Multi-Variable Modeling - Interactions

NRES 710

Last compiled: 2024-10-25

Any questions before we begin?

## Interactions

Today we are going to learn about **interactions**!

Last week I suggested that collinearity is critically important to understand, because it is a nuisance that we must account for to get unbiased estimates, and so we spent an entire week on the topic. This week, I think interactions are also critically important, so we will again spend a week on this topic. But interactions are actually really interesting and can greatly enhance the interest in our research and what we are learning.

What are interactions…? Let’s start with an example.

* **Y – reproductive output**
* **– Placebo / Pill** – a binomial variable describing groups in a clinical trials of whether the ‘estrogen/progesterone/birth-control pill’.

**Q:** What is the effect of The Pill on reproductive output? ….does it cause reproduction to be ~0? *It depends….*

* **– Sex**

The Pill reduces reproductive output of females to zero, but it does not influence the reproductive output of males. This is an interactions!

**interactions – the effect (beta) of one X-variable depends on the value of another X-variable**

Sometimes people will implicitly assume there is an interaction, and separate out their data to analyze the effect of the pill on two groups separately: males and females. This is okay! BUT…

1. You are implicitly assuming there is an interaction, and,
2. if there is a difference, you cannot say the difference is statistically significant or not, because you did not test that.

Let’s say that we find that the effect of the pill on males is some very small number close to ~0 (e.g., beta = 0.001), and then we get a large beta for the effect of the pill on females. How would we know that these effects are statistically different? Well, we can’t know that if we analyzed these effects separately.

If you want to know if the effect is different between two groups (sexes, species, etc.), you need to test for an interaction.

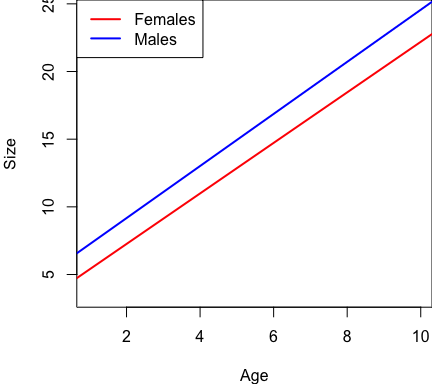
## Example

Let’s again consider an example we have seen before, that I like, because it is easy…

* **Y – Size**
* **– Age**
* **– Sex**’

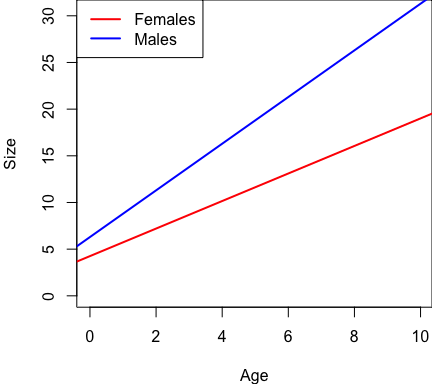
Previously we ran this multi-variable model:

This model is implicitly assuming that lines for each group are parallel – they have the same slope:



The effect of sex is the same for all ages, and the effect of age is the same for both sexes. This is an assumption that we have made. But this is *not* an assumption that we *have* to make…

What if we assume that these effects are different…?



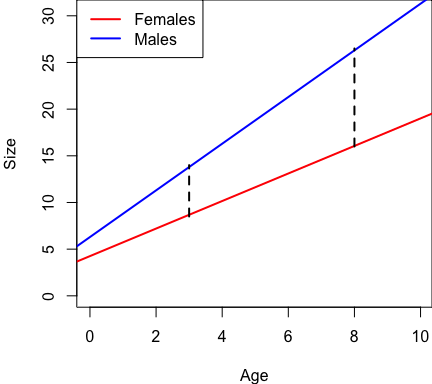
All individuals start out at about the same size, but then males grow faster and reach a larger size at an earlier age than females.

This is an interaction! The effect of age depends on sex. We get a different beta (slope) for the age-size relationship for males then we do for females.

Another interesting thing I want to point out is that for interactions, perspective does not matter:

* The effect of age depends on sex.
* The difference between the sexes also depends on age.

For example:



At a young age, the difference between sexes is small; at an older age, the difference is larger.

You can take whatever perspective is most valid for you, and it depends on the questions you are asking. A little open to interpretation.

## Interactions in the linear model

How do we add an interaction into the linear model?

1. Like I said before, you could just do this analysis separately: one regression for males, one for females. But you will get two betas and you won’t know if they are statistically significant.
2. Add in another effect to our linear model () and then multiply the two X-variables by eachother (times eachother).

So now the linear model looks like:

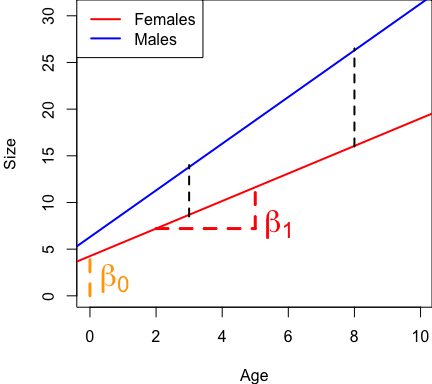
As we have done before, let’s break this down for what the interaction does for both males and females! Let’s consider females as the reference group (0) and males as the treatment group (1)

**Female = 0**

**Male = 1**

*Brian cancel out the effects of and*

The effect of beta2 disappears, and so does the effect of beta3! So the female line is simply: the intercept (beta0) plus the effect of age (beta1)!



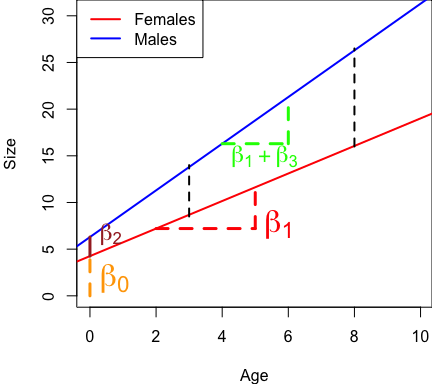
Now let’s examine this for males.

Intercept, plus effect of age, then sex gets a 1, and then sex \* age gets a 1…

We can rearrange this to make it less messy:

, or:

So now our new intercept is , and our our new slope is :



**Q:** Questions?

Let’s break this down and look at the meaning of our individual betas.

Before, was the intercept of the female line. This remains unchanged; it is the average Y-value when all other X-variables are zero! **The average size of females at the age of zero.**

What is ? **The difference in average size of males and females at age = 0.** Graphically, it’s that little distance between males and females at age zero.

If we didn’t have the interaction, what would mean? It would simply be the difference between males and females at all ages. With the interaction, it restricts this meaning of this parameter to be at zero for the continuous X-variable.

What is ? Before it was the effect of age for all individuals, everybody. Now, it is **the effect of age for females.**

Finally, what is ? This is a little more *tricky*. It is not the effect of age for males. Instead, it is the *difference* in the effect of age between females and males. The difference in the slope.

So, if we want to know if the slopes are different for males and females, that’s what tells us!

If is zero, what does that mean? …There is no interaction, there is no differnece in the slopes for females and males, and the slopes are the same.

For this reason, I argue that if you get a non-significant interaction, it would be best to take this parameter out of the model. The main reason is that:

* By leaving the insignificant interaction in, it changes the meaning of the other betas, and makes this more confusing and messy. If it’s not significant, you are justified to **take it out** for your results to be more simple and clear.

Alternatively, you might want to leave it in:

* If you have a small sample size and your interaction beta is large, maybe you might leave it in to report it so that maybe will encourage future researchers to collect more data to potentially report that effect.
* And if you are specifically testing for this interaction, you would **not want to take it out** – you would want to leave it in to **report it**! “I hypothesized this to be different, but I didn’t find support for that – but here are the results.”

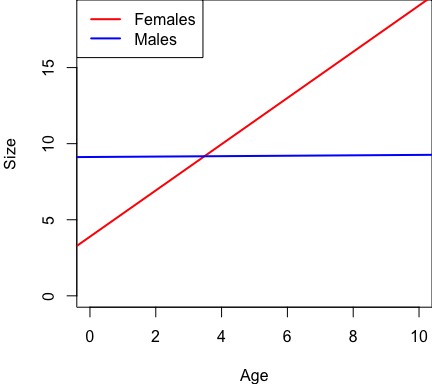
The reason why we test for interactions this way, rather than having a separate slope for males, is that:

* If , there is no interaction, and the slopes are the same.
* If , there is an interaaction, the slopes are different, and this equation accompanies that.

**Q:** As I’ve drawn this here, would be positive or negative?

Positive! It causes the slope to increase for males relative to females.

Alternatively, could be negative. Let’s assume we have a positive and a negative :



In this case, the negative causes the slope of the male line to be flat – or zero. If we did the male analysis separate from females, there would probably be no statistically significant effect here. However, the full model with the interaction effect **would recover a statistically significant interaction term**! In this case, would probably be pretty close to , but just negative as such, canceling eachother out.

There is no hard-and-fast rule for (1) whether to test for an interaction or not, or (2) whether to leave a non-significant interaction term in your model or not. It depends on your **needs**. The purpose of statistics is not to tell you the answer. Rather, it is a tool to support or reject hypotheses that you have about the world. Statistics are a way to test hypotheses. You should only test for an interaction if you have a specific hypothesis suggesting that there might be an interaction effect.

Note: this is a simple interaction model, a two-way interaction. What if we had another X-variable that might also interact with age and sex… e.g., Mother Size? That will get very complicated, very fast. And it will be very difficult to explain what those effects mean. So while you could start testing for three- or four-way interactions, it will be impossible to explain what those results mean! Point is: think about what your interactions are before putting them in the model.

## Reporting results

Without interactions, we learned to report results with two sentences:

* For the continuous variable, age: “For each unit increase in age, we observed [beta] increase in size.”
* For the categorical variable, sex: “We found that males were [beta] more/less than females.”

We now need to modify the sentence for the continuous variable. Our sentences will be:

* Continuous variable: “We found that, *in females*, for each one unit increase in age, we observed a [beta] increase in size.”
* Categorical variable: “We found that males were [beta] units larger than males *at age zero*.”
* Interaction variable sentence is a bit tricker: “We found a significant interaction, such that for males, the difference in the age-size relationship was [beta] [Y-units per X-units].”

We have to add in these *qualifiers* when we have an interaction, which makes it clunkier.

This third sentence for the interaction ~works, but it is clunky, ugly, and not exactly elegant!

Before I said that I didn’t like analyzing males separate from females because you couldn’t do a test to demonstrate that the male and female slopes were significantly different. However, describing the interaction effect is one situation where I will support (1) splitting your data, (2) re-running the simple model (e.g., males-only), and then (3) reporting that more simple beta to your readers. You can justify this approach by saying:

* “We found a significant interaction between males and females ( = […]; [p-value]). Therefore, we analyzed males and females separately. We found that, for females, for each one unit increase in age, we observed a [beta] unit (95% CI) increase in size (P-value). We also found that for males, for each one unit increase in age, we observed a [beta] increase in size.”

This all works when you have a categorical variable in the model. If you have a categorical variabel in the model, it is often easiest to figure a way to break them apart.

If you have two continuous variables, it becomes more confusing, and I recommend using a graph – specifically a three-dimensional one.

The plane is not flat! The effect of latitude is *steeper* at low elevations than it is at high elevations. Conversely, the effect of elevation is *steeper* at low latitudes than it is at high latitudes. You might describe these features qualitatively to your readers.

## Analysis in R

Let’s analyze some data that we simulated in R. Here is truth:

### Dataset 1: age + sex + age\*sex  
# This is similar to the Age, Sex, and Size data we simulated for in Lecture 12.  
# There is no collinearity between Age and Sex, but now there is an interaction  
# between Sex and Age.  
  
# First dataset  
# X variable  
n <- 50  
x1 <- c(rep("Female", n), rep("Male", n))  
x2 <- runif(n \* 2, 1, 10)  
dummy <- data.frame(model.matrix(~ x1 - 1))  
colnames(dummy) <- c("Female", "Male")  
  
# Simulate error  
Error <- rnorm(n \* 2, 0, 0.8)  
  
# Predict Y  
Response <- 1 + 2 \* x2 + 4 \* dummy$Male + 1 \* x2 \* dummy$Male + Error  
  
# Dataframe  
datum <- data.frame(Age = x2, Sex = x1, Male = dummy$Male, Size = Response)  
  
# These data were saved as 'lecture\_13\_dataset5.csv'

What is **Truth** for these data – and what do the effects mean?

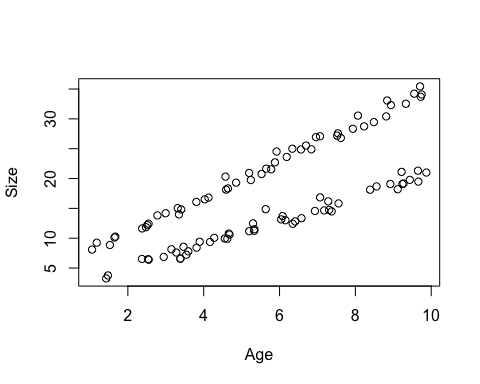
* **beta0 = 1** – average size of females at age 0
* **beta1 = 2** – effect of age in females
* **beta2 = 4** – difference between males and females at age 0
* **beta3 = 1** – difference in slope between males and females

**Q:** Given these numbers, what is the slope of the age-size relationship for males?

3! = 2 + 1 = 3

Let’s do the analysis in R.

# Load the data  
datum <- read.csv("lecture\_13\_dataset5.csv")  
head(datum)  
  
# Plot the data  
plot(Size ~ Age, data = datum)



If your data look like this, you likely have an interaction!!

# Fit a lm() without the interaction  
results <- lm(Size ~ Age + Sex, data = datum)  
summary(results)

##   
## Call:  
## lm(formula = Size ~ Age + Sex, data = datum)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.0042 -1.2144 0.0797 1.1571 3.1852   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -2.20797 0.41113 -5.371 5.38e-07 \*\*\*  
## Age 2.55612 0.06095 41.935 < 2e-16 \*\*\*  
## SexMale 9.65201 0.30203 31.957 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.509 on 97 degrees of freedom  
## Multiple R-squared: 0.9652, Adjusted R-squared: 0.9644   
## F-statistic: 1344 on 2 and 97 DF, p-value: < 2.2e-16

Is there any indication that we missed an interaction? Nope. There are symptoms of collinearity, but there aren’t symptoms of interactions.

The effect of age is ~2.5, which is basically the average of the male and female slopes. Which, isn’t too bad.

The difference between males and females is ~9.5. However, for our data, the difference between males and females at young ages is smaller, and at old ages is larger. So this is again an average. It’s not bad… But we are missing an opportunity to learn about interesting information in our data.

# Fit a lm() without the interaction  
results <- lm(Size ~ Age + Sex + Age:Sex, data = datum)  
summary(results)

##   
## Call:  
## lm(formula = Size ~ Age + Sex + Age:Sex, data = datum)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.2818 -0.5160 -0.1589 0.4397 2.5955   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.97179 0.28952 3.357 0.00113 \*\*   
## Age 2.00454 0.04640 43.200 < 2e-16 \*\*\*  
## SexMale 3.79496 0.39304 9.655 8.19e-16 \*\*\*  
## Age:SexMale 1.03077 0.06343 16.250 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7833 on 96 degrees of freedom  
## Multiple R-squared: 0.9907, Adjusted R-squared: 0.9904   
## F-statistic: 3413 on 3 and 96 DF, p-value: < 2.2e-16

The age-size relationship for females is 2! The effect of being male is 3.79, and truth (4) is contained within the confidence interval. The difference in slope between male line and female line is 1.03 – very close to truth.

You could report your results using these results. “We found that for each one year increase in age, females had a 2.00 cm increase in size.” – with p-values and confidence intervals. Same sentence as before, but we have the qualifiers *in females*. “We found that males were 3.79 [CI] cm larger at age 0 [P].” Again, need the qualifier of male. “We found that the slope of the age-size relationship was 1.03 [CI] cm per year larger in males than females [P].”

Or, you could state that you found a significant interaction between the X-variables: “We found a significant age-sex relationship (P < 2e-16); therefore, we analyzed males and females separately.”

How do you do that?

# Fit a lm() for males only  
resultsMale <- lm(Size ~ Age, data = subset(datum, datum$Sex == "Male"))  
summary(resultsMale)

##   
## Call:  
## lm(formula = Size ~ Age, data = subset(datum, datum$Sex == "Male"))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.1057 -0.5300 -0.2032 0.4265 1.7832   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.76675 0.25703 18.55 <2e-16 \*\*\*  
## Age 3.03531 0.04182 72.58 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7574 on 48 degrees of freedom  
## Multiple R-squared: 0.991, Adjusted R-squared: 0.9908   
## F-statistic: 5268 on 1 and 48 DF, p-value: < 2.2e-16

When we run that, we get a slope of 3.03 – which is equal to the sum of the age effect and the age:sex effect from our interaction model. Just adding them together and getting the same result – but now we have a confidence interval for our simple slope also:

# Confint for males  
confint(resultsMale)

## 2.5 % 97.5 %  
## (Intercept) 4.249954 5.283540  
## Age 2.951226 3.119389

And we can do this for females also:

# Fit a lm() for females only  
resultsFemale <- lm(Size ~ Age, data = subset(datum, datum$Sex == "Female"))  
summary(resultsFemale)

##   
## Call:  
## lm(formula = Size ~ Age, data = subset(datum, datum$Sex == "Female"))  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.2818 -0.4188 -0.1302 0.5029 2.5955   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.97179 0.29879 3.252 0.0021 \*\*   
## Age 2.00454 0.04789 41.860 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8084 on 48 degrees of freedom  
## Multiple R-squared: 0.9733, Adjusted R-squared: 0.9728   
## F-statistic: 1752 on 1 and 48 DF, p-value: < 2.2e-16

confint(resultsFemale)

## 2.5 % 97.5 %  
## (Intercept) 0.371030 1.572547  
## Age 1.908256 2.100820

This will slightly influence your confidence intervals, but that’s okay.

## Using the \*

Here’s some other stuff to know about:

# Fit a lm() with interactions using a \*  
results <- lm(Size ~ Age\*Sex, data = datum)  
summary(results)

##   
## Call:  
## lm(formula = Size ~ Age \* Sex, data = datum)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.2818 -0.5160 -0.1589 0.4397 2.5955   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.97179 0.28952 3.357 0.00113 \*\*   
## Age 2.00454 0.04640 43.200 < 2e-16 \*\*\*  
## SexMale 3.79496 0.39304 9.655 8.19e-16 \*\*\*  
## Age:SexMale 1.03077 0.06343 16.250 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7833 on 96 degrees of freedom  
## Multiple R-squared: 0.9907, Adjusted R-squared: 0.9904   
## F-statistic: 3413 on 3 and 96 DF, p-value: < 2.2e-16

The \* will fit all possible combinations of effects between the variables included in the term. This can be a useful shortcut. BUT, it might be a little sloppy and might not clearly convey what is going on. I prefer to be explicit and spell out the terms individually.

## Summary

Interactions are not that complicated! An interaction means that your betas for a given variable change depending on another variable. If you have groups, the beta depends on the group; different betas for each group. If you have interactions between continuous variables, the effect of each variable depends on the other variable, and you can change perspective depending on either variable (e.g., the second 3-D plot above).

I encourage you to always tests for interactions, rather than splitting your data and running separate tests for your groups. That way you can get an explicit test that your data are different, according to that interaction term! However, once you find that significant interaction term, I then *do* encourage you to break those groups apart, so that the explanation is much more simple.

## Truth

################### 'Truth' ####################  
### Lecture 13: code to simulate data for class  
  
# Set the seed for reproducibility  
set.seed(123)  
  
### Dataset 1: age + sex + age\*sex  
# This is similar to the Age, Sex, and Size data we simulated for in Lecture 12.  
# There is no collinearity between Age and Sex, but now there is an interaction  
# between Sex and Age.  
  
# First dataset  
# X variable  
n <- 50  
x1 <- c(rep("Female", n), rep("Male", n))  
x2 <- runif(n \* 2, 1, 10)  
dummy <- data.frame(model.matrix(~ x1 - 1))  
colnames(dummy) <- c("Female", "Male")  
  
# Simulate error  
Error <- rnorm(n \* 2, 0, 0.8)  
  
# Predict Y  
Response <- 4 + 1.5 \* x2 + 2.5 \* dummy$Male + 1 \* x2 \* dummy$Male + Error  
  
# Dataframe  
datum <- data.frame(Age = x2, Sex = x1, Male = dummy$Male, Size = Response)  
  
# Save the data  
write.csv(datum, "lecture\_13\_dataset1.csv", row.names = FALSE)  
  
  
### Dataset 2: age + sex + age\*sex, with a negative interaction effect  
# First dataset  
# X variable  
n <- 50  
x1 <- c(rep("Female", n), rep("Male", n))  
x2 <- runif(n \* 2, 1, 10)  
dummy <- data.frame(model.matrix(~ x1 - 1))  
colnames(dummy) <- c("Female", "Male")  
  
# Simulate error  
Error <- rnorm(n \* 2, 0, 0.8)  
  
# Predict Y  
Response <- 4 + 1.5 \* x2 + 5 \* dummy$Male + -1.5 \* x2 \* dummy$Male + Error  
  
# Dataframe  
datum <- data.frame(Age = x2, Sex = x1, Male = dummy$Male, Size = Response)  
  
# Save the data  
write.csv(datum, "lecture\_13\_dataset2.csv", row.names = FALSE)  
  
  
### Dataset 3:  
# Set the seed for reproducibility  
set.seed(123)  
  
# Sample size  
n <- 90  
  
# Simulate X-variables  
# Continuous variables: Latitude and Elevation  
Latitude <- runif(n, 0, 1) \* 30  
Elevation <- runif(n, 0, 1) \* 30  
  
# Response variable: Size  
Size <- 3.33\*Latitude + 1.6\*Elevation  
  
# Create dataframe  
datum <- data.frame(Latitude=Latitude, Elevation=Elevation, Size=Size)  
  
# Save the CSV file  
write.csv(datum, "lecture\_13\_dataset3.csv")  
  
  
### Dataset 4: size ~ latitude + elevation + latitude\*elevation  
# Set the seed for reproducibility  
set.seed(123)  
  
## Dataset 4  
# Sample size  
n <- 90  
  
# Simulate X-variables  
# Continuous variables: Latitude and Elevation  
Latitude <- runif(n, 0, 1) \* 30  
Elevation <- runif(n, 0, 1) \* 30  
  
# Response variable: Size  
Size <- 3.33\*Latitude + 1.6\*Elevation - 0.08 \* Latitude \* Elevation  
  
# Create dataframe  
datum <- data.frame(Latitude=Latitude, Elevation=Elevation, Size=Size)  
  
# Save the CSV file  
write.csv(datum, "lecture\_13\_dataset4.csv", row.names = FALSE)  
  
  
### Dataset 5: age + sex + age\*sex  
# This is similar to the Age, Sex, and Size data we simulated for in Lecture 12.  
# There is no collinearity between Age and Sex, but now there is an interaction  
# between Sex and Age.  
  
# First dataset  
# X variable  
n <- 50  
x1 <- c(rep("Female", n), rep("Male", n))  
x2 <- runif(n \* 2, 1, 10)  
dummy <- data.frame(model.matrix(~ x1 - 1))  
colnames(dummy) <- c("Female", "Male")  
  
# Simulate error  
Error <- rnorm(n \* 2, 0, 0.8)  
  
# Predict Y  
Response <- 1 + 2 \* x2 + 4 \* dummy$Male + 1 \* x2 \* dummy$Male + Error  
  
# Dataframe  
datum <- data.frame(Age = x2, Sex = x1, Male = dummy$Male, Size = Response)  
  
# Save as CSV  
write.csv(datum, "lecture\_13\_dataset5.csv")

[–go to next lecture–](lecture_14.html)