Effect Sizes for Calculation of Second Differences After Logistic and Probit Regression: Alternative Quantities of Interest[*]

Daniel Lempert[†]

September 2, 2020

Abstract

Social scientists are often interested in how the effect of some covariate on an outcome varies across groups. A common approach is to calculate a second difference, the difference between two group-specific effect sizes. I propose four alternative quantities of interest for calculating second differences after estimation of logistic or probit regression. While the conventional approach is based on setting covariate values that are constant across groups, the alternatives seek to compare effects that are more directly comparable across groups, in two specific senses. The approaches allow different substantive questions to be addressed and reduce dependence on the nonlinear functional form of logit and probit. I compare the proposed alternatives with the conventional approach in an application testing whether effects of election day registration on turnout are conditional on education, showing that the method chosen to calculate second differences can substantially affect the inferences drawn.

[*] [acks]

[†] Associate Professor of Politics, SUNY Potsdam. email: dalempert@gmail.com.

Social scientists are often interested in how the effect of some covariate on an outcome varies across groups (or contexts, etc.). When the outcome is a binary variable, a standard approach is to estimate a logit or a probit regression wherein one or more covariates are interacted with a variable indicating group membership (Berry, DeMeritt and Esarey 2016; Rainey 2016). In the usual situation, where the analyst is interested in the probability of a positive outcome, $\Pr(Y)$—as opposed to the latent, unobserved outcome Y^*—there is now consensus among methodologists that the relevant quantity of interest is *not* the sign, magnitude, and significance of the coefficient attending the interaction term(s), but rather the second difference in predicted probabilities of a positive outcome across groups (Berry, DeMeritt and Esarey 2016; Long and Mustillo 2019; Rainey 2016). Letting Z be the covariate of interest, b and r (for *baseline* and *reference*) two values of interest that Z takes on, G a binary group indicator, and $\boldsymbol{X}$ the vector of all other covariates, this second difference is defined as:

$$\begin{aligned}\Delta\Delta[\Pr(Y)] \equiv {} & [\Pr(Y \mid G=0,\, Z=r,\, \boldsymbol{X}=\boldsymbol{x}_0) - \Pr(Y \mid G=0,\, Z=b,\, \boldsymbol{X}=\boldsymbol{x}_0)] \\ & - [\Pr(Y \mid G=1,\, Z=r,\, \boldsymbol{X}=\boldsymbol{x}_1) - \Pr(Y \mid G=1,\, Z=b,\, \boldsymbol{X}=\boldsymbol{x}_1)]. \qquad (1)\end{aligned}$$

Because the inverse logit function and standard normal cdf are nonlinear, each of the four terms in Eq. 1 depend not just on the values of G and Z, but also on the values at which the other covariates $\boldsymbol{X}$ are set. In political science applications, the typical approach when calculating second differences, following King, Tomz and Wittenberg (2000, 355), has been to set $\boldsymbol{x}_1 = \boldsymbol{x}_0$; i.e., set $\boldsymbol{X}$ at fixed values that are constant *across groups*—commonly, the overall medians or means.[1]

[1] Notably, see Berry, DeMeritt and Esarey (2010, 249) and Rainey (2016, line 210 in replication file `or.R`). Setting $\boldsymbol{X}$ as constant across groups is seen as so uncontroversial that it is sometimes not even included in the formula defining the second difference (e.g., Berry, Demeritt, and Esarey 2010, 249; Rainey 2016, 622).

This conventional approach (i.e., setting $\boldsymbol{x}_1 = \boldsymbol{x}_0$) may be challenged on two grounds: (1) It does not necessarily yield the quantity of substantive interest and (2) The quantity it yields may be consequentially dependent on logit and probit's functional form. Here, I propose several alternative quantities of interest and illustrate how they may be preferable to the conventional approach.

Terminology and Assumptions

Consider postestimation after fitting a logit or probit model where some covariate of interest Z, and potentially some or all of the other covariates $\boldsymbol{X}$, are interacted with a binary or categorical variable indicating group membership. Assume that no higher-order terms for Z are included in the regression (see footnote 2). Since the goal is to compare the effect of Z across *two* groups, if the group membership variable is not binary, create a binary variable G that is coded 0 and 1 respectively, for the two groups being compared, and missing otherwise. If the group membership variable is binary, then call it G without transformation. Possible quantities of interest are then given by Eq. 1; the focus here is on how these quantities are affected by the choice of $\boldsymbol{x}_1$ and $\boldsymbol{x}_0$.

Alternative Quantities of Interest

Second Difference Based on Comparable Baseline Probabilities. An intuitive alternative to setting $\boldsymbol{x}_1 = \boldsymbol{x}_0$ is to set $\boldsymbol{x}_1$ and $\boldsymbol{x}_0$ so that the second and fourth terms on the rhs of Eq. 1—the "baseline probabilities" for each group—are (approximately) equal. Substantively, the conventional approach asks, "how does some covariate affect the probability of a positive outcome in two units that are in different groups but otherwise have the *same attributes*?" To the contrary, the substantive question the alternative addresses is "how does some covariate affect the probability of a positive outcome in two units that are in different groups but have the *same baseline probability of a positive outcome*?"

To preview the application below, suppose one is interested in how the effect of election-day registration (EDR) on turnout differs between the poorly educated and the highly educated (see the line of studies cited in Rainey 2016, dating back to Wolfinger and Rosenstone 1980). Consider two ways of specifying a question of interest. One, how does EDR affect the probability that an individual with a (specified) low level of education and a given gender, race, age, income (etc.) will vote, compared to its effect on an individual with a high level of education, and the same given values of gender, race, age, and income? Two, how does the presence of EDR affect the probability of voting for two individuals—one highly educated and one poorly educated—who, in the absence of EDR, have the same probability of voting? The answer to the first question is given by the conventional approach and the answer to the second question is given by the alternative based on comparable baseline probabilities.

The argument for this alternative is modest: the point is simply that, on substantive grounds, for some applications, it may be a quantity of interest in addition to, or instead of, the conventional second difference calculation. There is also another advantage to the alternative approach based on comparable baseline probabilities: it tends to, though it is not guaranteed to, reduce dependence on logit or probit's functional form, compared to the conventional approach, in a sense to be made more precise immediately below. I now turn to quantities of interest whose central goal is to reduce functional form dependence.

Second Difference Based on Comparable Effects. It is well understood that logit and probit's functional forms are such that their slopes increase as they approach 0.5 and decrease thereafter. In other words, the inverse logit and standard normal cdf are sigmoid functions with inflection points at 0.5. An immediate consequence is that the *instantaneous* change in a given covariate is greatest when the other covariates are fixed so that the probability of a positive outcome is 0.5, and monotonically decreases as the distance from 0.5 increases (Nagler 1991).[2] This is not problematic for researchers who are highly confident that the

[2] This is true if, as assumed above, no higher order term for the covariate of interest, Z, is included in the regression. For models with quadratic Z terms, it may be the case that there

logit or a probit model correctly identifies their data-generating process (DGP). However, in practice, this will rarely be the case; thus, it is advisable to limit the extent to which results depend on this assumption.

The analogous implication for the case of a *discrete* change in the context of second differences, given the assumptions and notation above, is as follows. The estimated effect size of the change in Z from r to b for group g is the greatest when $[\Pr(Y \mid G = g,\, Z = r,\, \boldsymbol{X} = \boldsymbol{x}_g) + \Pr(Y \mid G = g,\, Z = b,\, \boldsymbol{X} = \boldsymbol{x}_g)]/2$ is 0.5, and decreases as this quantity moves further away from 0.5. Equivalently, this effect size is a strictly decreasing function of

$$\left| 0.5 - \frac{\Pr(Y \mid G = g,\, Z = r,\, \boldsymbol{X} = \boldsymbol{x}_g) + \Pr(Y \mid G = g,\, Z = b,\, \boldsymbol{X} = \boldsymbol{x}_g)}{2} \right| \equiv d(g, \boldsymbol{x}_g). \tag{2}$$

To minimize the impact of functional form on the comparison of effect sizes across groups, then, it is desirable to set $\boldsymbol{X}$ in Eq. 1 so that the $d(g, \boldsymbol{x}_g)$ are equal across groups (or nearly so). Setting $\boldsymbol{X}$ so $|d(0, \boldsymbol{x}_0) - d(1, \boldsymbol{x}_1)| < \varepsilon$ ensures *comparable effects* across groups;[3] colloquially, this is because the shape of the function over the range for which the effect sizes (first differences) are calculated are as similar as possible.

Of course, because $d(g, \boldsymbol{x}_g)$ depends on how $\boldsymbol{X}$ is set, the set of possible $d(g, \boldsymbol{x}_g)$ is large, though finite if restricted, as will be the case here, to values of $\boldsymbol{X}$ actually observed in-sample.[4] Correspondingly, there will typically be many pairs $(d(0, \boldsymbol{x}_0),\, d(1, \boldsymbol{x}_1))$ such that $|d(0, \boldsymbol{x}_0) - d(1, \boldsymbol{x}_1)| < \varepsilon$, thereby yielding several different comparable effects across groups.

are two inflection points, with the global maximum or minimum between them. The resultant modifications to the formulas below are relatively minor, but they are omitted here due to space constraints. The current version of the associated Stata command `sdcompefx` does not yet allow quadratic Z, but this feature will be added shortly.

[3] ε should be understood as a small positive number, subjectively chosen by the analyst.

[4] To avoid out-of-sample predictions, I recommend considering only $\boldsymbol{x_g}$ that are actually present in the sample. If one disregards this recommendation, it is easy to set $d(0, \boldsymbol{x}_0) = d(1, \boldsymbol{x}_1)$ at any given value.

How should one choose among these pairs?

One obvious choice is to calculate the second difference based on **maximum comparable effects**. Very roughly, this entails selecting the smallest pair of $(d(0, \boldsymbol{x}_0),\ d(1, \boldsymbol{x}_1))$ such that $|d(0, \boldsymbol{x}_0) - d(1, \boldsymbol{x}_1)| < \varepsilon$. To be precise, let g and g' denote two different groups (i.e., $g' \equiv |g - 1|$), let $d_{(k)}$ refer to the kth order statistic, calculated over all d, and let $\underline{k}$ denote the smallest k for which it is true that $|d_{(k)}(g, \boldsymbol{x}_g) - d_{(k+1)}(g', \boldsymbol{x}_{g'})| < \varepsilon$. Then, the second difference based on maximum comparable effects sets $\boldsymbol{X}$ to the values given by the pair $\big(d_{(\underline{k})}(g, \boldsymbol{x}_g),\ d_{(\underline{k}+1)}(g', \boldsymbol{x}_{g'})\big)$. As the name suggests, this second difference compares the largest effect observed in each group, given that the effects are comparable.

The second difference based on **minimum comparable effects** is defined analogously: set $\boldsymbol{X}$ to the values given by the pair $\big(d_{(\bar{k})}(g, \boldsymbol{x}_g),\ d_{(\bar{k}+1)}(g', \boldsymbol{x}_{g'})\big)$, where $\bar{k}$ denotes the largest k for which it is true that $|d_{(k)}(g, \boldsymbol{x}_g) - d_{(k+1)}(g', \boldsymbol{x}_{g'})| < \varepsilon$. This, then, compares the smallest effect observed in each group, given that the effects are comparable.

Finally, we have the second difference based on the **median of comparable effects**. Approximately speaking, this selects the median pair of $(d(0, \boldsymbol{x}_0),\ d(1, \boldsymbol{x}_1))$, such that $|d(0, \boldsymbol{x}_0) - d(1, \boldsymbol{x}_1)| < \varepsilon$. To be precise, consider now only the N possible $d(g, \boldsymbol{x})$ for which there exists at least one $d(g', \boldsymbol{x}_{g'})$ such that $|d(g, \boldsymbol{x}_g) - d(g', \boldsymbol{x}_{g'})| < \varepsilon$; let $d_{(k)}$ be the kth order statistic, calculated over all such d. Define

$$\tilde{d}(g, \boldsymbol{x}_g) \equiv \begin{cases} \left| d(g, \boldsymbol{x}_g) - d_{\left(\frac{N+1}{2}\right)}(g, \boldsymbol{x}_g) \right| & \text{if } N \text{ odd}, \\ \left| d(g, \boldsymbol{x}_g) - \frac{1}{2}\left(d_{\left(\frac{N}{2}\right)}(g, \boldsymbol{x}_g) + d_{\left(\frac{N+2}{2}\right)}(g, \boldsymbol{x}_g) \right) \right| & \text{if } N \text{ even}. \end{cases} \tag{3}$$

Then, the second difference based on the median of comparable effects sets $\boldsymbol{X}$ to the values given by the pair $\Big(\tilde{d}_{(\underline{k})}(g, \boldsymbol{x}_g),\ \tilde{d}_{(\underline{k}+1)}(g', \boldsymbol{x}_{g'})\Big)$ where, to be clear, $\tilde{d}_{(k)}$ refers to the kth order statistic calculated over all $\tilde{d}$ and $\underline{k}$ denotes the smallest k for which it is true that $\left|\tilde{d}_{(k)}(g, \boldsymbol{x}_g) - \tilde{d}_{(k+1)}(g', \boldsymbol{x}_{g'})\right| < \varepsilon$. This second difference, then, compares the observed effect in each group that is closest to the median of all comparable effects.

If a single summary measure of second differences must be selected, this—as effectively the average comparable effect—will often be the most appropriate one, but the second differences based on maximum and minimum comparable effects are often additionally of interest. And, as discussed above, depending on the substantive question addressed, the second difference based on comparable baseline probabilities may be the appropriate choice, even though it is not guaranteed to minimize dependence on functional form as do the second differences based on comparable effects.

Which Quantity to Present?

In which applications should one or more of the proposed alternatives be used instead of the conventional approach? When choosing between the conventional approach and the second difference based on comparable baseline probabilities, the central factor is fit with the substantive question, as suggested above. For an example of a substantive question where the alternative based on comparable baseline probabilities is particularly appropriate, suppose one wants to examine the effect that the United States Solicitor General as petitioner has on the Supreme Court deciding to hearing a case in two eras: before and after the Warren Court's "rights revolution" [cite/expand]. [...]

For the alternatives based on comparable effects, no hard-and-fast rule is possible, since this would require—at the very minimum—quantifying the probability that the logit or the probit describes the DGP in question; this is of course unknown in the typical application.[5] What is clear, though, is that the alternative quantities of interest based on comparable effects are more conservative than the conventional approach: conventional second differences may be an artifact driven by a functional form that does not describe the DGP; this risk is effectively eliminated by the alternatives based on comparable effects.

[5] Even were this quantity knowable, any decision rule would be essentially arbitrary, because balancing the (arguable!) tradeoff between robustness to dependence on functional form and ease of interpretation is so subjective.

Perhaps it is most prudent to present both the conventional second difference and one or more of the alternatives based on comparable effects. If the substantive conclusions from both sets are similar, the empirical result is bolstered. If—as in the application below—the substantive conclusions are consequentially different, the analyst should recognize that in order to credit the conclusion implied by the traditional second difference, one must be extremely confident that the DGP is closely described by the logit or probit.

Application

Consider the effect of EDR on turnout, as a function of education in the 1984 elections (Berry, DeMeritt and Esarey 2010; Nagler 1991; Rainey 2016; Wolfinger and Rosenstone 1980). I estimate a logit predicting turnout, as a function of various standard covariates, including an eight-level education variable, which is interacted with a covariate indicating the number of days before an election that registration is required in the respondent's state. The data are as posted in the replication materials for Potoski and Urbatsch (2017). I compare how those with a high (post-graduate) versus a low (some high school) level of education are affected by being in a state with EDR, compared to one that has a registration deadline that is 25 days before the election. I calculate second differences based on the conventional approach, setting $\boldsymbol{X}$ to the mean (for continuous covarites) and median (for binary covariates) values in the sample as a whole, as well as the four alternative quantities of interest I have proposed. I use `margins` in Stata for the conventional approach, and my user-written Stata commands [`sdcasepick` and `sdcompefx`] for the proposed alternatives. (For further explanation of modeling details and choices see the Appendix.)

Table 1 gives the first differences and second difference for the conventional approach and the four proposed alternatives. The conventional approach implies that the effect of EDR is significantly greater, by .046, for those who are less-educated. But for all four alternatives, the 95% CIs for the second difference overlap with 0 (and the point estimates for the effect on the better-educated are higher). Thus, clearly, the quantity of interest chosen—and implicitly, the analyst's faith in the logit describing the DGP—affects the substantive inferences to be drawn.

Method	High Education	Low Education	Second Difference
Conventional	.021 [.012, .030]	.067 [.037, .096]	.046 [.017, .075]
Comparable Baseline	.049 [.027, .071]	.040 [.024, .056]	-.009 [-.034, .016]
Maximum CE	.084 [.044, .124]	.067 [.039, .095]	-.017 [-.062, .028]
Median CE	.073 [.039, .107]	.058 [.030, .087]	-.015 [-.055, .026]
Minimum CE	.040 [.022, .058]	.032 [.019, .045]	-.008 [-.028, .013]

Table 1. Five approaches to calculating effects of EDR, as a function of education, on the probability of turnout, 1984 general election. 95% CIs in brackets.

Related Approaches

I now discuss some related approaches to calculating differences in effects across groups. Perhaps most popular in economics, the linear probability model (LPM) uses least squares to predict binary outcomes. Summarizing (let alone resolving) the debate about whether the LPM is generally preferable to logit or probit is beyond the scope of this research note (see e.g., Horrace and Oaxaca 2006); suffice it to say that non-linearity imposed by logit and probit is not an issue for the LPM since, of course, it assumes a linear relationship between the outcome and covariates (and covariates' interactions). Still, using logit or probit for binary outcomes remains the dominant practice in political science; accordingly, improving the interpretation of interactions in those models remains relevant.

Similarly, various matching-based approaches for detecting effect heterogeneity reduce dependance on functional form (for an overview, see Brand and Thomas 2013), but have not displaced the logit and probit in political science. Machine learning-based approaches to detecting interactive effects (e.g., Imai and Ratkovic 2013; Green and Kern 2012) exist as well, but are used relatively rarely in applied work.

The most similar alternative approach to those I consider here is that advocated by Hanmer and Kalkan (2013), extended to comparisons of effects across groups, as described in Long and Mustillo (2019). After estimating a logit or probit, rather than setting $\boldsymbol{X}$ to representative or interesting values, as the traditional approach and my proposed alternatives do, Long and Mustillo (2019) suggest calculating effects for each group by averaging over all observed $\boldsymbol{X}$, weighted by the frequency with which they appear in the group. Letting N_g denote the number of observations in group g, this gives second difference,

$$\begin{aligned}\Delta\Delta[\Pr(Y)] \equiv \frac{1}{N_0}\sum_{i:g_i=0}&\left[\Pr(Y_i\,|\,G=g_i,\,Z=r,\,\boldsymbol{X}=\boldsymbol{x}_i)-\Pr(Y_i\,|\,G=g_i,\,Z=b,\,\boldsymbol{X}=\boldsymbol{x}_i)\right]\\ -\frac{1}{N_1}\sum_{i:g_i=1}&\left[\Pr(Y_i\,|\,G=g_i,\,Z=r,\,\boldsymbol{X}=\boldsymbol{x}_i)-\Pr(Y_i\,|\,G=g_i,\,Z=b,\,\boldsymbol{X}=\boldsymbol{x}_i)\right]. \quad (4)\end{aligned}$$

While this approach may have its advantages, observe that if the distribution of the $d_i(g_i, \boldsymbol{x}_i)$ is significantly different across groups, the problem of comparing incomparable effects across group arises, very much as in the conventional approach. One *partial* remedy is to calculate second differences based only those observations i for which there exists at least one observation j, $g_i \neq g_j$, such that $|d_i(g_i, \boldsymbol{x}_i) - d_j(g_j, \boldsymbol{x}_j)| < \varepsilon$. `sdcompefx` can create a variable that identifies such observations, which can then be used to restrict the postestimation sample. But this solution is not truly satisfactory, since the distribution of the $d_i(g_i, \boldsymbol{x}_i)$ may still remain very different across groups, though less different than if all observations were included in the postestimation. I elaborate on these points in the Appendix.

Conclusion

I have proposed alternative quantities of interest for calculating second differences to compare effects across groups, after estimation of logit and probit. While the conventional approach sets covariate values that are constant across groups, the alternatives seek to compare effects that are comparable across groups, in terms of baseline probability or location on the logit/probit curve. The approaches allow different substantive questions to be addressed and reduce dependence on the nonlinear functional form of logit and probit. I compared the proposed alternatives with the conventional approach in an application calculating the conditional effects of EDR, showing that the method chosen to calculate second differences can substantially affect the inferences drawn.

References

Berry, William D., Jacquline H.R. DeMeritt and Justin Esarey. 2010. "Testing for Interactions in Binary Logit and Probit Models: Is a Product Term Essential?" *American Journal of Political Science* 54(1): 248–266.

Berry, William D., Jacquline H.R. DeMeritt and Justin Esarey. 2016. "Bias and Overconfidence in Parametric Models of Interactive Processes." *American Journal of Political Science* 60(2): 521–539.

Brand, Jennie E. and Juli Simon Thomas. 2013. Causal Effect Heterogeneity. In *Handbook of Causal Analysis for Social Research*, ed. Stephen L. Morgan. Dordrecht, Germany: Springer.

Green, Donald P. and Holger L. Kern. 2012. "Modeling Heterogeneous Treatment Effects in Survey Experiments with Bayesian Additive Regression Trees." *Public Opinion Quarterly* 76(3): 491–511.

Hanmer, Michael J. and Kerem Ozan 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.

Horrace, William C. and Ronald L. Oaxaca. 2006. "Results on the Bias and Inconsistency of Ordinary Least Squares for the Linear Probability Model." *Economics Letters* 90(3): 321–327.

Imai, Kosuke and Marc Ratkovic. 2013. "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation." *Annals of Applied Statistics* 7(1): 443–470.

King, Gary, Michael Tomz and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44(2): 347–361.

Long, J. Scott and Sarah A. Mustillo. 2019. "Using Predictions and Marginal Effects to Compare Groups in Regression Models for Binary Outcomes." *Sociological Methods and Research* 00(0): 0–00.

Nagler, Jonathan. 1991. "The Effect of Registration Laws on Voter Turnout." *American Political Science Review* 85(4): 1393–1405.

Potoski, Matthew and R. Urbatsch. 2017. "Entertainment and the Opportunity Cost of Civic Participation: Monday Night Football Game Quality Suppresses Turnout in US Elections." *Journal of Politics* 79(2): 424–438.

Rainey, Carlisle. 2016. "Compression and Conditional Effects: A Product Term is Essential When Using Logistic Regression to Test for Interaction." *Political Science Research and Methods* 4(3): 621–639.

Wolfinger, Raymond E. and Steven J. Rosenstone. 1980. *Who Votes?* New Haven: Yale University Press.