Population marginal means

Terry Therneau

September 28, 2023

1 PMM

This is an attempt to explain ”population marginal means” (PMM), also called ”marginal” estimates, also called ”g-estimation”, also called by at least a dozen other names over the history of our profession. Underneath the concept are the two grand notions of statistics: balance with respect to confounders, and put $E()$ around everything. (A favorite quote of mine is that ”statistics is the art of clever averaging”.) PMM use both of these fundamental ideas, so it is no surprise that it keeps being re-invented. A fascinating (to me) review paper by Keiding and Clayton (2014) traces a particular one of these, comparison of death rates, back for over 300 years, including at least a dozen rediscoveries.

In any case, assume that we have a study with four factors (covariates) that we are interested in, call them Z, X1, X2, and X3, with Z the one we are most interested in at the moment, and say Z has three levels of low, moderate and high. To get the best estimate of the effect of Z, ignoring all others, one would do a balanced trial, one where the levels of Z had exactly the same distribution of X1, X2, and X3 in each group. We could then compute an “overall” effect for Z as a simple average. That is,

1. Fit your favorite model, with all 4 covariates
2. Get the n predicted values $\hat{y}_i$
3. Compute the mean predicted $\hat{y}$, for each of the 3 levels of Z.

The predictions can be simple or complex. For example a linear model plus the simple linear predictor, $\sum_j x_{ij}\hat{\beta}_j$, or a Cox model fit plus all n predicted survival curves. Such an estimate is called a *marginal* estimate, since you are taking an average. The name comes from thinking about the row or column margins of a table.

Of course, we don’t very often get data from a nicely balanced trial. This leads to the following variation of the algorithm.

1. Fit your favorite model, with all 4 covariates
2. Create a balanced target population, i.e., a data set that has m observations, such that each level of Z in that data has exactly the same distribution for X1, X2, X3. Usually m is bigger than n, often by a lot.
3. Get the m predicted values $\hat{y}_i$
4. Compute the mean predicted $\hat{y}$, for each of the 3 levels of Z.

I call this a population marginal mean. What you get depends first of all on the distribution (or population) you choose for X1, X2, and X3. There are 3 main ones in use. In all cases we are mimicing the balanced clinicl trial that we wanted to do, but could not.

- The data set in hand. This leads to a prediction population which is a data set with 3n observations: The first n rows are a copy of X1, X2, X3 as found in the data, all with Z=low, the next n rows have Z=medium (repeat X1, X2, X3), and the last n have Z = high.
- A known population. This is mostly used in epidemiologic studies where X = age, sex, race; and from a national census we know the overall population of the country wrt those three factors. Create a population data set with m rows along with case weights, so that this synthetic population represents the nation well, e.g., every possible combination of age/sex/race along with the fraction. Then make a larger data set with 3m rows, get the 3m predictions, etc.
- Factorial. This goes back to a 1934 paper by Yates, which is the era when agricultural trials were informing the development of statistics. Make a population data set that corresponds to the ideal agricultural trial. For X1, X2, X3 = sex, apoe, CMC this will lead to 2*2*8 = 32 unique combinations, and our population data set will have 3*32 rows. SAS calls the results a "least squares mean" estimate, and the test for equality of LSM a type III test.

I think that most would agree that number 2 is best, though it is only rarely possible outside of epidemiology and survey sampling. Number 1 has been rediscovered by the causal modeling crowd, who call it a g-estimator. Number 3 used to dominate but thankfully is fading. One problem is that the people who use 1, 2, and 3 don't read each other's literature and so don't see the commonality. Number 1 people think they have invented something new, and type III seem to think it came down from Mt Sinai — ignoring the fact that it is often a dumb estimate. (Can we imagine an actual population where all 32 combinations are equally likely?)

2 Problems

One of the long term problems with PMM estimates comes from another time honored traditions in statistics, which is to find a short, efficient shortcut calculation for anything that is interesting. The problem comes when we forget about the big picture and talk about the shortcut as though it were the target. Three examples from linear models

- When doing ANOVA calculations pre-computer, an intermediate quantity called the "mean square" was an important way station in the computation. MS actually have no meaning, they are just an intermediate sum, but for 40 years any paper that reported an anova result needed to include them in the output table.
- When using a factorial population, there is a shortcut way to test for (PMM for low amyloid = PMM for moderate = PMM for high) by removing selected rows/cols from X'X followed by a Cholesky decomposition. The algorithm is both very clever and completely opaque. SAS calls this a "type 3" test and has enshrined it. Not one statistician in 1000 knows what it really is, including those who work for SAS.
- If there are no interactions, then (PMM moderate - PMM high) = the difference between the coefficient for moderate and high. No derivations are needed, you can read things right off the page.

The PMM algorithm that always works is brute force: calculate the whole stack of predicted values and then take an average. For standard errors use a jackknife or bootstrap. With modern computing power this is not a serious impediment.

3 Application

All this brings me to Dr. Jack's paper. For the lifetime risk we have to use the brute force method. For the 12 relative risk estimates of ND:dementia we can average over a subpopulation (we can ignore CMC and education). The hazard estimates that do not involve an interaction can be read directly from the Cox model coefficients. In my opinion, there is not sufficient reason to reprise the discussion found in this document for the medical paper, i.e., to tell them that we computed things in 3 different ways. I welcome discussion of this last point, though.