FES 524: Natural Resources Data Analysis

Reading 5.2

**Contents**

[Other readings to do before class 1](#_Toc63068464)

[Continuous explanatory variables 2](#_Toc63068465)

[Intercepts and slopes 2](#_Toc63068466)

[Linear relationship 3](#_Toc63068467)

[Linearity and scale 4](#_Toc63068468)

[Confidence interval around relationship 5](#_Toc63068469)

[Spurious correlations 6](#_Toc63068470)

[Physical units of different sizes 7](#_Toc63068471)

[Working with continuous and categorical variables 7](#_Toc63068472)

[Parallel lines 7](#_Toc63068473)

[Separate lines 8](#_Toc63068474)

[Continuous by continuous interactions 9](#_Toc63068475)

[Analysis with many variables 10](#_Toc63068476)

[Hard thinking 10](#_Toc63068477)

[How many variables can I use? 10](#_Toc63068478)

[Model complexity 11](#_Toc63068479)

[Approach to explore the data 11](#_Toc63068480)

[Interpretation and added variable plots 13](#_Toc63068481)

[Model selection 15](#_Toc63068482)

[Known problems with model selection 15](#_Toc63068483)

[Explain vs predict 16](#_Toc63068484)

[Competing models 16](#_Toc63068485)

Class 5.2 will be about continuous explanatory variables. This is our main chance to discuss these this quarter.

This ends up being a fairly long reading because we are lightly touching on some very big topics. You may end up having more questions than usual to discuss in class. The last two sections, in particular, are to give you a starting place to jump off from if you have many explanatory variables or are considering doing model selection in your own work.

# Other readings to do before class

Make sure you’ve taken the time to look at the two GIF files that go along with the section in this reading about confidence intervals, Wilke\_CI\_gif\_bluejays.gif and Ross\_CI\_GAM\_animation.gif.

Read the description of the study that assignment 5 is based on in Handout 5.2 and come prepared to discuss all the listed questions.

# Continuous explanatory variables

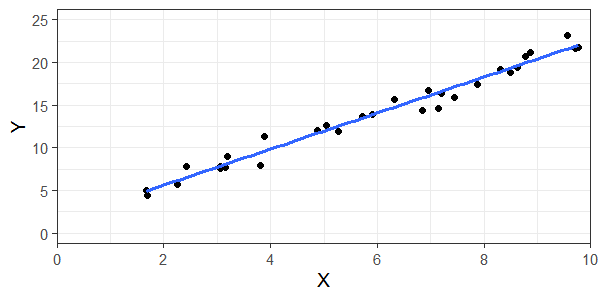
Most of this quarter we are going to continue to work with categorical explanatory variables for assignments. Many studies, however, involve continuous explanatory variables. You may have heard analyses working with continuous explanatory variables referred to as *regression*.

Regression is a special term used to indicate a linear model or a generalized linear model with continuous explanatory variables. This term also may be used when the model contains a mix of continuous and categorical explanatory variables. All of these models fall under the umbrella of linear models, though, and that is the term I will use in this class.

## Intercepts and slopes

Remember that when looking at relationships between continuous variables we are interested in *intercepts* and *slopes*.

Below is a plot showing a fitted line between some variable X and mean Y.



The statistical model is:

Yt = β0 + β1Xt + εt

εt ~ N(0, σ2), εt and εt’ are independent

Intercept

The *intercept*, represented in the statistical model by β0, is the estimate of mean Y when X is 0. There are many instances where the intercept is not interesting, but we must have an intercept in order to draw a line.

Slope

The *slope*, represented in the statistical model by β1, is the estimated change in mean Y from some change in X. We always need to be explicit about what change in X we are interested in. By default, software will give estimates of the slope for a 1-unit change in X.

We will talk more about intercepts and slopes and calculate them for the example shown in the above plot in class.

## Linear relationship

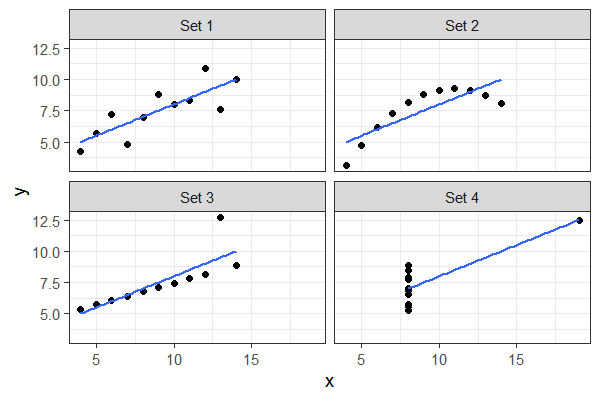
The assumption that the relationship between two variables is linear is a strong assumption. The investigator will need to consider how reasonable that assumption is before they are at the analysis stage. This could be based off prior research or other scientific expertise.

There are plenty of alternative models that don’t assume linear relationships in modern statistics. This includes generalized additive models (GAM’s), cubic splines, or polynomials.

I often see investigators choosing alternative models only after exploring the observed data. In particular, I see investigators adding in higher order polynomials because they saw a pattern after data collection. While data exploration is a vital part of analysis, deciding on a relationship only after looking at the data is poor statistical practice. The researcher is spending degrees of freedom but not accounting for it when making inference. If you are doing confirmatory work, the observed data shouldn’t dictate how you model a relationship because you already have a good idea of what you expect the relationship to be based on scientific theories. If you see something surprising in the observed data, you’ll likely want to add some exploratory analyses to discuss the unusual pattern while also reporting on your originally planned analysis. As always, if doing exploratory research you have more freedom.

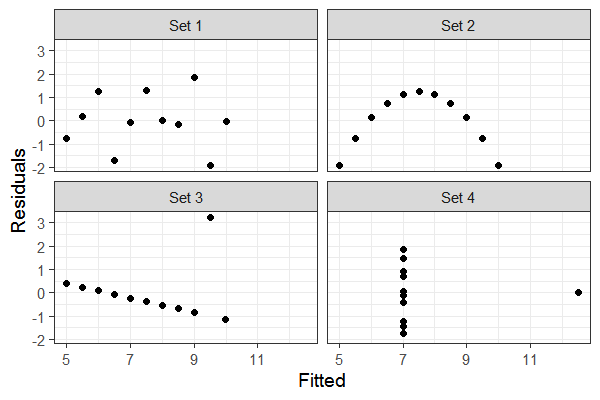
We will touch on spending degrees of freedom and allowing for more complex relationships between variables later, although we will not go into it in depth.

Below is Anscombe’s famous quartet. The four datasets show identical linear fits even though the underlying shapes of the relationships are very different.



We would be able to see issues of model fit in residuals plots. Looking for patterns that indicate lack of fit is one of the reasons we make residuals vs fitted and residual vs variable plots.

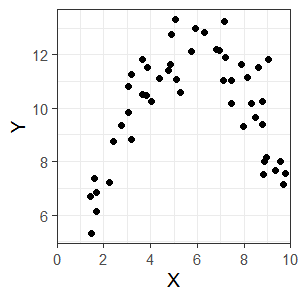
Here are the residuals plots from linear models fitted to each dataset in Anscombe’s quartet. You can see clear issues in all plots other than “Set 1”.



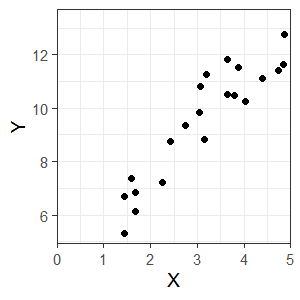
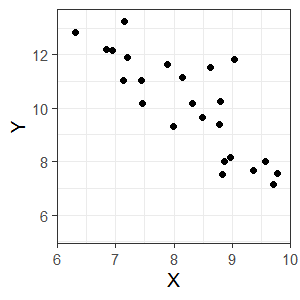
## Linearity and scale

The assumption of linearity and how reasonable it is can depend on the scale of the relationship.

For example, below is a scatterplot between two variables, X and Y. The measurements of Y were taken across a wide range of the X variable and there is clearly a non-linear relationship between the two variables.



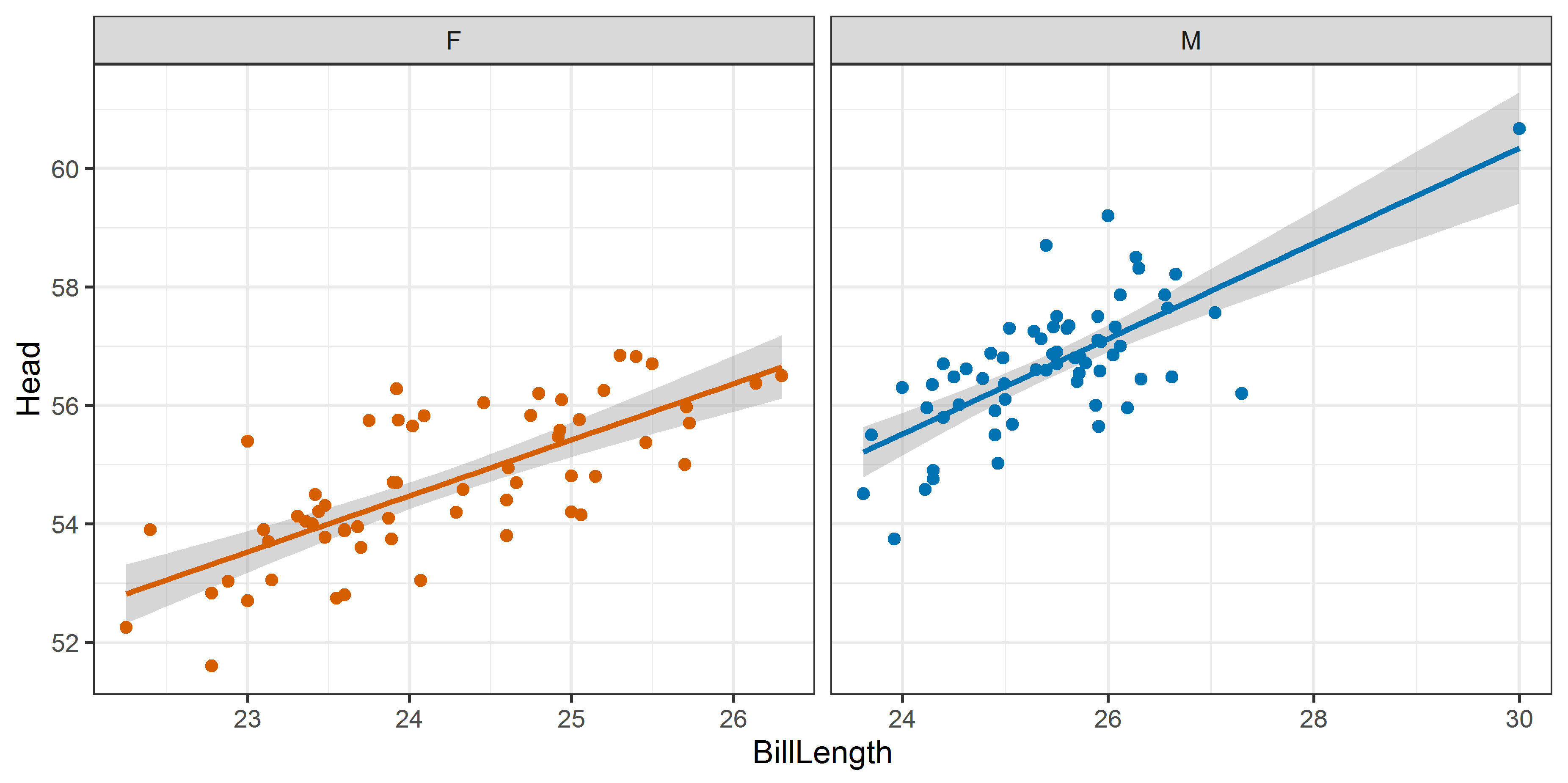
However, the assumption of linearity could still be reasonable for a narrower range of X. Things are often reasonably linear at small scales even if we “know” they won’t be at larger scales. The assumption of a linear relationship looks pretty reasonable if we measured X only from 0 to 5 or only from 6 to 10.

## Confidence interval around relationship

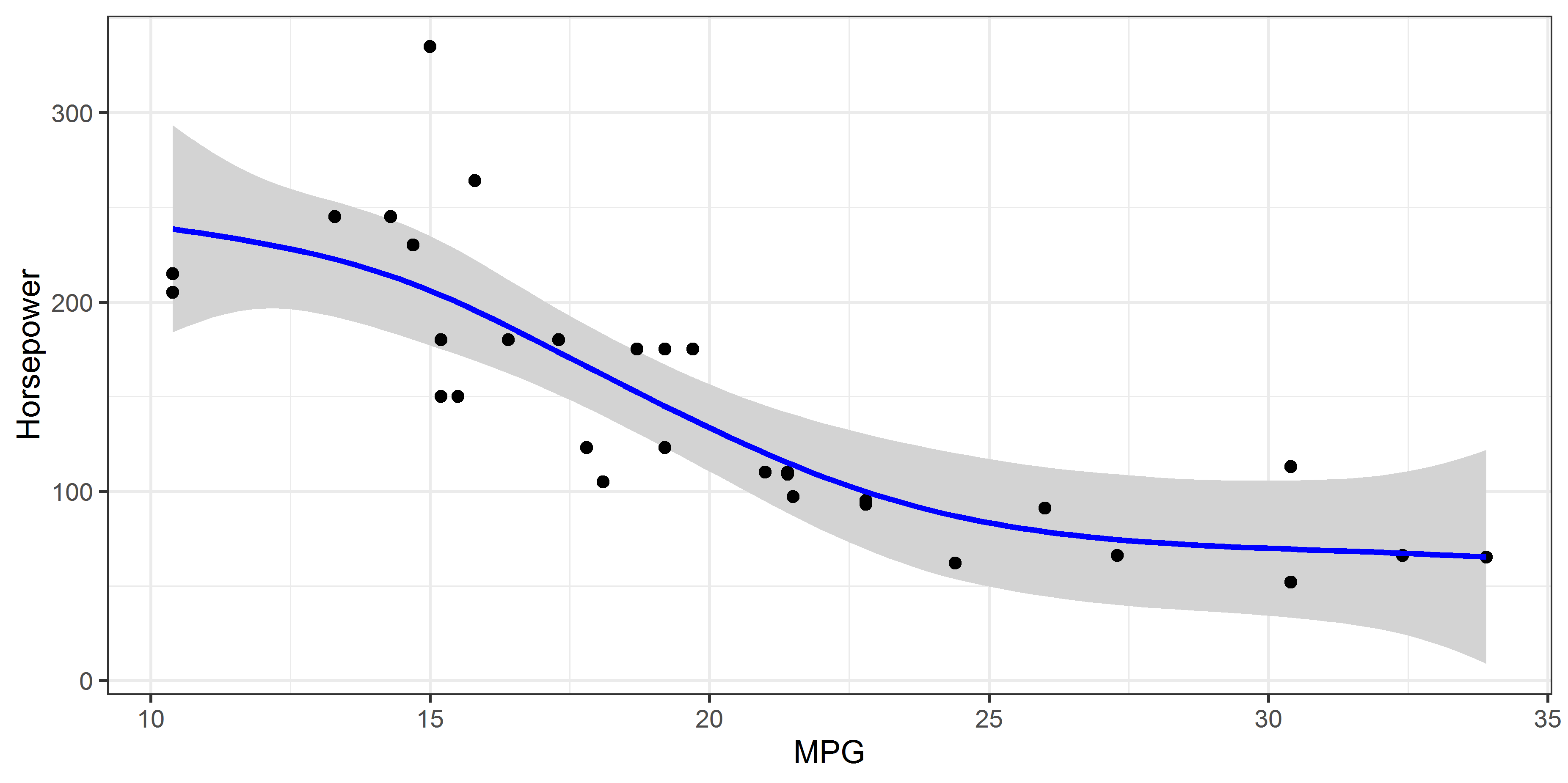
The way we think about confidence intervals around a fitted line is a little bit different than how we think about confidence intervals around an estimated mean or difference in means, and I’ve found people don’t always know how to interpret them. When we draw CI around a fitted line we should be thinking about the plausible values for the “true” *relationship*.

Here is an example of fitted lines with confidence intervals for two groups, made from data taken from a GitHub repository (<https://github.com/wilkelab/ungeviz>).



It is easiest to see the possible relationships these confidence intervals encompass through simulation. Make sure you look at the GIF version of the above plot on the GitHub repository, which I saved as Wilke\_CI\_gif\_bluejays.gif and provided to you on Canvas.

The possible relationship between two variables can be nonlinear. Below is a scatterplot showing a nonlinear fitted relationship (fit using a GAM) from a gist on GitHub: <https://gist.github.com/noamross/8bf1fc5b2f629b3a7e1eb6b4572e8388>.

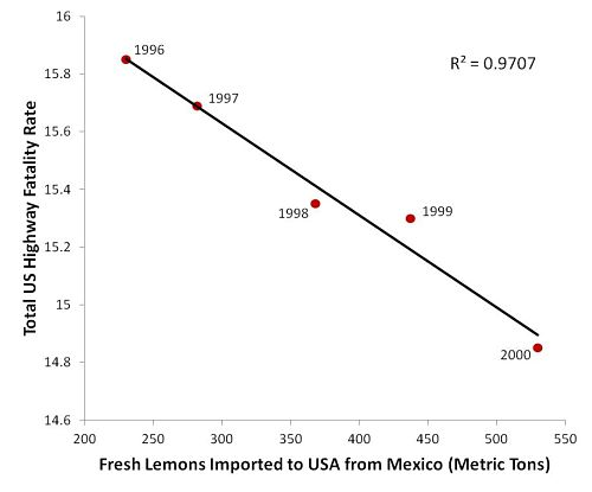


Make sure you look at the animated version I provided to you at Ross\_CI\_GAM\_animation.gif and think about what it’s showing you.

## Spurious correlations

A saying that often comes up when talking about fitted relationships is “correlation is not causation”. There are many examples out there of spurious correlations. In fact, there are entire websites that were set up to show some of these spurious correlations.

For example, here is a graph from <https://channels.theinnovationenterprise.com/articles/hilarious-graphs-and-pirates-prove-that-correlation-is-not-causation> that shows a strong relationship between the US fatality rate and the amount of lemons from Mexico.



Since it doesn’t make sense that increasing lemon imports actually leads to a decrease in highway fatality rates, this is an example of a spurious correlation. It is likely that both variables are actually related to something else that happened in the same time frame.

While there are many spurious correlations out there, when working in research we should understand that correlation is not *necessarily* causation. Remember that causation can be justified based on a strong scientific theory or other information even for observational data.

## Physical units of different sizes

We can collect continuous data from units of different sizes just like we did for factors in the motivating example this week. You would still use some sort of mixed model for analysis of such data, using the physical units as random effects.

For example, say you were measuring some response variable in stream reaches and you measured this variable at multiple reaches in a variety of different watersheds in a single year. Reaches are *nested* in watersheds.

In this hypothetical study, you collected information on watershed area, measured at the watershed level, and stream gradient, measured at the reach level, because you are interested in estimating a relationship between each of these variables and the mean response variable. For the analysis you would potentially use a linear mixed model with watershed area and stream gradient as fixed effects and watershed and reach as random effects. Reaches would be represented by the residual error term, since they are the level that the response variable was measured at.

# Working with continuous and categorical variables

You saw an example of working with continuous and categorical variables in the ST 512 prerequisite. This is the bat example in the Statistical Sleuth, which you may want to review.

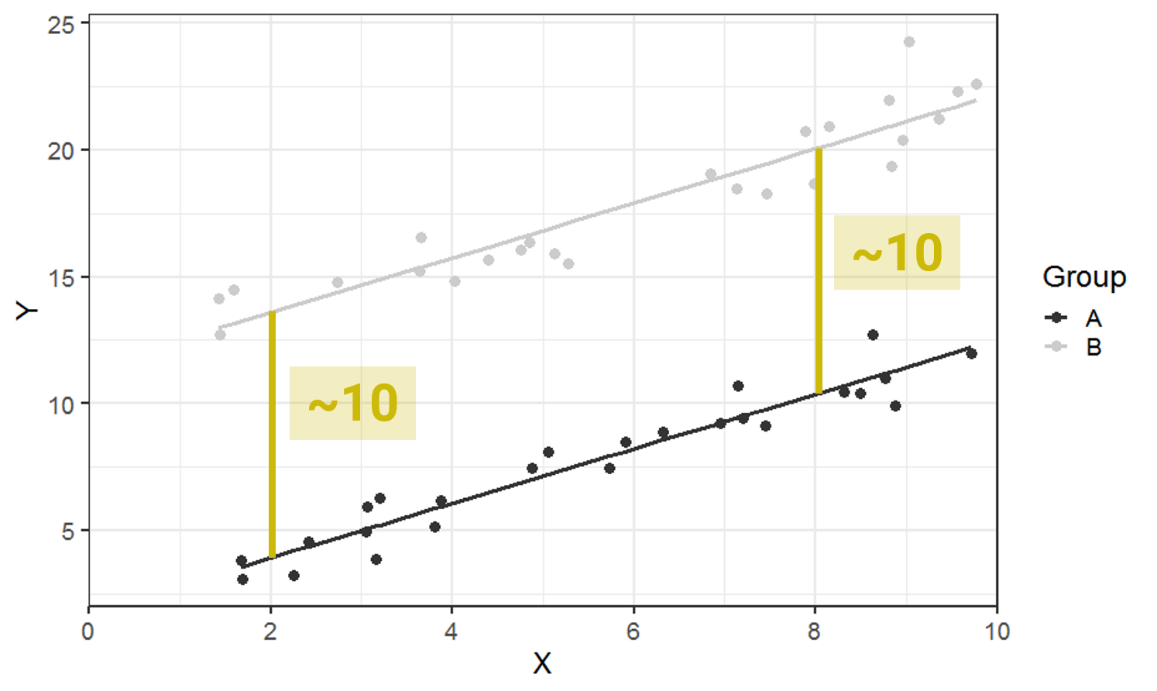
As always, the model you fit when working with continuous and categorical variables depends on your research question.

## Parallel lines

One possible model of interest is the *parallel lines* model. In this model, the primary interest is estimating a difference in mean response among groups, after accounting for some continuous variable.

Since the lines are parallel, the differences in mean response among groups does not depend on the value of the continuous variable. The factor and the continuous explanatory variable don’t *interact*.

Below is a picture of a parallel lines model from an analysis of two groups with a continuous variable X. The difference in mean Y among the two group is ~10 across the entire range of X.

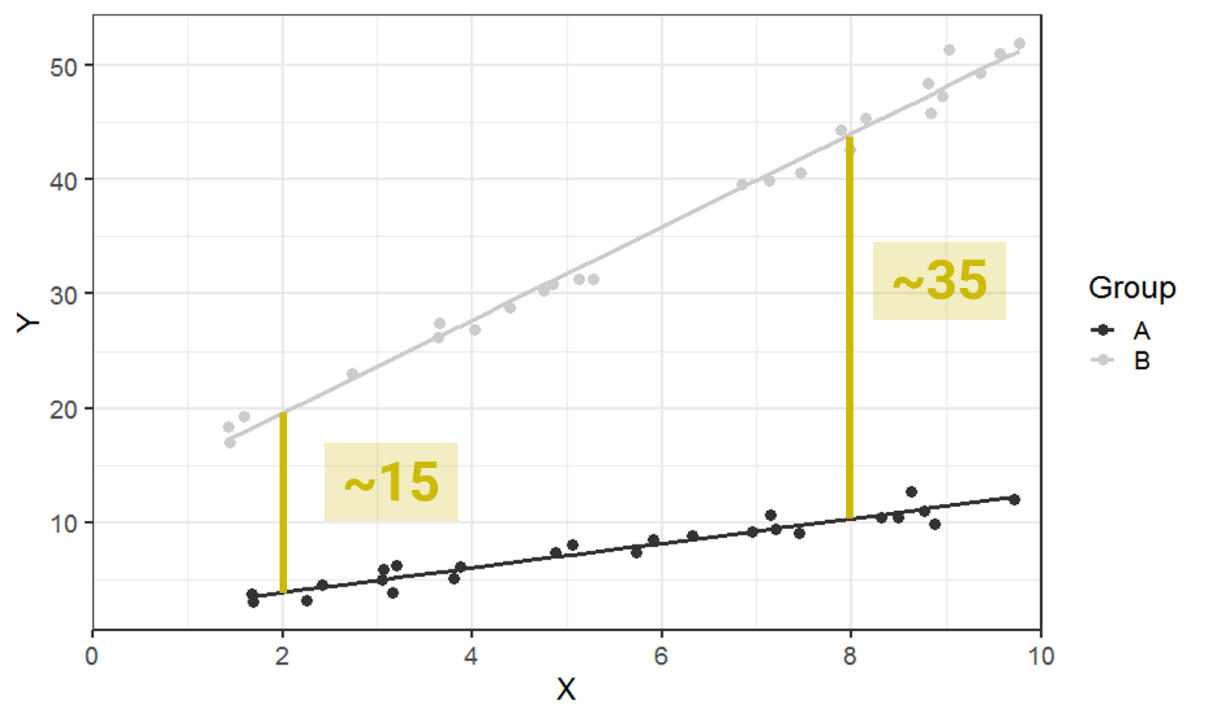


The parallel lines analysis is what people used to mean when they use the term ANCOVA (which stands for *analysis of covariance*). Ideally you’ll be able to skip using this term all together. However, be aware that some people use this term today to mean a linear model with both continuous and categorical variables in it.

## Separate lines

The *separate lines* model is one with an interaction between the categorical and continuous explanatory variable. Remember we’ve defined an interaction as when “the effect of one variable depends on the value of another variable”.

An interaction between a categorical and continuous variable indicates we are interested in differences among the *slopes* for different groups. The difference in mean response now differs for different values of the continuous explanatory variable, as shown below in a two-group example. The difference in mean Y between groups is ~15 when X is 2 but is ~35 when X is 8.



Using an interaction indicates an *a priori* interest in differences in slopes among groups. When that is the case, the interaction should be kept in the model regardless of any statistical results. You will report an estimate of the difference in slopes and likely make a graph like the one above to graphically display the results (although you may add confidence ribbons). You should not take the interaction out and treat it as if the interaction term is exactly 0 if your research question was specifically about a different relationship for different groups.

As discussed last week, if there is some interest in estimating the overall relationship we can get the “average” of slopes. However, calculating the standard error for this overall relationship may be a problem. Unlike when working with factors, the simplest thing to do is to remove the interaction to talk about an overall relationship. Do this with great caution, being really honest with yourself about the goals of your research, your *a priori* assumptions and whether you are doing confirmatory or exploratory research. Another option is bootstrap confidence intervals for the estimated overall slope with the interaction still in the model. The function emtrends() from package emmeans will also do overall estimates for the slope with CI in the presence of an interaction, although I haven’t looked out how the confidence intervals are calculated.

## Continuous by continuous interactions

One difficult topic, which we aren’t going to talk much about, is how to work with continuous by continuous interactions. Such an interaction would indicate that the relationship of one continuous X variable and mean Y is affected by the value of a second continuous X variable. The result would be a *plane* (drawn in 3D) that has curvature.

I’ve been seeing more and more researchers considering such interactions in their models. Certainly in complicated systems we think these interactions would exist. Unfortunately, a lot of data is needed to estimate a continuous by continuous interaction well. If you are planning on this you should have the full range of one variable for ~every value of the second variable. Essentially, you want the two continuous variables to be *crossed*. Without careful planning during the design phase to make sure you get well crossed variables or if you end working with very small samples, continuous by continuous interactions are not realistic to fit even when the concept of them is realistic. Note the computer won’t stop you from fitting them, but results can have some of the same problems we discussed for lower power studies in week 3.

# Analysis with many variables

Some researchers end up in a situation where they have many, many explanatory variables to contend with. This is a difficult problem to tackle.

First, some language: Models with multiple explanatory variables should be referred to as *multivariable* models and not *multivariate* models. A multivariate model is one that has multiple response variables, not multiple explanatory variables.

## Hard thinking

There isn’t some easy statistical answer to what to do when we have many variables. Even if you’ve heard of some amazing new tool that supposedly requires absolutely no work on your part to use, you should take such tools with a grain of salt. What we’ve learned over and over in statistics is that there’s no such thing as a free lunch.

The best thing you can do when working with many variables is to spend a lot of time thinking hard about the problem. This involves doing a lot of work figuring out exactly which variables are available and which are actually interesting to your research. You generally will need to focus on the most important variables, not all possible variables that you can calculate.

As always, there is more leeway to working with many variables and fitting many different models when doing exploratory work. However, I’ve started seeing a situation where investigators are doing similar research in a new place but they still claim the work is totally exploratory and they have absolutely no information on which variables could be important in an analysis. At some point the investigators will need to start using information from previous research and stop treating a study replication in a new place as only exploratory.

If I haven’t made it clear so far this quarter, make sure you understand that confirmatory work, based off previous research and scientific theory, and exploratory work can be done on the same data. There’s no need to limit yourself; just be honest about what is confirmatory that you can fit into the greater research field you are in and what is exploratory and opens up new research ideas.

## How many variables can I use?

See Harrell 2015 “Regression Modeling Strategies” book Chapter 4 for more coverage on this topic and as a citation for where some of the numbers below come from. His course notes are available online at <http://hbiostat.org/doc/rms.pdf>, although they don’t go into a lot of detail.

The discussion of the number of variables and model complexity is based off the assumption you are fitting some parametric model like a linear or additive model.

The number of observations you need to estimate each parameter is roughly about 15. This number is for pretty noisy studies, which is likely standard in a lot of ecological studies but may not be for more controlled studies. In a lot of ecological studies the *signal to noise* ratio is usually relatively low, meaning there is a lot of variability. If you have high signal to noise ratio you can likely justify having fewer samples per parameter estimated.

Writing this mathematically, we would say that p < m/15, where m = sample size and p = # of parameters in the *fullest model* or p = # *candidate* *parameters* in model selection approach. It is not the number of parameters in some final model (if you did model selection).

To complicate matters, you need to be able to figure out how to calculate the limiting sample size. For continuous data, this is relatively easy: it’s the number of rows in the dataset. But things aren’t as simple for discrete data, which contain less information than continuous data.

Below is information taken from Table 4.2 in the Regression Modeling Strategies book for a few response variables. You can see that for binary data, you calculate your limiting sample size by taking the number of observations in the smallest group (either 0 or 1). This is going to be much smaller than the number of rows in the dataset.

**Type of Response Variable Limiting Sample Size (m)**

Continuous *n* (total sample size)

Binary min(*n*1*, n*2)

Failure (survival) time number of failures

## Model complexity

You can now see that there is a limit in the number of parameters we can estimate based on the sample size. This means we only have a certain number of degrees of freedom we can use to estimate parameters. Part of the hard work when we have many variables is to figure out how we want to spend those degrees of freedom *before we collect any data and fit models*.

One thing you will need to decide is the relative of importance of different variables. For important variables (i.e., ones you think have a large effect) with unknown relationships with the response variable you will likely want to “spend” more degrees of freedom. For example, you might want to allow for nonlinear relationships for certain important variables where you have very little *a priori* information on what the relationship will look like. Depending on sample sizes, you can often justify spending no more than 3-5 degrees of freedom to allow for a nonlinear relationship. For less important variables or variables that are known to have simpler relationships you can spend fewer degrees of freedom and use linear relationships for those variables.

The extra degrees of freedom for nonlinear relationships count in your estimate of p, which is why the first step above was to figure out how many parameters you can justifiably estimate in one model. You first figure out how many parameters you can estimate in total and then you figure out how you want to divvy up that number.

## Approach to explore the data

Once you’ve decided on a reasonable number of candidate variables and have collected data, you need to explore your data. This is not to look for possible relationships between your response variable and your many explanatory variables. That is a form of data snooping. Instead, use the data exploration to get a better handle on what the explanatory variables look like in your particular dataset.

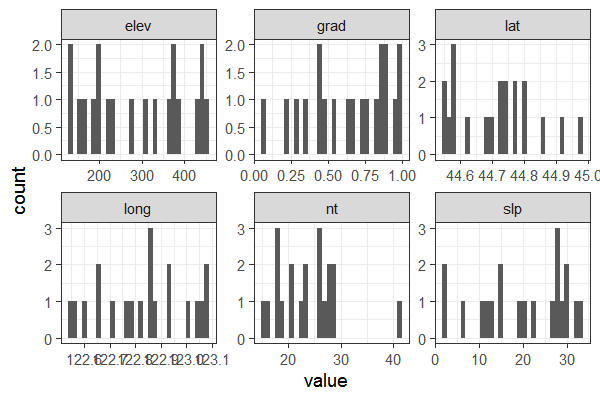
First, you will summarize and plot each explanatory variable on its own. You will look at the range of each variable to make sure the range is large enough to lead to a practically meaningful change in the response variable. Sometimes even if we “know” a variable should matter, it turns out we didn’t sample widely enough to get a range of values where we would expect the response variable to change.

Here’s a couple of examples.

First, I once worked with a student who had a question about how elevation was related to the presence of a certain tree species. When exploring that data, though, it turned out the elevation range in the sample was too small to expect a change in presence and the student ended up not doing the analysis. This decision was based on the expert knowledge of the student from past research and not some arbitrary statistical measure.

Second, in the past I worked on an analysis where it was well established that the amount of sedimentary rock would matter to the response variable. In the actual sample, though, all except two observations were 100% sedimentary rock. That variable simply didn’t have enough information in it in the sample to be useful. This was noted, since it can be considered an issue of scope of inference, and the variable was dropped from consideration for that particular analysis.

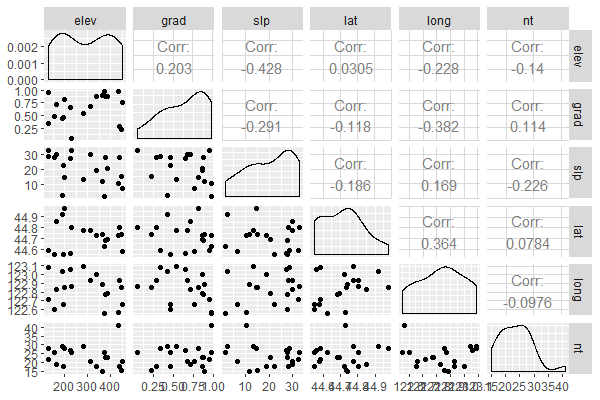
Below is an example of looking at histograms for six explanatory variables for an invented dataset. For each one I would examine the range of the data as well as any patterns of where I have observations for that variable. This looks pretty good, although I’d likely want to look more closely at the single value of 40 in “nt”.



After looking at each explanatory variable by itself, it’s time to look at simple relationships among the explanatory variables to think about correlations. For example, maybe two variables essentially measure the same thing and you’ll decide to consider only one of them in analysis.

Again, this step does not involve the response variable in any way. This is purely thinking about the explanatory variables.

Here is a plot of each variable in my hypothetical example in a scatterplot matrix made with package GGally. None of these variables are highly correlated.



How much correlation is too much correlation is not something that has been established. By the time two variables are 95% correlated it is pretty easy to justify using just one of the variables or combining them in some way. For less correlation things aren’t as clear; certainly a variable that is 75% correlated with another variable can still provide useful additional information even though standard errors could get inflated. Part of the decision on what to do in this case may depend on if the goal is prediction or estimation and how many degrees of freedom you have to spend on estimating parameters.

When you need to explore many variables things can get pretty time consuming. This is a point where you may want to learn how to automate the process in R. To get an idea of how you might approach such automation see the blog post <https://aosmith.rbind.io/2018/08/20/automating-exploratory-plots/>.

All of the work in this section is done “blinded to Y”, so what relationships of variables to Y look like in the observed data aren’t influencing any of your decisions. It may be that even with all the hard thinking you did it turned out some variables were redundant or lacked variability and needed to be removed from the analysis. This is an OK thing to do. What’s not OK is to look at plots of variables vs Y and then decide to no longer consider them based on what you saw there.

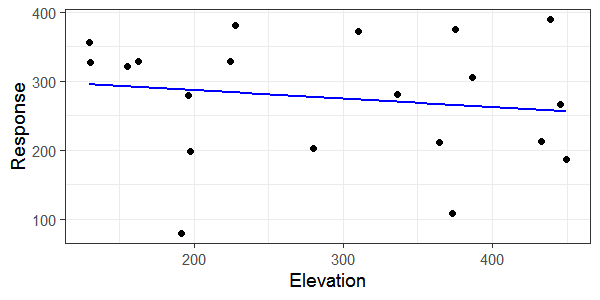
## Interpretation and added variable plots

Interpretation from multivariable models must be done carefully. Effects are interpreted *after accounting for other variables in the model*. This is different than a relationship without accounting for other variables.

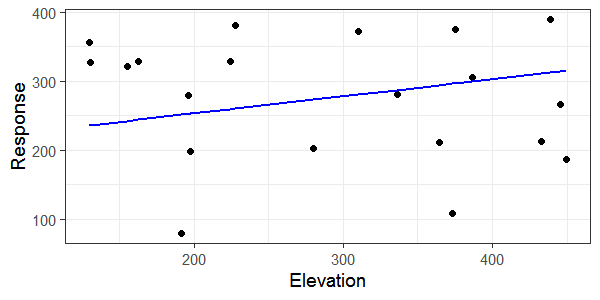
In order to display your results graphically you will likely want to make what is called an *added variable plot*, where you estimate the relationship between a variable and the mean response with all other variables fixed to some value. These are also known as *partial regression plots*.

Below I’ll go through an example of why we need added variable plots when showing fitted lines from multivariable models.

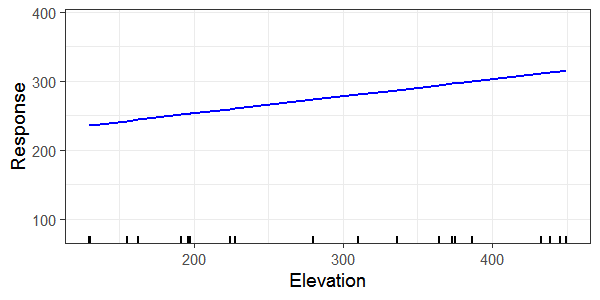
Say we had a model with two explanatory variables in it, elevation and slope. The observed relationship in the raw data between elevation and the mean response variable is slightly negative.



The estimated relationship between elevation and mean response after accounting for the effect of slope is positive. Below is an added variable plot of the relationship plotted on top of the original data. The slope variable was fixed to its mean value to make this plot.



The problem with showing the estimated relationship from a multivariable model on top of the raw data is it can look like the line doesn’t fit right. You can see this in the plot above. This is because the raw data points haven’t “accounted for” the other variable in the model. It can be confusing for your reader to include the raw data points along with a fitted line from multivariable models. This is a good case for a using rug plot to show where you have values of elevation but not show the raw data for the response variable.



If you need to make many added variable plots from complex multivariable models you may want to automate the process. See <https://aosmith.rbind.io/2018/01/31/added-variable-plots/>.

# Model selection

The last topic we are going to touch on is model selection. Unlike the impression you may have been given by the literature in your field, model selection is not a standard part of fitting a model. In many cases, we can fit a single model and report results from it. There is no reason to be taking things out of the model. Taking an effect out of the model amounts to setting effects to exactly 0; this is a huge assumption and one you should think about seriously if you hypothesized there would be some relationship.

## Known problems with model selection

There are many known problems with model selection, whether done using p-values or AIC or using some other criterion. A few of the most discussed ones are listed below. Notice that the 9th is that using model selection lets us skip the hard thinking step that is so important.

<https://www.stata.com/support/faqs/statistics/stepwise-regression-problems/>

<https://stats.stackexchange.com/questions/20836/algorithms-for-automatic-model-selection>

1. It yields R-squared values that are badly biased to be high.
2. The *F* and chi-squared tests quoted next to each variable on the printout do not have the claimed distribution.
3. The method yields confidence intervals for effects and predicted values that are falsely narrow; see Altman and Andersen (1989).
4. It yields *p*-values that do not have the proper meaning, and the proper correction for them is a difficult problem.
5. It gives biased regression coefficients that need shrinkage (the coefficients for remaining variables are too large; see Tibshirani [1996]).
6. It has severe problems in the presence of collinearity.
7. It is based on methods (e.g., *F* tests for nested models) that were intended to be used to test prespecified hypotheses.
8. Increasing the sample size does not help very much; see Derksen and Keselman (1992).
9. It allows us to not think about the problem.
10. It uses a lot of paper.

## Explain vs predict

Whether model selection is something you should be doing at least partially depends on the goal of your analysis. It is more likely we can justify some model selection for predictive models.

In addition, if you work is totally exploratory you may be able to justify using some model selection methods. However, you will need to be cautious about how you interpret the results. Remember that statistics alone can’t discover ecological truths. Any model selection where you report only some “final” model needs to be accompanied by some analysis of the uncertainty for the variables in that model. This will usually involve some sort of model internal validation, most often based on bootstrapping or cross-validation techniques. For example, in Regression Modeling Strategies chapter 5 researchers wanted to rank the “most important” predictors in a model and so uncertainty in ranks was captured via bootstrapping: http://hbiostat.org/doc/rms.pdf#section.5.4.

Fieberg and Johnson’s 2015 paper “MMI: Multimodel inference or models with management implications?” offers a good example of competing goals in modeling and different ways to tackle them. You may find the other topics that the paper hits on interesting, as well. You can find the paper here, <https://wildlife.onlinelibrary.wiley.com/doi/full/10.1002/jwmg.894>.

If your goal is to make predictions you may find the book “Introduction to Statistical Learning” (ISL) by James et al. a useful place to start, since predictive modeling is very different than the modeling we are doing in this class. ISL is written for applied scientists and available as a PDF online, <http://faculty.marshall.usc.edu/gareth-james/>.

## Competing models

Finally, another approach that generally falls under the “model selection” header is the Burnham and Anderson style approach of competing models and making multimodel inference (MMI).

In multimodel inference you carefully define different models to represent different hypotheses. This does not mean you blindly fit all possible models; instead each model means something practically important in your field.

You can see an example of this approach in Burnham et al. 2011, “AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons”, which you should read if you are considering using this approach. The paper also goes through some common issues with how people use MMI that is a must-read for practitioners. You can find the paper at <http://www.ericlwalters.org/Burnham_etal_2011.pdf>. Note issues with some types of model averaging, though (not required in MMI), in Cade’s 2015 “Model averaging and muddled multimodel inferences” (<https://esajournals.onlinelibrary.wiley.com/doi/abs/10.1890/14-1639.1>).

Burnham and Anderson’s 2002 book contains a lot of great information if you want to use MMI. One of the good things about it is how much they focus on the hard thinking part of this analysis. A lot of up-front work is needed to even attempt such an approach. I believe you can find this book online, but it’s also available at the OSU library.