FES 524: Natural Resources Data Analysis

Reading 5.2: Continuous explanatory variables and variable selection

Contents

1 Continuous explanatory variables

Most of this quarter we are going to continue to work with categorical explanatory variables for assignments, but you will also likely use continuous explanatory variables frequently in your work. 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.

1.1 Linear relationships

Recall that the statistical model for a simple linear regression is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where y_i is the observed response or *dependent variable* for observation i, β_0 is the *intercept*, or the mean response when $x = 0$, and β_1 is the *slope*, or the expected change in the mean of the response with a one-unit increase in x. As usual, we assume $\epsilon_1, \epsilon_2, ..., \epsilon_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. In other words, we are just modeling a line on the (x, y) plane (remember the ol' rise over run stuff?). The line falls along the mean of the response variable Y, as in Figure 1.

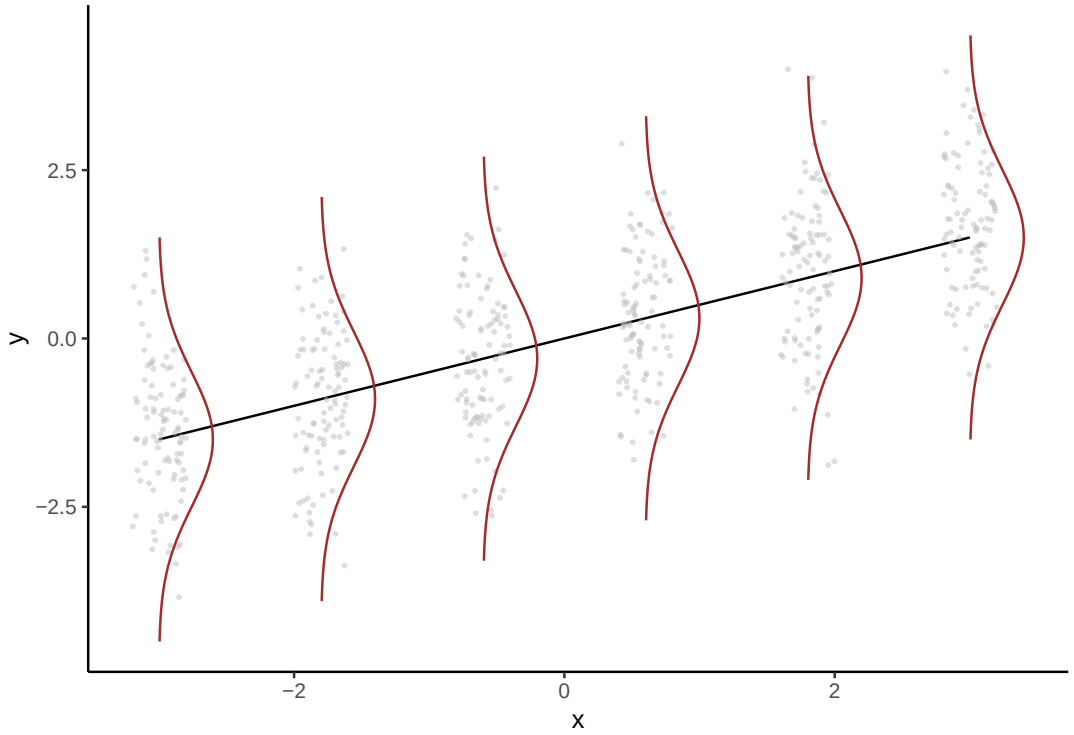

Figure 1: Conceptual drawing of linear regression. The line models the mean response, while the errors are normally distributed about the line with constant variance.

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. Statistical approaches for non-linear relationships can be broadly housed under the umbrella of *generalized additive models* (GAMs), but, of course, many machine learning models can also model non-linear relationships. Machine learning approaches are focused on prediction often at the expense of interpretation, so I consider them something of a different beast from GAMs.

It is common for investigators to choose 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 reason for this is that we are assuming a certain statistic model is an appropriate model of the data-generating process. In this model, only the data are random variables. If I decide on a model structure only after seeing the data, then both the data and the model structure

are random since the model structure is determined based on the data. However, I usually do not see folks acknowledging this uncertainty in the model structure and adjusting there inference for the extra variability.

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, it is appropriate 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.

Figure 2 is Anscombe's famous quartet. The four datasets show identical linear fits even though the underlying shapes of the relationships are very different.

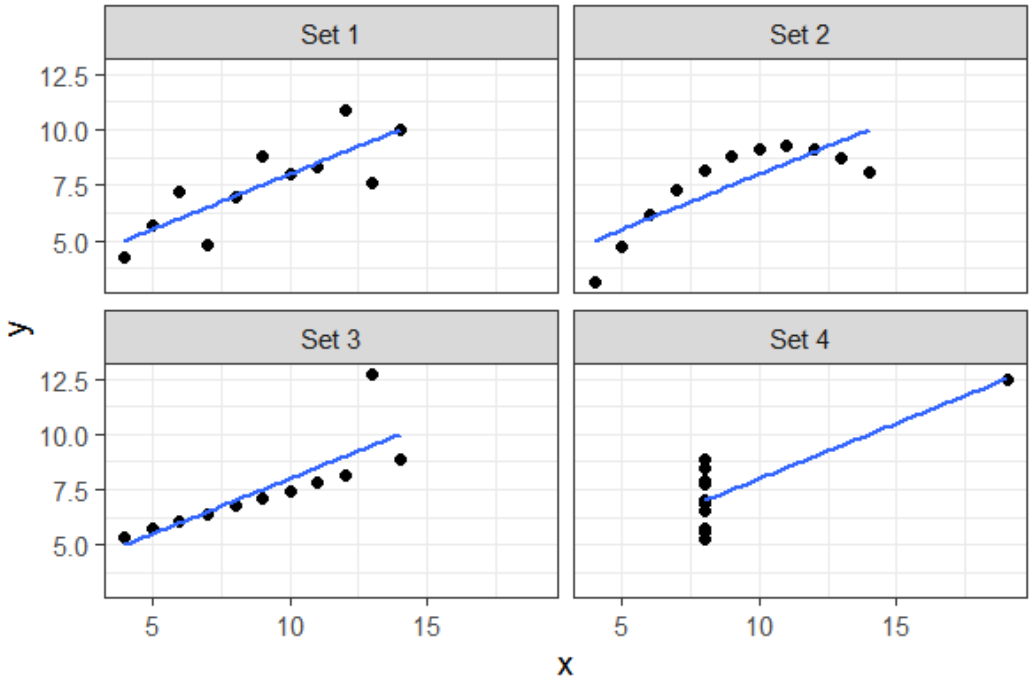

Figure 2: Anscombe's quartet. Four datasets that result in identical linear model estimates, but vary dramatically in the relationships between X and Y.

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 versus fitted and residual versus explanatory variable plots.

Figure **??** shows 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.

1.1.1 Linearity and scale

Remember the ol' definition of the derivative of a function $f(x)$ from calculus?

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

Parsing this out, the numerator is the change along the vertical axis with a given change in the horizontal axis, h, while the denominator is the change in the horizontal axis. In other words, *rise over run*! That means, we can describe *any continuous function*, $f(x)$, with a line, as long as we zoom in close enough on the function.

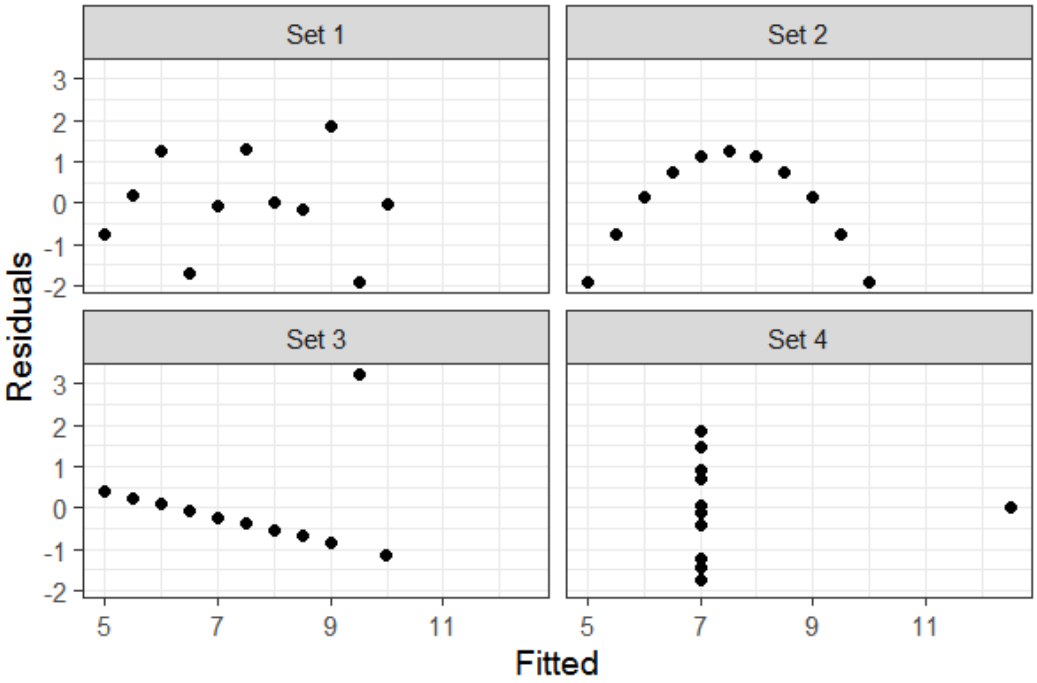

Figure 3: Residuals from simple linear models fit to each of Anscombe's datasets.

What I am getting at is that the assumption of linearity and how reasonable it is can depend on the scale of our measurements. For example, Figure 4 shows a scatterplot between two variables, X and Y. The measurements of Y were taken across a wide range of the X variable. Clearly, there is a non-linear relationship between the two variables.

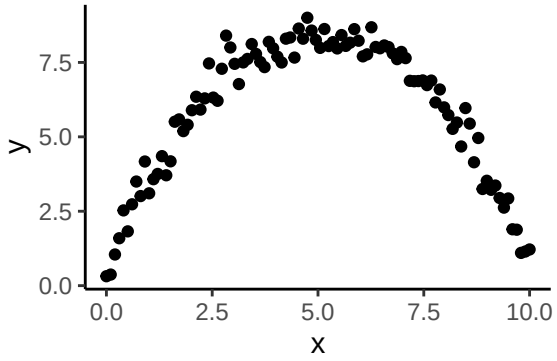

Figure 4: Scatterplot of fictitious data on X and Y.

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. Figure @ref{fig:scatter2} shows the same data as in Figure 4 but plotted over the range of $X \in (0, 2)$ and $X \in (2, 3)$, which look reasonably linear at this scale of X.

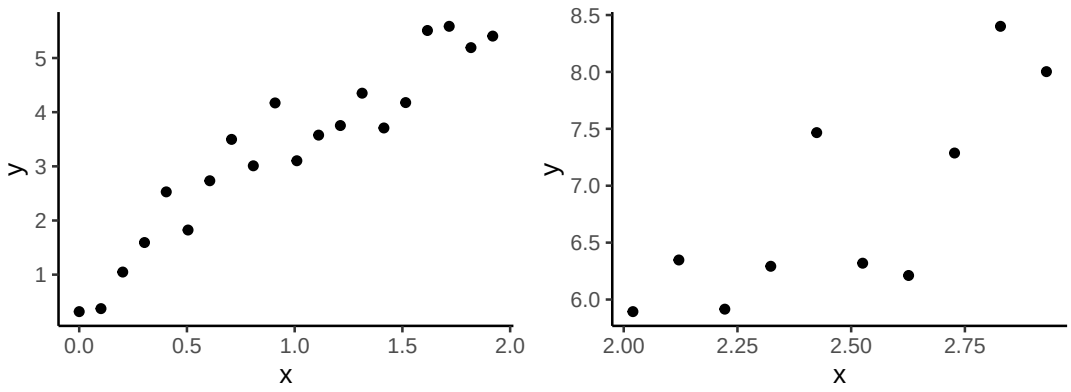

Figure 5: Scatter plots of X and Y assuming we only measured Y for smaller ranges of X.

2 Continuous and categorical variables

As always, the model you fit when working with continuous and categorical variables depends on your research question. Below are some different options that serve as a reminder from ST 512 as well as the lecture on interactions in Week 04.

2.1 Parallel lines model

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.

Figure 6 shows an example of a parallel lines model from an analysis of two groups with a continuous variable X. The difference in the mean of Y among the two group is roughly 10 units across the entire range of X.

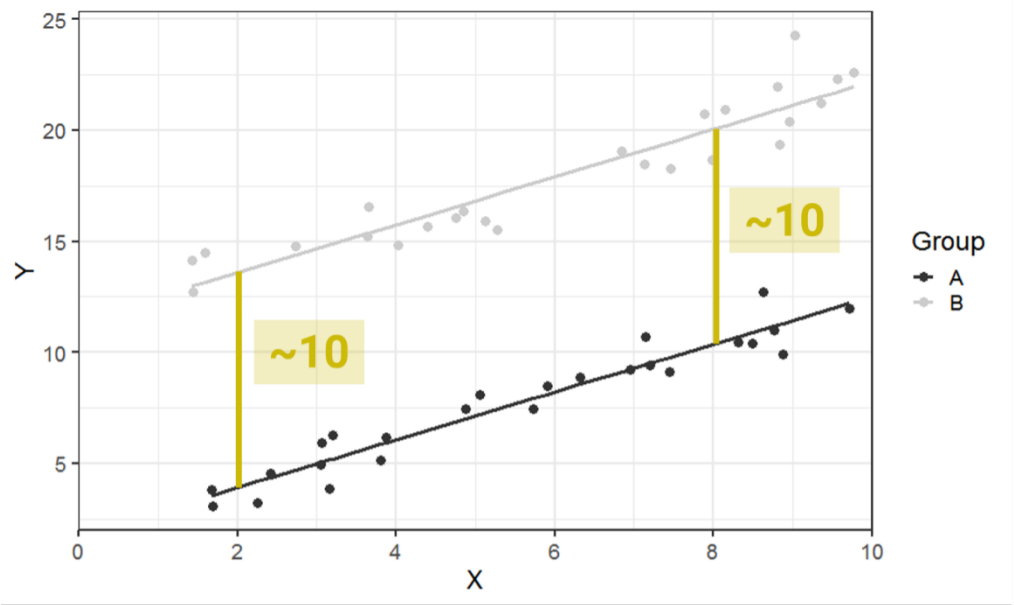

Figure 6: Example of a dataset generated using the parallel lines model.

The parallel lines analysis is what people used to mean when they use the term ANCOVA (which stands for analysis of covariance). We will just refer to it as a linear model, but, be aware that some people use this term today to mean a linear model with both continuous and categorical variables in it.

2.2 Separate lines model

The separate lines model is one with an interaction between the categorical and continuous explanatory variable. Remember from lecture that we 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 between groups now depends on the value of the continuous explanatory variable, as shown Figure 7. The difference in mean Y between groups is around 15 when $X = 2$, but about 35 when $X = 8$.

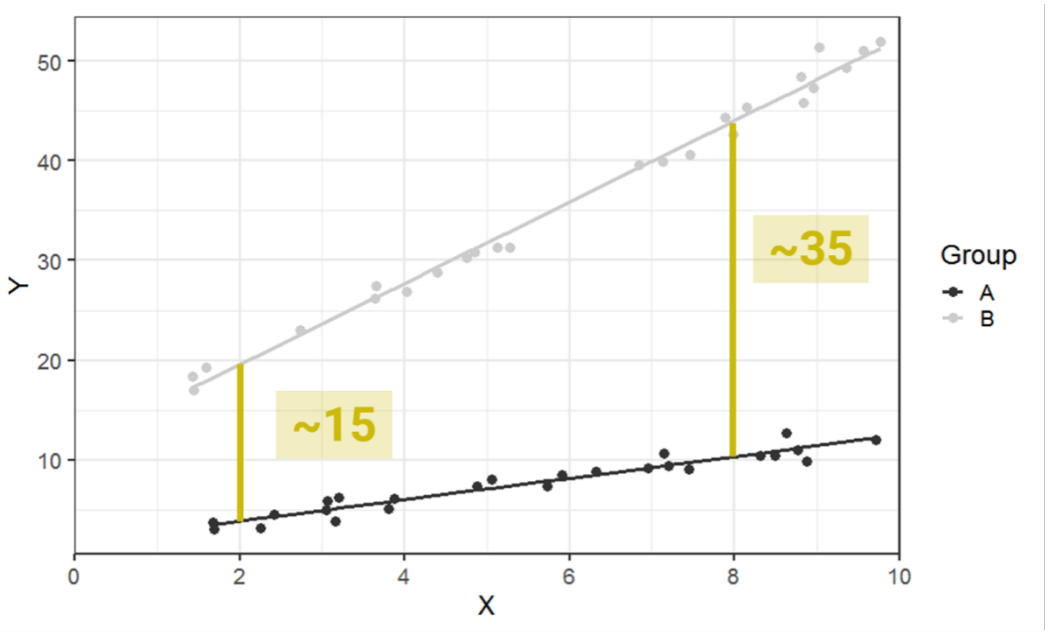

Figure 7: Example of a dataset generated using the separate lines model.

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 CIs in the presence of an interaction, although I haven't looked out how the confidence intervals are calculated.

2.3 Continuous by continuous interactions

One difficult topic is how to work with continuous by continuous interactions. Such an interaction would indicate that the relationship of one continuous variable, X_1, and the mean of Y is affected by the value of a second continuous variable, X_2. The result would be a plane (drawn in 3D) that has *curvature*, as discussed in lecture. For a reminder, see the "For Fun" section of the Week 04 module to find an interactive 3D graphic to play with.

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 (or close to it) of one variable for roughly 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 up working with very small samples, continuous by continuous interactions are not realistic to fit even when the concept of them is realistic. Note that 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.

3 Analyses with many variables

At some point during your scientific career, you will likely end up in a situation where you have many, many explanatory variables to contend with. In some cases, you may have more control variables than you have observations! This is known as the *high-dimensional setting* and is a difficult problem to tackle.

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

3.1 Option 1: Think really hard

There isn't some easy statistical answer to what to do when we have many variables. Even if you have 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 seem to learn 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 based on prior knowledge in your field. 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, it is common in ecology these days for investigators who are doing similar research in different places claim that 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.

3.2 How many variables can I use?

See Chapter 4 of the Harrell 2015 "Regression Modeling Strategies" book 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 on the assumption you are fitting some parametric model like a linear or additive model.

You will often hear rules of thumb that the number of observations you need in order to estimate each parameter is roughly 10-15. For example, if you are trying to estimate the coefficients for a multiple regression with 4 explanatory variables, you should shoot for a minimum of 50 observations (10 for each slope parameter and 10 for the intercept). This target 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 $p < n/15$, where n is the sample size and p is the number of parameters in the fullest model.

This is not to say that we *can't* estimate parameters with a smaller sample size, these are just rules of thumb about how to get reliable estimates of model parameters. In fact, with continuous variables, we can *saturate* a model and estimate up to as many parameters as we have observations, but then we are basing inference on those parameters on a single measurement. This is almost never a good idea.

3.3 Model complexity

Unsurprisingly, 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, the number of parameters to estimate, 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.

4 Model (variable) 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 because their effect estimates or relationship estimates are not "significant". 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.

4.1 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.

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 for example 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 😉.

4.2 Inference versus prediction

Whether model selection is something you should be doing at least partially depends on the goal of your analysis. For predictive models, we care less about estimating relationships and more about *useful predictors*, so the issues with post-selection inference discussed above are less of an issue.

In addition, if your 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 as well as uncertainty around including those variables over others in the first place. This can be done using 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. Additionally, Tredennick et al. (2021) and Fieberg and Johnson (2015) discuss these concepts in an approachable and practical way.

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/.

4.3 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) (Burnham and Anderson 2004). Structural equation modeling also fits inside this umbrella of competing models.

In multimodel inference you carefully define different models to represent different *a priori* hypotheses. This does not mean you blindly fit all possible models and say "hypothesis 1: Y is related to X_1; hypothesis 2: Y is related to X_1 and X_2" and so on. Instead each model means something practically important in your field and should be based on theory.

You can see an example of this approach in Burnham, Anderson, and Huyvaert (2011), 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 (2015).

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.

References

Burnham, Kenneth P., and David R. Anderson. 2004. "Multimodel Inference: Understanding AIC and BIC in Model Selection." *Sociological Methods & Research* 33 (2): 261–304. https://doi.org/10.1177/0049124104268644.

Burnham, Kenneth P., David R. Anderson, and Kathryn P. Huyvaert. 2011. "AIC Model Selection and Multimodel Inference in Behavioral Ecology: Some Background, Observations, and Comparisons." *Behavioral Ecology and Sociobiology* 65 (1): 23–35. https://doi.org/10.1007/s00265-010-1029-6.

Cade, Brian S. 2015. "Model Averaging and Muddled Multimodel Inferences." *Ecology* 96 (9): 2370–82. https://doi.org/10.1890/14-1639.1.

Fieberg, John, and Douglas H. Johnson. 2015. "MMI: Multimodel Inference or Models with Management Implications?" *The Journal of Wildlife Management* 79 (5): 708–18. https://doi.org/10.1002/jwmg.894.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1): 267–88. https://www.jstor.org/stable/2346178.

Tredennick, Andrew T., Giles Hooker, Stephen P. Ellner, and Peter B. Adler. 2021. "A Practical Guide to Selecting Models for Exploration, Inference, and Prediction in Ecology." *Ecology* 102 (6): e03336. https://doi.org/10.1002/ecy.3336.