**Statistical model building (http://www.theanalysisfactor.com/7-guidelines-model-building/)**

Model  Building–choosing predictors–is one of those skills in statistics that is difficult to teach.   It’s hard to lay out the steps, because at each step, you have to evaluate the situation and make decisions on the next step.

If you’re running purely predictive models, and the relationships among the variables aren’t the focus, it’s much easier.  Go ahead and run a stepwise regression model.  Let the data give you the best prediction.

But if the point is to answer a research question that describes relationships, you’re going to have to get your hands dirty.

It’s easy to say “use theory” or “test your research question” but that ignores a lot of practical issues.  Like the fact that you may have 10 different variables that all measure the same theoretical construct, and it’s not clear which one to use.

Or that you could, theoretically, make the case for all 40 demographic control variables.  But when you put them all in together, all of their coefficients become nonsignificant.

So how do you do it?  Like I said, it’s hard to give you step-by-step instructions because I’d need to look at the results from the each step to tell you what to do next.  But here are some guidelines to keep in mind.

**1. Remember that regression coefficients are marginal results.**

That means that the [**coefficient for each predictor**](http://www.theanalysisfactor.com/interpreting-regression-coefficients/) is the *unique* effect of that predictor on the response variable.  It’s not the full effect unless all predictors are independent.  It’s the effect after controlling for other variables in the model.

So it matters what else is in the model.  Coefficients can change quite a bit, depending on what else is in the model.

If two or more predictors overlap in how they explain an outcome, that overlap won’t be reflected in either regression coefficient.  It’s in the overall model [**F statistic and the R-squared**](http://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/), but not the coefficients.

**2. Start with univariate descriptives and graphs.**

Always, always, [**always start with descriptive statistics**](http://www.theanalysisfactor.com/the-11-steps-for-statistical-modeling-in-any-regression-or-anova/).

It will help you [**find any errors**](http://www.theanalysisfactor.com/steps-to-take-when-your-regression-or-other-statistical-results-just-lookwrong/)that you missed during cleaning (like the 99s you forgot to declare as missing values).

But more importantly, you have to know what you’re working with.

The first thing to do is univariate descriptives, or better yet, graphs.  You’re not just looking for bell curves.  You’re looking for interesting breaks in the middle of the distribution.  Values with a huge number of points.  Surprising values that are generally much higher or with less variation than you expected.

Once you put these variables in the model, they may behave funny.  If you know what they look like going in, you’ll have a much better understanding why.

**3. Next, run bivariate descriptives, again including graphs.**

You also need to understand how each potential predictor relates, on its own, to the outcome and to every other predictor.

Because the regression coefficients are marginal results (see #1), knowing the bivariate relationships among variables will give you insight into why certain variable lose significance in the bigger model.

I personally find that in addition to correlations or [**crosstabs**](http://www.theanalysisfactor.com/chi-square-test-proportions/), scatterplots of the relationship are extremely informative.  This is where you can see if [**linear relationships are plausible**](http://www.theanalysisfactor.com/regression-modelshow-do-you-know-you-need-a-polynomial/) or if you need to deal with nonlinearity in some way.

**4. Think about predictors in sets.**

In many of the models I’ve been working with recently, the predictors were in theoretically distinct sets.  By building the models within those sets first, we were able to see how related variables worked together and then what happened once we put them together.

For example, think about a model that predicts binge drinking in college students.  Potential sets of variables include:

* demographics (age, year in school, socio-economic status)
* history of Mental Health (diagnoses of mental illness, family history of alcoholism)
* current psychological health (stress, depression)
* social issues (feelings of isolation, connection to family, number of friends)

Often, the variables within a set are correlated, but not so much across sets.  If you put everything in at once, it’s hard to find any relationships. It’s a big, overwhelming mess.

By building each set separately first, you can build theoretically meaningful models with a solid understanding of how the pieces fit together.

**5. Model building and interpreting results go hand-in-hand.**

Every model you run tells you a story.  Stop and listen to it.

Look at the coefficients.  Look at R-squared.  Did it change?  How much do coefficients change from a model with control variables to one without?

When you pause to do this, you can make better decisions on the model to run next.

**6. Any variable involved in an interaction must be in the model by itself.**

As you’re deciding what to leave in and what to boot from the model, it’s easy to get rid of everything that’s not significant.

And it’s usually a good idea to eliminate non-significant interactions first (the exception is if the interaction was central to the research question, and it’s important to show that it was not significant).

But if the interaction is significant, you can’t take out the terms for the component variables (the ones that make up the interaction).  The [**interpretation of the interaction**](http://www.theanalysisfactor.com/clarifications-on-interpreting-interactions-in-regression/) is only possible if the component term is in the model.

**7. The research question is central.**

Especially when you have a very large data set, it’s very easy to step off the yellow brick road and into the poppies.  There are so many interesting relationships you can find (and they’re so shiny!).  Months later, you’ve testing every possible predictor, categorized every which way.  But you’re not making any real progress.

Keep the focus on your destination–[**the research question**](http://www.theanalysisfactor.com/what-makes-a-statistical-analysis-wrong/).  Write it out and tape it to the wall if it helps.

All of these guidelines apply to any type of model–linear regression, ANOVA, logistic regression, mixed models.  Keep them in mind the next time you’re doing statistical analysis.

[Bookmark and Share](http://addthis.com/bookmark.php?v=250&pub=kgracemartin)

If you want to learn all the ins and outs of interpreting regression coefficients, check out our 6-hour online workshop [Interpreting (Even Tricky) Regression Coefficients.](http://theanalysisinstitute.com/irc/)  This workshop will teach you the real meaning of coefficients for all the tricky regression terms: correlated predictors, dummy variables, interactions, polynomials, and more.