The Journal of Mathematical Sociology

ISSN: 0022-250X (Print) 1545-5874 (Online) Journal homepage: http://www.tandfonline.com/loi/gmas20

Understanding and interpreting generalized ordered logit models

Richard Williams

To cite this article: Richard Williams (2016) Understanding and interpreting generalized ordered logit models, The Journal of Mathematical Sociology, 40:1, 7-20, DOI: 10.1080/0022250X.2015.1112384

To link to this article: http://dx.doi.org/10.1080/0022250X.2015.1112384

Published online: 29 Jan 2016.

THE JOURNAL OF MATHEMATICAL SOCIOLOGY
2016, VOL. 40, NO. 1, 7–20
http://dx.doi.org/10.1080/0022250X.2015.1112384

Routledge
Taylor & Francis Group

Understanding and interpreting generalized ordered logit models

Richard Williams

Department of Sociology, University of Notre Dame, Notre Dame, Indiana, United States

ABSTRACT

When outcome variables are ordinal rather than continuous, the ordered logit model, aka the proportional odds model (ologit/po), is a popular analytical method. However, generalized ordered logit/partial proportional odds models (gologit/ppo) are often a superior alternative. Gologit/ppo models can be less restrictive than proportional odds models and more parsimonious than methods that ignore the ordering of categories altogether. However, the use of gologit/ppo models has itself been problematic or at least sub-optimal. Researchers typically note that such models fit better but fail to explain why the ordered logit model was inadequate or the substantive insights gained by using the gologit alternative. This paper uses both hypothetical examples and data from the 2012 European Social Survey to address these shortcomings.

ARTICLE HISTORY
Received 21 August 2014
Accepted 27 July 2015

KEYWORDS
Generalized ordered logit model; ordered logit model; partial proportional odds; proportional odds assumption; proportional odds model

1. Overview

Techniques such as Ordinary Least Squares Regression require that outcome variables have interval or ratio level measurement. When the outcome variable is ordinal (i.e., the relative ordering of response values is known but the exact distance between them is not), other types of methods should be used. Perhaps the most popular method is the ordered logit model, which (for reasons to be explained shortly) is also known as the proportional odds model.[1]

Unfortunately, experience suggests that the assumptions of the ordered logit model are frequently violated (Long & Freese, 2014). Researchers have then typically been left with a choice between staying with a method whose assumptions are known to be violated or switching to a method that is far less parsimonious and more difficult to interpret, such as the multinomial logit model which makes no use of information about the ordering of categories.

In this article, we present and critique a third choice: the Generalized Ordered Logit/Partial Proportional Odds Model (gologit/ppo). This model has been known about since at least the 1980s (e.g., McCullagh & Nelder, 1989; Peterson & Harrell, 1990), but recent advances in software (such as the user-written gologit and gologit2 routines in Stata) have made the model much easier to estimate and widely used (Fu, 1998; Williams, 2006).[2] The gologit/ppo model selectively relaxes the assumptions of the ordered logit model only as needed, potentially producing results that do not have the problems of the ordered logit model while being almost as easy to interpret.

Unfortunately, while gologit/ppo models have seen increasing use, these uses have themselves frequently been problematic. Often it is simply noted that the model fits better and avoids violating the assumptions of the ordered logit model (see, e.g., Cornwell, Laumann, & Shumm, 2008; Do &

CONTACT Richard Williams rwilliam@nd.edu 810 Flanner Hall, Department of Sociology, University of Notre Dame, Notre Dame, IN 46556, USA

[1]The ordered probit model is a popular alternative to the ordered logit model. The terms "Parallel Lines Assumption" and "Parallel Regressions Assumption" apply equally well for both the ordered logit and ordered probit models. However the ordered probit model does not require nor does it meet the proportional odds assumption.

[2]According to Google Scholar, Williams (2006), which introduced the gologit2 program for Stata, has been cited more than 800 times since its publication. Similarly, various papers by Hedeker (e.g. Hedeker & Mermelstein, 1998) on the similar "stages of change" models have been cited hundreds of times.

Farooqui, 2011; Kleinjans, 2009; Lehrer, Lehrer, Zhao, & Lehrer, 2007; Schafer & Upenieks, 2015). However, papers often fail to explain why the proportional odds model was inadequate. Even more critically, researchers often pay little attention to the substantive insights gained by using the gologit/ppo model that would be missed if proportional odds were used instead. That does not mean that such papers are not making valuable contributions but it could mean that authors are overlooking other important potential contributions of their work. These failings may reflect a lack of understanding of what the assumptions of these different models actually are and what violations of assumptions tell us about the underlying reality of what is being investigated.

This article therefore explains why the ordered logit model often fails, shows how and why gologit/ppo can often provide a superior alternative to it, and discusses the ways in which the parameters of the gologit/ppo model can be interpreted to gain insights that are often overlooked. We also note several other issues that researchers should be aware of when making their choice of models. By better understanding how to interpret results, researchers will gain a much better understanding of why they should consider using the gologit/ppo method in the first place. Both hypothetical examples and data from the 2012 European Social Survey are used to illustrate these points.

2. The ordered logit/proportional odds model

We are used to estimating models where a continuous outcome variable, *Y*, is regressed on an explanatory variable, *X*. But suppose the observed *Y* is not continuous – instead, it is a collapsed version of an underlying unobserved variable, *Y** (Long & Freese, 2014). As people cross thresholds on this underlying variable their values on the observed ordinal variable *Y* changes. For example, Income might be coded in categories like $0 = 1, $1–$10,000 = 2, $10,001–$30,000 = 3, $30,001–$60,000 = 4, $60,001 or higher = 5. Or, respondents might be asked, "Do you approve or disapprove of the President's health care plan?" The options could be 1 = Strongly disapprove, 2 = Disapprove, 3 = Approve, 4 = Strongly approve. Presumably there are more than four possible values for approval, but respondents must decide which option best reflects the range that their feelings fall into. For such variables, also known as limited dependent variables, we know the interval that the underlying *Y** falls in, but not its exact value. Ordinal regression techniques allow us to estimate the effects of the *X*s on the underlying *Y**.

However, in order for the use of the ordered logit model to be valid, certain conditions must hold. Tables 1-1 through 1-3 present hypothetical examples that clarify what these conditions are and why they may not be met. Each of these tables presents a simple bivariate relationship between gender and an ordinal attitudinal variable coded Strongly Disagree, Disagree, Agree, and Strongly Agree. In each table, a series of *cumulative logit models* are presented; that is, the original ordinal variable is collapsed

Table 1-1. Hypothetical example of perfect proportional odds/parallel lines*.

	Attitude				
Gender	SD	D	A	SA	Total
Male	250	250	250	250	1,000
Female	100	150	250	500	1,000
Total	350	400	500	750	2,000
		1 versus 2, 3, 4	1 & 2 versus 3 & 4		1, 2, 3 versus 4
OddsM		750/250 = 3	500/500 = 1		250/750 = 1/3
OddsF		900/100 = 9	750/250 = 3		500/500 = 1
OR (OddsF/OddsM)		9/3 = 3	3/1 = 3		1/(1/3) = 3
Betas		1.098612	1.098612		1.098612
Ologit Beta (OR)		1.098612 (3.00)			
Ologit χ^2 (1 d.f.)		176.63 (p = 0.0000)			
Gologit χ^2 (3 d.f.)		176.63 (p = 0.0000)			
Brant Test (2 d.f.)		0.0 (p = 1.000)			

into two categories and a series of binary logistic regressions are run. First it is category 1 (SD) versus categories 2, 3, 4 (D, A, SA); then it is categories 1 & 2 (SD, D) versus categories 3 & 4 (A, SA); then, finally, categories 1, 2, and 3 (SD, D, A) versus category 4 (SA). In each dichotomization the lower values are, in effect, recoded to zero, while the higher values are recoded to one. A positive coefficient means that increases in the explanatory variable lead to higher levels of support (or less opposition), while negative coefficients mean that increases in the explanatory value lead to less support (or stronger opposition).

If the assumptions of the ordered logit model are met, then all of the corresponding coefficients (except the intercepts) should be the same across the different logistic regressions, other than differences caused by sampling variability. The assumptions of the model are therefore sometimes referred to as the parallel lines or parallel regressions assumptions (Williams, 2006).

The ordered logit model is also sometimes called the proportional odds model because, if the assumptions of the model are met, the odds ratios will stay the same regardless of which of the collapsed logistic regressions is estimated (hence the term proportional odds assumption is also often used). A test devised by Brant (1990; also see Long & Freese, 2014) is commonly used to assess whether the observed deviations from what the proportional odds model predicts are larger than what could be attributed to chance alone.

The tables were constructed so that in Table 1-1, the proportional odds/parallel lines assumption would be perfectly met. In Tables 1-2 and 1-3 we then shifted the distribution of the female responses so that the assumption would not hold. Although these are hypothetical examples and data, they are

Table 1-2. Hypothetical example of proportional odds violated-I*.

	Attitude				
Gender	SD	D	A	SA	Total
Male	250	250	250	250	1,000
Female	100	300	300	300	1,000
Total	350	550	550	550	2,000
		1 versus 2, 3, 4	1 & 2 versus 3 & 4		1, 2, 3 versus 4
OddsM		750/250 = 3	500/500 = 1		250/750 = 1/3
OddsF		900/100 = 9	600/400 = 1.5		300/700 = 3/7
OR (OddsF/OddsM)		9/3 = 3	1.5/1 = 1.5		(3/7)/(1/3) = 1.28
Betas		1.098612	.4054651		.2513144
Ologit Beta (OR)		.4869136 (1.627286)			
Ologit χ^2 (1 d.f.)		36.44 (p = 0.0000)			
Gologit χ^2 (3 d.f.)		80.07 (p = 0.0000)			
Brant Test (2 d.f.)		40.29 (p = 0.000)			

Table 1-3. Hypothetical example of proportional odds violated-II*.

	Attitude				
Gender	SD	D	A	SA	Total
Male	250	250	250	250	1,000
Female	100	400	400	100	1,000
Total	350	650	650	350	2,000
		1 versus 2, 3, 4	1 & 2 versus 3 & 4		1, 2, 3 versus 4
OddsM		750/250 = 3	500/500 = 1		250/750 = 1/3
OddsF		900/100 = 9	500/500 = 1		100/900 = 1/9
OR (OddsF/OddsM)		9/3 = 3	1/1 = 1		(1/9)/(1/3) = 1/3
Betas		1.098612	0		−1.098612
Ologit Beta (OR)		0 (1.00)			
Ologit χ^2 (1 d.f.)		0.00 (p = 1.0000)			
Gologit χ^2 (3 d.f.)		202.69 (p = 0.0000)			
Brant Test (2 d.f.)		179.71 (p = 0.000)			

*The tables were constructed so that in Table 1-1, the proportional odds/parallel lines assumption would be perfectly met. In Tables 1-2 and 1-3 we then shifted the distribution of the female responses so that the assumption would not hold. Although these are hypothetical examples and data, they are typical of what is often encountered in practice.

typical of what is often encountered in practice. In Table 1-1, looking at the column labeled 1 versus 2, 3, 4, we see that men are three times as likely to be in one of the higher categories as they are to be in the lowest category, so the odds for men are 3, i.e. 750/250. Women, on the other hand, are nine times as likely to be in one of the higher categories, so the odds for women are 9, or 900/100. The ratio of the odds for women to men, that is, the odds ratio, is 9/3 = 3.

Similarly, for the column labeled 1, 2 versus 3, 4, men are equally likely to be in either the two lowest or the two highest categories, yielding odds of 1. Women are three times as likely to be in one of the two higher categories as they are to be in one of the two lowest categories, yielding odds of 3. The odds ratio for women compared to men is therefore once again 3.

Finally, for the 1, 2, 3 versus 4 logistic regression/cumulative logit, only 1/3 as many men are in the highest category as are in the 3 lowest categories, yielding odds of 1/3. Women are equally likely to be in the highest as opposed to the three lowest categories, yielding odds of 1. The odds ratio is therefore 1/(1/3), which is equal to three.

If the parallel lines assumption holds, then (subject to sampling variability) the coefficients should be the same in each of the cumulative logistic regressions, and (as the row labeled Betas shows) indeed they are (1.098612; this is also the same as the beta coefficient when a single ordered logit model is estimated). Similarly, if the proportional odds assumption holds, then the odds ratios should be the same for each of the ordered dichotomizations of the outcome variable. Proportional Odds works perfectly in this model, as the odds ratios are all 3. The Brant test reflects this and has a value of 0.

Table 1-2 presents a second example. In this case, women are again clearly more likely to agree than men, and yet the assumptions of the ordered logit model are not met.

Gender has its greatest effect at the lowest levels of attitudes; as the odds ratio of 3 indicates, women are much less likely to strongly disagree than men. But other differences are smaller; in the 1 & 2 versus 3 & 4 cumulative logit, the odds ratio is only 1.5, and in the last cumulative logit, 1, 2, 3 versus 4, the odds ratio is only 1.28. Nonetheless, as the Betas show, the effect of gender is consistently positive, i.e. the differences in the coefficients across the different dichotomizations of the outcome variable involve magnitude, not direction. Similarly, the odds for women are consistently greater than the odds for men (and hence the odds ratios are consistently greater than 1). But, because the odds ratios are not the same across the different regressions, the Brant test is highly significant (40.29 with

Table 2. Proportional odds and partial proportional odds models for government should reduce differences in income levels*.

	Model 1: Proportional odds		Model 2: Partial proportional odds**				
Explanatory variables	*P* Value	Coef	Overall *P* Value***	SD vs D, N, A, SA	SD, D vs N, A, SA	SD, D, N vs A, SA	SD, D, N, A vs SA
Life is getting worse	.000	.322	.000	.329			
Feelings about household income	.000	.234	.000	.227			
Member of ethnic minority	0.843	.037	.867	.032			
Age (in decades)	.065	–.042	.001	–.172	–.102	–.071	.042
Gender (1 = female, 0 = male)	.287	.096	.018	.484	.304	.217	–.182
Satisfaction with state of economy	.052	–.049	.000	.111	.047	–.043	–.109

*Data are from the European Social Survey. The European Social Survey (ESS) is a cross-national study that has been conducted every two years across Europe since 2001. For this example we use the 2012 ESS survey for Great Britain (ESS Round 6: European Social Survey Round 6 Data, 2012). The study has 2,286 respondents, of which 2,123 (92.8%) had complete data for the variables used in this analysis. Because cases have unequal probabilities of selection, sampling weights are used. The Stata user-written program gologit2 (Williams, 2006) is employed for the analysis.

**Only one set of coefficients is presented for explanatory variables that meet the proportional odds assumption. SD = Strongly Disagree, D = Disagree, N = Neither Agree Nor Disagree, A = Agree, SA = Strongly Agree

***The overall p value is based on a test of the joint significance of all coefficients for the variable that are in the model. For variables that meet the proportional odds assumption there is one coefficient; for variables that do not meet the assumption there are four coefficients.

2 d.f.). Comparing the coefficients of the binary logistic regressions with the ordinal logistic regression, the ordinal beta coefficient (.4869) underestimates the impact of gender on moving people away from the lowest category while also overstating gender's impact in moving people towards the highest category. It is clear that women are more supportive than men, but the ordered logit model (whose assumptions are violated in this case) fails to accurately reflect the nature of the influence.

Finally, Table 1-3 presents one last hypothetical example:

The effect of gender varies in both sign and magnitude across the range of attitudes. Basically, women tend to have less extreme attitudes in either direction. They are less likely to strongly disagree than are men, but they are also less likely to strongly agree. The ordered logit beta of 0 implies gender is unrelated to attitudes, but the binary logistic regressions suggest a very different story. Perhaps the current coding of attitudes is not ordinal with respect to gender; for example, coding by intensity of attitudes rather than direction may be more appropriate. Or suppose that, instead of attitudes, the categories represented a set of ordered hurdles, or achievement levels. Women as a whole may be more likely than men to clear the lowest hurdles (e.g., get a high school diploma) but less likely to clear the highest ones (e.g., get a PhD). If men are more variable than women, they will have more outlying cases in both directions. Use of an ordered logit model in this case, at least with the current coding of the outcome variable, would be highly misleading.

Every one of the above models represents a reasonable relationship involving an explanatory variable and an ordinal outcome variable; but only the model presented in Table 1-1 passes the Brant test. The use of an ordered logit model when its assumptions are violated creates a misleading impression of how the outcome and explanatory variables are related.

Further, keep in mind that these are simple bivariate models. When there are multiple explanatory variables, the situation can get much more complicated. For example, there could be a dozen variables in a model, 11 of which meet the parallel lines/proportional odds assumption and only one of which does not. Nonetheless, the one problematic variable could cause the entire model to fail the Brant test.

We want a more flexible model that can deal with situations like the above, a model whose assumptions are not violated but at the same time does not include a lot of extraneous and unnecessary parameters such as a multinomial logit model might. Perhaps even more critically, we want the model to yield substantive insights that the ordered logit model does not.

3. The gologit model

For an ordinal outcome variable with M categories, the Generalized Ordered Logit model (Williams, 2006) can be written as

$$P(Y_i > j) = \frac{\exp(\alpha_j + X_i\beta_j)}{1 + [\exp(\alpha_j + X_i\beta_j)]}, \mathrm{j} = 1, 2, \ldots, \mathrm{M} - 1$$

For example, if the outcome variable has four possible values, the gologit model will have three sets of coefficients; in effect, three equations are estimated simultaneously. An unconstrained gologit model gives results that are similar to what we get with the series of binary logistic regressions/cumulative logit models such as we presented earlier and can be interpreted the same way.[3] The ordered logit model is a special case of the gologit model where the betas are the same for each j; that is, the j subscripts are unnecessary in the above formula.

In between these two extremes is the partial proportional odds model (PPO). With the PPO, some of the beta coefficients are the same for all values of *j*, while others can differ. For example, in the following PPO model the betas for *X*1 and *X*2 are constrained to be the same across values of J but the betas for *X*3 are not:

[3]Small differences are typically found because the gologit model estimates all the parameters simultaneously whereas the separate logistic regressions estimate them one cumulative logit at a time.

$$P(Y_i > j) = \frac{\exp(\alpha_j + X1_i\beta 1 + X2_i\beta 2 + X3_i\beta 3_j)}{1 + [\exp(\alpha_j + X1_i\beta 1 + X2_i\beta 2 + X3_i\beta 3_j)]}, \mathrm{j} = 1, 2, \ldots, \mathrm{M} - 1$$

An unconstrained gologit model and a multinomial logit model will both generate many more parameters than an ordered logit model does. This is because, with these methods, all variables are freed from the proportional odds constraint, even though the assumption may only be violated by one or a few of them. With a partial proportional odds model, however, it is possible to relax the parallel lines/proportional odds assumption only for those variables where it is violated. Our next section uses real data to illustrate how this can be done.

4. A multivariate example

The European Social Survey (ESS) is a cross-national study that has been conducted every two years across Europe since 2001. For this example we use the 2012 ESS survey for Great Britain (ESS Round 6: European Social Survey Round 6 Data, 2012). The study has 2,286 respondents, of which 2,123 (92.8%) had complete data for the variables used in this analysis. Because cases have unequal probabilities of selection, sampling weights are used. The Stata user-written program gologit2 (Williams, 2006) is employed for the analysis.[4]

Respondents were asked the extent to which they agreed or disagreed with the following statement: "The government should take measures to reduce differences in income levels." The possible responses were 1 = Strongly Disagree, 2 = Disagree, 3 = Neither Agree nor Disagree, 4 = Agree, and 5 = Strongly Agree. We use this as our response variable.

The explanatory variables are the responses to the following questions.

- "For most people in this country life is getting worse rather than better." (Again coded 1 = Strongly Disagree to 5 = Strongly Agree)
- "Which of the descriptions on this card comes closest to how you feel about your household's income nowadays?" (1 = Living comfortably on present income, 2 = Coping on present income, 3 = Finding it difficult on present income, 4 = Finding it very difficult on present income)
- "Do you belong to a minority ethnic group in this country?" (1 = Yes, 0 = No)
- Age of respondent (In decades, e.g., a value of 3.4 means 34 years old)
- Gender of respondent (1 = Female, 0 = Male)
- "On the whole how satisfied are you with the present state of the economy in this country?" (11 point scale where 0 = extremely dissatisfied and 10 = extremely satisfied).

The analyses of these data are given in Table 2. Model 1 presents the coefficients for the proportional odds model. Several results immediately stand out. The first two variables—life is getting worse and feelings about household income—have highly significant and positive effects. Those who feel that life is getting worse and/or are dissatisfied with their household income are more likely to believe that the government should try to reduce differences in income levels.

The next four variables, however, all fail to achieve the .05 level of statistical significance, although age, and satisfaction with the economy, come close. If the .10 level of significance was used instead, the results would suggest that older people and those who are more satisfied with the economy are somewhat less likely to believe that the government should act to reduce income inequality. The other

[4]When survey weights are used, several conventional measures of model fit—BIC, AIC, and Likelihood Ratio Chi-square—are not appropriate. Similarly, the Brant test is not appropriate either. However, the Wald tests used by gologit2 to test the proportional odds assumption can still be used. We used gologit2 with the autofit option set to .025; that is, the assumption is rejected if the observed deviations from it would only be expected to 25 times out of 1,000 if the assumption is true. This is consistent with Williams' (2006) advice that the default .05 level of significance may not be stringent enough when multiple variables are being tested.

two variables in the model, ethnicity and gender, both have positive estimated effects, implying that women and ethnic minorities are more supportive of governmental action, but the estimates fall far short of statistical significance.

However, statistical tests of the proportional odds assumption reveal that three variables fail to meet it. Model 2 therefore presents the estimates for the partial proportional odds model.[5] There are clear similarities with the earlier results but also striking and important differences.

The first three variables—life is getting worse, feelings about household income, and ethnicity—all meet the proportional odds assumption. Their coefficients and p values are virtually identical to before and can be interpreted the same way.

The remaining three variables—age, gender, and satisfaction with the economy—all violate the proportional odds assumption; and once the assumption is relaxed for them, all three now have highly significant effects. Further, an examination of their coefficients makes clear why the proportional odds model does not work well for these variables.

In the proportional odds model, the effect of age was estimated at –.042. In the PPO model, however, it is seen that the effect of age differs greatly across the cumulative logits, starting at –.172 and then declining, actually becoming slightly positive in the last cumulative logit. This is similar to the pattern found in Table 1-2. Age clearly has an effect on attitudes but that effect does not conform to the rigid pattern assumed by the proportional odds model.

Gender shows similar results. In the PO model, its effect was estimated as a weak and statistically insignificant .096. But in the PPO model, the estimated effect starts at .484, declines substantially across each cumulative logit, and again actually reverses sign in the final cumulative logit. Estimating a single coefficient of .096 disguises and distorts this variability in effects.

Satisfaction with the economy shows perhaps the most interesting differences from the PO model. The original PO effect of –.049 suggested that the less satisfied someone is with the economy, the more likely they are to support government action. This is not an unreasonable finding, but the PPO model indicates that the relationship is much more complicated than that. In the PPO model, the coefficients in the first two cumulative logits are positive while the last two are negative. This is similar to Model 1-3. The results suggest that greater satisfaction with the economy leads to people taking less extreme positions in either direction.

In short, estimating a proportional odds model rather a partial proportional odds would lead to serious errors in this case. The PO model says that four variables in the model do not have statistically significant effects, possibly leading to the conclusion that these variables are not important for explaining how people feel about government intervention on income inequality. The PPO model shows that three of those variables have highly significant effects. The PO model says that the last three variables in the model have the same and somewhat weak (or nonexistent) effects across all the cumulative logits. The PPO model actually shows that their effects differ considerably. To some extent the differences across models are a matter of degree; that is, signs tend to be the same across the cumulative logits but the magnitudes of coefficients differ. But in the case of satisfaction with the economy, the PPO model suggests a much more complex relationship where those with more extreme feelings on the economy actually tend to have more middle-range feelings on government intervention.

Clearly, the partial proportional odds model has key differences with the proportional odds model. But what substantive interpretations can we attach to such differences? The next section deals with that issue. We do not claim to offer a precise set of guidelines for when each interpretation should be preferred; but we do offer several examples of the conditions under which a possible interpretation should at least be considered.

[5]Craemer (2009) offers excellent examples of how to format tables that present results from partial proportional odds models. We have adapted his approach here.

4.1 *Interpreting the gologit model*

Empirically, the gologit/ppo model can work very well, providing a substantially better fit to the data than the ordered logit model does while at the same time being much more parsimonious than other alternatives. However, the interpretation and justification for the gologit model is less straightforward than it is for the ordered logit model. Unfortunately, many, perhaps most, researchers note only the superior fit of the gologit/ppo model and say little about what the results might mean.

As noted earlier, one way to motivate the ordered logit model is to say that there is an underlying latent variable, Y^*, and that as people cross thresholds on this underlying variable their values on the observed ordinal variable Y changes. This rationale is viable because there is a single equation for Y^*. The idea of an underlying Y^* becomes problematic, however, once a model allows for more than one equation, since, in effect, each equation comes up with a different estimate of Y^*. How then can the results of the gologit/ppo model be interpreted and justified? There are at least five possible approaches.

4.2 *Interpretation 1: The model is misspecified*

Whenever the assumptions of any model appear to be violated, it is tempting to quickly turn to more advanced techniques. However, researchers should first consider simpler alternatives. Have key variables been omitted? Do squared terms need to be included in the model? Model misspecification could cause the proportional odds assumption to appear to be violated when a better specified model might not show such a violation. For example, Williams (2010) gives an example where the simple addition of a squared term to a model is both theoretically reasonable and leads to tests showing that assumptions of the model are met. As useful as the gologit model is, researchers should examine whether a modified ologit model is simpler, valid, and easier to understand. In the case of our ESS example, as striking as the results are, researchers should consider whether important variables have been omitted from the model, and if so see whether the inclusion of those variables causes the proportional odds assumption to no longer be violated. We tried several different models and sets of dependent and independent variables with the ESS data and found that variations in model specification could indeed change which variables were or were not found to violate the proportional odds assumption. For example, in the current model, after adding additional measures on political ideology and perceived place in society, only satisfaction with the state of the economy continued to violate the proportional odds assumption (although this may be partly because missing data reduced the sample size). In a model in which attitudes toward gays and lesbians was the dependent variable, only feelings about household income failed to meet the proportional odds assumption.

4.3 *Interpretation 2: Gologit as nonlinear probability model*

As Long and Freese (2006, p. 187) point out, "the ordinal regression model can also be developed as a nonlinear probability model without appealing to the idea of a latent variable." By way of extension, the simplest thing may just be to also interpret gologit as a nonlinear probability model that lets you estimate the determinants and probability of each outcome occurring. There is no need to rely on the idea of an underlying Y^* that accounts for the observed values of Y.

For employing this approach, Long and Freese (2014) note that the substantive effects of variables can be assessed via such things as adjusted predictions, marginal effects, and the examination of prototypical cases. They further show how such calculations can easily be done (for the current example their mtable program was used). For example, according to the ESS gologit model, a nonethnic minority woman with average values on the other explanatory variables has a 52.5% predicted probability of agreeing that government should try to reduce economic inequality. The corresponding predicted probability for a nonethnic minority male is only 45.1%. However, the predicted probability of strongly agreeing is slightly higher for nonethnic minority men, 17.3%, than it is for nonethnic minority women, 14.9%. Thus, according to the model, women are more

likely than men to agree or strongly agree that government should take action (67.4% vs. 62.4%). But men who agree tend to be a bit more intense in their feelings, with more of them saying they strongly agree. Thus, by computing and comparing the predicted probabilities of outcomes under different conditions, the effect of gender is made much clearer than it might be if only the coefficients were focused on[6]; indeed the coefficients do not need to be examined at all.

While simple and convenient, always adopting this approach could be a serious mistake, causing someone to overlook important insights that might be offered by the other alternatives.

4.4 *Interpretation 3: The effect of* x *on* y *is asymmetrical and is not the same across each of the cumulative logits*

The proportional odds model assumes that, for each cumulate logit model that can be estimated (e.g., 1 versus 2, 3 4; 1, 2 vs 3, 4; 1, 2, 3 vs 4), the effect of *X* on *Y* is the same. Both our real and hypothetical examples illustrated why such an assumption is often unreasonable and too restrictive. In our ESS example, there are no compelling reasons for believing that age and gender must have the same effects in each of the cumulative logits, e.g. just because increases in a variable decrease the likelihood of strongly disagreeing with government action does not mean that those increases will have equally strong effects on making people more likely to strongly agree.

Fullerton and Dixon (2010) refer to this as *asymmetrical effects*. In their work they find that several key determinants of attitudes toward government spending on welfare have much stronger effects on opposition to it than on support for it. They further argue (p. 649) that the generalized ordered logit model has key advantages over other techniques when such asymmetries exist: "Traditional OLS models do not allow for the possibility that age, period, and cohort may affect support for education spending but not opposition to it. Similarly, other models used in previous research—such as binary logistic regression—do not allow for this possibility and may even obfuscate it, given that collapsing categories [of an ordinal outcome variable] results in a loss of information."

Hedeker and Mermelstein (1998) offer another good example of how the effects of *X* on *Y* could differ across the various cumulative logit models. The categories of the dependent variable may represent stages, such as precontemplation, contemplation, and action. An intervention might be effective in moving people from precontemplation to contemplation, but be ineffective in moving people from contemplation to action. If so, the effects of an explanatory variable will not be the same across the K-1 cumulative logits of the model.

Boes and Winkelman (2004, p. 2) offer yet another plausible example where the effect of *X* on *Y* could reasonably be expected to vary across the different cumulative logits:

> Completely missing so far is any evidence whether the magnitude of the income effect depends on a person's happiness: Is it possible that the effect of income on happiness is different in different parts of the outcome distribution? Could it be that "money cannot buy happiness, but buy-off unhappiness" as a proverb says? And if so, how can such distributional effects be quantified?

Counter to what the proverb says, Boes and Winkelman (2004, p. 21) found that there was some evidence that money can buy happiness.

Fullerton and Dixon (2010) offer yet another argument for why effects may be asymmetric. They suggest that an observed ordinal variable may actually reflect two different underlying latent variables: support for the policy in question, and opposition to it. The determinants of each need not be the same or have effects that are equal in magnitude. They note that this may explain why determinants of attitudes toward welfare spending have much stronger effects on opposition to it than on support for it. As noted before, they contend that a gologit/ppo model is better suited to deal with such asymmetries than are other widely used alternatives.

[6]Such computations and comparisons can of course be done even when the proportional odds assumption is not violated. But, the approach may be especially useful with more complicated models where multiple effects for the same variables are estimated.

It is easy to see how asymmetrical effects might apply to our ESS findings. Fullerton and Dixon found that that several key determinants of attitudes toward government spending on welfare have much stronger effects on opposition to it than on support for it; perhaps the same applies to age and gender when it comes to opposition and support for government intervention on reducing income inequality. It is less obvious how the other possible reasons for asymmetrical effects would apply to the ESS but they could be relevant for other research topics.

As these examples illustrate, asymmetric relationships often make good theoretical sense and reveal substantive insights into the underlying relationships between variables. Yet many, perhaps most, studies simply report the coefficients from the gologit model without explaining why asymmetric relationships exist or what they mean. Even worse, other techniques, such as OLS regression, logistic regression with a collapsed ordinal variable, and the ordered logit model itself may obscure asymmetric relationships completely.

The next two interpretations offer additional reasons as to why relationships may be asymmetric.

4.5 *Interpretation 4: State-dependent reporting bias*

As we have argued, one way to motivate the ordered logit model is to argue that there is an unobserved continuous variable Y^* that gets collapsed into the limited number of categories for the observed variable y. In some cases, respondents will be told how to do that collapse, e.g. they will be asked to report their income within specified ranges. However, in many other situations, respondents have to decide for themselves how the collapsing should be done; for example, they have to decide whether their feelings cross the threshold between "agree" and "strongly agree," whether their health is "good" or "very good," etc.

Respondents do not necessarily need to do this collapsing the same way. For example, respondents may not use the same frame of reference when answering. For example, the elderly may use a different frame of reference than the young do when assessing their health. Or, some groups may be more modest in describing their wealth, IQ or other characteristics.

In these cases the underlying latent variable may be the same for all groups; but the thresholds/cut points used may vary. For example, an estimated gender effect could reflect differences in measurement across genders rather than a real gender effect on the outcome of interest. Lindeboom and van Doorslaer (2004; see also Schneider, Pfarr, Schneider, & Ulrich, 2012) note that this has been referred to as state-dependent reporting bias, scale of reference bias, response category cut-point shift, reporting heterogeneity, heterogeneous reporting behavior, and differential item functioning.

Note that this interpretation is not unique to gologit/ppo models, although such models may make the alternative interpretation stand out more. If the difference in thresholds is constant (index shift), proportional odds will still hold. For example, women's thresholds could all be a half point higher than the corresponding male thresholds. An ordered logit model could be used in such cases. However, if the difference is not constant (cut point shift), proportional odds will be violated. For example, men and women might have the same thresholds at lower levels of pain but have different thresholds for higher levels. A partial proportional odds model can capture this. For example, if gender is coded 1 = female and 0 = male, for women we can add the gender coefficients to the threshold estimates, thus giving us one set of threshold estimates for women and a different set for men.

If apparent effects reflect differences in measurement rather than real differences in effects, then the idea of an underlying Y^* is preserved; determinants of Y^* are the same for all, but reporting thresholds and the ways in which people classify themselves differ across individuals and groups.

Schneider et al. (2012) argue that reporting heterogeneity does indeed occur and that the consequences of it can be severe. In their study of self-assessed health, they note (p. 251) that "classification into a response category (good, fair, low, etc.) may systematically differ across population subgroups resulting in reporting heterogeneity and estimation problems." They further note (p. 253) that earlier studies "find evidence that subgroups of the population might use systematically different

thresholds in classifying their health into a categorical measure even if the underlying true health is at the same level." They warn (p. 251) that "neglecting these possible sources of heterogeneity may lead to an over- or underestimation of health effects making policy recommendations unreliable." Indeed, in their own study they found that the answering behavior of men and women differed significantly. They concluded (p.261) that "evaluating questionnaires from panel data surveys based on population subgroups or questionnaires from different countries comes with the risk of comparing apples with oranges if the problem of reporting heterogeneity is not adequately taken into account."

Given that Schneider et al. (2012) found evidence of reporting heterogeneity by gender for self-assessed health, researchers might want to consider whether the same might be true for the ESS respondents when reporting on their feelings concerning government intervention and income inequality. The failure of the gender coefficients to meet the proportional odds assumption could be due to the fact that men and women who actually share the same attitudes nonetheless report different answers. Perhaps one gender uses a higher threshold than the other before choosing to say that it strongly supports intervention as opposed to merely supporting it. Similarly, different age groups may have different definitions of what it means to strongly oppose something rather than just oppose it. If this were true, then the differences in the gender and age coefficients could reflect differences in measurement across group groups rather than real differences in the effects.

A key advantage of this interpretation, if it is valid, is that it could greatly improve cross-group comparisons, getting rid of artifactual differences caused by differences in measurement. For example, women might appear to have higher scores than men only because women are using different cut points when reporting their values on the observed Y. A key concern, however, is how can you really be sure the coefficients reflect differences in measurement and not real effects, or some combination of real and measurement effects? Theory may help: if your model strongly claims the effect of gender should be zero then any observed effect of gender can be attributed to measurement differences. Schneider et al. (2012) also suggest that, in the case of self-assessed health, including objectified indicators of health (e.g., health related behaviors such as smoking and experience with severe or chronic illness) can help to detect and correct for biases in reporting behavior. In any event, researchers should at least consider the possibility that apparent differences in effects could just be measurement artifacts caused by different groups classifying themselves differently.

4.6 *Interpretation 5: Some explanatory variables affect the direction of responses while others affect their intensity*

A variable that is ordinal in some respects may not be ordinal or else be differently-ordinal in others. Variables could be ordered either by direction (Strongly Disagree to Strongly Agree) or intensity (Indifferent to Feel Strongly).

For example, suppose women tend to take less extreme political positions than men. The example presented in Table 1-3 might reflect such a relationship. Using the first (directional) coding, an ordinal model might not work very well, whereas it could work well with the second (intensity) coding.

But suppose that for every other explanatory variable the directional coding works fine in an ordinal model. As in Table 1–3, the assumption of the ordered logit model will be violated because of the one problematic variable. Our choices in the past have either been to (a) run ordered logit, with the model really not appropriate for the gender variable, or (b) run multinomial logit, ignoring the parsimony of the ologit/po model just because one variable does not work with it. With gologit/partial proportional odds models, we have option (c)—constrain the variables where it works to meet the parallel lines assumption, while freeing up other variables (e.g., gender or, in the case of our ESS example, satisfaction with the economy) from that constraint.

Williams (2006) offers another example of this type of differing relationships between variables. In a study of self-reported health status, he finds that women are both less likely to say they are in poor health and also less likely to report that they are in very good health. This is similar to the pattern found in Table 1-3. This

might mean that women tend to give less extreme answers in either direction when reporting their health. Alternatively, it might mean that men are more variable in their health, causing more men to be found at either extreme. Other variables in his model either meet the proportional odds assumption or have effects that differ in magnitude but not direction. Similarly, in the ESS study, it may be that satisfaction with the economy is related more to the intensity of opinions concerning government intervention in the economy than it is to the direction of the attitudes. Those who are satisfied with the economy may feel that, so long as they themselves are okay, they do not care that much either way what the government does about income inequality. Those who are less satisfied, however, may have more intense feelings on the subject.

5. Other issues with the gologit model

There are several other issues to be aware of with the gologit/partial proportional odds model. First, it probably works best when relatively few of the variables in the model violate the proportional odds assumption. If several variables violate the assumption, then the gologit model offers little in the way of parsimony and more widely known techniques such as multinomial logit may be superior.

Second, unlike many other methods for the analysis of categorical data, the gologit model can produce negative predicted probabilities. McCullagh and Nelder discuss this in *Generalized Linear Models* (1989, p. 155):

> The usefulness of non-parallel regression models is limited to some extent by the fact that the lines must eventually intersect. Negative fitted values are then unavoidable for some values of x, though perhaps not in the observed range. If such intersections occur in a sufficiently remote region of the x-space, this flaw in the model need not be serious.

The support page for gologit2 (Williams, 2014) reports that such problems are apparently rare. If the problem does occur, sometimes combining categories of the response variable (especially if the *N*s for some categories are small) and/or simplifying the model helps. The imposition of parallel lines constraints on variables may also help because it reduces the likelihood of non-parallel lines intersecting. More stringent p values (which Williams (2006) recommends anyway since multiple tests are being conducted) will tend to increase the number of variables that meet the parallel lines constraint. If the number of negative predicted probabilities is still non-trivial, a different statistical technique may be preferable.

Third, when sample sizes are large, even small violations of the proportional odds assumption can be statistically significant. The researchers may wish to assess whether the deviations from proportionality are substantively important enough to warrant moving away from the more parsimonious ordered logit model. For example, parameter estimates across cumulative logits of .73, .77, and .80 may not be worth discussing even if they are significantly different from each other. Indeed, some examples that were considered for this paper were ultimately rejected because, even though the proportional odds assumption was violated, the substantive impact on the conclusions seemed minor. Conversely, larger variations may be worth noting, especially if they lead to the additional insights suggested by one or more of the possible interpretations of the gologit/ppo model discussed earlier. It may be especially important to focus on effects that differ in both magnitude and direction across cumulative logits.

Alternatively, other criteria can be used for assessing which model is best. In particular, information measures such as BIC (Raftery, 1995) and AIC (Hardin & Hilbe, 2007) can be used to compare the relative plausibility of two models rather than to find the absolute deviation of observed data from a particular model. Bayesian Information Criterion (BIC), Akaike's Information Criterion (AIC), and other information measures have penalties for including parameters that do not significantly improve fit. Particularly with large samples, the information measures can lead to more parsimonious but adequate models.[7]

[7]The ideal situation, of course, is when BIC, AIC, and likelihood ratio tests all lead to the same conclusion. However, as Dziak, Coffman, Lanza, and Li (2012) point out, this will not always be the case. Based on their simulations, they argue that AIC often risks choosing too large a model while BIC sometimes leads to selecting a model that is too parsimonious. But ultimately, when criteria conflict, they recommend going with what seems most important in a given situation: Is it worse to have a model that is too parsimonious, or a model that is not parsimonious enough?

Fourth, the researcher must somehow decide which variables should have the proportional odds constraint imposed and which should not. Ideally, researchers should have strong theoretical rationales to guide them. But since such theory rarely exists, empirical means are often used instead. For example, the Brant test can also be used to identify individual variables that violate assumptions. The gologit2 routine in Stata (Williams, 2006) uses a stepwise procedure called autofit to identify variables where proportionality constraints should be relaxed. Schneider et al. use the same procedure in their regoprob2 program (which was adapted from gologit2) arguing that "as long as no underlying theory would suggest which variables violate the parallel-lines assumption, the best solution is to use an iterative fitting procedure" (2012, p. 259). Such claims are debatable, but remember that other alternatives have their problems too, for example, sticking with a model whose assumptions are known to be violated (the proportional odds model) or switching to a model that has far more parameters than is necessary (the multinomial logit) are not ideal solutions either. Like all empirical stepwise procedures, caution should be used to avoid capitalizing on chance, e.g. more stringent p values can be used or the sample can be divided into two parts to see whether results are consistent across the subsamples.

Finally, there are other alternatives to the ordered logit model. For example, Williams (2009, 2010) notes that the ordered logit model also assumes that errors are homoscedastic, and that differences across groups in residual variability may cause tests of the proportional odds assumption to fail. The heterogeneous choice model addresses those problems by explicitly modeling the causes of the differing variability. Williams shows that a heterogeneous choice model can sometimes fit the data about as well as a gologit/ppo model and may be easier to explain and justify theoretically. Fullerton (2009) offers several other alternative models for ordinal data that may be appropriate given the nature of the data and the problem.

6. Conclusions

Few people would use any statistical method if they only had a vague idea of what the results meant; and even if they did have a vague idea, they might well overlook many of the useful insights the method could offer. The gologit/ppo method is somewhat popular now because it offers a "fix" to the problem of violated assumptions in the proportional odds model. This article has argued that the method has the potential to do much more than that. Several ideas have been offered on how gologit/ppo models can be interpreted. In the process, the article has also made a case for why those models should be used in the first place and why, when they are used, their implications should be considered more carefully than they typically have been.

The ordered logit model is one of the most popular methods for analyzing ordinal outcome variables (Long & Freese, 2014). But because its assumptions are often violated, the gologit/partial proportional odds model will often be a desirable alternative. When the gologit model works well, the effects of those variables that meet the proportional odds assumption will have the same interpretation as they do in the ordered logit model. For other variables, differences in effects will sometimes just be matters of degree. In some instances, such as when Y can be ordered in different ways (e.g., Strongly Disagree to Strongly Agree, Indifferent to Feels Strongly) the model can accommodate Xs that differ in both the magnitude of their effects and in their direction.

The gologit model may also help researchers to avoid serious errors concerning statistical significance that could lead them to erroneously conclude that an explanatory variable has little or no effect on the outcome variable being studied. In addition, it can sometimes address problems of reporting heterogeneity where groups differ in the ways they respond to questions, which can potentially causes differences in measurement across groups to be mistaken for differences in effects.

Now that software is available that can readily estimate gologit/ppo models, researchers should consider whether such models may better meet their needs than the ordered logit and multinomial

logit models do. Researchers should further consider whether such models offer important insights that they have been overlooking.

Acknowledgments

The author wishes to thank Richard Campbell, Sarah Mustillo, J. Scott Long, and the anonymous reviewers and the editor for their many helpful comments.

References

Boes, S., & Winkelmann, R. (2004). *Income and happiness: New results from generalized threshold and sequential models*. IZA *discussion paper* 1175. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=561724.

Brant, R. (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, *46*, 1171–1178. doi:10.2307/2532457

Cornwell, B., Laumann, E. O., & Schumm, L. P. (2008). The social connectedness of older adults: *A national profile. American Sociological Review*, *73*, 185–203. doi:10.1177/000312240807300201

Craemer, T. (2009). Psychological 'self–other overlap' and support for slavery reparations. *Social Science Research*, *38*, 668–680. doi:10.1016/j.ssresearch.2009.03.006

Do, Y. K., & Farooqui, M. A. (2011). Differential subjective responsiveness to a future cigarette price increase among South Korean youth smokers. *Nicotine and Tobacco Research October* 14, 209–216. doi:10.1093/ntr/ntr187

Dziak, J. J., Coffman, D. L., Lanza, S. L., & Runze, L. (2012). *Sensitivity and specificity of information criteria. The Pennsylvania State University Technical Report Series #12-119*. Retrieved from http://methodology.psu.edu/media/techreports/12-119.pdf. University Park, PA: The Methodology Center.

ESS Round 6: European Social Survey Round 6 Data. (2012). *Data file edition 2.1. Norwegian social science data services, Norway – data archive and distributor of ESS data*. Retrieved from http://www.europeansocialsurvey.org/

Fu, V. (1998). *Sg88: Estimating generalized ordered logit models. Stata technical bulletin 44: 27–30*. In Stata technical bulletin reprints (Vol. 8, pp. 160–164). College Station, TX: Stata Press.

Fullerton, A. S. (2009). A conceptual framework for ordered logistic regression models. *Sociological Methods & Research*, *38*, 306–347. doi:10.1177/0049124109346162

Fullerton, A. S., & Dixon, J. C. (2010). Generational conflict or methodological artifact? Reconsidering the relationship between age and policy attitudes in the U.S., 1984–2008. *Public Opinion Quarterly*, *74*(4), 643–673. doi:10.1093/poq/nfq043

Hardin, J. W., & Hilbe, J. M. (2007). *Generalized linear models and extensions* (2nd ed.). College Station, TX: Stata Press.

Hedeker, D., & Mermelstein, R. J. (1998). A multilevel thresholds of change model for analysis of stages of change data. *Multivariate Behavioral Research*, *33*(4), 427–455. doi:10.1207/s15327906mbr3304_1

Kleinjans, K. J. (2015). Do gender differences in preferences for competition matter for occupational expectations? *Journal of Economic Psychology*, *30*, 701–710. doi:10.1016/j.joep.2009.03.006.

Lehrer, J. A., Lehrer, E. L., Zhao, Z., & Lehrer, V. L. (2007). *Physical dating violence among college students in Chile (April 2007)*. IZA Discussion Paper No. 2753. Retrieved from http://ssrn.com/abstract=982623

Lindeboom, M., & van Doorslaer, E. (2004). Cut-point shift and index shift in self-reported health. *Journal of Health Economics*, *23*(6), 1083–1099. doi:10.1016/j.jhealeco.2004.01.002

Long, J. S., & Freese, J. (2006). *Regression models for categorical dependent variables using stata* (2nd ed.). College Station, TX: Stata Press.

Long, J. S., & Freese, J. (2014). *Regression models for categorical dependent variables using stata* (3rd ed.). College Station, TX: Stata Press.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). London, UK: Chapman and Hall.

Peterson, B., & Harrell, F. E. Jr. (1990). Partial proportional odds models for ordinal response variables. *Applied Statistics*, *39*(2), 205–217. doi:10.2307/2347760

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, *25*, 111–163. doi:10.2307/271063

Schafer, M. H., & Upenieks, L. (2015). Environmental disorder and functional decline among older adults: *A layered context approach. Social Science and Medicine*, *124*, 152–161. doi:10.1016/j.socscimed.2014.11.037

Schneider, U., Pfarr, C., Schneider, B. S., & Ulrich, V. (2012). I feel good! Gender differences and reporting heterogeneity in self-assessed health. *The European Journal of Health Economics*, *13*(3), 251–265. doi:10.1007/s10198-011-0301-7

Williams, R. (2006). Generalized ordered logit/partial proportional odds models for ordinal dependent variables. *Stata Journal*, *6*, 58–82.

Williams, R. (2009). Using heterogeneous choice models to compare logit and probit coefficients across groups. *Sociological Methods & Research*, *37*(4), 531–559. doi:10.1177/0049124109335735

Williams, R. (2010). Fitting heterogeneous choice models with oglm. *The Stata Journal*, *10*(4), 540–567.

Williams, R. (2014). *Gologit2/OGLM troubleshooting*. Retrieved from http://www3.nd.edu/~rwilliam/gologit2/tsfaq.html