Multi-variable modeling - collinearity

NRES 710

Last compiled: 2024-10-15

## Collinearity

Today’s class is an important one. The topic is on something that very few people understand well. So the plan is to spend 1-2 days on this topic so that you will understand it well. Many people have a decent idea of what **collinearity** is – but tend to make uninformed decisions about how to handle it.

**Collinearity – when X-variables are associated (correlated) with eachother**

Let’s start with a question for the class. We often say that “*correlation does not equal causation*”. More specifically, you might run a regression between X and Y and find a relationship, but that does not mean that X causes Y.

**Q:** Why not? Discuss this a bit.

Let’s consider an example. There was a study that came out a few years ago that generated some interest from the press. This study suggested that: for each 1 additional hour of TV that you watch per day, it increases the risk of mortality by 11%.

Let’s think about this for a minute. Do you think the actual act of watching TV increases the risk of mortality? How would that happen? Is the TV going to fall out of the wall and potentially injure you? No, probably not. Watching TV itself is not going to cause this relationship.

What’s the problem here? Well, there are a lot of other things that are **correlated** to watching TV that may have their own effect on the Y variable (mortality).

**Confounding factors – an X-variable that is correlated with the X-variable of interest and has its own effect on Y**

What is correlated with watching TV?

* You aren’t being active (e.g., exercising)!
* You might be snacking, and your snacks might not be something healthy. You probably aren’t eating carrots if you are watching TV. I like eating chocolate covered pretzels!
* You aren’t working, which might relate to your socioeconomic status.
* You may be less likely to have healthcare.
* *Et cetera*

These features might all be correlated to how much an individual watches TV, and all might have their own effects on mortality.

**Q:** Here’s the key idea If we really want to know what the true effect of TV watching on mortality is, what would would we do?

**A MANIPULATIVE EXPERIMENT**. In an ideal world, we would do a manipulative experiment, which is the only true way to demonstrate causation.

* Option 1: We would try to control for all of the above factors: everyone has the same job, goes to the gym the same amount of time, no smoking, etc. Let’s assume this class is our sample! We would assign half of the class to watch TV for 1 hour a day, while the other half of the class watches for 10 hours per day. **Control Everything!**
* Option 2: We wouldn’t try to control for any of those factors, but we would randomly assign for 1 vs. 10 hours of TV per day. This is a lower-quality experiment, because all of those other factors may cause noise that *swamps out* our ability to detect an effect of watching TV. We are not controlling for anything, so all of these other factors may create *noise*. **Randomize Everything!**
  + But, we could collect data on those factors, and include those data in the models, and then we could avoid the swamping issue.
  + Of course, this experimental design is super impractical… There’s no way we could get our sample to watch TV for 10 hours a day – or probably even 1 hour per day in many cases.

So what do we do? Researchers often resort to performing an **observational study**, where they collect data on all of these other features that might also influence the response variable.

* Diet, exercise, socioeconomic status, smoking, alcohol, and maybe a whole slew of other things that we might hypothesize also influence the response variable (mortality). - Put all those variables into the model and it will *statistically control* for them (decrease the swamping).
* When we include correlated variables in a model, it ***statistically controls*** for them – **and we’ll demonstrate this today**.

However, ultimately we may not be able to show causation, because there might be some other confounding variable(s) in the world that we did not collect data for and include in our model.

**The whole point of this is:** When you are conducting studies and experiments for your thesis or your career, you need to *think carefully* about what are the potential confounding factors – the other variables – that may be correlated to the X-variable you are interested in and that might influence your results.

* This is less of a concern in a **well-designed manipulative study**, where random assignment of treatment removes collinearity.
* For example, if I randomly assign to the class who watches 1 vs. 10 hours of TV, then there won’t be any collinearity between exercise and TV watching. Presumably people that go the gym often will be randomly assigned to 10 hours of TV, and people who never go to the gym will be made to watch 1 hour of TV. **This random assignment removes potential collinearity between these variables.**
* Again, manipulative experiments are the strongest approach to inference in science… but they can be **impractical**, and instead many of us resort to observational studies.

My goal with this lecture today is to demonstrate to you all that: **if there are confounding variables and you don’t include them in your model/analysis, the estimates from your analysis will be biased (incorrect)!**

**If confounding variables are not included in model/analysis –> biased (incorrect) estimates (s)**

People kind of understand this… but I want you to really understand it. If there are variables that are correlated to your X-variable of interest, and those other correlated X-variables have an effect on your response variable, and you don’t include data on those confounding variables in your analysis, then your estimates will be biased (incorrect). This is scary!! That’s why this is one of the more important topics we will cover this semester.

Let’s see how this works.

## Biased estimates without confounding variables

Let’s reconsider an example that we examined in the last class.

* **Y = size**
* **= age**
* **= sex**

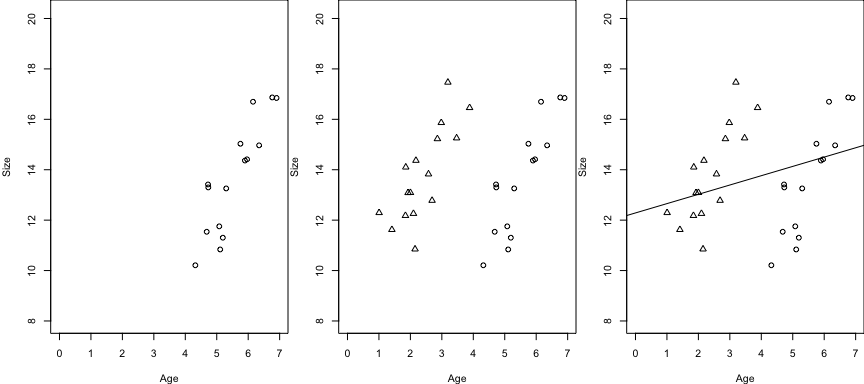
*Brian draw data with two clusters for males and females, with same slope, a sex effect, and complete sampling across all ages*

Last class we discussed what would happen if we built this model:

It would create a line down the middle, which would have an acceptable estimate of the effect of age on size.

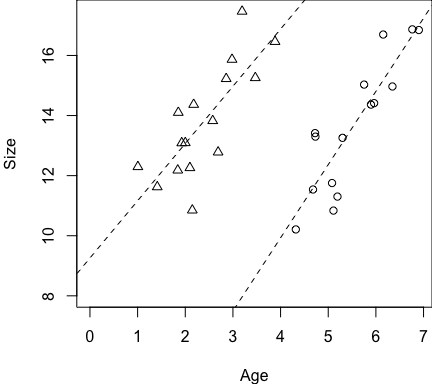
However, this time around when you were sampling the animals of interest, for some reason you had a hard time collecting data on young females, and also on old males.

*Brian erase young females and old males to create collinearity between age and sex*



**In your sample, sex is correlated with age!** Males are young and females are old.

**Q:** If I ran a regression, where would the line go?

For these data, it would be ~horizontal across the data. This is a biased estimate. This estimate **incorrectly estimates the effect of age on size of these animals**.

However, if we run this model:

We get two lines! And these lines will look like:

Now, we get *unbiased* estimates of the effect of Age on Size for each Sex.

What’s happening here is… **If we leave out Sex in the first model, some of that Sex effect is being transmitted to the effect of Age.**

* Females are smaller than males, and the difference in size is transmitted to the effect of Age on Size
* Age has a positive effect on size, female has a negative effect on size, and these two counter effects sort-of balance each other out when we don’t include Sex in the model.
* This causes a **biased estimate**.

The same thing might be happening in the TV-effect on mortality study. There were many other variables that were correlated to watching TV that were really influencing mortality. **Those other effects, which were left out of the statistical model, were being transmitted to the effect of watching TV.**

**Q:** Does that make sense? Questions?

Reiterating the important point: **if we leave confounding variables out of our model, we get biased estimates of the variables that we do leave in the model.**

* So, we want to collect data on confounding variables and include them in the model.
* And we want to do this even if they are not statistically significant.
* **By leaving potentially confounding variables in the model, we will *statistically control* for confounding effects.**

### Tradeoff

But, there is a tradeoff…!

**If confounding, collinear variables *are* included in model –> Variance Inflation**

When we run a model with multiple X-variables, the computer package is trying to figure out which of the two variables are having an effect on Y. But they are correlated to eachother, which makes it tough for the software to figure out out which X-variable is having what effect. And the more strongly they are associated with eachother, it becomes more difficult for the software to disentangle these effects.

A consequence of this is that running a model with multiple, collinear variables causes greater uncertainty in the effects (s). Increased uncertainty is reflected in:

* **increased SE**
* **increased CI**
* **increased P-values**

This can lead to problematic results.

I mentioned during the last class that, generally, if we add a variable to a model, the p-values for the variables in a model go down. **However, collinearity is an exception to that rule!** And in fact, it’s a red flag.

My recommendation is that, before you even start your analysis, you need to know what the collinearity is between your X-variables! Plot your X-variables, run some linear models (e.g., between Age and Sex), and learn about what the collinearity is.

* Some people fail to do this and problems occur. For example, they might build a model where Age was significant… and then they add Sex, and Age is no longer significant…
* P-values are supposed to go down when we add variables into models. If the P-value goes up a lot when you add a variable to a model, this is a symptom of collinearity! This is happening because of collinearity between X-variables.

## Symptoms of collinearity

1. Collinearity between independent variables

* High values between X-variables
* Statistically-significant relationships between X-variables

1. High variance inflation factors (VIF) of variables in model
2. Variables significant in simple regression, but not in multi-variable regression
3. Individual variables not significant in multi-variable regression model, but the overall multi-variable regression model is significant
4. Large changes in coefficient estimates between full and reduced models
5. Large SE in multi-variable regresion models, despite high power

Breaking this list down a bit…

* 1. The easiest way to know if you have collinearity is to examine fit a linear model between your two independent variables and see if they have **high values** or if there is a **statistically-significant relationship** between them.
  2. High variance inflation factors (VIF) of variables in model. Next class I will show you how to examine VIF for variables in a model.
  3. As I said before, if you have significant variables in a simple linear regression, but then you add another variable(s) to create a multi-variable regression and the individual variable(s) is no longer significant, then you have collinearity.
  4. If you build a multi-variable model where (i) none of the individuals variables are significant, but (ii) the overall whole model is significant (the p-value at the bottom of the model), then you have collinearity. The whole model is significant, but none of the individual variables is significant.
  + How does that happen? All of the X-variables are explaining a lot of the Y-variable, but the software is having a hard-time figure out which X-variable is explaining the response. And the collinearity causes the p-values for individual variables to go up.
  1. Generally, when you add variables to a model, the betas *should not change too much*. What do I mean by ‘too much’? E.g., when you have a simple model with beta and CI, and then you add another X-variable to the model to create a multi-variable model, then the new estimate for your beta should be within the 95% CI of the original estimate. **If it moves to be outside of the 95% CI of the original estimate, that’s a considerable change and is a symptom of collinearity of the X-variables.**
  2. What if you have 25,000 samples, and you have huge standard errors. Why are the SE so big, when we have such a large sample size? Collinearity in the system can increase SE, increase CI, and decrease p-values.

**Questions?**

## Simulation Exercise 1

When confronted with collinearity, a common approach taken by many folks in our field is to **take variables out of the model**. However, this has a consequence that they do not understand – it biases our estimates! – so this is not the approach I recommend.

I’m going to show you a simulation exercise that demonstrates:

1. Collinearity influences parameter **estimates**, **uncertainty**, and **p-values** when confounding variables are not included in the model, and
2. When all the necessary confounding variables are included, collinearity increases uncertainty and p-values, but **does not bias estimates**.

### Truth

* – an estimate of how much correlation (collinearity) there is between and

For the simulation exercise, I simulated 1,000 datasets with varying degrees of collinearity (correlation) between two X-variables. Here is truth:

* Simulations:
* – both X variables have effects on Y!
* For each simulation, I used a different value of *z* from a uniform distribution: .

**Low values** of *z* (e.g., *z* = 0.5) caused **high collinearity** between the X-variables, because only a little bit of noise gets added to to calculate . Alternatively, **high values** of *z* (e.g., *z* = 20) had **low collinearity**, because large noise was added to to calculate . In other words, the low values of *z* caused very little noise (and high collinearity) between and , while the high values caused great noise (and low collinearity) between the variables.

For each simulation, I did a few things:

* Fit a **simple model** () and measured the estimate, SE, and p-value for
* Fit a **multi-variable model** () that included both of the collinear, confounding variables, and measured the effect, SE, and p-value for .

Most importantly, I measured how collinearity between and (i.e., ) influenced the the Variance Inflation Factor from the multi-variable model.

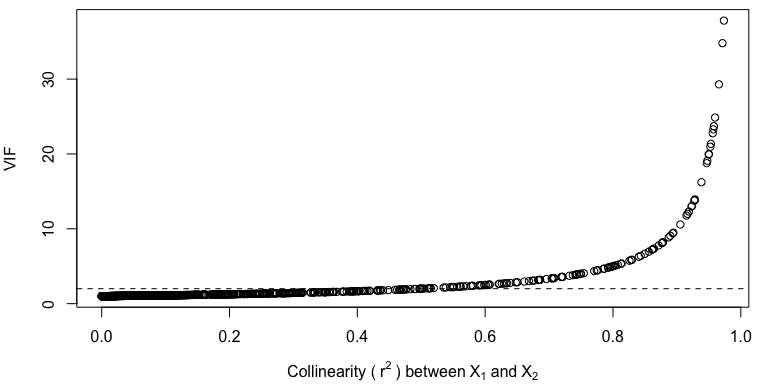
I’m going to show a number of graphs, and for all of these graphs the X-axis is a measure of collinearity (correlation; ) between and . If it’s collinearity equals 1, then the variables are perfectly correlated, whereas low values show weak/no correlation.

The **Variance Inflation Factor** is a metric used to diagnose whether there may be collinearity between X-variables in a multiple-variable model. The first graph I am going to show you describes how the **Variance Inflation Factor** score is influenced by collinearity.

**Variance Inflation Factor (VIF) – the amount (in *times*) that the variance () in the increases due to collinearity**

* E.g., if VIF = 4, then the variance is 4 times what it would be if you didn’t have collinearity – and the SE would be doubled ().

First, let’s see how the VIF is influenced by collinearity:



This graph shows how VIF scores increase with increasing collinearity.

Some people say that if the VIF > 10, you need to worry about collinearity. In this case, the SE would be triple what it would otherwise be without collinearity. This happens when , and the two X-variables are almost perfectly correlated. This is too high, in my opinion.

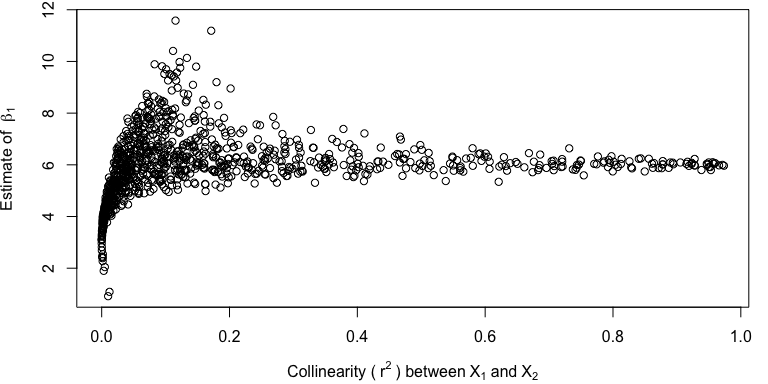
Rule-of-thumb: **VIF > 2, recognize that you have collinearity in your data.**

VIF = 2 ~ = 0.6 – so another rule of thumb is that an may also indicate collinearity.

### Simple regression model:

Now, let’s consider results from the simple model () across the simulations.

The most important thing I want to emphasize to you all is here in this first graph. How is the **beta for**  influenced by collinearity in the simple model?



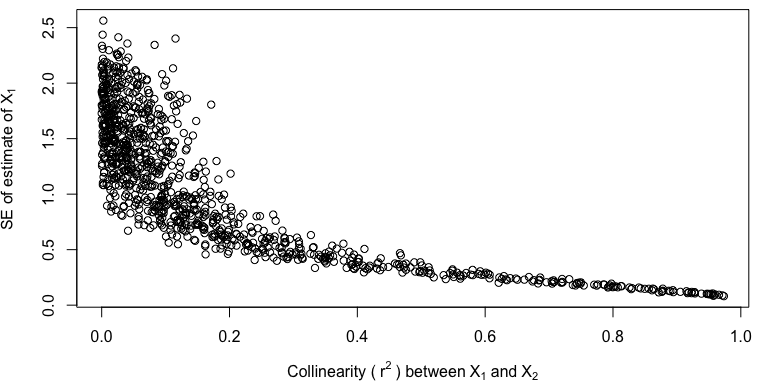
This graph shows, again, the between and (how correlated they are) on the X-axis, and the Y-axis shows the estimate for the beta for . If there is no collinearity, the estimate is 3 – which is truth! Bottom-left-hand side of the graph.

However, even with a very little bit of collinearity (), the parameter estimate is biased! As collinearity increases, the parameter estimate jumps pretty quickly up to 6!

**Q:** Why 6? 6 = 3 + 3. Because the two X-variables are correlated, the estimation model starts to **assign** the effect of onto .

Not including the confounding variable introduces bias in our estimates and negatively influences our inference.

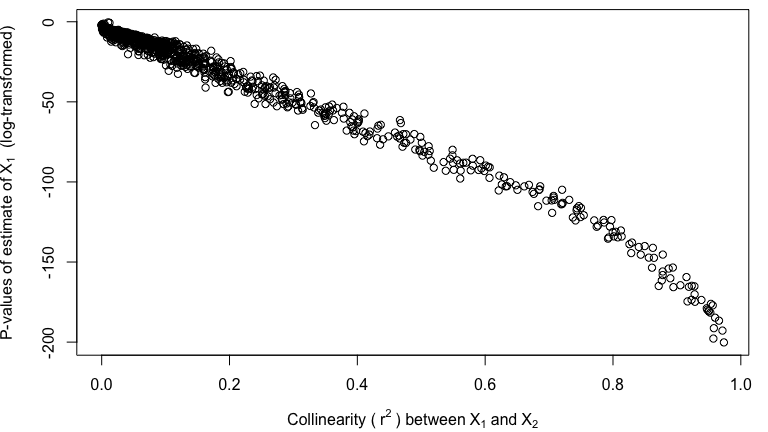
Now, often in this class we ask: how do different approaches influence effects, uncertainty, and p-values. Next we might ask ourselves: how is **standard error (SE)** influenced by collinearity in the simple model?



This is where it gets a little scary.

Not only is the parameter estimate biased, the standard error gets smaller and smaller. As collinearity increases, your estimates are getting more and more precise…

How are **p-values** influenced by collinearity in the simple model?



As uncertainty gets smaller and smaller, our p-values also get smaller and smaller, because the effect of collinearity gets bigger and bigger…

So, collinearity is **really dangerous**! If you have collinearity in your data, and you leave out some confounding variable, you are going to get biased estimates, and it’s going to seem like it’s a really precise estimate with a clear result. Unless…

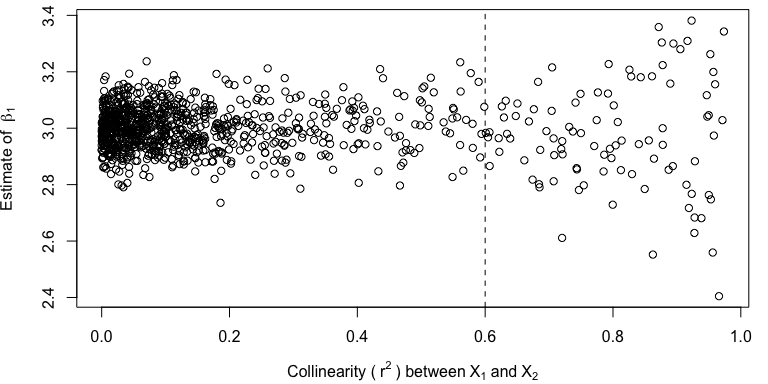
**Q:** What if the effect of was -3? The two effects might cancel eachother out, and our estimate would be zero, with a small SE, and *false confidence that there is no effect*.

If the effect of the unaccounted confounding variable, , has an effect in the opposite direction, the collinear effects may partially or wholly cancel eachother out, and you may fail to observe any true effect of X on Y.

### Multi-variable model:

Now, let’s consider results from the multi-variable model () across the simulations. This is the model where we have *statistically controlled* for that collinearity.

Let’s start here by asking: how does collinearity influence the **beta for**  in the multi-variable model?

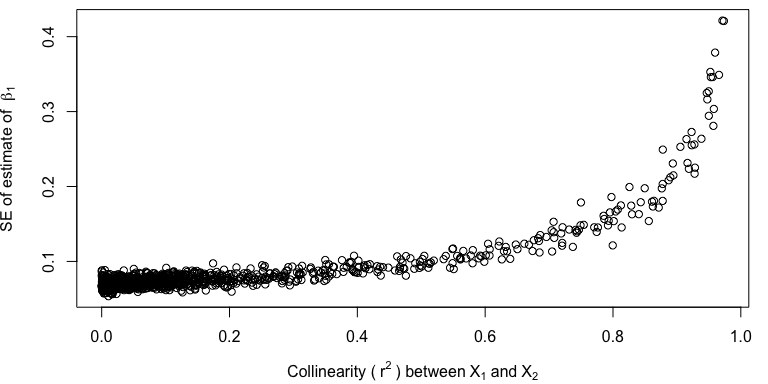


On the Y-axis, we have the coefficient estimate for , and the X-axis is again the correlation between the two X-variables.

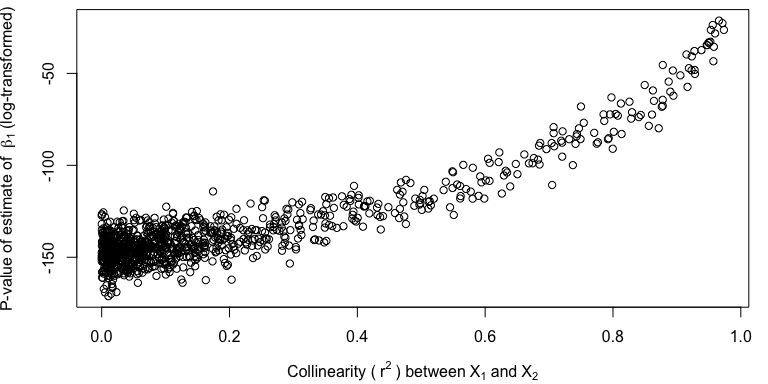
**The estimate is good**, right around 3 (truth)!! But as collinearity increases, the estimate can be more variable. So by including the confounding variables, your estimate is good, but the bad news is that due to the collinearity, you have less certainty in your estimate. But at least the estimate is unbiased, which is the most important thing.

Remember the rule of thumb about indicating collinearity? Previously I said that when , collinearity should be on your radar. In this graph, we can see that simulations with collinearity greater than 0.6 start to have parameter estimates that are more variable. However, the mean is still centered on ~3 (Truth) after this rule-of-thumb. So no matter how much collinearity there is, on average we will continue to get an unbiased estimate of the effect of X-variables in the model. In statistics, any one estimate may be bad – but on average, for this multi-variable model, the estimates will be unbiased.

Another thing that we can see from this graph is the **variance inflation** itself! As collinearity increases along the X-axis, the cone of simulation results gets bigger, because there is more and more uncertainty in what the parameter estimate really is.



And we see this here as well with the SE estimates. As collinearity increases, SE of the estimate increases, again suggesting increasing uncertainty due to collinearity.



And again, as uncertainty increases, p-values increase as well.

But hopefully the takehome message is clear. **Failing to account for confounding variables in a model will bias estimates**. Accounting for those confounding variables will produce unbiased estimates, although the uncertainty may increase.

**Q:** So how do we deal with collinearity? **We include all confounding variables in a multi-variable model!**

### Confounding variables

**Confounding variable – a variable that will bias results if you leave it out**. There are two key features for confounding variables:

* Correlated with another X-variable
* Has it’s own effect on Y

To avoid negative effects of confounding variables, I recommend:

1. **Sample in a manner that eliminates collinearity**. Collinearity may be real, or it may be due to sampling artifact. E.g., in the first example above, collinearity was introduced due to non-random sampling of the different sexes. Collect your data in a way that eliminates collinearity.
2. **Use multi-variable regression**. If you have collinear X-variables, it may cause your SE to be large, but your estimates will be unbiased. If you fail to include confounding variables, then estimates will be biased.
3. **Include confounding variables, even if they are non-significant**.

Scenario: two collinear variables (Age and Sex), and you run the full model. VIF is really high (e.g., 10). SE and CI are very inflated, but everything is still significant. Do you do anything? No.

* Your estimates are unbiased, because you ran the full model.
* Your CI are large, but still statistically significant.
* You have identified that you have collinearity, which caused the large CI. But the betas are unbiased.

But sometimes you run the analysis but X-variables are not significant. Maybe the model is significant, but none of the X-variables are. What do you do then?

1. **Get more data!** This decreases SE and VIF.

Most problems in statistics can be improved by getting more data. Increasing N decreases SE, CI, and P-values… If none of your data are significant due to collinearity, maybe you just need to collect more data, reduce uncertainty, and increase power.

### Redundant variables

If collinearity between X-variables results in biased estimates, why do so many people throw out their variables in favor of reduced models that will suffer from collinearity and thus have biased estimates?

Well, you can do this *sometimes* – primarily when your collinear X-variables don’t have an effect on Y.

These are sometimes called **redundant variables – collinear X-variables that don’t have an effect on the Y-variable**.

1. Correlated to another X-variable, but
2. Do not have an effect on Y

They are similar to confounding variables by being correlated to another X-variable, but the **key difference is that they do not have an effect on Y**.

* If they don’t really have an effect (i.e., = 0), there is no problem when you take them out. Leaving out a redundant variable does not cause bias, because it does not have an effect to add onto another variable included in the model!
* If you leave out a confounding variable, it assigns it’s effect to the other collinear variable you left in. In the case of redundant variables, leaving them out assigns the effect to the other X-variable – but that effect was zero!
* If you include a redundant variable, it will cause **variance inflation**.

A useful way to think about confounding or redundant variables might be with the s.

* If the , it’s a confounding variable.
* If the , it’s a redundant variable.

You don’t need to leave in redundant variables, and leaving them in will generate variance inflation.

What’s an example of a redundant variable? Watching TV and it’s effect on mortality. If you leave this in your analysis, you will measure a after you include all of the other important things that truly do influence mortality. And by leaving it in the model you will increase SE, CI, and p-values of all of the other important variables.

## Redundant vs. Confounding Variables

Like I said before, many scientists tend to throw out correlated X-variables because they are afraid of ‘collinearity’, and they want to reduce the model to avoid variance inflation.

This is fine – if you are 100% certain that those variables are redundant.

If they are confounding variables, they must be included – because removing them will cause biased betas, reduced uncertainty, and small p-values.

So, how do we know whether variables are **redundant** or **confounding** variables?

**Redundant vs. Confounding Variable**

### 1. Logic!

Put some thought into your variables. Statistics is a thinking exercise and should involve critical thinking. Is it likely that your X-variable has it’s own effect on Y? If so, include it.



Example: Gopher Tortoises. The gopher tortoise is a burrow-digging turtle species in the southeastern United States that is a species of conservation concern. Reproduction is important for populations to grow and persist, and researchers are interested in what causes reproduction in tortoise populations. Male tortoises are known to compete with eachother for access to females, and they often get in wrestling matches for access to females. Males sometimes have large shell extensions under their neck called ‘gular scutes’ that are used to flip other males during male-male combat. People have wondered what are the factors that influence reproductive success of tortoises: age, body size, or gular scute size? These factors are all correlated (collinear): older tortoises get larger bodies and may grow larger gular scutes. However, there is likely variation in growth rates such that sometimes younger tortoises may be larger and/or have larger gular scutes than older tortoises. The question is: are these variables redundant or confounding variables?

There may be an argument that they all of these variables are measures of age, and shell size and gular scute size are redundant variables. Alternatively, it could be that they each have their own effect, where age, size, and gular scute length have different effects on reproductive success. Maybe: (1) older tortoise are more experienced at winning wrestling matches, (2) larger animals are better at winning wrestling matches, and (3) animals with larger gular scutes are better at winning wrestling matches or provides a signal of fitness to females. So each one could have a different mechanism, such that you might have an instance where a young, small animal but with big gular scutes performs well at reproducing.

So, think about mechanisms for how different X-variables might influence Y. If those mechanisms are different, then they may have different effects on Y, and thus may be **confounding variables**.

A helpful way to do this is to build a conceptual model describing hypotheses for how all variables might be related in your system and then use that model to guide what variables you might want to measure. I recommend this to most students and collaborators I work with!

### 2. What happens between ‘full’ and ‘reduced’ models?

A full model has all of the X-variables, whereas a reduced model has a subset of X-variables. How does reducing the model influence the betas and p-values? If you reduce the model and the betas change a lot, this is an indication of collinearity with confounding variables.

* **Confounding variable – removing it changes the betas for the other variables**
* **Redundant variable – removing it does not change other betas**

(‘Change’ is a subjective term, but our usual rule of thumb would be that it caused effect to move outside of the original confidence limits.)

Conversely, we can think about this in terms of adding variables to a model.

**When adding other variables, redundant s tend to go to zero**

A good example of this is the ‘watching TV study’! It looks like watching TV has an effect when it’s all by itself in the model. But when you add the confounding variables that truly have an effect on Y, the effect of TV will probably go to zero, because it is a redundant variable. So, when you add other variables to a model, the beta for the redundant variable will tend to go to zero.

In other words, if a variable had an effect but then you added in a bunch of other variables and the effect disappeared – then it probably never had an effect to begin with, it was just a redundant variable.

**When adding other variables, confounding s will still be ‘substantial’**. They may not still be statistically significant because we have added more variables and contributed to variance inflation. However, if we look at the beta and it still looks like a biologically-significant effect, then it may be a confounding variable that lost significance due to variance inflation.

In general, betas should not change when you add additional variables to a model. If they do, you have collinearity.

This is a point in the class where things are starting to get pretty nuanced and we need to be pretty strategic and careful about how we approach our work.

### 3. between variables

This is not a rule, but rather a guideline. If you examine the correlation between two X-variables and see an extremely high value, it’s most likely that those variables are redundant. Very, very strong collinearity tends to imply redundancy.

Rule-of-thumb: high tends to suggest redundancy.

### 4. Not sure? Treat as confounding.

If you are not sure, treat it as confounding.

**Q:** Why?

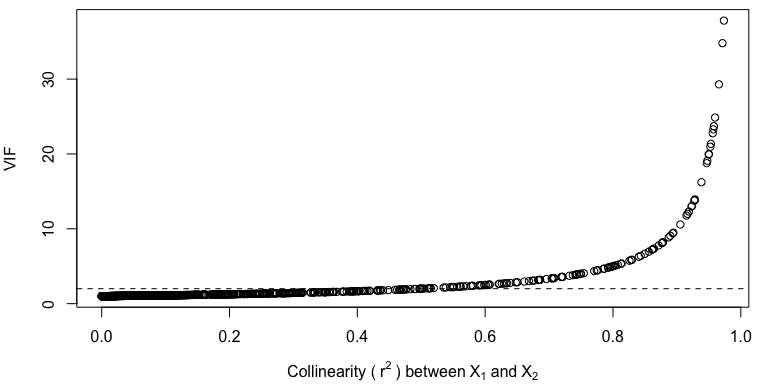
Because then you will include it in your model, account for its statistical effect, if one exists. If it’s actually a redundant variable, then there is no effect! Your parameter estimates will be unbiased and the only consequence will be a little bit of variance inflation and reduced uncertainty. This is much better than leaving out an important, confounding variable and having consequences on your betas!

General takeaway: if adding variable(s) causes the beta of a variable to change to zero, then it’s probably a redundant variable, and this should be pretty obvious.

## Simulation Exercise 2

This simulation exercise is similar as before, but now only has an effect, and is a redundant variable.

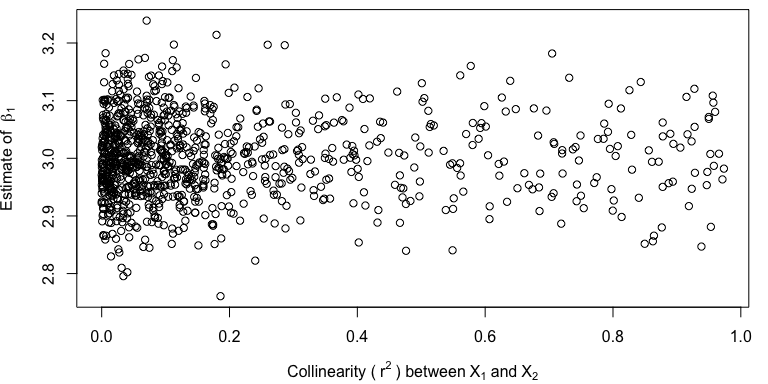
Let’s look at how correlation between and influences the Variance Inflation Factor.



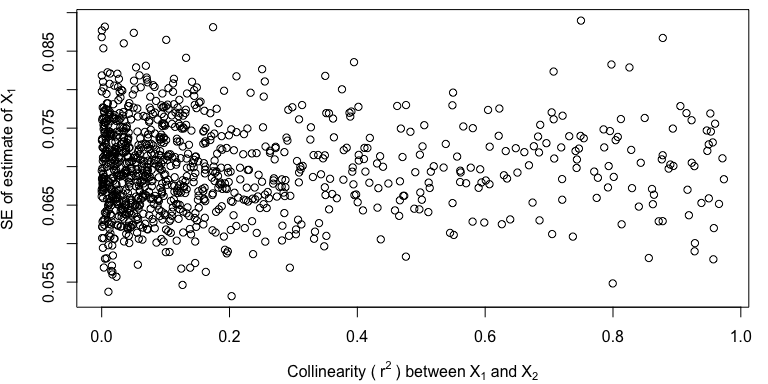
The relationship between collinearity and Variance Inflation does not change if it’s a confounding or redundant variable. Adding a redundant variable will still cause Variance Inflation, as I mentioned before.

### Simple regression model:

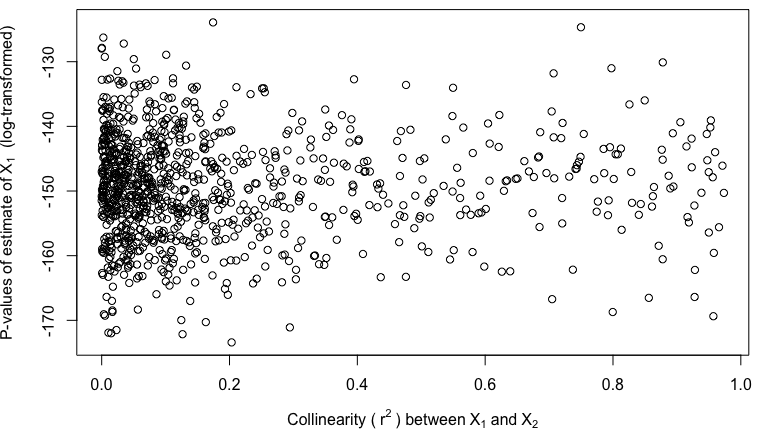
Now, we see that the existence of a redundant variable does not influence our estimation of using a simple linear model! This is good.



There is some variation in the parameter estimates due to random chance, but the estimates are unbiased along the gradation of correlation with the redundant, collinear variable.



SE estimates are also unbiased and do not change due to the existence of the redundant variable.

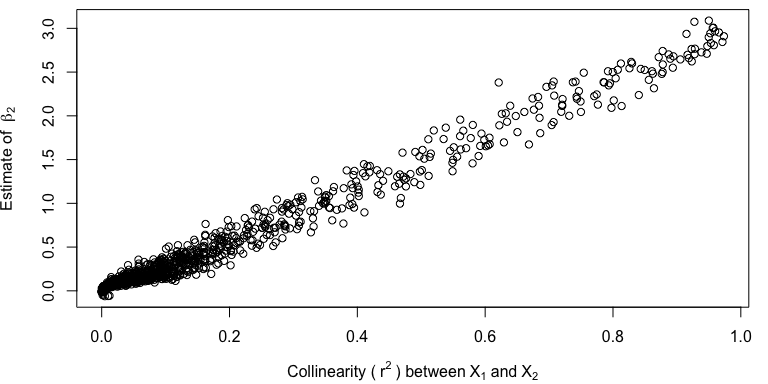


Nor are p-values influenced!

This simulation supports my previous suggestion that: redundant variables do not influence estimation, error, or p-values in our simple linear model.

### Simple linear model w/ redundant var:

What if I now run a simple linear model but with as the only X-variable? Remember, is collinear with – but does not have an effect on Y.



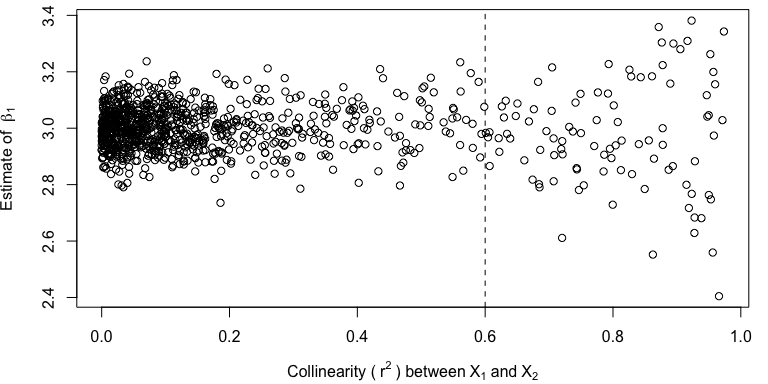
So this is interesting. When has low correlation with , the effect is measured as zero – because the two variables are not collinear, and truly does not have an effect. However, when correlation between and increases, all of the sudden the true effect of (the variable not included in the model), is now assigned to ! The stronger the relationship between X1 and X2, the more the effect of X1 is assigned to X2.

This is what is happening in the TV-mortality study. A redundant, collinear variable is being assigned the effect(s) of other true, confounding variables.

### Multi-variable model:

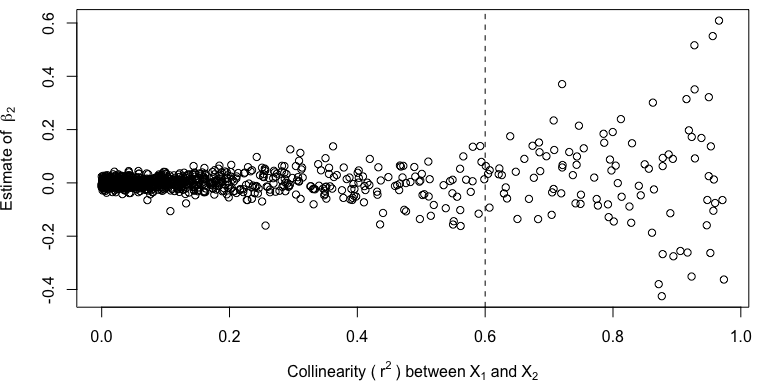
Now, what happens with the multi-variable model? Let’s include the redundant variable.

Estimates of :



Our estimate of is unbiased and close to truth! But the variance increases due to the presence of the collinear, redundant variable in the model. The cone shape suggests more uncertainty due to the variance inflation and the presence of the redundant variable. With two collinear variables in the model, R has a harder time figuring out which was has the effect.

Let’s see what it estimates for :



It estimates the effect of the redundant variable to be zero! As collinearity increases, the estimate remains unbiased and correct at around zero – although the variance inflation again causes less certainty in the estimate, on average.

## Practical guidance for examining and dealing with collinearity

These lists mostly review what I have talked about during class so far today - quite a bit of content! *Show lists on the projector.*

**Do you have collinearity in your data or system?**

1. Be careful to identify potential confounding variables prior to data collection. Use logic and try to identify all confounding variables and measure these.
2. Calculate collinearity and VIF among independent variables – before you start your analysis. High collinearity between X-variables tends to imply redundancy.
3. Pay attention to how coefficient estimates and variable significance change as variables are removed or added.

**Is a variable redundant or confounding?**

1. Think! Use logic.
2. If there is extreme collinearity, there’s likely a **redundant** variable
3. Large changes in coefficient estimates of **both variables** between full and reduced models: variables are likely **confounding**.
4. Large changes in coefficient estimates of **one variable** between the reduced and full model, and the full model estimates a variable to be close to zero: **redundant**
5. Not sure whether it’s redundant or confounding? **Assume confounding & include it.** Multi-variable regression also produces unbiased estimates (on average) regardless of the type of collinearity.

What to do with **redundant variables**?

1. Determine which variable best explains the response using P-values from regression and changes in coefficient estimates with variable addition and removal
2. Do not include redundant variable in final model (to reduce VIF)

What to do with **confounding variables**?

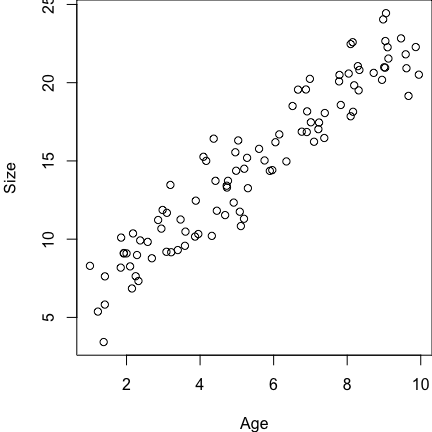
1. Sample in a manner that eliminates collinearity, which can be due to real collinearity or sampling artifact.
2. Use multi-variable regression; may have large SE if collinearity is strong.
3. Include confounding variables, even if non-significant.
4. Get more data! Decrease SE due to variance inflation.

## Analysis in R

Let’s examine some data I simulated in R. The dataset is located [here](lecture_12_dataset1.csv).

These data are similar to the Age, Sex, and Size data we simulated for last class. There is no collinearity between Age and Sex. However, now there are two additional X-variables, MotherSize and FatherSize, and both of those variables are a function of Age. All of these variables influence size, so they are all examples of confounding variables.

par(mar=c(4,4,0,0))  
  
# Read in the data  
datum <- read.csv("lecture\_12\_dataset1.csv")  
  
# Examine the data  
head(datum, 10)  
  
# Plot the data  
plot(Size ~ Age, data = datum)



It’s useful to write Truth up on the white board so that we can refer back to that while running the analysis in R…

Truth:

* Age = 1.5
* Sex = 2.5
* MotherSize = 0.2
* FatherSize = 0.2

We want to start by testing if there is collinearity among our data. The easiest way to do this is by running linear models among our X-variables.

# Run lm() for different pairs of X-variables  
results <- lm(Age ~ Sex, data = datum)  
summary(results)

##   
## Call:  
## lm(formula = Age ~ Sex, data = datum)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4.459 -2.293 -0.285 2.430 4.571   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.681 0.364 15.63 <2e-16 \*\*\*  
## SexMale -0.388 0.514 -0.75 0.45   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.57 on 98 degrees of freedom  
## Multiple R-squared: 0.00577, Adjusted R-squared: -0.00438   
## F-statistic: 0.568 on 1 and 98 DF, p-value: 0.453

The p-value is high (non-significant) and value is low, suggesting there is no collinearity. (And there isn’t, because I didn’t simulate there to be any.)

# Run lm() for different pairs of X-variables  
results <- lm(Age ~ MotherSize, data = datum)  
summary(results)

##   
## Call:  
## lm(formula = Age ~ MotherSize, data = datum)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.2585 -0.5142 0.0213 0.6388 2.1960   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.1045 0.0927 55.1 <2e-16 \*\*\*  
## MotherSize 0.8828 0.0339 26.1 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.915 on 98 degrees of freedom  
## Multiple R-squared: 0.874, Adjusted R-squared: 0.873   
## F-statistic: 679 on 1 and 98 DF, p-value: <2e-16

is now 0.874 – definitely collinearity here! (And we know this to be the case.)

And we can do this for other pair of X-variables as well, as much as we want or need to.

However, an alternative way to do this is to create a ‘correlation matrix’ using the function ‘cor()’, which calculates the correlation among all pairs of continuous X-variables your provide to the function.

* ‘cor()’ won’t work for our dataframe, because it requires all continuous variables and Sex is a categorical variable. So drop Sex and Size.
* This will run a simple ‘lm()’ between all pairs of X-variables and calculate *r* value for each pair (measure of correlation)

# Create new dataframe without categorical Sex  
datum2 <- subset(datum, select = -Sex)  
datum2 <- subset(datum2, select = -Size)  
  
# Examine new subsetted data  
head(datum2)  
  
# Correlation matrix - runs a simple 'lm()' between each pair of X-variables  
cor(datum2)

## Age Male MotherSize FatherSize  
## Age 1.00000 -0.07593 0.9349 0.93823  
## Male -0.07593 1.00000 -0.1458 -0.08094  
## MotherSize 0.93485 -0.14576 1.0000 0.86632  
## FatherSize 0.93823 -0.08094 0.8663 1.00000

[Squaring this would produce the usual metric we are familiar with.]

This is a quick-and-easy way to assess correlation between X-variables.

Another way is to run a full linear model with all of the X-variables and ask for the Variance Inflation Factors! The VIF function exists in the package ‘car’, which you will have to download to your computer one time. It has a lot of ‘dependencies’ (other necessary R packages) that will have to be downloaded…

# Download 'car' by uncommenting this next line and running it through your console  
# install.packages("car", dependencies=TRUE)

Now we can fit a linear model with all X-variables and then run the VIF function.

# Full linear model with all X-variables  
results <- lm(Size ~ Age + Sex + MotherSize + FatherSize, data = datum)  
  
# Use the VIF function from 'car'  
# using 'car::vif()' will tell R to call the function 'vif()' from the package 'car' specifically  
car::vif(results)

## Age Sex MotherSize FatherSize   
## 17.063 1.055 8.381 8.437

These are the Variance Inflation Factors for each X-variable. These effects describe how much (in *times*) the variance of each variable increased due to collinearity. So, the variance of Age was 17 times increased what it would have been in the absence of collinearity…

# Square-root this to get VIF in terms of Standard Error  
car::vif(results)^0.5

## Age Sex MotherSize FatherSize   
## 4.131 1.027 2.895 2.905

SE of Age is 4 times larger than what it would have been without collinearity. Sex is not influenced (times ~1), but the SE and 95% CI are 4 or 3 times larger than what would happen without collinearity.

Note: the VIF is the variance inflation due to collinearity of all other variables combined. So collinearity with multiple other X-variables can contribute to variance inflation of each variable.

# Simple linear model of Size ~ Age  
results <- lm(Size ~ Age, data = datum)  
summary(results)

##   
## Call:  
## lm(formula = Size ~ Age, data = datum)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.740 -1.236 -0.033 0.945 3.683   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.5977 0.3829 12.0 <2e-16 \*\*\*  
## Age 1.8627 0.0633 29.4 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.61 on 98 degrees of freedom  
## Multiple R-squared: 0.898, Adjusted R-squared: 0.897   
## F-statistic: 867 on 1 and 98 DF, p-value: <2e-16

This estimate is wrong – not close to truth – and the confidence intervals are small.The effect size and confidence intervals were misleading because we did not account for confounding variables!

# Simple linear model of Size ~ MotherSize  
results <- lm(Size ~ MotherSize, data = datum)  
summary(results)

##   
## Call:  
## lm(formula = Size ~ MotherSize, data = datum)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.623 -1.825 -0.075 1.803 5.196   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 14.1121 0.2451 57.6 <2e-16 \*\*\*  
## MotherSize 1.6304 0.0895 18.2 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.42 on 98 degrees of freedom  
## Multiple R-squared: 0.772, Adjusted R-squared: 0.77   
## F-statistic: 332 on 1 and 98 DF, p-value: <2e-16

Again, same problem here.

# Simple linear model of Size ~ MotherSize  
results <- lm(Size ~ Age + MotherSize, data = datum)  
summary(results)

##   
## Call:  
## lm(formula = Size ~ Age + MotherSize, data = datum)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.733 -1.246 -0.015 0.943 3.790   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.044 0.927 4.36 3.2e-05 \*\*\*  
## Age 1.972 0.179 11.04 < 2e-16 \*\*\*  
## MotherSize -0.111 0.169 -0.66 0.51   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.62 on 97 degrees of freedom  
## Multiple R-squared: 0.899, Adjusted R-squared: 0.897   
## F-statistic: 431 on 2 and 97 DF, p-value: <2e-16

Truth for Age is 1.5, so this is closer – but the 95% CI around this estimate of age still do not overlap truth. And MotherSize is not significant, which it should be.

Last, the full model with all of the confounding variables:

# Simple linear model of Size ~ MotherSize  
results <- lm(Size ~ Age + Sex + MotherSize + FatherSize, data = datum)  
summary(results)

##   
## Call:  
## lm(formula = Size ~ Age + Sex + MotherSize + FatherSize, data = datum)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.3964 -0.7442 -0.0976 0.8520 2.5024   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.061 0.690 5.89 5.9e-08 \*\*\*  
## Age 1.486 0.195 7.63 1.8e-11 \*\*\*  
## SexMale 2.187 0.247 8.85 4.7e-14 \*\*\*  
## MotherSize 0.151 0.129 1.18 0.242   
## FatherSize 0.263 0.130 2.02 0.046 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.2 on 95 degrees of freedom  
## Multiple R-squared: 0.945, Adjusted R-squared: 0.943   
## F-statistic: 411 on 4 and 95 DF, p-value: <2e-16

We now get unbiased estimates that overlap with truth, and all estimates are ~significant. MotherSize is narrowly insigificant, but it’s CI overlap truth. And the p-value for the whole model is highly significant.

## Summary

* If you add variables to a model, p-values should go down, because you have added parameters, more error is explained, and you have **reduced swamping**.
* Multivariable models are generally useful, because they eliminate issues related to swamping & collinearity.