Y_i be the response,
- 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

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

- β_0 is the expected response for category A and type 0
- β_1 is the change in response for moving from type 0 to type 1
- β_2 is the change in response for moving from category A to category B
- β_3 is the change in response for moving from category A to category C

The means are then

	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

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

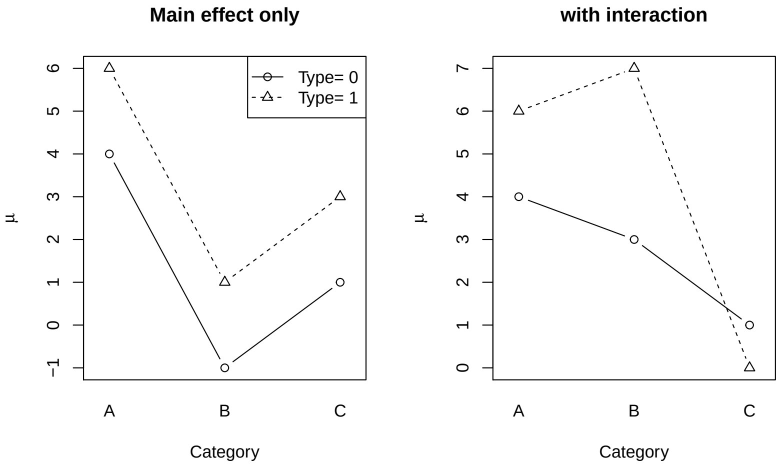
- β_0 is the expected response for category A and type 0
- β_1 is the change in response for moving from type 0 to type 1 for category A
- β_2 is the change in response for moving from category A to category B for type 0
- β_3 is the change in response for moving from category A to category C for type 0
- β_4 is the difference in change in response for moving from category A to category B for type 1 compared to type 0
- β_5 is the difference in change in response for moving from category A to category C for type 1 compared to type 0

The means are then

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

This is referred to as the cell-means model.

Visualizing the models



SAS code and output - main effects only

```
DATA case1301;
  INFILE 'case1301.csv' DSD FIRSTOBS=2;
  INPUT Cover Block $ Treat $;

PROC GLM DATA=case1301;
  WHERE Block IN ('B1','B2') AND Treat IN ('L','Lf','LfF');
  CLASS Block Treat; /* reference levels default to 1st alphabetically */
  MODEL Cover = Block Treat / SOLUTION;
```

```
                                        Sum of
      Source                      DF   Squares     Mean Square    F Value     Pr > F
      Model                        3   32.08333333  10.69444444      6.04     0.0188
      Error                        8   14.16666667   1.77083333
      Corrected Total             11   46.25000000

                     R-Square     Coeff Var      Root MSE    Cover Mean
                     0.693694      31.31121      1.330727      4.250000

                                                     Standard
           Parameter           Estimate                 Error    t Value    Pr > |t|
           Intercept        4.666666667 B          0.76829537       6.07      0.0003
           Block     B2     2.166666667 B          0.76829537       2.82      0.0225
           Block     B1     0.000000000 B           .                .         .
           Treat     Lf    -1.500000000 B          0.94096582      -1.59      0.1496
           Treat     LfF   -3.000000000 B          0.94096582      -3.19      0.0128
           Treat     L      0.000000000 B           .                .         .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve
      the normal equations.  Terms whose estimates are followed by the letter 'B' are not
      uniquely estimable.
```

SAS code and output - with an interaction

```
PROC GLM DATA=case1002;
  WHERE Block IN ('B1','B2') AND Treat IN ('L','Lf','LfF');
  CLASS Block Treat;
  MODEL Cover = Block|Treat / SOLUTION;
```

```
                                        Sum of
      Source                      DF       Squares     Mean Square    F Value    Pr > F
      Model                        5   36.75000000      7.35000000       4.64    0.0443
      Error                        6    9.50000000      1.58333333
      Corrected Total             11   46.25000000

                    R-Square     Coeff Var      Root MSE     Cover Mean
                    0.794595      29.60719      1.258306       4.250000

                                                      Standard
         Parameter                   Estimate            Error     t Value     Pr > |t|
         Intercept                4.000000000 B     0.88975652        4.50       0.0041
         Block         B2         3.500000000 B     1.25830574        2.78       0.0319
         Block         B1         0.000000000 B      .                 .          .
         Treat         Lf         0.000000000 B     1.25830574        0.00       1.0000
         Treat         LfF       -2.500000000 B     1.25830574       -1.99       0.0941
         Treat         L          0.000000000 B      .                 .          .
         Block*Treat B2 Lf       -3.000000000 B     1.77951304       -1.69       0.1428
         Block*Treat B2 LfF      -1.000000000 B     1.77951304       -0.56       0.5945
         Block*Treat B2 L         0.000000000 B      .                 .          .
         Block*Treat B1 Lf        0.000000000 B      .                 .          .
         Block*Treat B1 LfF       0.000000000 B      .                 .          .
         Block*Treat B1 L         0.000000000 B      .                 .          .

NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve
      the normal equations.  Terms whose estimates are followed by the letter 'B' are not
```

R code and output - main effects only

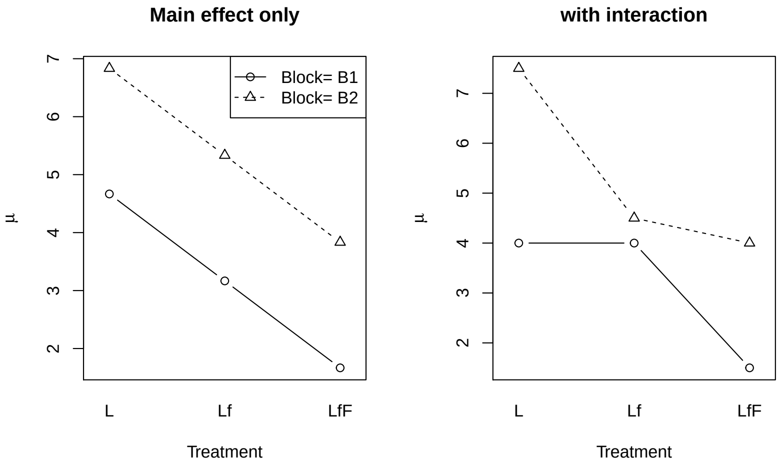
```
# Set the reference levels
case1301$Block = relevel(case1301$Block, ref='B1')
case1301$Treat = relevel(case1301$Treat, ref='L' )
summary(mM <- lm(Cover~Block+Treat, case1301, subset=Block %in% c("B1","B2") & Treat %in% c("L","Lf","LfF")))



Call:
lm(formula = Cover ~ Block + Treat, data = case1301, subset = Block %in%
    c("B1", "B2") & Treat %in% c("L", "Lf", "LfF"))

Residuals:
   Min     1Q Median     3Q    Max
-2.333 -0.667  0.000  0.792  1.833

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.667      0.768    6.07   0.0003 ***
BlockB2        2.167      0.768    2.82   0.0225 *
TreatLf       -1.500      0.941   -1.59   0.1496
TreatLfF      -3.000      0.941   -3.19   0.0128 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.33 on 8 degrees of freedom
Multiple R-squared:  0.694, Adjusted R-squared:  0.579
F-statistic: 6.04 on 3 and 8 DF,  p-value: 0.0188
```

R code and output - with an interaction

```
summary(mI <- lm(Cover~Block*Treat, case1301, subset=Block %in% c("B1","B2") & Treat %in% c("L","Lf","LfF")))
```

```
Call:
lm(formula = Cover ~ Block * Treat, data = case1301, subset = Block %in%
    c("B1", "B2") & Treat %in% c("L", "Lf", "LfF"))

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)       4.00e+00   8.90e-01    4.50   0.0041 **
BlockB2           3.50e+00   1.26e+00    2.78   0.0319 * 
TreatLf          -2.72e-16   1.26e+00    0.00   1.0000   
TreatLfF         -2.50e+00   1.26e+00   -1.99   0.0941 . 
BlockB2:TreatLf  -3.00e+00   1.78e+00   -1.69   0.1428   
BlockB2:TreatLfF -1.00e+00   1.78e+00   -0.56   0.5945   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.26 on 6 degrees of freedom
Multiple R-squared:  0.795, Adjusted R-squared:  0.623 
F-statistic: 4.64 on 5 and 6 DF,  p-value: 0.0443
```

Visualizing the models



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 explanatory variables

The possibilities for explanatory variables are

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

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

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