Module 4: Multiple Regression

Ellen Bledsoe

2023-11-14

Multiple Regression

Sometimes (often!), we have multiple independent variables that we think might be influencing our dependent variable. In order to take into account more than one variable, we can perform what is called a "multiple regression."

Multiple regressions can become complicated beasts, but we are going to focus on adding in just one additional variable and how it can impact our interpretations.

Let's use some penguins nesting site data to demonstrate.

Set-Up

As usual, we want to load our packages and our data before we begin.

```
library(tidyverse)
site_data <- read_csv("../data/site_changes.csv") %>%
  mutate(year = as.factor(year)) # don't worry about this code too much
# we want the year column treated as a category, not a number
```

We want to understand what factors influence penguin nesting—we wouldn't want to build our road through important penguin nesting territories!

Let's quickly take a look at the data we have.

```
head(site_data)
```

```
## # A tibble: 6 x 6
##   site_id year  tussocks dist_to_water stone_size num_nests
##     <dbl> <fct>    <dbl>         <dbl>      <dbl>     <dbl>
## 1       1 1971      4.95          26.6       45.0        25
## 2       2 1971      3.60          28.2       45.1        18
## 3       3 1971      3.71          24.1       46.3        27
## 4       4 1971      5.14          29.6       43.3        16
## 5       5 1971      6.71          27.9       46.6        15
## 6       6 1971      5.16          26.7       47.1        29
```

```
tail(site_data)
```

```
## # A tibble: 6 x 6
##   site_id year  tussocks dist_to_water stone_size num_nests
##     <dbl> <fct>    <dbl>         <dbl>      <dbl>     <dbl>
## 1      95 2011      7.46          25.8       42.1        19
## 2      96 2011      6.29          21.9       43.7        31
## 3      97 2011      7.10          21.7       45.5        25
## 4      98 2011      6.61          21.9       42.3        18
## 5      99 2011      6.74          21.7       41.7        29
## 6     100 2011      7.50          18.6       40.1        31
```

Okay, it looks like we have data for 100 different sites collected in two different years. We have the number of penguin nests, which is the variable that we are interested in. We also have a few other variables that might be influencing the number of nests at each site: the number of tussocks of grass, the stone size, the distance to water, and the year that data were collected.

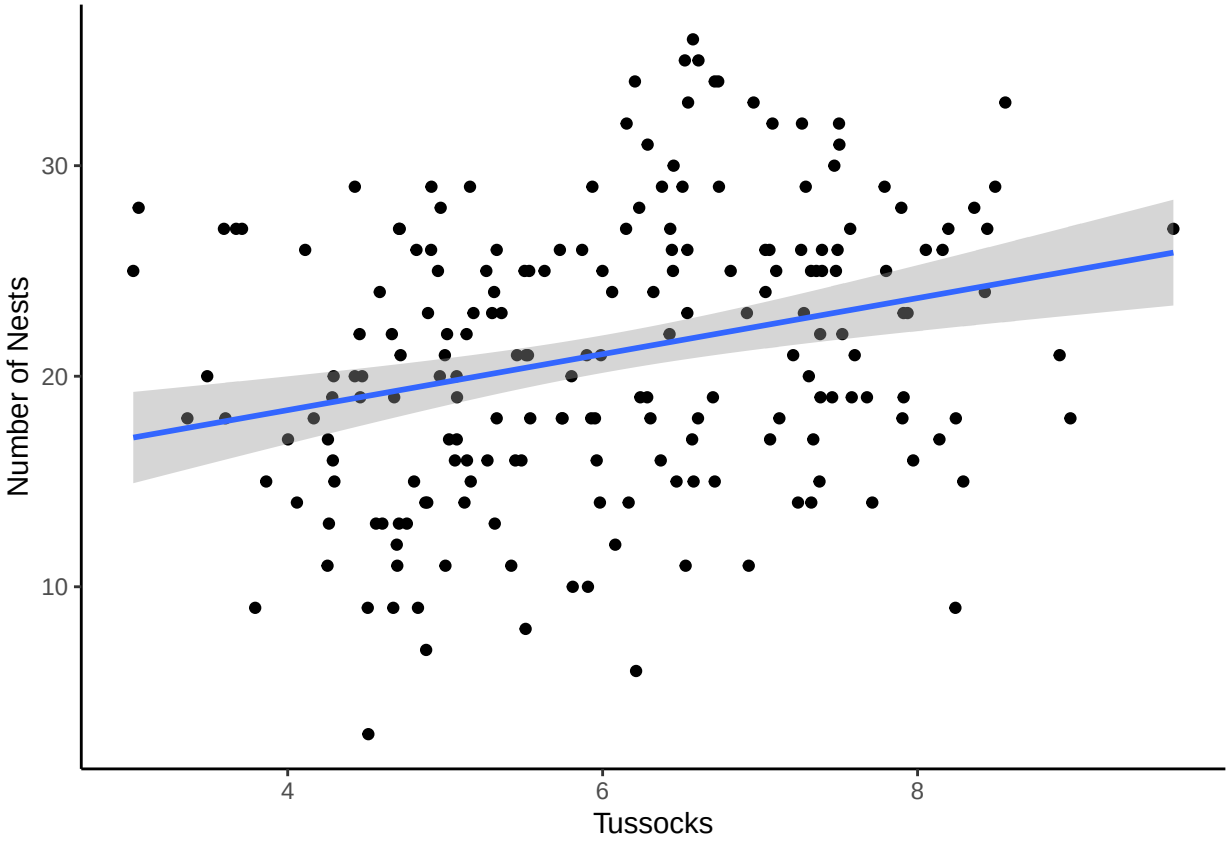
Let's start with looking at how the number of grass tussocks affects nest numbers.

Regression Model (1 Independent Variable)

Let's start by plotting our tussocks data and run a model.

```
ggplot(site_data, aes(x = tussocks, y = num_nests)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Tussocks", y = "Number of Nests") +
  theme_classic()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



This model is showing us a positive relationship between the number of tussocks and number of nests. Let's run our linear regression model.

```
tussocks_model <- lm(num_nests ~ tussocks, data = site_data)
summary(tussocks_model)
```

```
##
## Call:
## lm(formula = num_nests ~ tussocks, data = site_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.0717  -4.5553   0.2871   4.4302  14.1867
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  13.0711     2.0555   6.359 1.37e-09 ***
## tussocks      1.3299     0.3322   4.004 8.81e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.384 on 198 degrees of freedom
## Multiple R-squared:  0.0749, Adjusted R-squared:  0.07023
## F-statistic: 16.03 on 1 and 198 DF,  p-value: 8.81e-05
```

Our model is telling us that the relationship between the number of nests and number of tussocks is highly

significant.

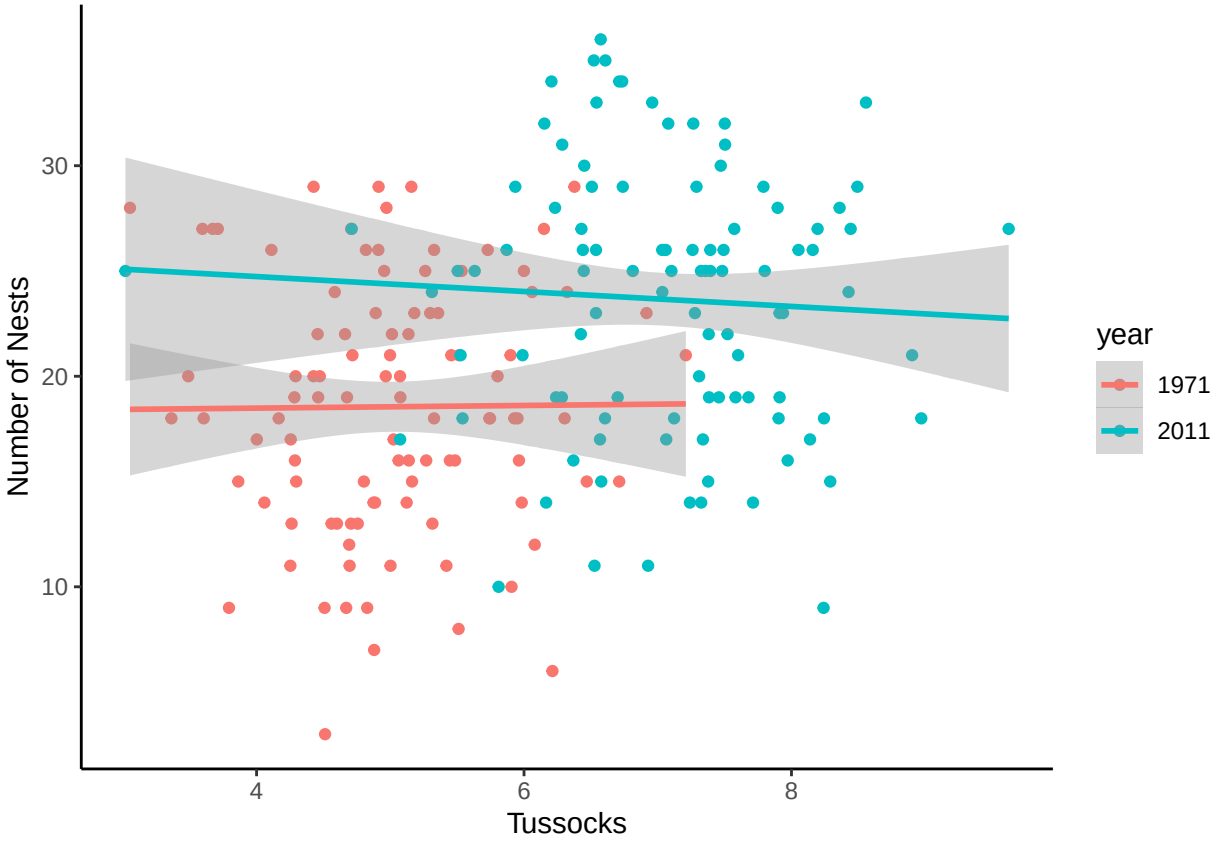
Running a Multiple Regression

What if we incorporate the year the data were collected into our model? Adding an additional independent variable to our model can change how we interpret the results.

First, let's plot a multiple scatterplot. Do we see the same patterns?

```
ggplot(site_data, aes(x = tussocks, y = num_nests, color = year)) +
  geom_point() +
  geom_smooth(method = "lm") +
  labs(x = "Tussocks", y = "Number of Nests") +
  theme_classic()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Huh, that looks a little bit different... we no longer seem to have much of a positive relationship. Instead, both lines are looking pretty flat! Let's see what happens when we add year into our regression model.

By adding the year variable into the model, we are also asking if the number of nests changes by year. Does it change our significance?

The way we add an independent variable is with a + in the equation.

```r
tussocks_model <- lm(num_nests ~ tussocks + year, data = site_data)
summary(tussocks_model)
```

```
##
## Call:
## lm(formula = num_nests ~ tussocks + year, data = site_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.644  -4.580   0.424   4.504  12.259
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.4867     2.5055   7.777 4.05e-13 ***
## tussocks     -0.1868     0.4844  -0.386      0.7
## year2011      5.4818     1.3167   4.163 4.69e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.136 on 197 degrees of freedom
## Multiple R-squared:  0.1497, Adjusted R-squared:  0.1411
## F-statistic: 17.34 on 2 and 197 DF,  p-value: 1.154e-07
```

We have a bunch of estimates now! Don't worry, I won't be asking you to put them into an equation :) We are only focusing on significance values.

- `tussocks`: the p-value for the tussocks row (p = 0.7) shows whether the number of tussocks is a significant driver of nest numbers (it isn't!)
- `year2011`: this p-value is significant (p = 4.69e-05)—year does seem to impact how many nests there are
  - (you don't need to know this, but if you are curious...): the reason this says `year2011` and not just `year` is because year is acting as a category, and we can get different coefficient estimates for year. The first option in the categorical variable (1971, in this case) is incorporated into the current intercept estimate, so `year2011` indicates the difference from the original.

Interactions

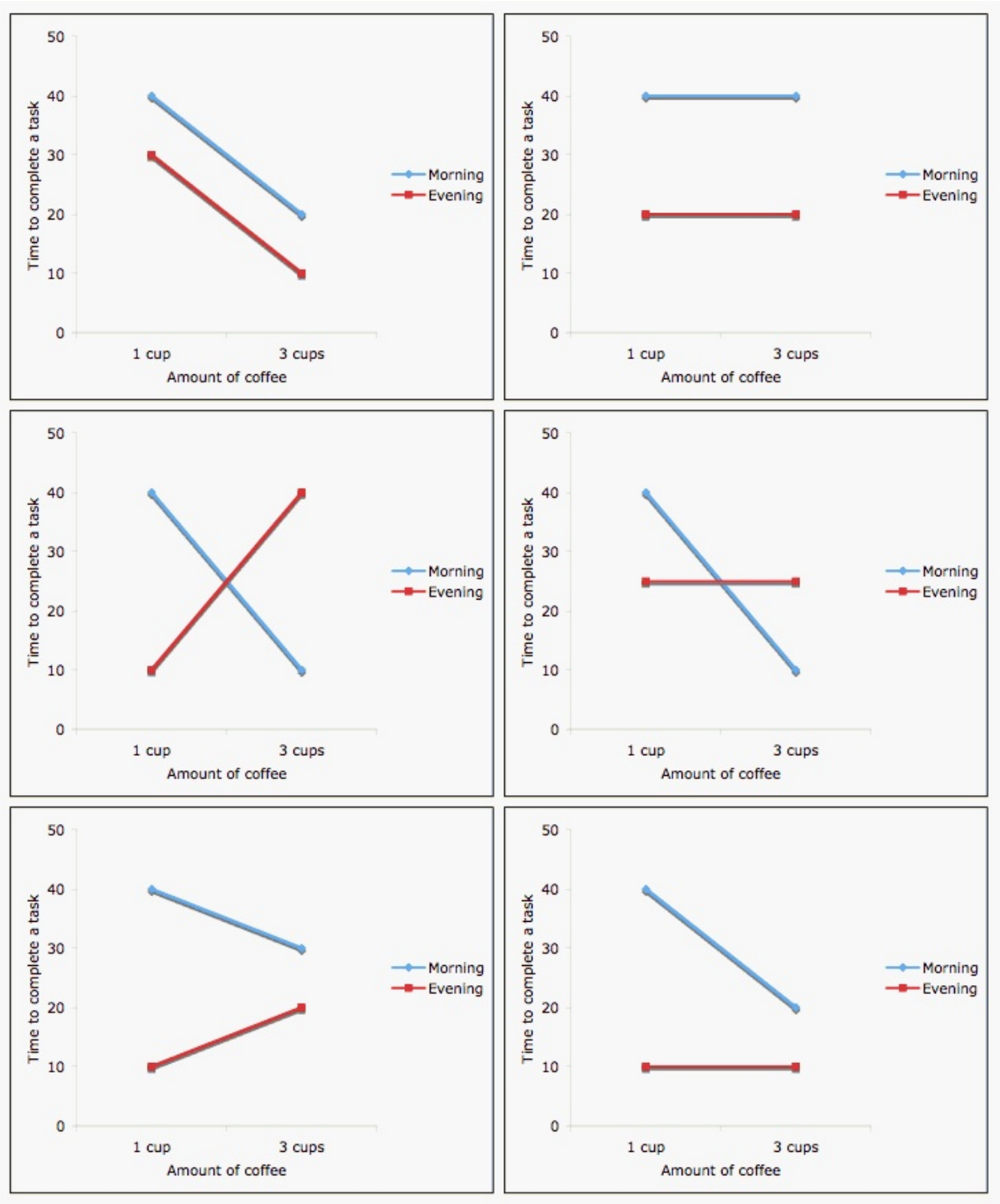
Sometimes, we might expect that there will be in interactive effect between the two independent variables we are including in our model. This means that the combinations of independent variables impact the dependent variable in different ways.

We can typically spot interactive effects in a plot when the slopes of the regression lines differ from each other a lot.

Let's talk through the visual demonstration below to hopefully get a better understanding of how interaction effects work. These plots are showing different possible relationships between two independent variables: how many cups of coffee someone has had and whether they drank them in the morning or the afternoon (categorical variable), and how those might impact the dependent variable, how long it takes to complete a task.

- The top two panels show examples of no interactive effect. It looks like there is an effect of time of day; in each, drinking coffee in the morning makes tasks take longer. However, the *slopes* between time of day are the same. In the first panel (top left), there also looks to be an impact from the number of cups of coffee.

- The bottom four panels demonstrate possible variations of *interactive* effects. Not only does time of day and/or number of cups of coffee impact how long it takes to complete a task, but those independent variable affect how long it takes to complete a task *in different ways*.



In our example above, an interactive effect between tussocks and year would mean that we think the number of tussocks impacts the number of nests *in different ways* between the two categories (year). Since the slopes

between the years don't look very different, we aren't really expecting there to be a significant interaction, but let's check anyway.

Let's run a model that includes the interaction term in our model. To do this, we will use an asterisk (*) instead of a plus sign.

```
tussocks_model <- lm(data = site_data, num_nests ~ tussocks * year)
summary(tussocks_model)
```

```
##
## Call:
## lm(formula = num_nests ~ tussocks * year, data = site_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5186  -4.5456   0.4779   4.4342  12.1779
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       18.23648    3.88911   4.689 5.13e-06 ***
## tussocks           0.06251    0.76571   0.082    0.935
## year2011           7.91428    5.92732   1.335    0.183
## tussocks:year2011 -0.41675    0.99005  -0.421    0.674
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.149 on 196 degrees of freedom
## Multiple R-squared:  0.1505, Adjusted R-squared:  0.1375
## F-statistic: 11.57 on 3 and 196 DF,  p-value: 5.12e-07
```

We can see that we still have estimates for both tussocks and year alone. However, we now also have something new: our interaction term, `tussocks:year2011`.

It turns out that adding in the year variable has changed the significance in our model, and the interaction term is also not significant.

- The positive relationship we originally observed is no longer present when we account for the year.
- There is no significant interaction, meaning that penguins respond to the number of tussocks the same way in each year.
- Interestingly (and maybe frustratingly), the significant effect of year that we saw before is also gone.
  - This is a confusing bit of statistics, and I won't go into depth here about why this happens. Similar things sometimes happen when we have a significant overall ANOVA but no significant pairwise comparisons.
  - What is happening behind the scenes has to do with how the information for each variable is being partitioned. Adding new variables (including an interaction) decreases the likelihood of any individual variable being significant.
  - I will never ask you to explain this! Only to correctly interpret the significance values.

Let's explore another example to get a better understanding.

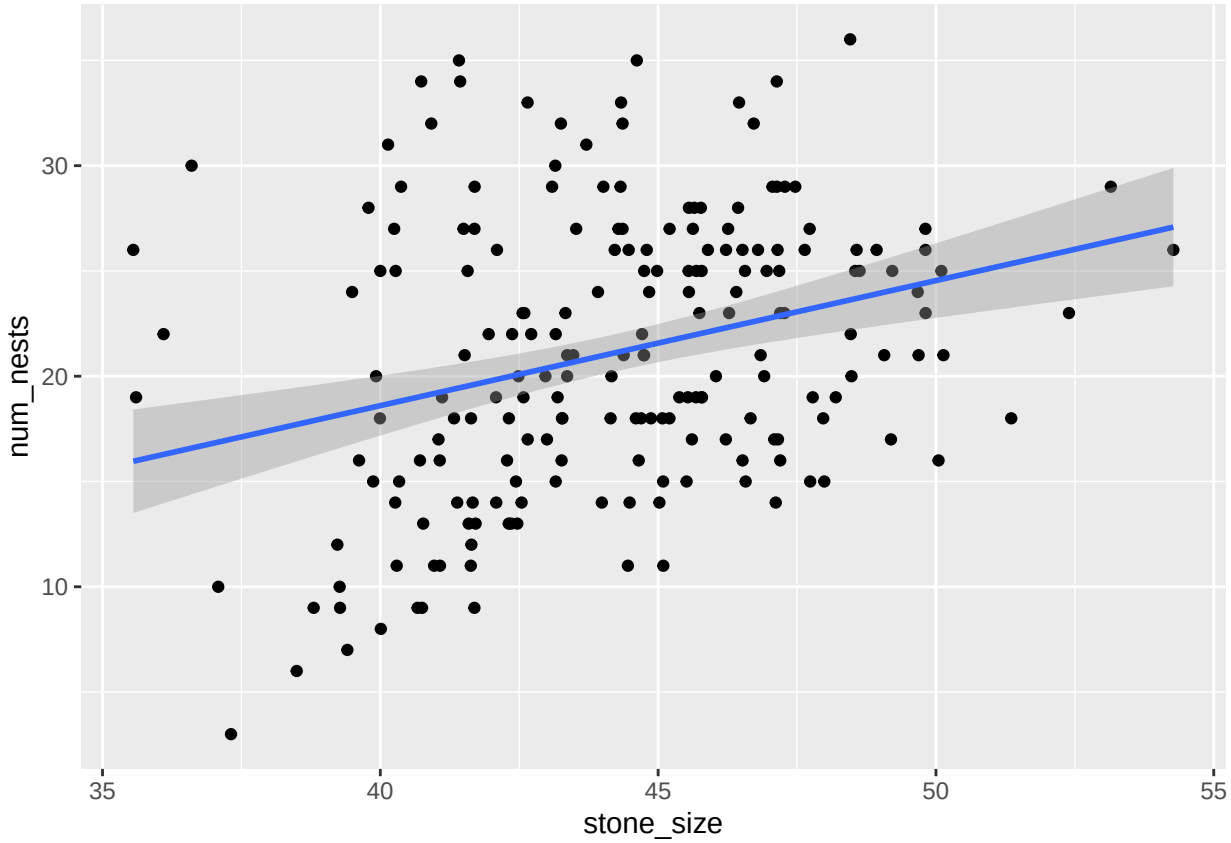
Practicing with Stone Size

Let's investigate how stone size influences nest numbers. Gentoo penguins build their nests out of stones (they actually court their penguin partner with stones, too)!

Start by plotting the relationship between stone size and the number of nests.

```
ggplot(site_data, aes(x = stone_size, y = num_nests)) +
  geom_point() +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



What does the regression model?

```
stone_size_mod = lm(num_nests ~ stone_size, data = site_data)
summary(stone_size_mod)
```

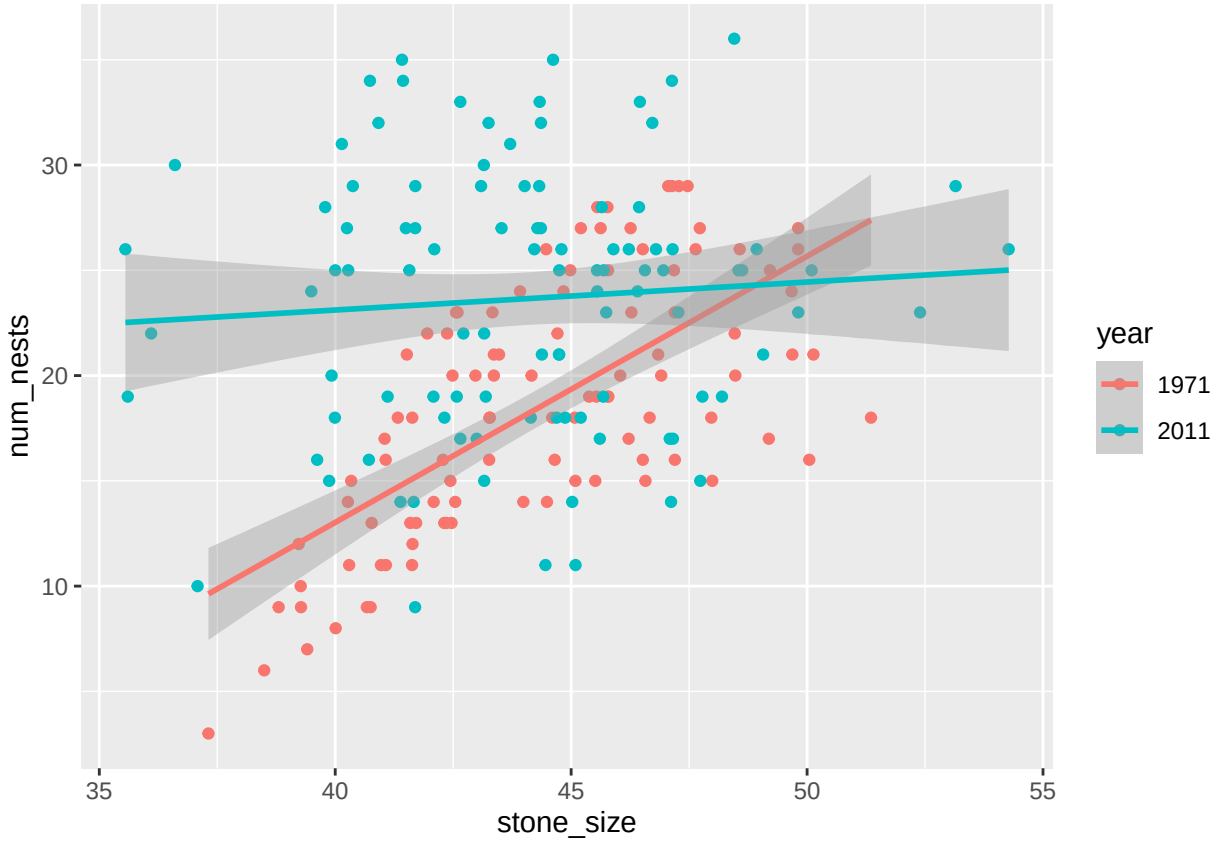
```
##
## Call:
## lm(formula = num_nests ~ stone_size, data = site_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.0069  -4.9641  -0.3939   4.0581  15.5562
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -5.1572     5.9598  -0.865    0.388
## stone_size    0.5940     0.1344   4.418 1.64e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.333 on 198 degrees of freedom
## Multiple R-squared:  0.08974,    Adjusted R-squared:  0.08515
## F-statistic: 19.52 on 1 and 198 DF,  p-value: 1.637e-05
```

How to we interpret this model?

Next, let's add year in the mix. First, plot the relationship between the number of nests and stone size by year.

```
ggplot(site_data, aes(x = stone_size, y = num_nests, color = year)) +
  geom_point() +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



This plot looks pretty different from the one with tussocks. First, let's see what happens when we add the year variable *without* the interactive term. To do that, we use a plus sign (+), not an asterisk (*).

```
stone_size_mod = lm(num_nests ~ stone_size + year, data = site_data)
summary(stone_size_mod)
```

```
##
## Call:
## lm(formula = num_nests ~ stone_size + year, data = site_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.3183  -4.3173  -0.4471   4.7328  13.0118
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  -9.5585     5.4622  -1.750   0.0817 .
## stone_size    0.6335     0.1224   5.175  5.6e-07 ***
## year2011      5.3092     0.8155   6.510  6.1e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.76 on 197 degrees of freedom
## Multiple R-squared:  0.2509, Adjusted R-squared:  0.2433
## F-statistic: 32.99 on 2 and 197 DF,  p-value: 4.387e-13
```

How do we interpret this multiple regression model?

From the ggplot above, we can see that the slopes for the two different years are quite different: one is pretty flat while the other is definitely positive. Different slopes often indicate that an *interaction* might be at work. Let's find out. Run another multiple regression model, this time adding in the interactive effect (*).

```
stone_size_mod = lm(num_nests ~ stone_size * year, data = site_data)
summary(stone_size_mod)
```

```
##
## Call:
## lm(formula = num_nests ~ stone_size * year, data = site_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.3384  -4.1293   0.1189   3.7493  11.7629
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -37.5446     7.7555  -4.841 2.61e-06 ***
## stone_size            1.2643     0.1744   7.251 9.35e-12 ***
## year2011             55.3424    10.3558   5.344 2.51e-07 ***
## stone_size:year2011  -1.1314     0.2335  -4.845 2.57e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.457 on 196 degrees of freedom
## Multiple R-squared:  0.331,  Adjusted R-squared:  0.3208
## F-statistic: 32.33 on 3 and 196 DF,  p-value: < 2.2e-16
```

Wow, a bunch of significant variables! What does that mean, exactly?

Not only are stone size and year significant, but so is the interaction between stone size and year. What does this mean? It means that penguins in different years seems to respond different to stone size. In 1971, the number of nests increases with stone size, but in 2011, the number of nests doesn't seem impacted by stone size. Therefore, there is an *interaction* between stone size and year.

Resources

Still feeling a bit confused about multiple regression and interactions? I recommend checking out this website. You can skip all the equation stuff at the beginning. Below that, though, I think they do a nice job of laying out what interactions are.

Also, if you are interested in how we select which model is the best model, this article is a decent place to start.