Correlated Component Regression: a powerful new technique for pharmaceutical and B2B research

Alex Vishney shares his experience using a new technique for building predictive models from small data sets.

One of the persistent challenges in ethical pharmaceutical and B2B quantitative research is small sample sizes, particularly in Australia, where markets tend to be smaller than in the UK and US. In some medical and commercial fields, it is not uncommon to have national populations of only several hundred individuals, resulting in survey samples as small as n=30, once response rates are taken into account. In these instances, when it is simply impossible to collect larger samples, it can be tempting to apply traditional analytical techniques such as regression, arguing (often in one's own mind!) that 'any model, even if less reliable, is better than no model'. However, many existing regression applications can produce unstable and misleading results when applied to small samples.

We have recently employed a new technique (collaborating with Logit Research) that addresses this issue, called Correlated Component Regression analysis, or CCR. We find that CCR yields more robust predictive models based on small samples than traditional regression techniques.

The performance of a regression model is generally assessed using R2. Many researchers think that higher R2 always equals a better model, but this is not always the case. As the ratio between the sample size (N) and number of predictors (P) becomes smaller, R2 tends towards 100%. For N=P, R2 always equals 100% even for completely random data! In other words, as we tend towards smaller samples, we find that the model 'over-fits' the sample. In fact, the model is reflecting 'noise' (sampling error) rather than 'signal' (population). To overcome this we need to differentiate between the 'in-sample' R2 – in the sample used to build the model – and 'out-of-sample' R2, which assesses how well a model predicts for new cases.

CCR deals with this sort of problem very effectively by employing some unique features:

1. Stabilising the model using a factor-analysis like approach. This effectively strengthens our ability to detect the 'signal' whilst at the same time reducing any 'noise' in the sample.
2. Selecting the best model based on how it performs on 'new cases' not used to build the model. It does this via a process called cross-validation, where the model is built many times holding back different randomly selected bits of the sample to test them on. We get a cross-validation (CV) r-squared which tells you how well it performs on new cases.
3. Using a stepping-down procedure to screen out irrelevant predictors using the coefficients which have been stabilised in (step 1).

We can contrast the results of a CCR analysis with that of a regular regression analysis in (predictive analytics software) SPSS, using real data from an employee research study, in which 76 employees of a pharmaceutical company took part in the survey, providing ratings on overall commitment to the company and a battery of 34 other statements (all measured on a 7-point scale) about various aspects of their job role. The objective was to identify the subset of statements which are drivers of overall commitment and determine their relative importance. For this comparison we used both the Stepwise and Backward elimination options in SPSS (with their default settings) and CCR. Table 1 compares model performance. CV R2 for each method was obtained by 1000 rounds of 10-fold cross-validation. In every round the sample is randomly partitioned into 10 portions, the model is built 10 times, each on a different 9/10 of the sample (training sample) with the remaining 1/10 held back to test it (validation sample). The effect is to re-use every piece of sample 1000 times for cross-validation, giving a very robust measure of performance out-of-sample.

We see from Table 1 that although the 'in-sample' R2 for CCR is lower than for the regression models, its 'out-of-sample' CV R2 is considerably higher. In other words, the CCR model did a better job of predicting new cases than the regression models (and, ultimately, predicting new cases should be the real test of any predictive model).

CCR tends to deliver the greatest benefits where the ratio of cases to predictors is ←10, making it ideal for use when researching small populations (e.g. medical specialists, senior or sector B2B, etc.), but also as a tool that can be used for small sub samples of larger populations (i.e. the usual situation where sample just doesn't stretch far enough).

Alex Vishney, partner, Hall & Partners | Open Mind

Table 1 - Comparison of in-sample and out-of-sample (CV) R2

Method	In-sample R2	CV (Out-of sample) R2
SPSS Stepwise	72.4%	44.0%
SPSS Backward	78.3%	40.0%
CCR*	68.0%	59.7%

* Based on a one-principle-component model. In a typical project the optimal number of components (and hence, models that need to be run) is between one and six.