ONLINE APPENDIX

Why the 'margins' Approach in Stata Does Not Sufficiently Solve the Weakness of the Simulation Method

Soon after Zelner (2009) was written, Stata introduced in July 2009 an updated version (version 11) of its software which superseded the simulation approach.[1] Previously, Stata did not have the functionality to automate both the calculation of interaction terms and their marginal effects. This means that one had to do the hundreds of lines of Stata programming that went behind the simulation method. Starting with July 2009, Stata began to offer such functionality with its 'margins' command.

The 'margins' command just takes the logistic regression and transforms it using the delta method (Wooldridge, 2010, p. 47) to get the change in probability. The simulation approach is supposed to give the same answer as what the simple 'margins' command produces. If there should be any difference in the results between the simulation method and the 'margins' command, it will be due to the randomness of the simulation approach and its simulated data. Yet this randomness should be controlled to be a very small randomness that does not affect the final output. With Stata's 'margins' command, one no longer needs to use user-written commands such as 'CLARIFY' and 'intgph' (Tomz, Wittenberg, & King., 2003; Zelner, 2009) for assessing interaction terms in nonlinear models anymore. This enables one to avoid many steps in the simulation that can go wrong because one has to make numerous assumptions for any simulation model behind the scenes.

Although the 'margins' command has an advantage over user-written commands for the simulation approach for its substantially improved simplicity, it should be noted that the 'margins' approach also needs to make highly consequential assumptions about the values of the covariates just like the simulation method does. In particular, what the 'margins' command does is to produce mean predicted probabilities, calculated across the observations in the estimation sample and subject to the

[1] https://www.stata.com/support/faqs/resources/history-of-stata/

'at' option in the 'margins' command where one can fix certain covariates at a chosen value or a set of values, and subject to integrating across all the other control variables (which will be explained further below). At this point, the 'margins' approach could have theoretically been designed to do either of two things: to assume the covariate values at their means or to integrate over them. The problem with assuming covariate values at their means is that often one has a series of categorical variables as controls. Yet it would be meaningless to have those categorical variables take on a value of 0.5, for example. In the case of an individual-level sample controlling for the often-important categorical variable of gender, for example, it would be simply erroneous to treat the entire sample as being the new entity of half-male/half-female. The problem with assuming categorical variables at their mean is that there is no such thing as half-and-half and thus the sample does not represent anyone. As a result, the output in this case would not generalize to the general population.

If one were to instead pursue the alternative of calculating margins by integrating over the control variables, which is the default for the 'margins' command, this also does not help get interaction results that can provide inference about the larger population of interest. Integrating means taking the original sample data and taking a group within the data like gender, which has a binary distribution, and assuming that the distribution in the sample in terms of, for example, gender is the same as the population distribution of interest. So if one looks at the whole population of interest, what the 'margins' command will produce will be the change in probability given the proportion of male and female staying the same in the general population as was in the sample. But as soon as the proportion of male and female in the general population turns out to be different, the results from the 'margins' command cannot be used to make inference to the general population. In fact, the potential for problems is yet more severe when we consider that the 'margins' command assumes that the distribution for every single continuous as well as categorical variable in the sample is the same as in the general population of interest. Note that even when just one assumption for one control variable

is wrong, there can be erroneous inference about the general population of interest. Yet quite often in applied research, a researcher does not know for sure whether the distribution of every covariate is the same as the distribution in the population of interest. And all it takes is for at least one assumption to be wrong for the study's conclusion to be perhaps wrong.

In the field of statistics, there is a commonly shared desire to use methods that deliver answers that are robust to alternative assumptions—or that do not depend on one perhaps questionable assumption. When focusing on the log odds ratio, one does not require any assumption about the proportion of male and female in the general population. In contrast, the 'margins' command, like the simulation method, requires one to have the correct assumption about the distribution of every subgroup in the data for the output to be generalizable. One would prefer a method that does not require this assumption about the distribution of every subgroup in the data, and we discuss below why this is the case using two illustrative examples.

In numerous real-world instances, there is something unknown about the true distribution proportions in the population of interest. One good illustrative example was the 2016 U.S. presidential election results. Nearly all the experts doing fine-grained statistical analysis predicting the election results on the day of the election based on polling and prior distributions of groups in the electorate were surprised that Trump got elected as president, but in statistical terms (a less than 80,000 vote difference in the three tipping states of Wisconsin, Michigan, and Pennsylvania), it was because of a small difference in the distribution between polled samples and actual voter population. There was very little bias in terms of the sample compared to the population. Yet the small difference in distribution results in a meaningful difference in terms of the final predicted outcome. But if you do a logistic regression and look at only the log odds ratio, then this small difference in proportions will not affect the log odds ratio answer and its generalizability.

To take another illustrative example, in the total population, one conjectures that the number of U.S. children exposed to lead would be much smaller than that of U.S. children not exposed to lead. To study the effect of exposure to lead on child development by gender, consider that one collected a sample whose ratio of lead exposure to non-exposure was 1:2. If in the whole true population the same ratio were 1:1,000, then the researcher using the 'margins' command on gender in a nonlinear setting would see a biased result (in terms of the difference in probability) that cannot be generalized to the true population. One may then wonder if the 'margins' command would generate an unbiased result should one collect a sample whose ratio of lead exposure to non-exposure is 1:1000, exactly the same as the true population. Theoretically one could, but practically it would not be feasible to collect a far greater number of samples to achieve the same testing power of the sampling design of 1:2. It would not be viable particularly when the study has a limited budget and acquiring a sample is costly. In contrast, examining the effect of exposure to lead on child development by gender using the log odds ratio will provide a consistent and robust answer. In fact, using the log odds ratio will provide a consistent and robust answer regardless of what the proportion turns out to be in the general population for the number of children exposed to lead relative to those not exposed to lead. In summary, these illustrative examples help show why one would prefer a method that does not rely on knowing every subgroup proportion in the general population to a method that critically relies on having the correct assumptions about every subgroup proportion.

REFERENCES

Tomz, M., Wittenberg, J., & King, G. (2003). CLARIFY: Software for interpreting and presenting statistical results, version 2.1. Stanford University, University of Wisconsin, and Harvard University. Available at http://gking.harvard.edu/

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data* (2nd ed.). Cambridge, MA: MIT Press.

Zelner, B. A. (2009). Using simulation to interpret results from logit, probit, and other nonlinear models. *Strategic Management Journal, 30*(12), 1335–1348.