Multiple Linear Regression

Michael Noonan

Biol 520C: Statistical modelling for biological data

Table of contents

1. Multiple Linear Regression
2. Multiple Linear Regression Pre-analysis
3. Multiple linear regression in action
4. Parameter Interactions
5. Parameter Interactions in Action

Multiple Linear Regression

Multiple Linear Regression

So far, we've been interested in answering the question: "Is there a relationship between X and Y?".

For some systems knowing the relationship between X and Y is sufficient, but in most biological systems there are typically multiple variables that influence outcomes (e.g., rainfall, sunlight, soil composition all influence plant growth)

Our verbal hypothesis in this case is 'Y is proportional to X_1, X_2, $\ldots X_m$'. With multiple factors, how do we approach the problem statistically?

Simple Linear Regression

Remembering that for simple linear regression our data is of the form:

$$d = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$$

X	Y
x_1	y_1
x_2	y_2
$\ldots$	$\ldots$
x_n	y_n

And our relationship is described by an intercept (β_0) and a slope (β_1): $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Multiple Regression Model

For *m* explanatory variables, the data are of the form:

Y	X_1	X_2	$\ldots$	X_m
y_1	X_{11}	X_{21}	$\ldots$	X_m1
y_1	X_{12}	X_{22}	$\ldots$	X_{m2}
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
y_n	X_{1n}	X_{2n}	$\ldots$	X_{mn}

A linear regression with multiple explanatory variables is described by an intercept (β_0) and a regression coefficients ($\beta_1, \beta_2, \ldots, \beta_m$):
$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_m x_{mi} + \varepsilon_i$

Linear regression and matrix notation

Given our dataset $(y_1, x_11, x_21), (y_2, x_12, x_22), \ldots, (y_n, x_1n, x_2n)$ we can re-write the problem in matrix notation:

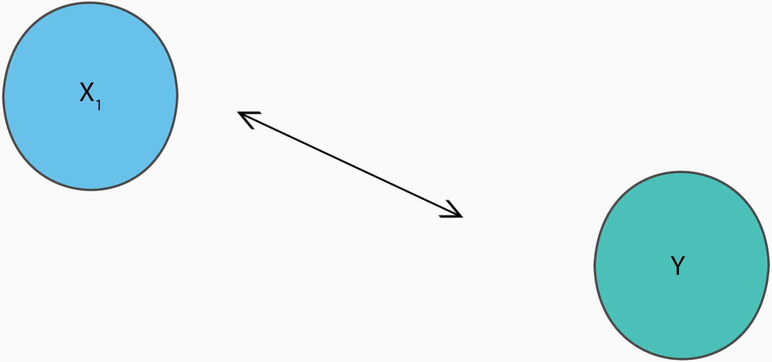
$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \qquad \mathbf{x} = \begin{pmatrix} 1 & x_11 & x_21 \\ 1 & x_12 & x_22 \\ \vdots & \vdots & \vdots \\ 1 & x_1n & x_2n \end{pmatrix}$$

Notice how the dimensions of the matrices all line up, so even though we've added new parameters we can still estimate their values using our old friend $\beta = (\mathbf{x}^\mathbf{T}\mathbf{x})^{-1}\mathbf{x}^\mathbf{T}\mathbf{y}$.

Multiple Linear Regression
Pre-analysis

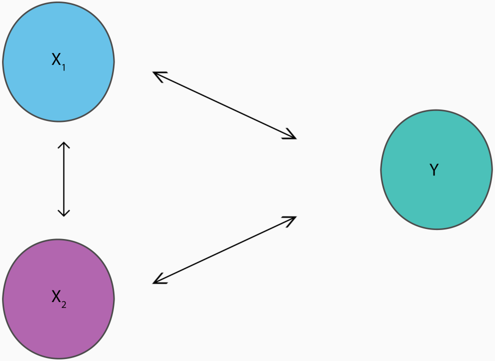
Mult. Parameters, Mult. Interactions

With only one parameter, there is only a single relationship that is possible



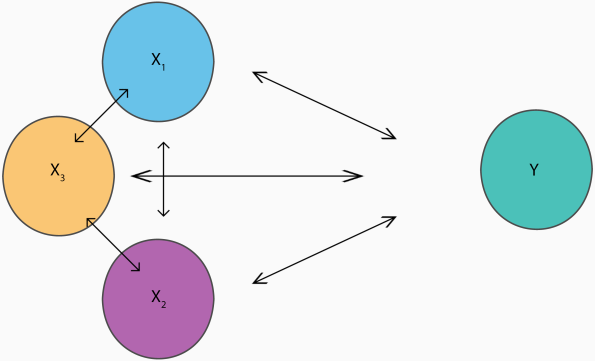
Mult. Parameters, Mult. Interactions

With two parameters, there are 3 relationship that are possible



Mult. Parameters, Mult. Interactions

With three parameters, there are 6 relationship that are possible



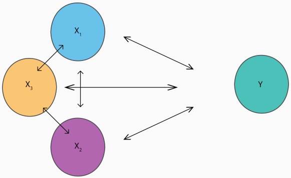
... and it only gets worse as you add more parameters

Collinearity

The challenge is that only some of the relationships are of interest, whereas others can introduce identifiability issues

Generally speaking we're only interested in the relationship between or predictors and our response variable

Relationships between the predictors introduce the problem of collinearity



Collinearity Cont.

Collinearity (also multicollinearity) is a condition in which some of the independent variables are highly correlated with one another.

For collinear predictors, the estimate of a single variable's impact on the dependent variable is less precise.

For example if X_1 and X_2 are correlated, all changes in X_1 produce changes in X_2, so we have an inaccurate estimate of the independent effect of X_1 on Y.

In other words, parameter values for collinear predictors can't be trusted, which means we can't trust any of the predictions that the model makes.

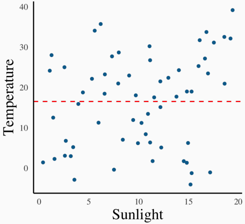
Collinearity Example

THE UNIVERSITY OF BRITISH COLUMBIA
Okanagan Campus

Imagine a situation where both sunlight and temperature affect plant growth.

Sunlight and temperature may be entirely unrelated

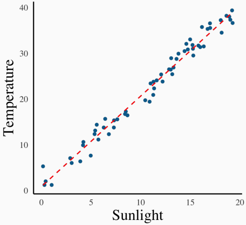
```
SUN <- runif(60, min = 0, max = 20)
TEMP <- runif(60, min = -5, max = 40)
```



...or sunlight can influence temperature

$$\text{temperature} = \beta_0 + \beta_1 \times \text{sunlight} + \varepsilon$$

```
TEMP <- function(sun) {
  B_0 <- 0
  B_1 <- 2
  sig <- 2
  temp <- B_0 + B_1*sun

  temp <- rnorm(n = length(temp),
                mean = temp,
                sd = sig)
  temp
}

sun <- runif(60, min = 0, max = 20)
temp <- TEMP(sun)
```

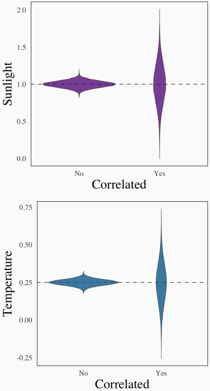


Collinearity Example Cont.

If the true relationship between sunlight, temperature, and plant mass is given by: $\text{mass}_i = 0.25 \times \text{temperature}_i + 1 \times \text{sunlight}_i + \varepsilon_i$

We can test for the influence of correlated data on parameter estimates:

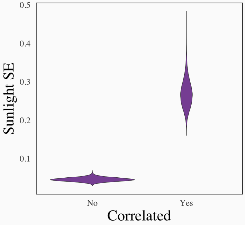
```
growth <- function(temp, sun) {
  B_0 <- 0
  B_1 <- 0.25
  B_2 <- 1
  sig <- 2
  mass <- B_0 + B_1*temp + B_2*sun

  mass <- rnorm(n = length(mass),
                mean = mass,
                sd = sig)

  mass
}


mass <- growth(temp, sun)

FIT <- lm(mass ~ temp + sun)
```

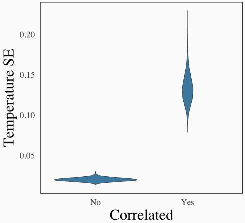


Note how some parameter estimates can even change signs!

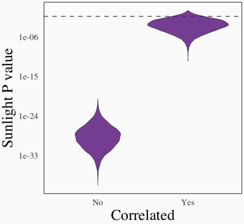
Collinearity and Standard Errors

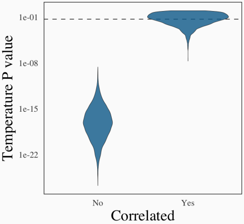
Collinearity also inflates standard errors





. . . which means that p values are untrustworthy



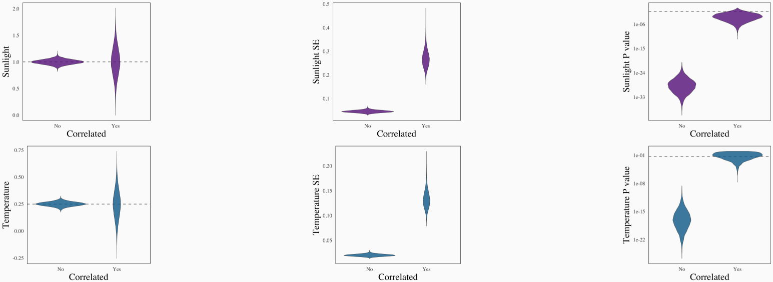


Correcting for collinearity

- Do nothing.
- Collect more data.
- Drop one of the correlated variables.
- Perform a Principal Component Analysis (PCA) on the data to reduce dimensionality.

Correction: Do nothing

Collinearity doesn’t necessarily mean your model is incorrect



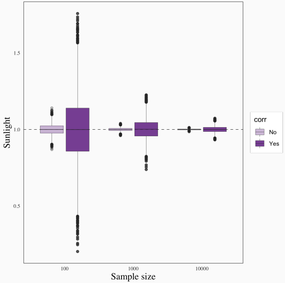
. . . but it does mean you need to be careful when making inferences.

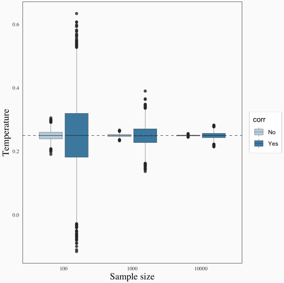
Rule of thumb:: If the correlation between predictors is < 0.7 you won’t run into major issues. Dormann et al. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography, 36(1), 27-46.

Correction: Collect more data

Collinearity doesn’t necessarily bias the parameter estimates (i.e., parameter estimates are correct on average), but it does increase standard errors and inflates the variance around the parameter estimates.

With enough data, the problem of collinearity eventually goes away (termed asymptotic convergence).





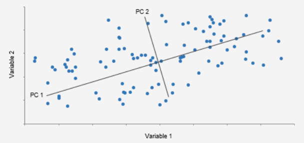
Correction: Drop parameters

Dropping collinear parameters can correct for the issue of collinearity, at a cost of reducing your R-squared and potentially reducing biological relevance.

Correction: PCA

Principal component analysis (PCA) is a common technique for correcting for collinearity.

We're not covering the details of PCAs, but, briefly, they reduce the dimensionality of a dataset by creating new uncorrelated variables that successively maximize variance.



Source: statistiXL

Instead of fitting a model to your original data, you use the uncorrelated principal components in your analysis.

Analyses are robust to collinearity in the original data, but results can be more difficult to interpret.

Multiple linear regression in action

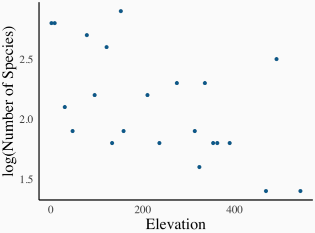
Example: Ant richness

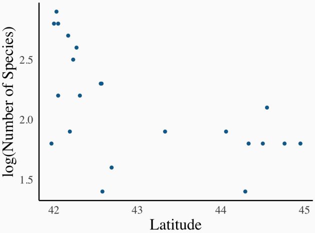
The Question: Do latitude and elevation have a measurable effect on ant species richness in forest plots in New England?

Gotelli, N.J. & Ellison, A.M. (2002). Biogeography at a regional scale: determinants of ant species density in bogs and forests of New England. Ecology, 83, 1604–1609

Ant richness: The data

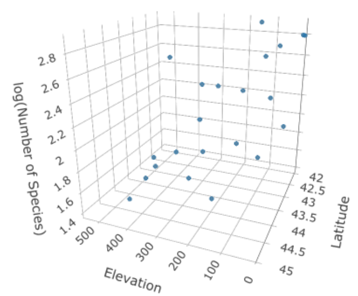
Number of Species	Latitude	Elevation
1.8	41.97	389
2.8	42	8
2.9	42.03	152
2.8	42.05	1
2.2	42.05	210
2.7	42.17	78
1.9	42.19	47
2.5	42.23	491
2.6	42.27	121
2.2	42.31	95
2.3	42.56	274
2.3	42.57	335
1.4	42.58	543
1.6	42.69	323
1.9	43.33	158
1.9	44.06	313
1.4	44.29	468
1.8	44.33	362
1.8	44.5	236
2.1	44.55	30
1.8	44.76	353
1.8	44.95	133





Ant richness: The data cont.

The relationships between richness, latitude, and elevation play out in multiple dimensions.

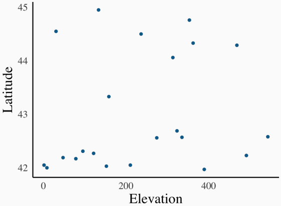


The first step is to check for any relationship between elevation and latitude

Ant Richness: Pre-analysis

We can also do this by using the `lm()` function:

```
data <- read.csv("Ant_Richness.csv")

preFIT <- lm(latitude ˜ elevation, data = data)

summary(preFIT)


Call:
lm(formula = latitude ˜ elevation, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-1.2357 -0.7383 -0.5589  0.9789  2.0485

Coefficients:
             Estimate Std. Error t value Pr(>t)
(Intercept) 42.743402   0.411929 103.764   <2e-16 ***
elevation    0.001189   0.001461   0.814    0.425
---

Residual standard error: 1.091 on 20 degrees of freedom
Multiple R-squared:  0.03205, Adjusted R-squared:  -0.01635
F-statistic: 0.6622 on 1 and 20 DF,  p-value: 0.4254
```



Ant richness: matrix notation

There was no significant relationship between our two predictors (Elevation and Latitude) so we can continue with our analyses.

The regression problem in matrix notation is:

$$\begin{pmatrix} 1.8 \\ 2.8 \\ \vdots \\ 1.8 \end{pmatrix} = \begin{pmatrix} 1 & 41.97 & 389 \\ 1 & 42.00 & 8 \\ \vdots & \vdots & \\ 1 & 44.95 & 133 \end{pmatrix} \cdot \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Because the dimensions of the matrices are all compatible we can estimate the parameters values using $\beta = (\mathbf{x}^\mathbf{T}\mathbf{x})^{-\mathbf{1}}\mathbf{x}^\mathbf{T}\mathbf{y}$.

Ant Richness: Est. the parameters

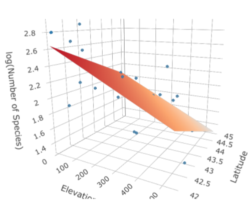
Again, this can easily done in R:

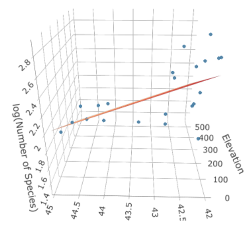
```
x <- matrix(c(rep(1, nrow(data)),
              data$latitude,
              data$elevation),
            nrow = nrow(data), ncol = 3)

y <- matrix(data$num_sp,
            nrow = nrow(data), ncol = 1)

xtx <- t(x) %*% x
xtx.inv <- solve(xtx)
xty <- t(x) %*% y

beta <- xtx.inv %*% xty

beta

             [,1]
[1,] 11.115818937
[2,] -0.201828913
[3,] -0.001372863
```

Ant Richness: Analysis

We can also do this by using the `lm()` function:

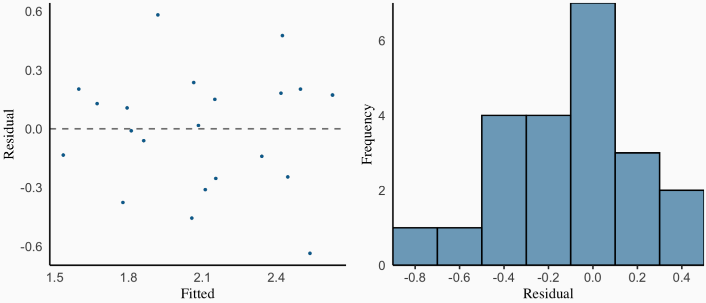
```
FIT <- lm(num_sp ~ latitude + elevation, data = data)



Call:
lm(formula = num_sp ~ latitude + elevation, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.63613 -0.21966  0.06166  0.17932  0.58149

Coefficients:
              Estimate Std. Error t value Pr(>t)
(Intercept) 11.1158189  2.7445783   4.050 0.000683 ***
latitude    -0.2018289  0.0641510  -3.146 0.005318 **
elevation   -0.0013729  0.0004259  -3.223 0.004473 **
---

Residual standard error: 0.3131 on 19 degrees of freedom
Multiple R-squared:  0.5653,  Adjusted R-squared:  0.5196
F-statistic: 12.36 on 2 and 19 DF,  p-value: 0.0003651
```





Ant richness: Residuals

After fitting the model it's important to check the residuals to check for the assumption of normality.



We'll cover residuals in more detail next lecture, but for now suffice it to say these look pretty good.

A note on adding parameters

Adding more variables to a model will always soak up some of the residual variation.

This can be good (because it means we've identified another key parameter and we're explaining more of the variation in our system)

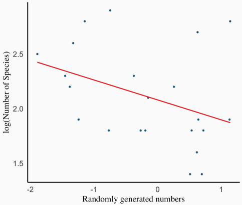
But it can also be bad (if the parameters are not adding a meaningful amount of information, have no underlying biological importance, introduce collinearity, and/or risk overfitting)

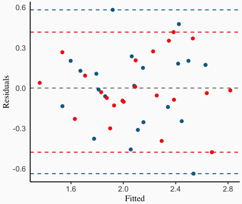
It's important to be very careful when adding parameters to a model.

A note on adding parameters cont.

```
set.seed(84)
NOISE <- rnorm(22)
FIT2 <- lm(num_sp ~ latitude + elevation + NOISE, data = data)
```

```
summary(FIT2)

Call:
lm(formula = num_sp ~ latitude + elevation + NOISE, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.47676 -0.09979 -0.03401  0.17771  0.41594

Coefficients:
              Estimate Std. Error t value Pr(>t)
(Intercept) 10.9552493  2.2723853   4.821 0.000137 ***
latitude    -0.1986631  0.0531102  -3.741 0.001497 **
elevation   -0.0014188  0.0003528  -4.021 0.000801 ***
NOISE       -0.1893679  0.0607058  -3.119 0.005922 **
---

Residual standard error: 0.2592 on 18 degrees of freedom
Multiple R-squared:  0.7179,  Adjusted R-squared:  0.6708
F-statistic: 15.27 on 3 and 18 DF,  p-value: 3.448e-05
```





Our previous R^2 was ∼0.57, so on the surface it looks like we have a better model, even though we know for a fact it's just junk.

Parameter Interactions

Parameter Interactions

We opened today's lecture by noting that biological systems are complex with multiple variables influencing outcomes.

Well, biological systems are complex, and variables can also interact with each other to influence outcomes (e.g., a breeze on a warm day is refreshing, a breeze in the middle of winter is not at all refreshing).

Interactions between explanatory variables are expressed as a product of the X's: $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i$

Parameter Interactions Cont.

For a model with two predictor variables and their interaction, $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \varepsilon_i$

we can re-write the problem in matrix notation:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \qquad \mathbf{x} = \begin{pmatrix} 1 & x_1 1 & x_2 1 & x_1 1 x_2 1 \\ 1 & x_1 2 & x_2 2 & x_1 2 x_2 2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_1 n & x_2 n & x_1 n x_2 n \end{pmatrix}$$

Parameter Interactions in Action

Parameter Interactions: Ant richness

The Question: We saw that latitude and elevation had a measurable effect on ant species richness. But what about the interaction between these?

It's conceivable that high elevation at high latitude is worse for ants than high elevation at lower latitudes.

Source: WallpaperAccess

Source: Wikipedia

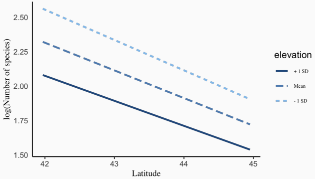
Ant Richness: Analysis

We can easily model interaction terms using the : operator:

```
FIT <- lm(num_sp ~ latitude + elevation + latitude:elevation, data = data)
```

```
summary(FIT)
```

```
Call:
lm(formula = num_sp ~ latitude * elevation, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.64627 -0.22755  0.07892  0.17200  0.60449

Coefficients:
                    Estimate Std. Error t value Pr(>t)
(Intercept)        12.2599423  5.1337948   2.388   0.0281 *
latitude           -0.2286208  0.1201511  -1.903   0.0732 .
elevation          -0.0064539  0.0190732  -0.338   0.7390
latitude:elevation  0.0001186  0.0004451   0.266   0.7929
---

Residual standard error: 0.3211 on 18 degrees of freedom
Multiple R-squared:  0.567, Adjusted R-squared:  0.4949
F-statistic: 7.858 on 3 and 18 DF,  p-value: 0.001471
```



A note on interactions in R

In R we can include interaction terms in two ways:

```
FIT1 <- lm(num_sp ~ latitude:elevation, data = data)
```

```
Call:
lm(formula = num_sp ~ latitude:elevation, data = data)

Coefficients:
                     Estimate Std. Error t value Pr(>t)
(Intercept)         2.499e+00  1.402e-01  17.826  9.6e-14 ***
latitude:elevation -3.833e-05  1.152e-05  -3.328  0.00335 **
```

```
FIT2 <- lm(num_sp ~ latitude*elevation, data = data)
```

```
Call:
lm(formula = num_sp ~ latitude * elevation, data = data)

Coefficients:
                     Estimate Std. Error t value Pr(>t)
(Intercept)        12.2599423  5.1337948   2.388   0.0281 *
latitude           -0.2286208  0.1201511  -1.903   0.0732 .
elevation          -0.0064539  0.0190732  -0.338   0.7390
latitude:elevation  0.0001186  0.0004451   0.266   0.7929
```

A note on interactions

It is common to read that interactions should only be included in the model when the corresponding main effects are also included, but there is nothing wrong with including interaction effects by themselves *per se*.

Your goal as a modeler is to build a model that is a reasonable description of the data and system, not merely following a recipe.

Figure out what each model means given the process you are modeling and whether a model with or without the main effects makes more sense given your theory or hypothesis.

. . . and always use objective measures to decide on what parameters should be included/discarded.