Regression models are an important tool in natural sciences. Effectively implementing regression modeling methods requires both basic knowledge of the underlying model and the practice of a consistent workflow. In this document I will cover a few key aspects of regression modeling that can be very useful to keep in mind while working on a problem, and outline a workflow that should consistently produce useful models.

Some basic ideas in regression modeling:

This may well be reviewed, but there are several points about regression models that are worth reiterating.

First, the basic goal of a regression model is to predict some value $y$using a set of variables x_i. This is summarized with a single equation

1) $$y = f(x_1, x_2, \ldots, x_3).$$

Our goal is to find the variables $x$and function $f$that best describe the observations$y$. This can be useful for predicting how future observations may change or for understanding how the independent variables cause changes in the observations.

In addition to describing the relationship between the dependent variables $x$and observations $y$we will want to describe how the observations may vary from this relationship. This can be done in a number of ways but a simple extension to equation 1 will suffice

2) $$y = f(x_1, x_2, \ldots, x_3) + \epsilon \ ,$$

where ϵ is a random variable that describes the deviations of the observations $y$from the predictions f.

One major difference between modeling approaches is how they represent the function f that describes the relationship between the observations and the dependent variables. Linear regression represents this function as a linear function, GAMs represent this function as a sum of simple nonlinear functions, and regression tree represent the function as a piecewise step function represented by a decision tree. Regardless of our choice of the functional form we are confronted with several problems related to the form of the independent variables x.

There are two main types of data we might use as independent variables, categorical and continuous. Categorical data are usually modeled by fitting group means, this is shown in figure 1.

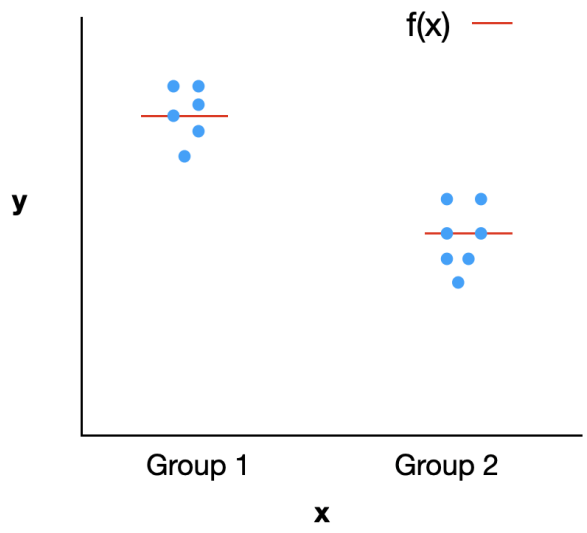

Figure 1: For categorical variable we typically represent the value of the function f(x) as a mean value for each level of the category

For continuous variables we fit some form of continuous function such as a line with the form $y = a + bx$. Both of these cases are relatively easy to manage, but what happens when we have both categorical and continuous variables (i.e. we have two groups of observations responding to a continuous variable)?

One approach is to fit one function to both groups and adjust the level of the function for each group. In the case of linear regression this would imply fitting unique intercept terms $a$to each group and one slope term. A graph of this type of solution is shown in figure 2 a. The other approach would be to fit a unique slope and intercept to each level of the categorical variable. More generally this would mean fitting different models to each group of the categorical variable. We say that this type of model includes an interaction between the categorical and continuous variable.

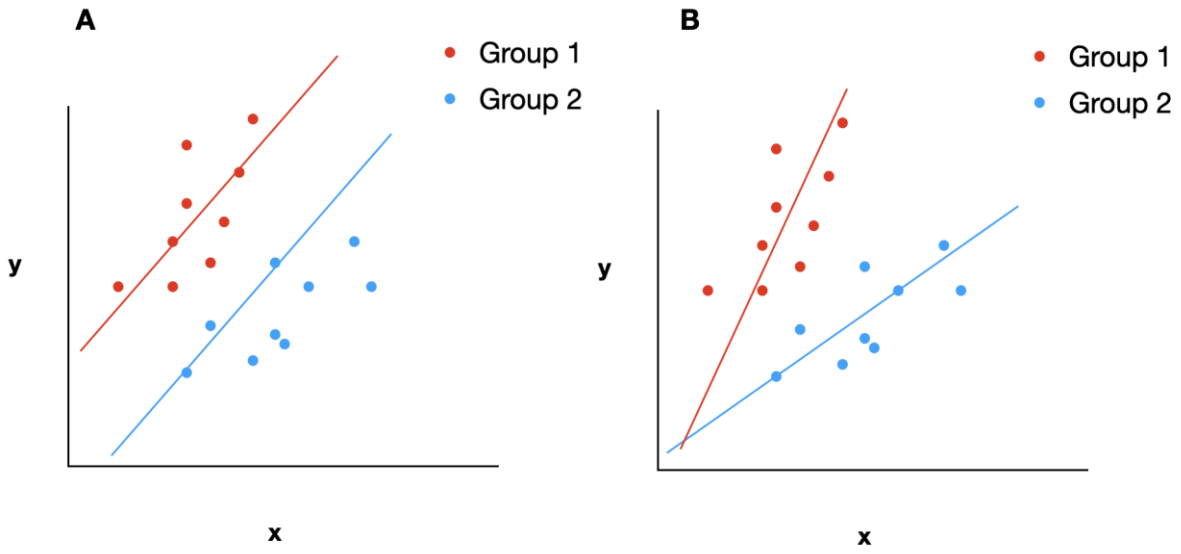

Figure 2: the graph of a linear model with fixed sloped for each group, but variable intercepts A) and the graph of a model that includes an interaction between the categorical and continuous variable B).

Although the example show are from a linear model the idea of the interaction between variables is relevant to all regression modeling tools, particularly methods like linear models and GAMs. Regression tree methods (e.g. random forest and boosted regression trees) do not require us to explicitly include interaction in our model because regression trees automatically account for them if they are appropriate, but the concept is still worth keeping in mind.

Workflow

An effective regression modeling workflow is an iterative process characterized by three steps

1) Select a model structure and independent variables
2) Fit the model
3) Check the fit of the model (cross validation)

These three steps are repeated until step three produces satisfactory results.

Step 1)
The step essentially requires us to decide where to start. In general on of two options will be best. Either start with the complicated model we can think of, that we predict will describe our system. This model may include a large number of variables and or complex functional forms. Such a model will likely not be the best model in practice. It may overfit the data, be too hard to fit to the data given our level of computational ability or two difficult to interpret the results from. All of these problem will require that we simplify the model in future iterations, but it will help structure our thinking about the problem and provide a useful starting point.

The other approach is to start simple and then build up (I would recommend this in most cases). I recommend starting with the smallest model that might teach you something about the data. This model will only include a few key variables, and simple functional forms. The model should be easy to fit and understand. If the model misses key patterns in the data or does not fully answer the question that motivated the analysis more detail can be added in future iterations.

A third approach is to initially construct the full complex model and a simple model and iterate on both of them. The goal will be for the two processes to meet somewhere in the middle with the ideal model

Step 2)
This will depend on the model we choose in step 1.

Step 3)
Assessing model performance is a key part of regression analysis. This can be done in a number of ways and the methods will likely depend on the precise model used. However, as a general rule it is important

to use cross validation. All cross validation entails is saving some subset of the full data set to uses as a reference, usually about 10%. The model can then be fit to the rest of the data set, and then predictions from the model can be tested against the reserved validation data set.

In addition to using cross validation methods it will also be important to check the model at this point to determine if it is interpretable and useful for answering the motivating questions.