easiest to start with a simple model and then build in additional complexity, taking care to check for problems along the way.

There are typically many reasonable ways in which a model can be constructed. Models may differ depending on the inferential goals or the way the data were collected. Key choices include how the input variables should be combined in creating predictors, and which predictors should be included in the model. In classical regression, these are huge issues, because if you include too many predictors in a model, the parameter estimates become so variable as to be useless. Some of these issues are less important in multilevel regression but they certainly do not disappear completely.

This section focuses on the problem of building models for prediction. Building models that can yield causal inferences is a related but separate topic that is addressed in Chapters 9 and 10.

General principles

Our general principles for building regression models for prediction are as follows:

1. Include all input variables that, for substantive reasons, might be expected to be important in predicting the outcome.
2. It is not always necessary to include these inputs as separate predictors—for example, sometimes several inputs can be averaged or summed to create a "total score" that can be used as a single predictor in the model.
3. For inputs that have large effects, consider including their interactions as well.
4. We suggest the following strategy for decisions regarding whether to exclude a variable from a prediction model based on expected sign and statistical significance (typically measured at the 5% level; that is, a coefficient is "statistically significant" if its estimate is more than 2 standard errors from zero):
   (a) If a predictor is not statistically significant and has the expected sign, it is generally fine to keep it in. It may not help predictions dramatically but is also probably not hurting them.
   (b) If a predictor is not statistically significant and does not have the expected sign (for example, incumbency having a negative effect on vote share), consider removing it from the model (that is, setting its coefficient to zero).
   (c) If a predictor *is* statistically significant and does not have the expected sign, then think hard if it makes sense. (For example, perhaps this is a country such as India in which incumbents are generally unpopular; see Linden, 2006.) Try to gather data on potential lurking variables and include them in the analysis.
   (d) If a predictor is statistically significant and has the expected sign, then by all means keep it in the model.

These strategies do not completely solve our problems but they help keep us from making mistakes such as discarding important information. They are predicated on having thought hard about these relationships *before* fitting the model. It's always easier to justify a coefficient's sign after the fact than to think hard ahead of time about what we expect. On the other hand, an explanation that is determined after running the model can still be valid. We should be able to adjust our theories in light of new information.