

### 4.3: The Backward Elimination Process

We are finally ready to develop the multi-factor linear regression model for the `int00.dat` data set. As mentioned in the previous section, we must find the right balance in the number of predictors that we use in our model. Too many predictors will train our model to follow the data's random variations (noise) too closely. Too few predictors will produce a model that may not be as accurate at predicting future values as a model with more predictors.

We will use a process called *backward elimination* [1] to help decide which predictors to keep in our model and which to exclude. In backward elimination, we start with all possible predictors and then use `lm()` to compute the model. We use the `summary()` function to find each predictor's significance level. The predictor with the least significance has the largest p-value. If this value is larger than our predetermined significance threshold, we remove that predictor from the model and start over. A typical threshold for keeping predictors in a model is  $p = 0.05$ , meaning that there is at least a 95 percent chance that the predictor is meaningful. A threshold of  $p = 0.10$  also is not unusual. We repeat this process until the significance levels of all of the predictors remaining in the model are below our threshold.

All of these approaches have their advantages and disadvantages, their supporters and detractors. I prefer the backward elimination process because it is usually straightforward to determine which factor we should drop at each step of the process. Determining which factor to try at each step is more difficult with forward selection. Backward elimination has a further advantage, in that several factors together may have better predictive power than any subset of these factors. As a result, the backward elimination process is more likely to include these factors as a group in the final model than is the forward selection process.

The automated procedures have a very strong allure because, as technologically savvy individuals, we tend to believe that this type of automated process will likely test a broader range of possible predictor combinations than we could test manually. However, these automated procedures lack intuitive insights into the underlying physical nature of the system being modeled. Intuition can help us answer the question of whether this is a reasonable model to construct in the first place.

As you develop your models, continually ask yourself whether the model “makes sense.” Does it make sense that factor *i* is included but factor *j* is excluded? Is there a physical explanation to support the inclusion or exclusion of any potential factor? Although the automated methods can simplify the process, they also make it too easy for you to forget to think about whether or not each step in the modeling process makes sense.

---

This page titled [4.3: The Backward Elimination Process](#) is shared under a [CC BY-NC 4.0](#) license and was authored, remixed, and/or curated by [David Lilja \(University of Minnesota Libraries Publishing\)](#) via [source content](#) that was edited to the style and standards of the LibreTexts platform.