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.

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

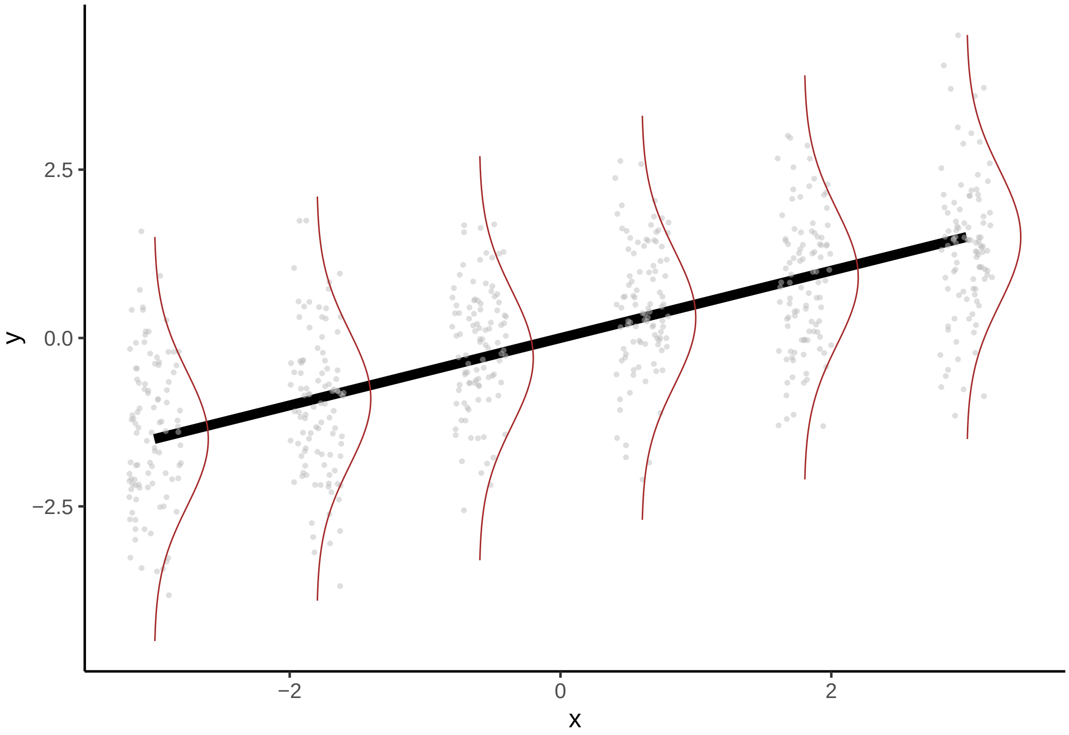

Figure 1: Conceptual drawing of linear regression. The line models the mean response, while the errors are normally distributed about the line with constant variance.

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.

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.

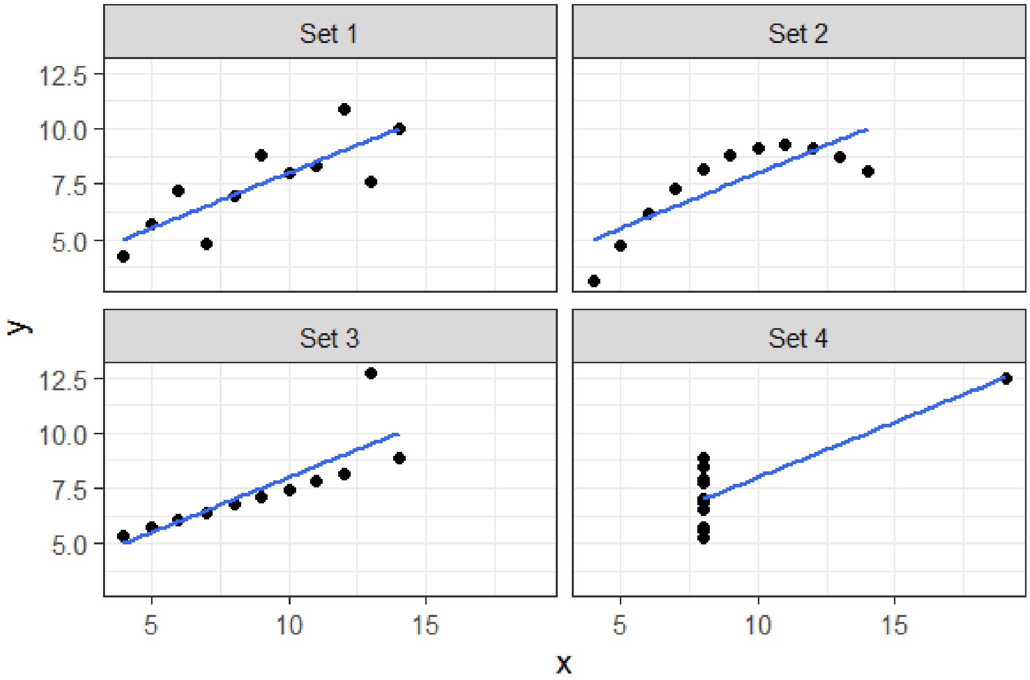

Figure 2: Anscombe's quartet. Four datasets that result in identical linear model estimates, but vary dramatically in the relationships between X and Y.

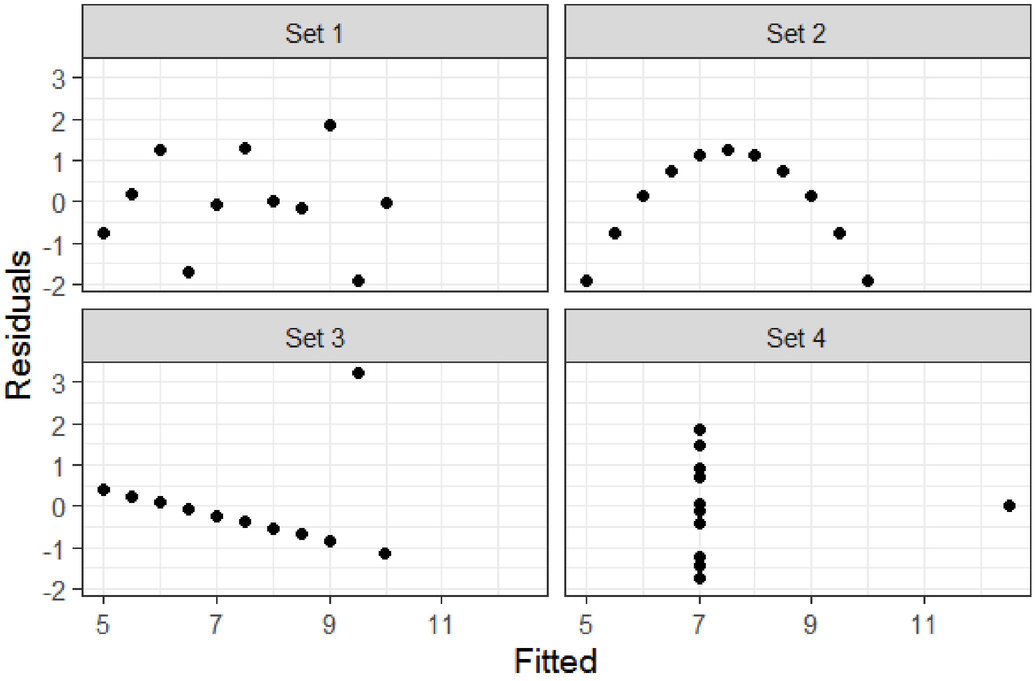

Figure 3: Residuals from simple linear models fit to each of Anscombe's datasets.

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.

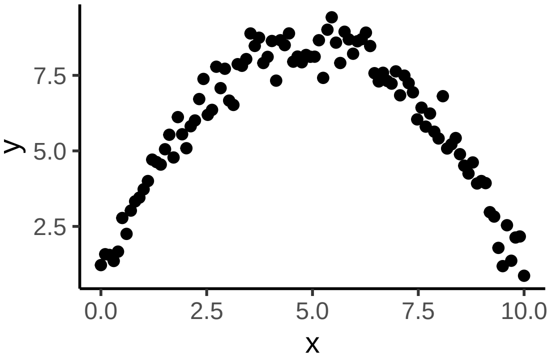

Figure 4: Scatterplot of fictitious data on X and Y.

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.

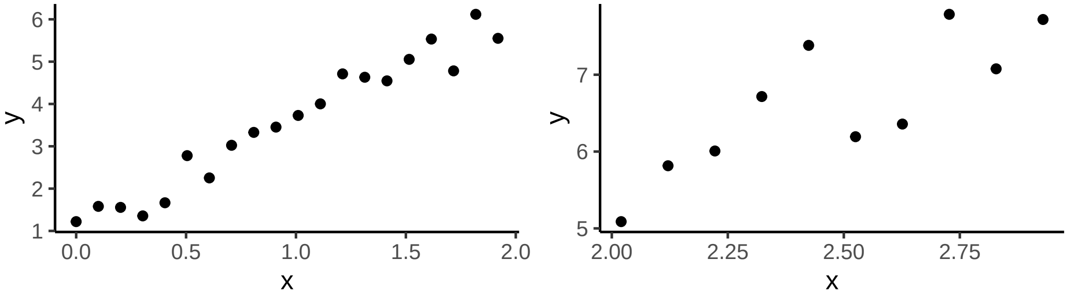

Figure 5: Scatter plots of X and Y assuming we only measured Y for smaller ranges of X.