Building Choice Models that Forecast Well

(Hint: It's Not What You Think)

by
George Boomer
StatWizards LLC

Years ago, fresh out of graduate school, I was hired by Data Resources, Inc., Otto Eckstein's econometric consulting firm. Our green, doe-eyed class went through DRI's training program and steeped ourselves in academic literature on building time-series regression models.

Yet after building dozens of models and seeing which ones forecast well and which did not, I reached a surprising conclusion: **Models that would pass even the most rigorous academic scrutiny were not necessarily the ones that produced the best forecasts.** In fact, some models that appeared weak from an academic perspective and could never get published were forecasting champs. After examining a number of models like this I realized that one thing separated winners from losers. More surprisingly, this key factor did not appear in any academic textbooks. What was it? Coefficient elasticities. Models with reasonable elasticities on the coefficients performed well; those that were too large or too small did not.

Years later, after building almost 200 choice models in commercial settings, I find a similar situation in this realm, too. Choice models are typically judged by performance on holdout samples. It's easy to see why. Predicting results using holdout samples is a comparatively easy thing to do and appears a reasonable approach. Journal editors, in particular, like this idea. For practitioners, though, the goal is very different, for they have to contend with the messy task of producing models that forecast well in the real world. Holdout samples, especially those using made-up data, avoid the messiness that prevails in the world's complexities.

Practitioners, though, are less interested in getting papers published. Real-world clients don't pay for forecasting holdout samples; they care about how well models forecast in a complex setting where extraneous variables often affect outcomes. A number of studies, including mine, show that models that predict well using holdout samples do not necessarily produce better forecasts in a complicated world.

So what <u>does</u> work? Once again it's not found in academic papers or conference proceedings. Based on years of observing successful models and unsuccessful models, this is what I have found:

Principles for building successful discrete-choice models

Some models forecast well and some don't. That's my only interest: forecast performance, not holdout samples. After seeing and building enough models, I see patterns emerging along with key principles that greatly increase (although nothing can guarantee) the probability of forecasting success. These are, of course, my own views and are based on experience rather than rigorous academic testing. I present them not as declarations, but as hypotheses to be considered.

1. Set up your experimental designs so they mimic the process customers undertake when choosing products and services in the marketplace.

One of the great things about discrete-choice models is that designs can replicate how consumers make choices in the real world. Motorola once went so far as to build replicas of stores and lay out mockup products on shelves to collect choice data. Going to this length doesn't guarantee success, but in general, designs that mimic how choices are actually made have a better chance of success than designs that do not. All of this presumes, of course, that researchers understand customers' choice processes to begin with.

1.1. If the choice process is complex, employ focus groups and/or prior research to identify the process or processes by which customers choose products in your space.

I have seen academics argue about whether choice exercises are too complex, too confusing or even bewildering. In my view, this is a pointless argument**. If the choices consumers face in the real world are complex, then there is no reason they can't handle choice exercises that are equally complex.** Do shoppers become paralyzed or hopelessly confused when faced with the decision of which cable company to choose? Of course not. They inevitably find a way to deal with the complexity, so arguing about limits on the complexity of choice exercises is, in my view, a useless diversion. Too many experimental designs devote too much attention to how hard or easy a questionnaire is to fill out. Proponents of this view sometimes forget that the principle objective is not to make the instrument easy, but to get useful information. Don't get me wrong. Facility is an important consideration (after all you want respondents to complete the questionnaire and answer questions honestly), but the predominant concern is collecting actionable, pertinent information. If in a given situation the choices consumers make are complex, there is no reason experimental designs cannot be equally complex. We have used such designs often in the past. They have worked—and continue to work—just fine, because they mimic actual choice processes. As long as the heuristics that customers use to complete choice exercises are the same as the ones they use when they shop, there is no need to worry. That doesn't mean you shouldn't remain vigilant, though, as the next section describes.

1.2. Expand pretests to vet your designs.

If a design is complicated because the inherent choices consumers make are complicated, it's a good idea to pretest the exercises, either with trusted colleagues and/or in focus groups. The feedback you get will be invaluable in honing the instrument. **Often, confusion arises not from the complexity of the instrument, but from its layout, formatting and wording.** Those problems need to be identified and fixed before you to go field. Even if your designs are not complex, it's still a good idea to devote resources to a pretest.

Another concern is choice sets that are dominated by a single product configuration, such as one with desirable attributes combined with a low price. While that's a legitimate consideration, we have found that **most dominated choice sets can be identified by simple**

inspection. There is no need to have design software jump through hoops in order to avoid them. Furthermore, a good pretest will identify any that might escape detection. A simple frequency distribution of responses will tell you whether complete domination or complete avoidance exists. In our experience, either situation rarely occurs, but it's a good idea to check anyway. This implies, of course, that you have a design generator that allows you to inspect an entire design. If you use such a generator, great; if you don't, well, good luck to you.

If your pretest is large enough, you may even want to build a preliminary model from the results. To the extent possible, you want to identify problems before you go to field. Building a model, no matter how small the sample, is a good way to do that. Just be sure you carry the exercise through to building a simulator, where you can calculate willingness to pay, examine price curves and see whether price elasticities look reasonable. See section 3.3 below for more on this topic.

1.3. If you can, don't leave impossible combinations of attributes out of your design; place restrictions on them in the simulator instead.

Conventional wisdom says that you shouldn't show respondents products with absurd or impossible combinations of attributes. In doing so, you throw away information and may confuse respondents. This approach, however, contains a hidden trap. Restricting combinations inevitably reduces the efficiency of the design and risks injecting unwanted correlations among variables.

Moreover, respondents may either be unaware that impossible combinations really *are* impossible, or they may not care. For example, automobile companies occasionally study tradeoffs between such features as wheelbase (*i.e.*, distance between front and rear wheels) and trunk space. From an engineering standpoint, a car with a small wheelbase doesn't have room for a large trunk, but most customers don't know that; they can easily conceive of a small car with a big trunk, even though it may be impossible to build.

So what do you do with a situation like this? **Leave impossible combinations in the design, if you can, but impose this and all other engineering restrictions in the simulator.** Of course, this presumes you're your simulator can be modified to incorporate restrictions. Most cannot, but Excel-based simulators can. That's just one reason why we regard Excel as the preferred platform for simulation.

2. Ensure you have an adequate sample.

This should go without saying, but recently we have heard suggestions that you can get useful information from small samples. By far, the worst models we have encountered have all come from inadequate samples. Now, sampling theory for choice models is complex and unfortunate in one respect: The equation for the variance of parameter estimates includes the parameter itself, so you can't perform the calculation until you have done the research. Moreover, unlike sampling for linear regression models, the optimum sample size can vary widely based on the value of the parameters.[1] For that reason, most practitioners avoid this calculation[2] and instead use rules of thumb that tend to work well in practice. **The absolute minimum sample size is 300, and for stratified samples (*i.e.,* those with quotas for demographics) the minimum per cell (such as one age and income cohort) should be 50.** These are absolute minimums; more is better.

3. Choose estimation software and model specification based on criteria shown in the past to work in the real world.

Once data return from the field, most academics and practitioners choose a single methodology along with associated software and then proceed with the analysis. I submit that a better approach is to **have a variety of methods and software available, then choose one that matches most closely the study's objectives.** If the resulting model does not satisfy the criteria discussed below, be prepared to take a completely different approach, either a different specification or different method. Method and specification are related. Some software applications like NLOGIT permit complex specifications, whereas others do not. A word to clients: especially for critical projects, don't limit yourself to practitioners who are familiar with only one discrete-choice methodology. It's always better to have more tools in the toolbox than less. Many of our engagements start with one methodology and finish with another.

One tool that helps in this regard is our Data Wizard, which lets you take a common data set from the field and quickly create input data sets for a variety of estimation programs. If one statistical package fails you, it's easy to build a dataset for another. The process takes a matter of seconds, eliminating a key barrier to employing multiple approaches.

From assembling data sets, we now turn to building and evaluating models. How do you evaluate a choice model?

3.1. If variables are continuous in the real world, model them as continuous.

This sounds obvious, but many packages treat continuous variables such as price as discrete. This results in discontinuous elasticities and price curves that simulate poorly. The same is true for other variables that are continuous in the real world. **Just as designs work best when emulating the real world, models work best when variables reflect real-world concepts.**

3.2. If you believe that variables describing respondents rather than choices explain some portion of heterogeneity, choose a technique that can model such variables explicitly.

Modeling heterogeneity is useful, but understanding the source of heterogeneity is even more useful. Often, differences in preferences are associated with different respondent groups. If you or your clients seek a deeper understanding of such preferences, I suggest you employ a

[1] See Ben-Akiva, Moshe and Steven Lerman, "Discrete-Choice Analysis: Theory and Application to Travel Demand" (Cambridge, Massachusetts: The MIT Press, 1985), pp. 245-247

[2] Ibid, pp. 162-164.

methodology that uncovers latent segments, such as Latent GOLD Choice[3], or a methodology that allows you to enter socio-demographic variables directly into your specification, such as NLOGIT[4]. Both programs allow you to test the significance of such variables in your model. That's preferable to methodologies that allow you to tack these variables onto individual-level estimates but provide no insight into whether the variables explain differences in heterogeneity.

3.3. Examine price curves and price elasticities for realism.

This is analogous to the story I was telling earlier about elasticities of coefficients for econometric models. Some choice models, even ones that pass statistical tests, fail to produce reasonable price curves and/or price elasticities. Unfortunately, especially with today's models that estimate individual-level coefficients, you can't know this until you build a simulator. That means that **the traditional cycle of specify model⇨estimate model⇨check statistics⇨revise specification⇨re-estimate model needs to be replaced with specify model⇨estimate model⇨check statistics⇨build simulator⇨check elasticities⇨revise specification and/or method⇨re-estimate model.** Few practitioners do this, yet in my view this process is essential to producing top-quality models that forecast well.

3.4. Scrutinize model results, including simulation results, for face validity

If a model cannot produce reasonable forecasts, it is likely flawed in some way. How do you assess this? Relying on direct evidence is impossible, because we're forecasting probabilities rather than scalars. For that reason, we must rely on indirect evidence. One approach is to share early model results with senior executives and those familiar with how markets for the product really work. Ask whether they think results are reasonable. Examine other relevant research on the product(s) or service(s) at issue. In a sense this is a Bayesian process, in which results are reviewed then updated by experience (*i.e.*, prior expert beliefs). **If something in a model doesn't seem realistic, the model needs to be reworked through the cycle described above.** This does not mean that external advice is sacrosanct. Sometimes, especially with new products, experience with the marketplace does not exist or is colored by misperception. However, in most cases, external judgment can improve model specification.

3.5. Where appropriate, use judgment to temper model results

In those cases, give more weight to model results. How much judgment should you use? At Data Resources, the econometric firm mentioned earlier, *all* model results, including those used to formulate U.S. policy, were adjusted using judgment, down to individual equations. Surprised? Don't be. That has been the practice ever since Lawrence Klein invented modern econometrics[5]. And there is academic support for this point of view. In the late 1970s an

[3] See Statistical Innovations, http://statisticalinnovations.com/products/lgchoice.html.
[4] See http://www.limdep.com/ for more information.
[5] For his contribution to the field, Dr. Klein won the Nobel Prize in Economics in 1980 (http://www.nobelprize.org/nobel_prizes/economic-sciences/laureates/1980/klein-facts.html)

economist named Steve McNees[6] at the Boston Federal Reserve compared forecasts produced by models alone, economists without models, and models tempered with judgment. **Models tempered with judgment outperformed others.** I submit the same is true with choice models as well. At DRI, freshly minted Ph.D. economists with world-class educational backgrounds were almost always empiricists when they arrived. That is, they believed models should stand untouched, in publishable form. An error in forecasting meant a problem with the model. Once they had been at the firm for a while and had been humbled by actual events, they became Bayesians, in the sense that they acknowledged that prior beliefs were important in developing models that forecast well. As Nate Silver has said, "The key to making a good forecast is not in limiting yourself to quantitative information."[7]

That is not to say that unadjusted models don't work. If well-constructed, they do, often surprisingly so. As just one example, in an international study of laptop computers, early choice models showed that in Japan, the part-worth of the Toshiba brand was virtually identical to that of a clone. This was the only country in which we found such an anomaly, so we investigated further, checking our work and re-translating the questionnaire. What we discovered was that in Japan, "Toshiba" was synonymous with "clone" (!), so the model had actually picked up a cultural phenomenon. We have a number of other similar examples in which models uncovered market subtleties, but that's not an excuse to ignore extra-model information.

4. Employ statistical tests, but use real-world tests, too.

4.1. Apply statistical tests to avoid over-fitting

Model tests are important, especially hypothesis testing, but sometimes we forget why we do them. Modern statistical techniques are so powerful, they can seductively fit data once thought impossible to model. But in so doing, they risk creating structures that don't represent the underlying relationships in the real world. They model noise, rather than signal.[8]

Statistical tests cannot eliminate the problem of over-fitting, but they can mitigate it. One obvious test is for variable significance; if a parameter estimate does not differ significantly from zero, and there is not a good reason for keeping it in a model's specification, it should be dropped. Another, not so obvious test is for heterogeneity in parameter estimates. **It is as much a mistake to assume that heterogeneity exists in parameter estimates as to assume heterogeneity does not exist.** The existence of heterogeneity should be a hypothesis, not an assumption. Mixed-logit models, for example, include hypothesis tests for heterogeneity. For each variable, mixed-logit models (or random-parameter models as they are sometimes described) include significance tests for the value of the standard deviation or other dispersion

[6] Stephen K. McNees, "The Role of Judgment in Macroeconomic Forecasting Accuracy," International Journal of Forecasting, 6, no. 3, pp. 287– 99, October 1990. http://www.sciencedirect.com/science/article/pii/016920709090056H.

[7] Nate Silver. BrainyQuote.com, Xplore Inc, 2014. http://www.brainyquote.com/quotes/quotes/n/natesilver467578.html, accessed June 7, 2014.

[8] See Nate Silver, *The Signal and the Noise* (New York: Penguin Group, 2012) for a discussion of this issue.

parameters as well as for the mean. We find that for real-world models many variables fail dispersion significance tests. To avoid over-fitting, we drop heterogeneity assumptions for any variable that fails this test.

Most of the time, remaining specifications include one or more heterogeneous variables, but on rare occasions, all dispersion tests fail. In these circumstances, models degenerate into either simple logit or nested logit. I suspect (but cannot prove) this is why such degenerate models occasionally forecast as well or nearly as well as mixed logit or hierarchical Bayes models.

4.2. Use real-world priors to evaluate simulator as well as model performance

Bayesian methods have led to landmark breakthroughs in choice model estimation, but the Bayesian approach applies to the modeling process, as well. Recall that at the beginning of this paper we talked about the surprising relationship between coefficient elasticities and time-series models that forecast accurately. What we didn't mention was the criteria we used to evaluate these elasticities. We employed prior beliefs in what we thought the elasticities should be. That may sound like a simplistic kluge, but actually it's not. By looking at elasticities across a wide range of regression models that forecast well, we gradually learned what values to expect. In a Bayesian sense, we updated our priors as we gained more experience.

I submit that the same approach can be used to improve choice models. With the latter, evaluation is more difficult, because choice models are probabilistic whereas observations are discrete. We can't evaluate choice models as we can regression models. However, we can impose our learned Bayesian priors in much the same way. We do that in two ways: by checking the reasonability of price elasticities and by the reasonability of willingness-to-pay calculations. Most simulators, including ours, calculate these values. **If either price elasticities or WTP calculations fail reasonability tests, we reject either the model's specification or the methodology we are using.** We will often switch between CBC/HB, Latent GOLD Choice, mixed logit and—yes—even nested logit to create models that pass these tests.

We recently encountered an example that illustrates the point. Using identical questionnaires, we sampled parents, teachers and children to uncover interest in novel teaching approaches. A concern was that students asked how much their parents would spend would have no common frame of reference. This was borne out in the simulators. For one attribute, students' willing-ness to pay was over $30,000, confirming our suspicions; for parents and teachers, though, WTP amounts all fell within an expected range. Price-related analyses in the students' model were thereby discounted, whereas pricing analyses in the parents/teachers group was not.

4.3. Pick evaluation criteria based on your study's objectives

The above recommendations must be applied with a large dose of common sense. If pricing is a minor consideration, down-weight the WTP test, or even ignore it. Focus instead on other post-estimation tests, such as the importance ranking of attributes.

There aren’t any guarantees in life, and the same is true of the risky business of forecasting. No less a forecasting guru than Alan Greenspan has said, “We really can't forecast all that well, and yet we pretend that we can, but we really can't.”[9] Yet practitioners have no choice. To us, the principle value in choice modeling lies in forecasting, so we are compelled to try. Limiting ourselves to the safe harbor of share prediction is not enough; the value lies in the turbulent, uncertain waters of forecasting. We owe our clients nothing less.

[9] Alan Greenspan. BrainyQuote.com, Xplore Inc, 2014.
http://www.brainyquote.com/quotes/quotes/a/alangreens613457.html, accessed June 7, 2014.