Predicted Probabilities and Inference with Multinomial Logit

Philip Paolino

Department of Political Science, University of North Texas, Denton, TX, USA. Email: paolino@unt.edu

Abstract

Multinomial logit (MNL) differs from many other econometric methods because it estimates the effects of variables upon nominal, not ordered outcomes. One consequence of this is that the estimated coefficients vary depending upon a researcher's decision about the choice of a reference, or "baseline," outcome. Most researchers realize this in principle, but many focus upon the statistical significance of MNL coefficients for inference in the same way that they use the coefficients from models with ordered dependent variables. In some instances, this leads researchers to report statistics that do not reflect the correct quantities of interest and reach flawed conclusions. In this note, I argue that researchers need to orient their approach to analyzing both the substantive and statistical significance of predicted probabilities of interest that match their research questions.

Keywords: multinomial logit, qualitative choice, maximum likelihood

1 Introduction

Multinomial logit (MNL) remains a common approach for researchers estimating models with nominal outcomes. Twenty-six papers published in *The American Political Science Review*, *The American Journal of Political Science*, and *The Journal of Politics* from 2014 to 2018 used MNL for some part of the analysis, compared to four that used multinomial probit. The vast majority of these papers come from American and comparative politics. Only two of these papers focus on international relations, but MNL is used extensively in international relations research. The corresponding numbers for the top IR journals, *International Organization*, *International Studies Quarterly*, and *The Journal of Conflict Resolution*, are 35 that used MNL and 3 that used multinomial probit.[1]

There is notable variation in the analysis and presentation of MNL results. Articles regularly present predicted probabilities of particular outcomes, but many base inferences on the statistical significance of MNL coefficients, even though the coefficients only capture the effect of observing an option relative to a "baseline" (or "reference") outcome. Fourteen of the 24 articles in the top 3 general journals and 15 of the 26 articles in the top IR journals using MNL for primary analysis base inferences on MNL coefficients. Researchers should recognize that analysis based upon the statistical significance of these coefficients can lead them to report nonsignificant effects as significant and to overlook significant effects. I argue that researchers using MNL should match the quantities analyzed with their question of interest and focus upon the statistical, as well as substantive, significance of predicted probabilities.

In this brief note, I outline the problem with relying upon MNL coefficients for inference. While the survey of the top general journals suggests that MNL is used (and sometimes misinterpreted) most commonly in American and comparative politics, my examples come from IR research to demonstrate how proper interpretation of MNL results concerns scholars in all quantitative subfields of political science. I present one example of how reliance upon MNL

1 The supplemental appendix provides a summary table of the articles in APSR, AJPS, JOP, IO, JCR, and ISQ from 2014 to 2018 that used MNL in any part of their analysis.

Political Analysis (2021)
vol. 29: 416–421
DOI: 10.1017/pan.2020.35

Published
16 November 2020

Corresponding author
Philip Paolino

Edited by
Lonna Atkeson

coefficients can lead researchers to overlook important results. I follow with another example of how researchers focusing upon significant MNL coefficients to draw inferences about the effects of covariates upon the relative probability of observing an outcome against a specific baseline can also lead researchers to misinterpret results. I conclude with recommendations for researchers using MNL.

2 Analyzing MNL Results

The fundamental aspects of MNL are covered elsewhere (e.g., Greene, 2012), and I refer to them only as needed in order to focus upon matters that arise in analyzing results, but a critical aspect is that a covariate's effect upon the probability of observing each outcome is a function of all of the estimated coefficients. Most researchers seem to comprehend this by reporting changes in the predicted probabilities, but many papers ignore this when evaluating statistical significance. A central part of analysis and presentation of MNL results should be the calculation of substantively meaningful differences in predicted probabilities and their standard errors, using either simulation (e.g., Tomz, Whittenberg, and King, 2003) or approximation (e.g., Fox and Andersen, 2006).

MNL coefficients are unreliable for assessing statistical significance because the coefficients depend upon the choice of a "baseline" outcome that determines which specific log odds ratio is estimated. Choose a different baseline and some set of coefficients and standard errors will change.[2] The standard errors of MNL coefficients also vary with different baselines because they are related to the number of observations in the two relevant categories. A baseline category with relatively few observations can produce elevated standard errors for all coefficients. The choice of baseline, however, *does not* affect predicted probabilities and their standard errors.

To see the difference, consider $Y = \{A, B, C\}$, where Y is the dependent variable and A, B, C are the possible outcomes, and $x_t = 1$ and $x_c = 0$ refer to treatment and control conditions whose effect on Y is examined. If $Y = A$ is the baseline, then the MNL coefficient for B is simply:

$$\ln\left[\frac{\Pr(B|X = x_t)/\Pr(A|X = x_t)}{\Pr(B|X = x_c)/\Pr(A|X = x_c)}\right]. \tag{1}$$

This might be useful if the relevant question is the treatment's effect upon $\Pr(B)$ relative to $\Pr(A)$, but not if the relevant question is the treatment's effect upon $\Pr(Y = B) = \Pr(B|X = x_t) - \Pr(B|X = x_c)$. The problem with using Equation (1) to evaluate the effect of a treatment upon $\Pr(B)$ is that a substantively or statistically significant change in the ratio can be entirely a function of a significant change in $\Pr(A)$. Conversely, a significant change in $\Pr(B)$ can be overlooked if the treatment has a similar effect upon $\Pr(A)$ or even a considerable effect upon $\Pr(B)$ that is proportionally equal to a very small effect upon $\Pr(A)$. For example, if the treatment changes $\Pr(B)$ from .25 to .50, while changing $\Pr(A)$ from .01 to .02, the MNL coefficient would be zero, despite the considerable effect of the treatment upon $\Pr(B)$. A coefficient and its standard error do not necessarily provide any information about the substantive or statistical significance of the covariate's effect upon the expected probability of particular outcomes.

2.1 Example 1: Problems with Focusing upon Coefficients

An example where focusing upon MNL coefficients can lead researchers to overlook significant effects that affect their conclusions comes from Gelpi (2017). Gelpi (2017) uses MNL to evaluate the effect of information from events and messages upon individuals' beliefs about the war in Iraq. Participants in an experiment receiving different events and cues information placed on a scale their opinions regarding whether the troop surge was a success, whether the US would succeed

2 One set of coefficients changes only with respect to sign and has the same standard errors.

in Iraq, and whether there should be a timetable for the withdrawal of troops.[3] Gelpi (2017, 1824) states plainly that he relies upon the statistical significance of the MNL coefficients for hypothesis testing, writing "multinomial logit coefficients can tell us which treatment effects are statistically significant" and presenting changes in predicted probabilities primarily for treatments with statistically significant MNL coefficients, though without confidence intervals for those changes.

Critically, the question of interest concerns the treatments' effects upon respondents' probabilities of choosing a category that reflects their attitudes toward the war. The significance of the coefficients cannot provide a meaningful value for the quantity of interest because none of the options provides an obvious baseline for comparing the other options. Gelpi (2017) uses "somewhat approve" as a baseline for analyzing attitudes toward troop withdrawal in Iraq, but this is no more relevant than "somewhat disapprove." The experimental research design, however, provides obvious values for the other variables when calculating the predicted effects of each treatment. Gelpi sets the values of all other treatments to zero, indicating no exposure to another stimulus.

After concluding his analysis, Gelpi (2017, 1832) writes, "the results reveal little evidence of any impact for elite opinion at all in this experiment," as the only instance where the MNL coefficient for elite opinion had a statistically significant effect in a sensibly interpretable direction was when respondents who "strongly approved" of President Bush exposed to a cautious message from Bush were more likely to "somewhat disapprove" of a timetable for withdrawal. Gelpi's (2017) focus upon the statistical significance of MNL coefficients, however, overlooks a significant effect of this message upon the attitudes of respondents who "somewhat approved" of President Bush. Evaluation of the predicted probabilities shows that these respondents had a .177 (se = .073) lower predicted probability of "strongly disapproving" and a .157 (se = .076) greater probability of "somewhat disapproving" of a timetable than those not exposed to a cautious message (Table 1). This means that the cautious message influenced respondents' attitudes about a timetable in both cases where Gelpi's (2017) "surprising opinions" hypothesis predicts they would and with substantively greater effects than from "surprising events." From this, we might instead conclude that elite opinion had greater influence than events upon participants' attitudes toward withdrawal. The supplemental results also provide analysis that suggests revising another conclusion about the effect of positive events upon attitudes toward the likelihood of success in Iraq.

2.2 Example 2: Problems with Interpretation Against the Baseline

While many researchers are concerned with the effect of a covariate upon the change in the probability of particular outcomes, others choose a specific baseline because their question concerns the change in the occurrence of one outcome rather than another. Researchers interested in relative odds may seem on safer ground using coefficients for inference, but even these researchers should calculate changes in the predicted probabilities of relevant outcomes in order to understand better the changes in the underlying probabilities. A significant log odds ratio resulting from a large change in the probability of only one outcome may produce a substantively different interpretation than one where both probabilities change significantly.

Greenhill and Oppenheim (2017) examine the effect of trust upon acceptance of rumors in conflict zones. They use MNL to assess the hypothesis that "as distrust of the implicated entity rises, the more likely it is that a rumor will be perceived as possibly or definitely true" against a baseline of disbelief (663). A similar hypothesis relates to threat of conflict. Greenhill and Oppenheim (2017, 668) present risk ratios, "which estimate the relative risk of being in the neutral (agnostic) or receptive (belief) category, relative to the baseline state of disbelief." Like

3 Ordered logit is often used for these dependent variables, but Gelpi makes a valid argument that using MNL avoids potential violations of the proportional odds assumption.

Table 1. Positive events, cautious cues, and a timetable for withdrawal in Iraq.

	Approval of President Bush			
Timetable	Strongly disapprove	Somewhat disapprove	Somewhat approve	Strongly approve
		Positive events treatment		
Strongly approve	.060 (.056)	−.058 (.080)	−.001 (.028)	−.063 (.053)
Somewhat approve	−.117 (.057)	−.120 (.078)	−.013 (.051)	.103 (.100)
Somewhat disapprove	−.011 (.044)	.083 (.099)	−.089 (.053)	−.047 (.037)
Strongly disapprove	.068 (.034)	.095 (.106)	.104 (.070)	.007 (.113)
		Cautious Bush treatment		
Strongly approve	−.034 (.052)	.008 (.088)	.006 (.024)	.025 (.089)
Somewhat approve	−.036 (.059)	.142 (.109)	.014 (.051)	−.036 (.037)
Somewhat disapprove	.081 (.054)	−.075 (.065)	.157 (.076)	.294 (.129)
Strongly disapprove	−.011 (.017)	−.076 (.089)	−.177 (.073)	−.283 (.130)
Observations	450	145	281	118

Note: Entries are the change in the predicted probability (and standard error) of each response for a respondent receiving the "positive events" or "cautious Bush" treatment compared with a respondent who did not receive any treatment.

Greenhill and Oppenheim (2017), researchers often argue that risk ratios have a more intuitive interpretation, as the percentage change in probabilities of one outcome relative to the baseline. Risk ratios, however, are only transformations of MNL coefficients and have similar problems with interpretation.

In Greenhill and Oppenheim's (2017) analysis, the coefficient for the effect of distrust or threat is often significant for either none or only one of the response categories and reflects the statistical significance of the marginal effect of the predicted probabilities; although, often at a lower level of significance. In two cases, the coefficients for distrust or threat are statistically significant for both response categories, and the authors interpret these instances as indicating higher odds of being in both categories compared to the baseline: "threat perception increase[s] the odds of being in both the agnostic and receptive categories" (668).

This interpretation of the risk ratio, however, can be misleading. Threat perceptions increase both the probabilities of a rumor of a Thai coup as being plausible and being accepted; although, at levels of statistical significance below the $p < .01$ reported for the coefficients and with little change in the probability of accepting the rumor (Figure 1a). The marginal effect of accepting the rumor at the lowest and highest levels of threat of conflict is .018 (se = .009), $p = .057$. Both coefficients for the effect of distrust in the Philippines upon beliefs about a rumor concerning government corruption are statistically significant, but examination of changes in the predicted probabilities reveals that there is no change in accepting the rumor is true given changes in respondents' trust in local officials (Figure 1b). The increase in the odds that people accept the rumor relative to denying the rumor as distrust increases is entirely a function of the decreasing probability that respondents deny the rumor. The difference of accepting the rumor at the lowest and highest levels of distrust is .004 (se = .052).

These examples illustrate an important point. Interpretations of odds against a baseline often imply that a significant coefficient indicates a change in the probabilities of *both* alternatives, but the change in the predicted probability of one alternative with a significant coefficient may be no different than the change in the predicted probability with a nonsignificant coefficient. Because we often rely upon significance levels for testing hypotheses, these cases demonstrate the importance of examining changes in predicted probabilities, even when researchers are interested only in the effect of a covariate upon the odds of one category against a baseline.

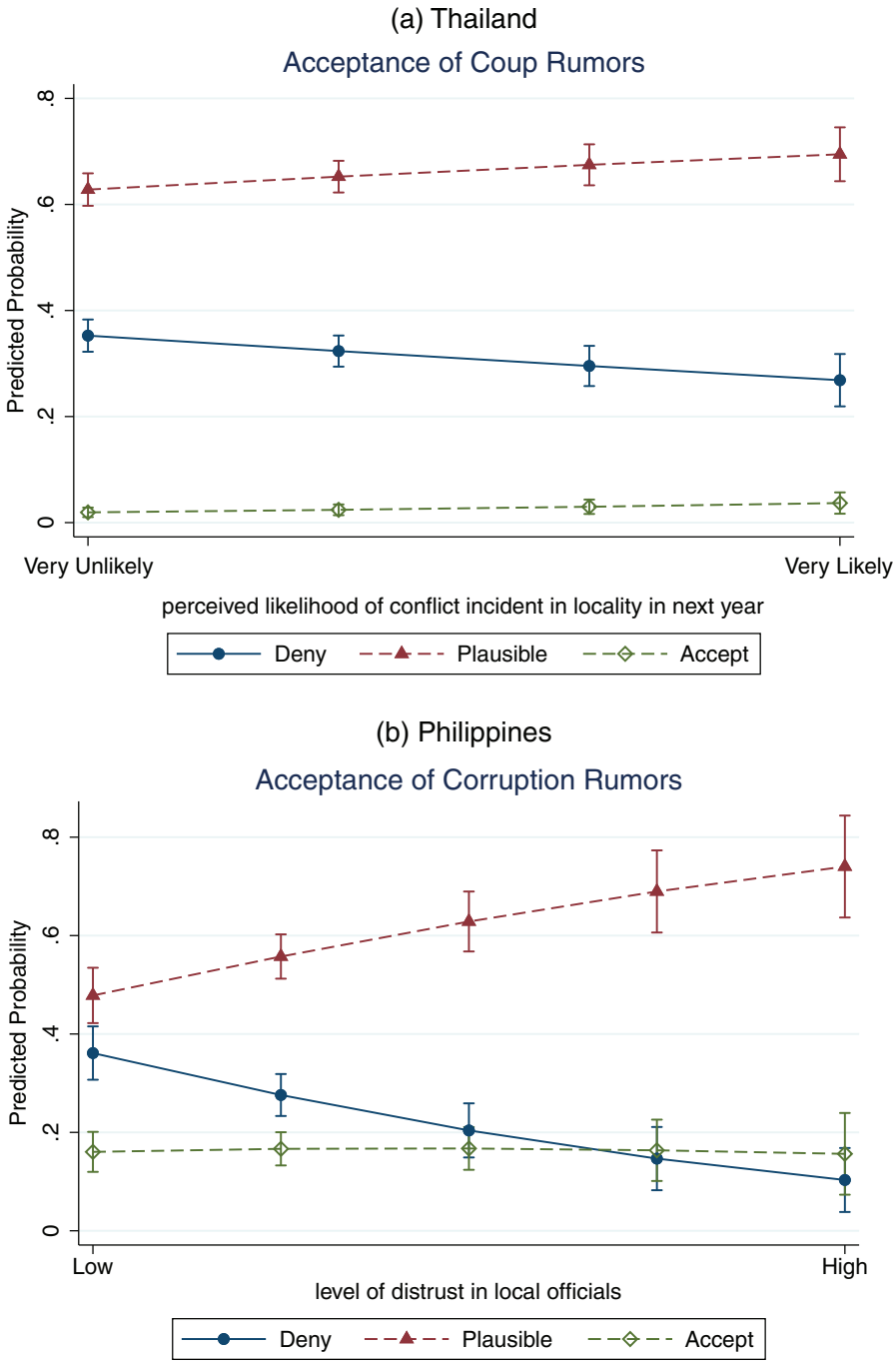

Figure 1. Interpretations against baseline. Note: Figures show the predicted probabilities of each response for respondents, with 95% confidence intervals for those probabilities.

3 Conclusion

Most researchers understand the basic points in this note, but do not always apply them. Researchers should think carefully about their primary quantities of interest when conducting analysis and recognize what specific outputs of their analysis can actually tell them. A focus upon the statistical significance of predicted probabilities and marginal effects provides several benefits. First, even researchers who present standard errors or confidence intervals for predicted probabilities or marginal effects often do so only after basing their conclusions upon the significance of MNL coefficients. Second, reviewers evaluating papers should reasonably question inferences based solely upon MNL coefficients. Inferences based upon the predicted probabilities

or marginal effects can be consistent with those drawn from MNL coefficients, but absent the information about precision of the predicted probabilities, reviewers cannot necessarily distinguish valid inferences from invalid ones. As political methodologists, we should encourage the use of methods that most accurately represent statistical relationships of interest. Finally, there is a practical matter that researchers often present lengthy tables of MNL coefficients, even when the analysis only addresses one or two covariates. Presenting those details in a supplemental appendix and drawing readers' attention in the main text only to the important quantities of interest would improve the communication of one's results.

The survey of articles using MNL in the three major general and three major IR journals indicates that researchers are becoming more attentive to these matters, but too many articles still overlook these points. It is worth mentioning that as researchers pay more attention to the statistical significance of predicted probabilities, significance levels can vary depending upon the values of the explanatory variables used to generate predicted probabilities. Hanmer and Kalkan (2013) show that differences in predicted probabilities and their standard errors can change when researchers hold other variables at the values observed in the data rather than at their means. People fixated on specific levels of statistical significance may find this disconcerting, but any problems of inference arising from these differences are likely to be much smaller than those that can arise from basing inferences on different baselines and nonrelevant quantities of interest. This concern notwithstanding, this note provides a friendly reminder to researchers to consider these larger issues when analyzing and presenting MNL results.

Acknowledgments

The author would like to thank Bora Jeong for her valuable research assistance and the anonymous reviewers at *Political Analysis* for their very useful suggestions.

Data Availability

Replication materials are provided in Paolino (2020).

Conflicts of Interest

There are no conflicts of interest to disclose.

Supplementary Materials

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2020.35.

Bibliography

Fox, J., and R. Andersen. 2006. "Effect Displays for Multinomial and Proportional-odds Logit Models." *Sociological Methodology* 36(1):225–255.

Gelpi, C. 2017. "The Surprising Robustness of Surprising Events: A Response to a Critique of "Performing on Cue"." *Journal of Conflict Resolution* 61(8):1816–1834. https://u.osu.edu/gelpi.10/files/2016/08/GelpiPOCReplication-2clvxdn.zip.

Greene, W. 2012. *Econometric Analysis*. 7th edn. Upper Saddle River, NJ: Prentice Hall.

Greenhill, K. M., and B. Oppenheim. 2017. "Rumor Has It: The Adoption of Unverified Information in Conflict Zones." *International Studies Quarterly* 61(3):660–676. https://doi.org/10.7910/DVN/KMRVWY, Harvard Dataverse, V1.

Hanmer, M. J., and K. O. Kalkan. 2013. "Behind the Curve: Clarifying the Best Approach to Calculating Predicted Probabilities and Marginal Effects from Limited Dependent Variable Models." *American Journal of Political Science* 57(1):263–277.

Paolino, P. 2020. "Replication Data for: Predicted Probabilities and Inference with Multinomial Logit." https://doi.org/10.7910/DVN/MVP6ID Harvard dataverse, V1, UNF:6:pOUJ1wqKmQbvbw83r/up2Q==[fileUNF].

Tomz, M., J. Whittenberg, and G. King. 2003. "Clarify: Software for Interpreting and Presenting Statistical Results." *Journal of Statistical Software* 8(1):1–30.