Hitting the Target: Effective Techniques for Market Share and Response Rate Modeling Using Demographics

Charles Schwartz
Principal, Intelligent Analytical Services

Demographic target marketing has come of age. Market researchers have produced enough response rate studies, demographic market share investigations, and penetration analyses to have motivated even a television documentary on the subject[1]. One thing all these studies have in common is that they use demographic measures to predict a response that is expressed as a proportion (responses as a proportion of mailings, sales as a proportion of the market), and apply those results to geographic areas like zip codes or census tracts. Despite violating a host of the technique's assumptions, they often use regression analysis to reach their conclusions.

Regression analysis is a robust technique. Even when its assumptions are violated it can produce useful descriptive results. Target marketing studies are a case in point. Even the most straightforward regression based target market model can lead to increased sales even though it may be useless for statistical inference. Still, regression analysis makes certain assumptions about reality. The closer the reality being modeled corresponds to those assumptions, the more powerful the results of the model and the stronger its predictive ability. By looking at how target market models relate to the assumptions of regression, it is easy to come up with some simple techniques to enhance the results we can get from our analyses. This article will look at some issues relating to dependent variables. A future article will look at independent variables.

The Dependent Variable

Typically, target market models use a proportion as the dependent variable. Response rates by zip code are proportions. Market penetration by census tract and market share by ADI are both proportions. This is not strictly kosher from the standpoint of regression analysis for two reasons. First, regression imposes a linear and unbounded model on the data. Second, regression requires that the variance of the dependent variable be unrelated to its value.

---

1. A fascinating survey of the power of demographics produced for PBS' Nova series.

Linearity and Unboundedness

Linearity is simply the assumption that a difference in the dependent variable is constantly proportional to a difference in the independent variable. Unboundedness signifies only that a prediction from the model can have any value regardless of the nature of the variable to be predicted. These assumptions mean that if your model shows that a $5,000 difference in median income corresponds to a 10% difference in market share, the relationship will hold at all levels of market share and all levels of income. If predicted market share is 40% in an area with a median income of $25,000, it will be 50% in an area with a median income of $30,000, 90% in an area with a median income of $50,000 and 140% in an area with a median income of $75,000

The most apparent problem with these models is that their predictions have no bounds - they can assume meaningless values like 140% or -10%. One obvious solution is to code all predictions above 100% as 100 and all below zero percent as zero. Unfortunately, doing this will bias the predictions even though the model they come from may not itself be biased. This may be more clear if you think of market share as a probability - the probability that a resident of a market area will become a customer. A prediction of 100% implies that you are absolutely certain that the person will buy your product if the product is offered. How certain are you really? Common sense says that you should be more certain of a sale when the prediction is 140% than when the prediction is only 100%. Statistics reflect common sense. You can calculate statistical confidence intervals for the predictions which, in the first case, would show very little chance that actual market share would be under 100%, while in the second, a very strong chance that market share would be less than 100%. By coding both areas the same, 100%, you are implying that the chances for a certain sale in the two areas are equal. Not only that, by doing so you will be introducing error correlated with the actual values that would be predicted on the basis of the independent variables in the model. If the prediction is 140% you add 40 percentage points to the statistical error by coding the prediction as 100%, if it is 150% you add 50 and so on. This is the very definition of statistical bias.

There are ways to surmount these biases. Non-linear mathematical programming techniques for linear probability models subject to inequality constraints are the most direct, but recommended only for the most adept practitioner. Even if you brave their complexity, these techniques do not overcome another kind of bias. It is inherent in selecting and using geographic areas as the subject of your modeling. This bias is known as selection bias and can effect any non-random sample. It is particularly damaging to a linear probability model. Again

consider areas where your model will predict market shares over 100%. Some of these areas will have very good demographics and lead to predictions near 100%. Some will have demographics to kill for and lead to predictions well over the 100% mark. What is certain is that all of the areas will have actual market shares under 100%. Furthermore, given the quality of their demographics, the areas will be bunched closely together in the upper end of the market share distribution. This is no problem if two conditions are met: First, customers have the same access to the product in all areas of the market or, at the least, access is not related to demographics. Second, your study covers all current and potential markets or a representative sample of them. If your current distribution system was planned with demographics in mind or if you plan to use the study to evaluate new markets, one or both of these assumptions will be violated.

What happens when the assumptions are violated? The consequences are glaringly obvious for market areas where market share is over 95%. No matter how different the demographics between two areas in this group, they can differ in market share by only five percentage points. On the other hand, two areas with poorer over all demographics but with the same relative demographic difference between them can differ in market share by a far higher margin. In the previous median income example we saw that an area with an income of $55,000 would have a market share of 95%; if income were $25,000 higher, actual market share could increase at most five percentage points. On the other hand, if median income in the first area were $10,000 and the corresponding market share were 20%, a $25,000 increase in income would lead to a market share of 70%, a 50 percentage point increase. In essence, the effects of the independent variables become compressed as you reach the extremes of zero and 100% simply because they bump up against an arbitrary floor or ceiling imposed by the data. The result is that areas at the extremes of the range will have high variability in the demographic variables, but very little variability in market share.

What happens if your study does not cover areas with the very, very best demographics? Your study will be biased. Both market share and average demographics will be lower in your sample of areas than in the universe of areas. But the bias will be differential - the omitted areas could easily be much more extreme on demographics than on market share simply because of the compression effect described in the last paragraph. This will lower the apparent market share of the high market share areas taken as a group, but it will change the group's apparent demographics much more seriously. The result is a biased slope - the model will estimate too high or too low

depending on the demographics of the market areas that are left out - regardless of what method you use to deal with predictions above 100% or below 0%.

You should note that this is a particular difficulty for linear probability models (no matter how they are estimated) as compared to other regression models. If market share had no ceiling, the omitted areas might have market shares of 140% or 200% instead of being limited to a 100% maximum. Both market share and demographics would have the same degree of downward bias if these areas were omitted. While this would bias their absolute levels, it would preserve the relationship between the variables. In this case, unlike our previous example, a $10,000 increase in income would lead to a ten point increase in the actual value of the dependent variable in any and all areas - and that is what your regression would show regardless of which areas were omitted. There may be substantive reasons that would make this untrue and bias your results, but they would not be biased for simple arithmetic reasons as they would be if your dependent variable had a 100% ceiling.

All this is not to say that substantive issues do not play a role in biasing probability models. They do - and generally in a way that will intensify the effects of selection bias. This is related to problems in cracking new markets and saturating old ones. Often the hardest market share points to gain are the five from zero to five percent and the five from 95 to 100 percent. In the first case you are entering a new market, you have little visibility and no word of mouth. There might be a threshold level of demographics you need to succeed. Other things being equal, you will need more "push" from your independent variables to go from nothing to five percent in a new or below threshold market than you will need to go from 20 percent to 25 percent in a market where you are well established and have proven drawing power for the population. Likewise at 95%, to be extreme, your market is saturated. Regardless of favorability, you may not be able to get that last five percent. Demographics or anything else will make little difference. Note that this substantive bias goes in the same direction as the selection bias - both cause the "distance" between extremely high or extremely low percentages to appear much greater than the distance between moderate percentages like 25 and 30 percent. This is well known in the statistical literature which suggests that things will be essentially linear between 25 percent and 75 percent, distances will be equal in other words, but that beyond those levels bias will cause distances to stretch more and more, until as they approach 100 percent or zero they appear infinite.

The same kind of problems appear in the analysis of contingency tables when looking at such factors as product preference. The solution there is the well known logistic model. That model treats a probability as an odds ratio - the probability of someone choosing brand A divided by the probability of choosing any of the competing brands. It then takes the logarithm of that ratio.[2] With some algebra you would be able see that this is a direct transformation of the market share or preference value itself - a transformation with just the qualities we want. It makes distances between percentages linear at moderate levels but stretches out the distance between percentages as they become more and more extreme. In effect it takes care of both the substantive and arithmetic biases of the linear probability model by reflecting the true nature of the response of a probability, such as market share, to independent predictive variables. By reflecting the nature of the dependent variable, models will predict better, especially with a less than ideal sample.

There is little in the literature about the application of logistic models to geodemographic market share data. I have found, however, that a simple adaptation of the grouped logistic model can be used and usually results in more robust, more predictive demographic models. The procedure is simple. I will use market penetration to illustrate. For each area, such as a zip code, obtain information on the population that is able or likely to purchase the product and the number of actual customers. Rather than computing the penetration, compute the logit with suitable corrections. That is:

$$\text{Logit} = \ln [(\text{Customers} + .5) / (\text{Population} - \text{Customers} + .5)].$$

Then, using proper weights (to be discussed in the next section), do a linear regression using the logit. To obtain a stable model, leave out areas with no customers or even just one or two. In those areas, the logit will be a function only of population.

---

2. That is, ln[P / (1 - P)]. Many books cover this model. See, for example, Pindyck, E.S. and D.L. Rubinfeld, **Econometric Models and Forecasts**. New York: McGraw Hill, 1976.

Since differences between logits are not very intuitive, you should assess the relative importance of the independent variables in your model by comparing their standardized regression coefficients rather than by looking at the raw coefficients. Standardized coefficients express the relationship between two variables in standard deviation units - that is, the expected change in standard deviations of the dependent variable per standard deviation of the independent variable. They can be compared to one another with the same kind of interpretation you would use with partial correlations. In fact, if there is only one independent variable, the standardized coefficient will be the correlation between the dependent and independent variables. Most statistical software packages offer the standardized coefficient as an option.

Once you are satisfied with the model, obtain predicted values for the logits and convert them to penetration by the formula:

$$\text{Penetration} = \exp(\text{logit}) / [\ 1 + \exp(\text{logit})\ ].$$

I have found that using these formulas for prediction results in lower root mean square error (RMSE), higher R^2, and less sensitivity to outliers than linear probability models. In particular, linear models appear extremely sensitive to outliers when compared to logistic models. In one notable case, I found a difference in R^2 of .20 when I left out one census tract from a universe of 2,000 in a linear market share model. That tract happened to be a large Marine base whose population was ineligible for the service being offered. There was no similar difference when the model was recalculated using logits.

The improvement over the linear model can be tested. Randomly divide your market into two parts. Produce models using one part and evaluate them by deriving predicted penetration rates for the other. Even when the linear model obtains a higher R^2 in the model population, the logistic model, after being converted to penetration, will typically show an equal or higher R^2 and lower RMSE than the linear probability model in the test group. Another procedure for comparatively evaluating the models is to estimate both the linear and transformed model on the entire sample and use Box and Cox's procedure for evaluating transformations. Recipes for the procedure are in many

books but a good one is by Maddala[3]. If you do compute linear and logistic models on your entire sample remember that you cannot compare them directly in terms of R^2 or RMSE. You must use a procedure like Box and Cox.

Variance

A basic assumption of regression analysis goes by the term homoscedasticity. That means only that the random error variance in the variable to be predicted is not correlated with its value. In other words, predictions are equally reliable whether their value is ten or ten thousand - the amount of error will be the same at either level. Violating the assumption will not bias your estimates or make them inconsistent, but it will make them inefficient. Coefficients that should be significant will appear insignificant and confidence intervals for predictions will be wider than they need to be given the data.

Correcting this problem is a simple matter if the variance is known. Use the reciprocal of the known variance to weight the observations in your regression analysis. Most statistical programs have a simple mechanism for doing just this.

But is the variance known? In geodemographic models it almost always is. In elementary statistics, we all learned that given a sufficient sample size the variance of a proportion is pq / n, where p is the proportion, q is 1 - p, and n is the sample size on which the proportion is based. In a geodemographic study, each area, such as a zip code, is both a single observation and a sample in its own right. The market share or penetration variable corresponds to p, and its denominator, population or market size within the geographic unit, is n. To attain more efficient estimates, weight each observation by the reciprocal of this variance or (for a penetration model):

$$1 / [\text{Penetration} * (1 - \text{Penetration}) / \text{Population}]$$

where penetration is coded as a proportion between zero and one and population is the population that is eligible or likely to purchase the product or service. It is a straightforward matter to adapt the equation to market share as well as penetration. The variance of a logit is also well known. For a logistic model weight each observation by:

---

3. Maddala, G.S., **Econometrics**. New York: McGraw Hill, 1977. Pp. 315-317.

1 / { [(Population + 1)*(Population + 2)] / [Population * (Customers + 1) * (Population - Customers + 1)] }.

These variance formulas represent two obvious substantive truths. First, observations will be more reliable, they will persist over time, if they are based on large populations. A penetration rate based on 5 people will change 20 percentage points if one person becomes a customer; based on 1000 people, it will change one tenth of a percentage point. Second, the more extreme the proportion, the less room it has to vary. We saw this in our discussion of linearity. By using these variance weights, we are implicitly saying that we will pay the most attention to our most reliable observations when we draw conclusions. In practice, I often find many of the smaller zip codes or census tracts to be outliers. Including them without weighting can, and often does, lead to high R^2 models that make absolutely no substantive sense and often lead to nonsensical predictions. Weighting will tend to lower your R^2, but in many cases will increase the predictive value of the models you produce.

A Final Word

Geodemographic market share and penetration models have proven their worth in the marketing research discipline. Compared to other methods, these regression models are cheap, easy to calculate and they get results. The data they use are readily available and the analysis programs are legion and easy to use. It is in the power of these models that their danger lies. From a statistical standpoint, geodemographics are not straightforward. They tend to be non-linear and demand that the analyst use procedures such as those I have just described. The independent variables themselves present difficulties. They are highly correlated to one another, and often dependent on geographic distances. That too is controllable, but the subject of another article.

It is possible to ignore these issues and still develop usable models if you pay very close attention to the substantive meaning of the results and make extensive use of regression diagnostics to identify outliers and overly influential data points. On the other hand, with proper mathematical transformations and weighting you will reduce the number of outliers and the influence of the odd zip code or census tract. It will be easier to diagnose your models and you will ultimately produce simpler models than you could with linear regression. More important, you will use more of the information your data offer. By making your models fit the reality of the data more closely, your predictions will be more accurate, more efficient and, most important, more valuable to those who use them.