A Simple Pullback Estimator For Local Markets

Steve Vail
Impiric
(formerly Wunderman Cato Johnson)

June 2000

We sometimes need to estimate response rate within each of several local markets. However, sparse data in some markets may lead to unstable estimates. We derive a simple method for obtaining reasonable local estimates under conditions of varying stability. The new rate estimates are essentially weighted averages of the local rate and the overall rate. Weights are chosen to minimize mean squared error of the estimator. We conclude by illustrating the method with an example involving direct mail response within postal regions.

Let there be I non-overlapping geographic market regions be indexed by $i = 1,...,I$, and let N_i and R_i indicate number mailed and responded in each region, respectively. Furthermore, let $N_\bullet = \sum_i N_i$ and $R_\bullet = \sum_i R_i$ denote the overall number mailed and responded. Let $\hat{\theta}_i = R_i / N_i$ and $\hat{\theta}_\bullet = R_\bullet / N_\bullet$ estimate the regional and overall response rates, respectively. The problem we intend to solve is how to reliably estimate local response rates θ_i when the number mailed, and the number of respondents, vary considerably among regions.

The usual estimate of response rate θ_i is $\hat{\theta}_i = R_i / N_i$, with $\hat{\sigma}_i^2 = \hat{\theta}_i(1-\hat{\theta}_i)/N_i$ as it's estimate of variance. The main problem we deal with here is that the N_i vary greatly, leading to the $\hat{\theta}_i$ having wildly different amounts of precision. At low levels of precision some markets will, by chance, have implausibly extreme response rates. A secondary problem is that several R_i are zero.

One approach to this problem is to set a threshold in the form of an upper limit on $\hat{\sigma}_i^2$. Whenever the variance exceeds this threshold then use the overall mean $\hat{\theta}_\bullet$ to estimate θ_i, instead of the local mean $\hat{\theta}_i$. This approach fails on two fronts:

1. Since the transition from $\hat{\theta}_i$ to $\hat{\theta}_\bullet$ is abrupt, estimator bias and variance change abruptly as well;
2. When R_i is zero, $\hat{\theta}_i$ and $\hat{\sigma}_i^2$ are zero as well, and while these estimates are unbiased, they are not very useful, and this situation arises frequently, particularly in small samples.

This approach is obviously a poor one and we do not explore it any further here.

Instead, we construct the estimator $\hat{\phi}_i = r_i\hat{\theta}_i + (1-r_i)\hat{\theta}_\bullet$, where $0 \le r_i \le 1$ is a weight specifying the degree to which $\hat{\phi}_i$ relies on the local rather than the overall estimate. We want $\hat{\phi}_i$ to have several properties:

1. $\hat{\phi}_i$ should transition smoothly from $\hat{\theta}_i$ to $\hat{\theta}_\bullet$, gradually 'borrowing strength' from $\hat{\theta}_\bullet$ as the variance of $\hat{\theta}_i$ grows large;
2. $\hat{\phi}_i$ should be more stable (have lower variance) than $\hat{\theta}_i$.

We choose r_i in a way which minimizes the mean squared error [MSE] of $\hat{\phi}_i$:

$$MSE = E\left[(\hat{\phi}_i - \theta_i)^2\right] = Bias^2(\hat{\phi}_i) + Var(\hat{\phi}_i)$$

Now we can show that[1] $Bias^2 = (1-r_i)^2(\theta_i - \theta_\bullet)^2$ and $Var = r_i^2\sigma_i^2 + (1-r_i)^2\sigma_\bullet^2$ so that MSE is minimized when

$$\frac{\partial}{\partial r_i} MSE = 0,$$

$$\Rightarrow r_i = \frac{\sigma_\bullet^2 + (\theta_i - \theta_\bullet)^2}{\sigma_i^2 + \sigma_\bullet^2 + (\theta_i - \theta_\bullet)^2}.$$

The weight is estimated by substituting the sample estimates for the population quantities in this last equation.

One property of $\hat{\phi}_i$ is that its expectation is $r_i\theta_i + (1-r_i)\theta_\bullet$, so $\hat{\phi}_i$ is a biased estimator of $\hat{\theta}_i$ except when $r_i = 0$. Another property is that as $\hat{\sigma}_i^2$ increases so does the weight, causing $\hat{\phi}_i$ to rely more and more on $\hat{\theta}_\bullet$. Statistics like $\hat{\phi}_i$ are sometimes called 'pullback estimators' since they draw back toward

---

[1] We have assumed that the naïve local and overall estimates are independent. This is a reasonable approximation so long as no one $\hat{\theta}_i$ dominates.

a central estimate as the local estimate becomes unstable. The philosophy behind the approach is that it is reasonable to loosen the requirement of unbiasedness in order to achieve a large reduction in variance.

Another problem mentioned earlier is that, when N_i is small, R_i may easily be zero. This gives $\hat{\theta}_i = 0$ and the usual variance estimate $\hat{\sigma}_i^2 = \hat{\theta}_i(1-\hat{\theta}_i)/N_i$ equals zero as well. This results in $\hat{\phi}_i = \hat{\theta}_i = 0$, which we don't really believe. In cases such as this it sometimes helps to employ a 'correction' factor. Here we use

$$\dot{\theta}_i = \frac{R_i + a/2}{N_i + a}, \quad 0 \le a \le 1$$

as the input to $\hat{\phi}_i = r_i\dot{\theta}_i + (1-r_i)\hat{\theta}_\bullet$. Working with correction factors is as much art as science, so we leave the choice of a for the reader to explore on her own.

A national direct marketer commissioned Impiric to build a response model consisting of response rate by postal region. US Postal codes are grouped into contiguous geographic areas called Sectional Center Facilities (SCF). A single SCF consists of all the zip codes starting with the same three digits. Impiric's client has mail history covering 876 SCFs. Mail history is sparse in a number of SCFs and response rates in these locations are unstable. Overall response rate is 1.3414% on total mail quantity of 2,724,025.

A sample calculation on one SCF using the correction $a = 1$ should go like this:

Sample Calculation

#Responses	Mail Qty	Resp Rate	Corrected Resp Rate	Weight	Pullback Resp Rate
16	2,131	0.75%	0.77%	0.8994	0.83%

The calculations get a bit messy, but a spreadsheet program can be used to keep things straight.

An idea of how key quantities are distributed can be seen from their order statistics:

Order Statistics of Key Quantities
Full Data Set

	# Responses	Mail Qty	Resp Rate	Corrected Resp Rate	Weight	Pullback Resp Rate
Maximum	265	23,315	12.50%	16.67%	0.9816	10.15%
95th %ile	115	8,722	2.42%	2.51%	0.9316	2.30%
3rd Quartile	55	4,141	1.71%	1.76%	0.8074	1.60%
Median	32	2,135	1.36%	1.39%	0.5921	1.34%
1st Quartile	16	1,169	1.08%	1.12%	0.2486	1.21%
5th %ile	5	382	0.73%	0.77%	0.0117	0.85%
Minimum	0	8	0.00%	0.32%	0.0006	0.38%

Several features are noteworthy. First, the pullback estimates shrink in toward the overall response rate of 1.34%, and are clearly less variable than the naïve estimates. More than half the weights are greater than 0.59, so most final estimates are weighted more toward $\dot{\theta}_i$ than $\hat{\theta}_\bullet$. Finally, the correction is significant only when the mail quantity is quite small.

One way to view this method is to observe that there are two sources of information available about a specific market. One source is internal to the market, and the other is external across markets. The combination follows the simple rule of weighting roughly inversely proportional to the variances. The minimum MSE criterion replaces minimum variance, and the requirement of unbiasedness is abandoned. Astute readers will recognize empirical Bayes ideas in this approach.

In closing, we note that this method can be generalized to other marketing parameters. For example, one might wish to estimate price elasticity, or the effects of various types of promotion by local market. Regression parameters can be shrunk as well as simple means, although the calculations become more difficult. Have fun!

Steve Vail is Group Director of Analytics in the Database Marketing Group at Impiric (formerly WCJ). Steve's responsibilities include evaluation and implementation of new targeting and segmentation technologies, planning and development of modeling resources, and support for client modeling and analytics across Impiric's client roster. Prior to joining WCJ, Steve was an Associate at Mitchell Madison Group, a management consulting firm, where he worked primarily with clients in the financial services industry. Steve holds two degrees in Statistics, a Master of Science from the University of Washington and a Bachelor of Arts from Princeton.