Regression Models for Nominal and Ordinal Outcomes [1]

J. Scott Long

Indiana University

2012-05-29

Forthcoming in Best and Wolf (editors), *Regression Models*, Sage Publications

Abstract

Advances in software make regression models for nominal and ordinal outcomes simple to estimate. The greatest challenge is finding a model that is appropriate for your application and interpreting the results to highlight the key findings from these often complicated, nonlinear models. When choosing a model it is important to realize that ordinal models restrict the relationship between regressors and the probabilities of the outcomes. The classic definition of ordinality assumes ranking on a single attribute, but many seemingly ordinal variables can be ranked on multiple dimensions. In such cases the constraints in ordinal models can lead to incorrect conclusions. Models for nominal outcomes do not impose ordinality, but at the cost of additional parameters. While it is tempting to reduce the number of parameters with stepwise procedures, this risks over-fitting the data. Interpretation of models for nominal and ordinal outcomes uses odds ratios and quantities based on predicted probabilities. Odds ratios do not depend on the values of the regressors, but the meaning of odds ratios in terms of probabilities depends on the values of the regressors. Changes in probabilities have a direct interpretation, but the magnitude of the change depends on the values of all regressors. There is no simple solution to the interpretation of nonlinear models. Interpretation requires detailed post-estimation analyses to determine the most important findings and to find an elegant way to present them.

[1] I thank Rich Williams, Andy Fullerton, Tom VanHeuvelen, and Mike Vasseur for their comments. They are not, of course, responsible for my ignoring their advice on some issues.

1 Introduction to the method

Ordinal and nominal outcomes are common in the social sciences with examples ranging from Likert scales in surveys to assessments of physical health to how armed conflicts are resolved. Since the 1980s numerous regression models for nominal and ordinal outcomes have been developed. These models are essentially sets of binary regressions that are estimated simultaneously with constraints on the parameters. With current software making estimation routine, the greatest challenge is interpretation. Finding an effective way to convey the results of models for nominal and ordinal outcomes is a vexingly difficult art that requires time, practice, and a firm grounding in the goals of your analysis and the characteristics of your model. Too often interpretation is limited to a table of coefficients with a brief discussion of signs and statistical significance. While the implications of a model are implicit in the parameters, postestimation computations of probabilities and related quantities are essential for understanding the substantive impact of the regressors.

The goal in selecting a model is to find a model that is parsimonious without distorting critical relationships. A too simple model risks bias, while an unnecessarily complex model is statistically inefficient. Models for nominal outcomes are sometimes avoided because of the number of parameters and perceived difficulty in their interpretation. Ordinal models have fewer parameters, but this simplicity is achieved by imposing constraints that potentially distort the process being modeled. While nominal models have more parameters to interpret, this complexity is transparent when probabilities are used for interpretation since software easily makes the computations. On the other hand, if stepwise procedures are used to find a simpler model, the resulting model can depend on peculiarities of the sample rather than the underlying process. Overall, a firm grounding in the substantive and theoretical context of your research accompanied by an evaluation of the robustness of your results to alternative specifications are fundamental to using regression models for nominal and ordinal outcomes.

1.1 What does ordinal or nominal mean?

S. S. Stevens (1946) provided the initial definitions of nominal and ordinal variables:

> *Nominal scales* assign numbers to categories as labels with no ordering implied by the numbers.
>
> *Ordinal scales* use numbers to indicate rank ordering on a single attribute.

Even though Stevens' taxonomy was hotly debated when it was proposed and has been critiqued since (see Velleman and Wilkinson 1993 for a review), it is firmly established in the methods of many disciplines and is often used to classify models. Many variables commonly thought of as ordinal do not meet Stevens' criterion since they reflect multiple attributes. Consider political party affiliation which is used as an example in this paper. Affiliation was collected from a survey using the categories *Strong Democrat* (1=*SD*), *Democrat* (2=*D*), *Independent* (3=*I*), *Republican* (4=*R*), and *Strong Republican* (5=*SR*). On the attribute of left-right orientation the categories are ranked from 1=*SD* to 5=*SR*. In terms of intensity of partisanship the categories are ordered 1=*I*; 2=*R&D*; and 3=*SR&SD*. Anticipating results from Section 3, age could increase intensity of partisanship so that both *SD* and *SR* increase with age, while income could affect left-right orientation but not intensity.

Ordinal models constrain the relationship between regressors and outcomes in specific way that was elaborated by Anderson (1984). Suppose that x has a positive effect. As x increases the lowest outcome category decreases in probability from 1 to 0 while the highest increases from 0 to 1. The probabilities for other categories are bell-shaped with modes that increase for higher categories. This is illustrated in Figure 1. While non-ordinal models can lead to predictions consistent with an ordinal model, they are not constrained to do so. Table 1 lists the models reviewed in this paper and indicates which models are ordinal (note that "ordinal" in the name does not make it ordinal!).

– Table 1 and Figure 1 here –

2 Mathematical foundations and advanced aspects

I begin with the multinomial logit model (MNLM) since it builds directly on the binary logit model of the last chapter. Next I consider two models that are closely linked to the MNLM. The adjacent category logit model (ACLM) constrains the MNLM parameters so that the effect of x_k on the odds is identical for all adjacent categories (e.g., 1 and 2, 2 and 3, and so on). Anderson's (1984) stereotype logit model (SLM) modifies the MNLM to reduce the number of parameters. Next I consider the most common ordinal model, often called *the* ordinal regression model (OLM), which is a set of logits on binary outcomes that divide the outcome into lower and higher categories (e.g., 1 versus higher categories; 1 and 2 versus higher categories). The effect of x_k is constrained to be equal in all equations, a constraint that is often unrealistic. In response to limitations of the OLM, the generalized ordinal regression model (GOLM) allows the effects of x_k to differ across equations. This model has as many parameters as the MNLM and is (I think) more complicated to interpret than the MNLM. Models that fall between the OLM and the GOLM are considered briefly. In most cases both logit and probit versions of these models are available and produce nearly identical predictions. I focus on logit models since they can be interpreted using odds ratios while probit models cannot.

These models can be interpreted using predicted probabilities and logit models with odds ratios. In this section I develop the formula for these quantities with examples of their use in Section 3. While the models can be parameterized in a several ways, I use parameterizations and notation that emphasize the similarities among models. The outcome y has J categories with K regressors x_1 through x_K. The intercept is β_0 with the linear combination of regressors and coefficients written as $\mathbf{x}'\boldsymbol{\beta} = \beta_1 x_1 + \cdots + \beta_K x_K$. For some models the β's have additional subscripts such as $\beta_{k,m|n}$. I introduce each model using three outcomes and two regressors before presenting the general form of the model.

2.1 Multinomial logit model (MNLM)

Multinomial logit, the most common model for nominal outcomes, is equivalent to a set of binary logits (BLM) for all pairs of outcome categories. To see this, let y equal D=*Democrat* and R=*Republican*. With two regressors the model is:

$$\ln \frac{\Pr(y = D \mid \mathbf{x})}{\Pr(y = R \mid \mathbf{x})} = \beta_{0,D|R} + \beta_{age,D|R} age + \beta_{income,D|R} income$$

If I add the outcome I=*Independent*, there are three binary comparisons:[2]

$$\begin{aligned}
\ln \frac{\Pr(y = D \mid \mathbf{x})}{\Pr(y = R \mid \mathbf{x})} &= \beta_{0,D|R} + \beta_{age,D|R} age + \beta_{income,D|R} income \\
\ln \frac{\Pr(y = D \mid \mathbf{x})}{\Pr(y = I \mid \mathbf{x})} &= \beta_{0,D|I} + \beta_{age,D|I} age + \beta_{income,D|I} income \\
\ln \frac{\Pr(y = R \mid \mathbf{x})}{\Pr(y = I \mid \mathbf{x})} &= \beta_{0,R|I} + \beta_{age,R|I} age + \beta_{income,R|I} income
\end{aligned} \tag{1}$$

Begg and Gray (1984) show that estimates of the binary logits are consistent but inefficient estimates of the MNLM. Software for the MNLM obtains efficient estimates by simultaneously estimating all equations while imposing mathematically necessary constraints that link the equations. These constraints can be seen in the mathematically necessary relationship:

$$\ln \frac{\Pr(y = D \mid \mathbf{x})}{\Pr(y = R \mid \mathbf{x})} = \ln \frac{\Pr(y = D \mid \mathbf{x})}{\Pr(y = I \mid \mathbf{x})} - \ln \frac{\Pr(y = R \mid \mathbf{x})}{\Pr(y = I \mid \mathbf{x})}$$

which implies that $\beta_{k,D|R} = \beta_{k,D|I} - \beta_{k,R|I}$. Accordingly, if I know the coefficients for any two of the binary logits I can determine *exactly* the coefficients for the remaining logit. The smallest set of parameters that implies the parameters for all comparisons is called a *minimal set*. Often the minimal set consists of all comparisons relative to one of the categories referred to as the *base category*. I assume the base category is J but other values could be used.

[2] I exclude the redundant comparisons R versus D, I versus D, and I versus R.

Defining the odds of category j versus base category J given $\mathbf{x}$ as $\Omega_{j|J}(\mathbf{x}) = \frac{\Pr(y=j|\mathbf{x})}{\Pr(y=J|\mathbf{x})}$ and $\mathbf{x}'\boldsymbol{\beta}_{j|J} = \beta_{1,j|J}x_1 + \cdots + \beta_{K,j|J}x_K$, the MNLM is:

$$\ln \Omega_{j|J}(\mathbf{x}) = \ln \frac{\Pr(y = j \mid \mathbf{x})}{\Pr(y = J \mid \mathbf{x})} = \beta_{0,j|J} + \mathbf{x}'\boldsymbol{\beta}_{j|J} \text{ for } j = 1, J$$

Since $\Omega_{J|J}(\mathbf{x}) = 1$, then $\beta_{0,J|J=0}$ and $\boldsymbol{\beta}_{J|J} = \mathbf{0}$. Taking the exponential:

$$\Omega_{j|J}(\mathbf{x}) = \exp\left(\beta_{0,j|J} + \mathbf{x}'\boldsymbol{\beta}_{j|J}\right)$$

with the odds ratio:

$$OR_{j|J,k} = \frac{\Omega_{j|J}(\mathbf{x}, x_k + 1)}{\Omega_{j|J}(\mathbf{x}, x_k)} = \exp\left(\beta_{k,j|J}\right)$$

The odds for any two categories m and n is:

$$\begin{aligned} \Omega_{m|n}(\mathbf{x}) &= \exp\left(\left[\beta_{0,m|J} + \mathbf{x}'\boldsymbol{\beta}_{m|J}\right] - \left[\beta_{0,n|J} + \mathbf{x}'\boldsymbol{\beta}_{n|J}\right]\right) \\ &= \exp\left(\beta_{0,m|n} + \mathbf{x}'\boldsymbol{\beta}_{m|n}\right) \end{aligned}$$

where $\beta_{0,m|n} = \beta_{0,m|J} - \beta_{0,n|J}$ with the corresponding odds ratio:

$$OR_{m|n,k} = \exp\left(\beta_{k,m|J} - \beta_{k,n|J}\right)$$

The OR can be interpreted as:

> For a unit increases in x_k, the odds of outcome m versus n change by a factor of $\exp\left(\beta_{k,m|J} - \beta_{k,n|J}\right)$, holding other variables constant.

If the OR is greater than one, you might say: "The odds are OR times larger"; if less than one, "The odds are OR times smaller." From the equations for the odds, we can derive the probability of outcome j as:

$$\Pr(y = j \mid \mathbf{x}) = \frac{\exp\left(\beta_{0,j|J} + \mathbf{x}'\boldsymbol{\beta}_{j|J}\right)}{\sum_{q=1}^{J} \exp\left(\beta_{0,q|J} + \mathbf{x}'\boldsymbol{\beta}_{q|J}\right)} \text{ for } j = 1, ..., J \tag{2}$$

Since there are $J-1$ coefficients for each regressor, if a variable has no effect $J-1$ coefficients must be simultaneously 0. In our example, the hypothesis that age has no effect is H_{age}: $\beta_{age,D|I} = \beta_{age,R|I} = 0$. H_{age} is *not* equivalent to the pair of hypotheses $H_{D|I}$: $\beta_{age,D|I} = 0$ and $H_{R|I}$: $\beta_{age,R|I} = 0$ since it is possible to reject H_{age} while *not* rejecting either $H_{D|I}$ or $H_{R|I}$. How? Suppose that age has a *nonsignificant positive* effect on D versus I and a *nonsignificant negative* effect on R versus I. D and I could be close enough politically that age increases D relative to I but not significantly so. Conversely, age could decreases R relative to I. Since D and R are further apart politically, age could significantly increase D relative to R. In general, the hypothesis that x_k has no effect is:

$$H_{x_k}\colon \beta_{k,1|J} = \cdots = \beta_{k,J-1|J} = 0$$

which can be tested with a Wald or a LR test with $J-1$ degrees of freedom.

Independence of Irrelevant Alternatives (IIA) IIA is the defining property of the MNLM that simplifies estimation and interpretation, but is potentially unrealistic. IIA implies that a person's choice between two outcomes (i.e., alternatives) is unaffected by the other choices. Suppose you add an alternative that is similar to an existing alternative. You would expect that individuals would split between the original alternative and the new, similar alternative while dissimilar alternatives would be unaffected. IIA requires that the probabilities of all alternatives be decreased proportionately, which is behaviorally unrealistic. Numerous tests of IIA have been proposed and for decades multinomial probit was considered a solution if computational problems were solved. While theoretically compelling, these solutions are limited in practice.

Tests of IIA assess how estimates change when the model is estimated with a restricted set of outcomes (e.g., compare estimates using J outcomes to those obtained using $J-1$ outcomes). If the test is significant, the assumption of IIA is rejected indicating that the MNLM is inappropriate.

The Hausman-McFadden test (1984) and the Small-Hsiao test (1985) are the most common IIA tests. Using Monte Carlo experiments, Fry and Harris (1996, 1998) and Cheng and Long (2005) found these and other IIA tests to have poor statistical properties in finite samples. They conclude that IIA tests are *not* useful for assessing violations of IIA. The best advice regarding IIA goes back to an early statement by McFadden (1973) who wrote that the MNLM should only be used when the outcomes "can plausibly be assumed to be distinct and weighed independently in the eyes of each decision maker." If you have two outcomes that are very similar in how they are evaluated as choices, such as riding a red bus and riding a blue bus, combine the categories. Care in specifying the model to include distinct alternatives that are not substitutes is reasonable, albeit ambiguous, advice.

Multinomial Probit Model (MNPM) The probit counterpart to the MNLM is based on the normal distribution. Unlike the logistic distribution, the normal distribution allows choices to be correlated in the sense that a person can be more likely to selection choice m and n after controlling for regressors. This avoids the IIA assumption. For years the MNPM was considered an ideal solution to IIA if computational problems in estimation could be resolved. Estimation using simulation is now practical even though it requires 100's or 1000's times more computation. Still, several factors limit the model's usefulness. First, identification requires alternative-specific regressors. These are variables whose values depend on the outcome (i.e., alternative).[3] For example, the choice of mode of transportation for commuting could depend on the time each mode requires, where travel time would varies by alternative. When alternative-specific variables are not available, the MNPM is not identified. Second, even with alternative-specific regressors identification requires constraints on correlations among errors. Substantive motivation for these constraints is often unavailable. Finally, even if the model is *formally* identified, Keane (1992) finds that identification is

[3] In this paper I do not consider models with alternative specific regressors such as the conditional logit model.

fragile which means that additional restrictions are necessary to avoid the risk of unreliable estimates. My experience confirms Keane's (1992) statement: "Given the lack of practical experience with [multinomial probit] models, however, there is a need to develop a 'folklore' concerning the conditions under which the model performs well." Full details on the MNPM are found in Train (2009).

MNLM summary The MNLM is a flexible model that imposes few restrictions on the relationships between regressors and outcomes. While it can lead to relationships between regressors and outcome probabilities that are consistent with Anderson's definition of an ordinal model, it is not constrained so that it must do so. Ordinal models restrict the nature of the relationships. If these constraints are appropriate, statistical efficiency is gained and interpretation is simpler. However, as illustrated in Section 3, when the constraints are unrealistic, incorrect conclusions can be drawn. Regardless of your assessment of the reasonableness of assumptions imposed by ordinal models, I recommend estimating the MNLM or the GOLM (discussed below) to evaluate your model. If results differ in substantively meaningful ways, carefully assess the appropriateness of the model you are using.

2.2 Adjacent category logit model (ACLM)

The *adjacent categories logit model* (Goodman 1983; Clogg and Shihadeh 1994:149-154) is an ordinal regression model that constrains the MNLM so that coefficients from adjacent ordinal categories equal. For example, here is the MNLM for outcomes ordered 1, 2, and 3 (excluding the redundant equation for outcomes 1 and 3:

$$\begin{aligned}\ln\frac{\Pr(y=1\mid\mathbf{x})}{\Pr(y=2\mid\mathbf{x})} &= \beta_{0,1|2}+\beta_{age,1|2}age+\beta_{income,1|2}income\\ \ln\frac{\Pr(y=2\mid\mathbf{x})}{\Pr(y=3\mid\mathbf{x})} &= \beta_{0,2|3}+\beta_{age,2|3}age+\beta_{income,2|3}income\end{aligned}$$

The ACLM constrains the effects to be equal for adjacent categories, as shown by the lack of subscripts for β_{age} and β_{income}:

$$\begin{aligned}\ln \frac{\Pr(y=1 \mid \mathbf{x})}{\Pr(y=2 \mid \mathbf{x})} &= \beta_{0,1|2} + \beta_{age} age + \beta_{income} income \\ \ln \frac{\Pr(y=2 \mid \mathbf{x})}{\Pr(y=3 \mid \mathbf{x})} &= \beta_{0,2|3} + \beta_{age} age + \beta_{income} income\end{aligned}$$

This model makes the distance between 1 and 2 the same as the distance between 2 and 3 in the sense that OR_k for outcomes 1 versus 2 is the same as OR_k for 2 and 3. This implies that the comparison of 1 and 3 is constrained by:

$$\ln \frac{\Pr(y=1 \mid \mathbf{x})}{\Pr(y=3 \mid \mathbf{x})} = \ln \frac{\Pr(y=1 \mid \mathbf{x})}{\Pr(y=2 \mid \mathbf{x})} + \ln \frac{\Pr(y=2 \mid \mathbf{x})}{\Pr(y=3 \mid \mathbf{x})}$$

so that:

$$\begin{aligned}\ln \frac{\Pr(y=1 \mid \mathbf{x})}{\Pr(y=3 \mid \mathbf{x})} &= \left(\beta_{0,1|2} + \beta_{age} age + \beta_{income} income\right) + \left(\beta_{0,2|3} + \beta_{age} age + \beta_{income} income\right) \\ &= \left(\beta_{0,1|2} + \beta_{0,2|3}\right) + \left(2\beta_{age}\right) age + \left(2\beta_{income}\right) income\end{aligned}$$

The ACLM imposes constraints of the form $\beta_{k,j|j+2} = 2\beta_{k,j|j+1}$.

More generally, the ACLM can be written as:

$$\ln \frac{\Pr(y=j \mid \mathbf{x})}{\Pr(y=j+1 \mid \mathbf{x})} = \beta_{0,j|j+1} + \mathbf{x}'\boldsymbol{\beta} \text{ for } j = 1, J-1$$

where the intercepts vary by j but the effects of x_k do not. Estimation is possible with software for the MNLM that allows constraints on the parameters. Taking exponentials,

$$\Omega_{j|j+1}(\mathbf{x}) = \exp\left(\beta_{0,j|j+1} + \mathbf{x}'\boldsymbol{\beta}\right) \text{ for } j = 1, J-1$$

with the odds ratios:

$$\begin{aligned}OR_{j|j+1,k} &= \frac{\Omega_{j|j+1}(\mathbf{x}, x_k + 1)}{\Omega_{j|j+1}(\mathbf{x}, x_k)} = \exp(\beta_k) \\ OR_{j|j+q,k} &= \frac{\Omega_{j|j+2}(\mathbf{x}, x_k + 1)}{\Omega_{j|j+2}(\mathbf{x}, x_k)} = \exp(q\beta_k)\end{aligned}$$

We can interpret the parameters as:

For a unit increases in x_k, the odds of adjacent categories change by a factor of $\exp(\beta_k)$, holding other variables constant.

For a unit increases in x_k, the odds of categories separated by q change by a factor of $\exp(q\beta_1)$, holding other variables constant.

Probabilities are:

$$\Pr(y = j \mid \mathbf{x}) = \frac{\exp\left(\beta_{0,j|j+1} + \mathbf{x}'\boldsymbol{\beta}\right)}{1 + \sum_{q=1}^{J-1}\left[\exp\left(\beta_{0,q|q+1} + \mathbf{x}'\boldsymbol{\beta}\right)\right]} \text{ for } j = 1, J-1$$

$$\Pr(y = J \mid \mathbf{x}) = 1 - \sum_{q=1}^{J-1}\Pr(y = q \mid \mathbf{x})$$

The critical issue is whether the OR's for adjacent categories are equal. In social science research it seems unlikely that this would be suggested by theory. To test if your data supports these constraints, a LR test comparing the ACLM to the MNLM can be used. In my experience, the hypothesis is usually rejected with a large chi-square.

2.3 Stereotype logit model (SLM)

The *stereotype logit model* (SLM) was proposed by Anderson (1984) in response to the restrictive assumption of parallel regressions in the ordered logit model (presented next) and to reduce the number of parameters in the MNLM.[4] The MNLM with base J is:

$$\ln\frac{\Pr(y = j \mid \mathbf{x})}{\Pr(y = J \mid \mathbf{x})} = \beta_{0,j|J} + \beta_{age,j|J}age + \beta_{income,j|J}income$$

[4]The name of this model causes confusion. First, Anderson only referred to the one-dimensional models as the stereotype model. Common usage refers to models for all dimensions as stereotype models. Second, why "stereotype"? In contrast to an ordinal variable created by dividing a continuous scale, Anderson discussed ordinal variables constructed by assessing multiple characteristics, such as results on medical tests, and assigning a category, such as poor health, based on the stereotype for that category .

with $J-1$ parameters for each regressor. The coefficient for comparing of m to n is $\beta_{k,m|n} = \beta_{k,m|J} - \beta_{k,n|J}$. To reduce the number of parameters, the SLM restricts the coefficients to vary by *scale factors* α_j and ϕ_j for $j = 1, J$. The α's define the intercepts $\beta^*_{0,m|n} = (\alpha_m - \alpha_n)\beta_0$ that reflect the proportion of cases in each outcome, while the ϕ's define the effects of x_k as $\beta^*_{k,m|n} = (\phi_m - \phi_n)\beta_k$. This leads to the one-dimensional SL1M:

$$\begin{aligned}\ln\frac{\Pr(y=j\mid\mathbf{x})}{\Pr(y=J\mid\mathbf{x})} &= (\alpha_j - \alpha_J)\beta_0 + \left(\phi_j - \phi_J\right)\beta_{age}age + \left(\phi_j - \phi_J\right)\beta_{income}income \\ &= \beta^*_{0,j|J} + \beta^*_{age,j|J}age + \beta^*_{income,j|J}income\end{aligned}$$

There is only one coefficient for each regressor with scale factors associated with outcomes that are the same for all regressors. Identification requires constraints on the α's and ϕ's (see Long and Freese 2006 for details). Commonly, it is assumed that $\phi_1 = \alpha_1 = 1$ and $\phi_J = \alpha_J = 0$. With J categories and K regressors there are $2(J-2) + K + 1$ parameters in the SLM compared to $(K+1)(J-1)$ for the MNLM. For example, with 4 outcomes and 6 regressors, the MNLM has 21 parameters compared to the SLM's 11. While there are fewer parameters, the effects still vary across comparisons, but not as freely as in the MNLM. To make the model ordinal, Anderson (1984) added the constraints $\phi_1 = 1 > \phi_2 > \cdots > \phi_{J-1} > \phi_J = 0$. Most software does not enforce these constraints so that if you rearrange the order of the outcomes (e.g., renumber category 1 to 5 and category 5 to 1) the values of the ϕ's switch. Substantively, the results are identical.

The general SL1M with base J is:

$$\begin{aligned}\ln\frac{\Pr(y=j\mid\mathbf{x})}{\Pr(y=J\mid\mathbf{x})} &= (\alpha_j - \alpha_J)\beta_0 + \left(\phi_j - \phi_J\right)\mathbf{x}'\boldsymbol{\beta} \\ &= \alpha_j\beta_0 + \phi_j\mathbf{x}'\boldsymbol{\beta}\end{aligned}$$

where the last equality follows from the constraints $\alpha_J = \phi_J = 0$. In terms of odds:

$$\Omega_{m|n}(\mathbf{x},x_k) = \frac{\Pr(y=m)}{\Pr(y=n)} = \exp\left[(\alpha_m - \alpha_n)\beta_0 + (\phi_m - \phi_n)\mathbf{x}'\boldsymbol{\beta}\right]$$

with the odds ratio:

$$OR_{m|n,k} = \frac{\Omega_{m|n}\left(\mathbf{x}{,}x_k + 1\right)}{\Omega_{m|n}\left(\mathbf{x}{,}x_k\right)} = \exp\left(\left[\phi_m - \phi_n\right]\beta_k\right)$$

which can be interpreted just as the OR for the MNLM. The probabilities are:

$$\Pr\left(y = j \mid \mathbf{x}\right) = \frac{\exp\left(\alpha_j\beta_0 + \phi_j\mathbf{x}'\boldsymbol{\beta}\right)}{\sum_{q=1}^{J}\exp\left(\alpha_q\beta_0 + \phi_q\mathbf{x}'\boldsymbol{\beta}\right)}$$

The two-dimensional model (SL2M) has two coefficients for each regressor:

$$\ln\frac{\Pr(y = j|\mathbf{x})}{\Pr(y = J|\mathbf{x})} = \alpha_j\beta_0 + \phi_j^{[1]}\mathbf{x}'\boldsymbol{\beta}^{[1]} + \phi_j^{[2]}\mathbf{x}'\boldsymbol{\beta}^{[2]}$$

where $\alpha_J = \phi_2^{[1]} = \phi_J^{[1]} = \phi_1^{[2]} = \phi_2^{[2]} = 0$ and $\alpha_1 = \phi_1^{[1]} = \phi_2^{[2]} = 1$ for identification. In the SL2M, you can have regressors that are significant on one dimension but not the other, or that have effects in opposite directions in the two dimensions. Consequently, the model is no longer ordinal in Anderson's sense. The model can be extended to add more dimensions until with $J - 1$ dimensions it is identical to the MNLM.

While the SLM model has fewer parameters than the MNLM, full interpretation requires you to evaluate all comparisons. Since most of us cannot look at the scale factors and automatically compute the coefficients for the implied odds ratios, the smaller number of parameters does not practically simplify interpretation. Further, it may be difficult to provide substantive justification for the number of dimensions.[5]

2.4 The ordinal regression model (OLM)

The ordinal regression model is the most common model for ordinal outcomes. The probit version was introduced by McKelvey and Zavoina (1976). McCullagh (1980) presented the logit version called the proportional odds model, sometimes called the cumulative logit model. These

[5] It is inappropriate to use a LR test to compare stereotype models with different dimensions (StataCorp 2011:2015), but AIC or BIC statistics can be used.

models are all known as the parallel regression model and the grouped continuous model. The model is so well known that it is often called simply *the* ordinal regression model.

The model can be derived from a regression on a unobserved, continuous variable y^*:

$$y_i^* = \beta_0 + \beta_{age} age + \beta_{income} income + \varepsilon_i$$

The ordinal probit model (OPM) assumes that ε is normal with mean 0 and variance 1, while the ordinal logit model (OLM) assumes that ε is logistic with mean 0 and variance $\pi^2/3$. Since the models provide nearly identical predictions, I only consider the OLM. The continuous y^* is divided into observed, ordinal categories using the *thresholds* τ_0 through τ_J:

$$y_i = j \quad \text{if } \tau_{j-1} \leq y_i^* < \tau_j \text{ for } j = 1 \text{ to } J$$

where $\tau_0 = -\infty$ and $\tau_J = \infty$. For party affiliation, y_i^* is a continuous measure of left-right orientation with observed categories determined by this measurement model:

$$y_i = \begin{cases} 1 \Rightarrow \text{SD} & \text{if } \tau_0 = -\infty \leq y_i^* < \tau_1 \\ 2 \Rightarrow \text{D} & \text{if } \tau_1 \leq y_i^* < \tau_2 \\ 3 \Rightarrow \text{I} & \text{if } \tau_2 \leq y_i^* < \tau_3 \\ 4 \Rightarrow \text{R} & \text{if } \tau_3 \leq y_i^* < \tau_4 \\ 5 \Rightarrow \text{SR} & \text{if } \tau_4 \leq y_i^* < \tau_5 = \infty \end{cases}$$

The simplest way to see the implied structure of the model is by using *cumulative probabilities* of being less than or equal to category j:

$$\begin{aligned} \Pr(y \leq j \mid \mathbf{x}) &= \Pr(y^* < \tau_j \mid \mathbf{x}) \\ &= \Pr\left(\varepsilon < \tau_j - \left[\beta_0 + \beta_{age} age + \beta_{income} income\right] \mid \mathbf{x}\right) \text{ for } j = 1, J-1 \end{aligned}$$

where I substituted the equation for y^* and simplified. With Λ as the CDF for the logistic:

$$\Pr(y \leq j \mid \mathbf{x}) = \Lambda\left(\tau_j - \beta_0 - \mathbf{x}'\boldsymbol{\beta}\right) \text{ for } j = 1, J-1$$

The probability of an individual category j is the probability that $y \leq j$ minus the probability that $y \leq j-1$:

$$\Pr(y = j \mid \mathbf{x}) = \Lambda\left(\tau_j - \beta_0 - \mathbf{x}'\boldsymbol{\beta}\right) - \Lambda\left(\tau_{j-1} - \beta_0 - \mathbf{x}'\boldsymbol{\beta}\right) \text{ for } j = 1, J$$

We cannot estimate the intercept β_0 and all thresholds. To see this, add $\delta - \delta = 0$ within the CDF, leading to $\Lambda([\tau_j + \delta] - [\beta_0 + \delta] - \mathbf{x}'\boldsymbol{\beta})$. We can add any δ to τ_j and subtract δ from β_0 without changing the probability. For identification, we fix the value of either one threshold or the intercept. Assuming $\beta_0 = 0$, the model is:

$$\Pr(y \leq j \mid \mathbf{x}) = \Lambda\left(\tau_j - \mathbf{x}'\boldsymbol{\beta}\right) \text{ for } j = 1, J-1$$

For each j, this a binary logit on an outcome dividing categories between lower and higher values. The similarity to the BLM is easier to see if I define $\beta^*_{0,j} = \tau_j$ and $\boldsymbol{\beta}^* = -\boldsymbol{\beta}$ so that $\Pr(y \leq j \mid \mathbf{x}) = \Lambda\left(\beta^*_{0,j} + \mathbf{x}'\boldsymbol{\beta}^*\right)$. For the $J-1$ ways I can divide the ordinal categories, I have binary logits with different intercepts but *identical slopes*. This is known as the *parallel regression assumption* which is shown by the parallel curves in Figure 2..

– Figure 2 here –

As a consequence of the identical slopes, you can combine adjacent categories of the outcome and obtain consistent but inefficient estimates of the β_k's. Precision is lost since information is lost through combining categories.

The odds of being less than or equal to j is:

$$\Omega_j(x) = \frac{\Pr(y \leq j \mid \mathbf{x})}{1 - \Pr(y \leq j \mid \mathbf{x})} = \frac{\Lambda(\tau_j - \mathbf{x}'\boldsymbol{\beta})}{1 - \Lambda(\tau_j - \mathbf{x}'\boldsymbol{\beta})} \text{ for } j = 1, J$$

Since $\Lambda(\tau_j - \mathbf{x}'\boldsymbol{\beta}) = \exp(\tau_j - \mathbf{x}'\boldsymbol{\beta}) / [1 + \exp(\tau_j - \mathbf{x}'\boldsymbol{\beta})]$, this simplifies to:

$$\Omega_j(\mathbf{x}) = \exp\left(\tau_j - \mathbf{x}'\boldsymbol{\beta}\right) \text{ for } j = 1, J$$

The OR for x_k is:

$$OR_k = \frac{\Omega_j\left(\mathbf{x},x_k+1\right)}{\Omega_j\left(\mathbf{x},x_k\right)} = \exp\left(-\beta_k\right) \text{ for } j = 1, J$$

which can be interpreted as:

> For a unit increase in x_k, the odds of being in a category less than or equal to j (compared to greater than j) change by a factor of $\exp\left(-\beta_k\right)$, holding other variables constant.

Since the odds ratio is the same for all j, I can say:

> For a unit increase in x_k, the odds of being in a lower category compared to a higher category change by a factor of $\exp\left(-\beta_k\right)$, holding other variables constant.

The odds ratio for a change of δ would be $\exp\left(-\delta\beta_k\right)$.

The parallel regression assumption leads to the elegant interpretation of the odds of higher and lower outcomes, but the assumption might be unrealistic. Score, LR, and Wald tests of the assumption are available. Essentially these tests compare the OLM estimates to those from binary logits where the β's are not constrained to be equal. The model without constraints is called the generalized ordered logit model (GOLM), considered next. In my experience tests of parallel regressions are usually rejected. Allison (1999:141) finds that the test usually is significant when there are many regressors or the sample is large. When the hypothesis is rejected the results from the OLM could be quite similar to those from the GOLM, but they could differ in substantively critical ways. It is prudent, I believe, to always compare the results of the ORM to those of the MNLM or the GOLM before accepting the conclusions from the ORM or before deciding the model is inappropriate based on the parallel regression test.

2.5 Generalized ordered logit model (GOLM)

The generalized ordered logit model allows the β_k's to vary by category, resulting in $J-1$ parameters for each regressor:

$$\begin{aligned}
\ln \Omega_{y\leq j}(\mathbf{x}) &= \tau_j - \mathbf{x}'\boldsymbol{\beta}_j \text{ for } j=1,J-1 \\
\Omega_{y\leq j}(\mathbf{x}) &= \exp\left(\tau_j - \mathbf{x}'\boldsymbol{\beta}_j\right) \text{ for } j=1,J-1 \\
OR_{j,k} &= \exp\left(-\beta_{k,j}\right) \text{ for } j=1,J
\end{aligned}$$

The OR's can be interpreted as:

> For a unit increase in x_k the odds of being less than or equal to j change by a factor of $\exp\left(-\beta_{k,j}\right)$ holding other variables constant.

Probabilities are:

$$\Pr(y=j \mid \mathbf{x}) = \frac{\exp\left(\tau_j - \mathbf{x}'\boldsymbol{\beta}_j\right)}{1+\exp\left(\tau_j - \mathbf{x}'\boldsymbol{\beta}_j\right)} - \frac{\exp\left(\tau_{j-1} - \mathbf{x}'\boldsymbol{\beta}_{j-1}\right)}{1+\exp\left(\tau_{j-1} - \mathbf{x}'\boldsymbol{\beta}_{j-1}\right)} \text{ for } j=1,J$$

Unless $\left(\tau_j - \mathbf{x}'\boldsymbol{\beta}_j\right) \geq \left(\tau_{j-1} - \mathbf{x}'\boldsymbol{\beta}_{j-1}\right)$ probabilities can be be negative for observations in the sample. Since software does not impose this constraint, it is prudent to check predictions. As noted by McCullagh and Nelder (1989:155), "If [negative probabilities] occur in a sufficiently remote region of the x-space, this flaw in the model need not be serious." Williams (2006) in the help file for gologit2, a Stata program for the GOLM, reports that negative probabilities tend to occur when "the model is overly complicated and/or there are very small N's for some categories of the dependent variable." In these cases he suggests combining categories or simplifying the model.

Letting β_k vary by j avoids the parallel regression assumption of the OLM. The resulting model has as many parameters as the MNLM and the model is no longer ordinal. Personally, I find interpretations of the GOLM to be more difficult than the MNLM as shown in Section 3. There are several related models that reduce the number of parameters in the GOLM. The partial generalize ordered logit model (PGOLM) lets the β's for some variables differ by j while others do not. Williams (2006) describes stepwise procedures for deciding which the effects should vary by j. This can be a useful diagnostic for discovering important patterns in the data. Models with partial proportionality constraints impose constraints that are similar to those in the SLM. See Fullerton (2009) for a review of these models. If stepwise procedures are used to select the model, it is important to let readers know what you did to use a small p-value in stepwise selection to minimize the chance that the reduction in parameters reflects peculiarities of the sample.

3 Example analysis: Modeling political attitudes[6]

To illustrate the interpretation of the models discussed above, I use data from the 1992 American National Election Study (ANES n.d.). The source variable for party affiliation had nine categories that were collapsed to *Strong Democrat* (1=*SD*), *Democrat* (2=*D*), *Independent* (3=*I*), *Republican* (4=*R*), and *Strong Republican* (5=*SR*). As a reflection of left-right political orientation, the categories are ordered from 1 to 5; as a reflection of intensity of partisanship they are ordered 3=*I*, (2=*D*, 4=*R*), (1=*SD*, 5=*SR*). The distribution of categories is shown in Figure 3. Six regressors are used: age, income, race indicated as black or not, gender, and education using dummies for completing high school and completing college with not completing high school as the excluded category. Descriptive statistics are given in Table 2. My analyses are used to illustrate methodological issues, not to make a substantive contribution.

– Figure 3 and Table 2 here –

[6] Models were estimated using Stata 12. The do-files and data can be obtained by entering findit cdaNOR in Stata while connected to the internet.

3.1 Approaches to interpretation

Models for nominal and ordinal outcomes can be interpreted using probabilities and odds ratios for logit models. Each approach to interpretation is illustrated, but not all methods are shown for all models. To highlight the consequences of assuming ordinality, I compare results from ordinal models to those from either the MNLM or the GOLM.

Odds ratios can be interpreted as:

> For a unit increase in x_k, the odds of A versus B change by a factor of OR_k holding other variables constant.

The value of OR_k is $\exp(\beta_k^*)$, where β_k^* depends on the specific model. The odds ratios for a change of δ in x_k is $\exp(\delta\beta_k^*)$. Sometimes it is convenient to talk about a percentage change in the odds which is computed as $ORPCT_k = 100\,(OR_k - 1)$ and is interpreted as:

> For a unit increase in x_k, the odds of A versus B increase [or decreases] by $ORPCT_k$ percent, holding other variables constant.

For MNLM, SLM, and ACLM, the odds are for a one category compared to another, such as the odds of *SD* compared to *D*. For the OLM and the GOLM, the odds compare being in lower ordinal categories compared to higher categories, such as *SD* or *D* compared to *I*, *R*, or *SR*. Odds ratios have the advantage that their interpretation does not depend on the value of the regressors. In this respect they are similar to coefficients from the linear regressions model. However, a given factor change in the odds implies different amounts of changes in the probability of the outcomes depending on the values of the regressors. For example, if the odds of *R* versus *D* are 100 to 1, doubling the odds to 200 to 1 is a small change in the probabilities. If the odds are 1 to 1, doubling the odds to 2 to 1 implies a much larger change in probabilities. Simply saying the odds double tells you little about the substantive process unless you know the value of the odds before they are doubled.

The second approach to interpretation uses probabilities and functions of these probabilities. Collectively, these quantities are called *predictive margins* (Graubard and Korn 1999). Predictions can be used in a many ways: look at the distribution of predictions in the sample, compute predictions at substantively interesting values of the regressors, plot probabilities over the range of a regressor, create tables to show how probabilities are affected by the levels of a few regressors, compute changes in probabilities for a discrete change in a regressor, or compute the rate of change (i.e., the derivative) with respect to a regressor. Unlike odds ratios, probabilities and changes in probabilities depend on where a regressor is at the start of the change, how much the regressor changes, and the levels of all other regressors.

3.2 Odds ratios

Table 3 shows factor changes in the odds of adjacent party affiliations (e.g., SD vs. D, I vs. R). To save space, I do not show odds ratios for other comparisons such as SD vs. SR. Column 1 lists ORs from the MNLM. Overall, increasing income decreases the odds of affiliations to the left relative to the adjacent category to the right. For example, a $10,000 increase in income decreases the odds of SD versus D by a factor of .93, holding other variables constant. This effect is significant at the .10 level for a two-tailed test. The odds ratios for the MNLM are plotted in Figure 4. The distance between adjacent categories corresponds to the magnitude of the $\widehat{\beta}$'s measured on the logit coefficient scale at the bottom of the figure. The largest ORs are for SD versus D and I versus R, with ORs for other adjacent categories being smaller. The pattern of ORs is consistent with party being ordinal with respect to income. The odds ratios for age, shown in Table 3 and Figure 4, are quite different. A ten year increase in age increases the odds of SD versus D by a factor of

1.27 and decreases the odds of *R* versus *SR* by a factor of 0.80 (or equivalently increases the odds of *SR* versus *R* by a factor of 1.24). The other *OR*s are not significant. Overall, age increases the intensity of partisanship, suggesting that party affiliation reflects both orientation and intensity and that age and income affect these dimensions is different ways. An ordinal model could not uncover this pattern of effects.

– Table 3 and Figure 4 here –

The ACLM forces the *OR*s for adjacent categories to be equal as shown in column 2 of Table 3. For a $10,000 increase in income the odds of being in a party to the left compared to the adjacent party to the right decrease by a factor of .96 for all adjacent parties. By comparing Figures 4 and 5 you can see how the *OR*s for adjacent categories are the same size for the ACLM while they differ in size for the MNLM.[7] Both models, however, arrange the categories in the same order from *SD* on the left to *SR* on the right. For age, with the *OR*s constrained to be equal the effect is not significant. The one-dimensional stereotype model (SL1M) does not force the *OR*s for adjacent categories to be equal, but still constrains them through scaling coefficients. For income, we find roughly the same pattern of *OR*s as for the MNLM. For age, the effects are not significant and the pattern is quite different. Recall, the SL1M model forces the relationship between regressors and outcomes to be ordinal.

The two-dimensional stereotype model has two effects for each regressor. Table 3 shows that the combined effects from dimension 1 and 2 for income have roughly the same effect on the odds as the MNLM with effects that are more significant. This suggests that the two dimensional model is capturing the findings from the MNLM. The two dimensions have effects that operate in different directions for three pairs of categories. For example, dimension one decreases the odds of *SD* versus

[7] The graph also illustrates that coefficients for categories that are two categories apart at twice as large and those for three categories are three times as large.

D by a factor of .84, while dimension two increases the odds by a factor of 1.10. It isn't clear what the substantive meaning of the dimensions is. For age dimension two increases the odds of more extreme partisanship, while dimension one has a weak effect on left-right orientation. Overall, the conclusions from SL2M and MNLM are very similar.

– Table 4 here –

The models in Table 4 predict cumulative odds of being in parties to the right versus to the left: $SR{+}R{+}I{+}D$ *vs.* SD; $SR{+}R{+}I$ *vs.* $D{+}SD$; $SR{+}R$ *vs.* $I{+}D{+}SD$; and SR *vs.* $R{+}I{+}D{+}SD$. Here I use percentage change in the odds rather than factor changes. For the OLM, each regressor has the same effect of the odds of being to the right compared to the left regardless of where you divide right and left. A $10,000 increase in income increases the odds of being to the right by 10.1 percent holding other variables constant ($p<.001$). A ten-year increase in age decreases the odds by 6.2 percent. A Brant test of the parallel regression assumption is significant ($X^2_{18}=89.84$, $p<0.001$) suggesting we should examine whether the effects differ by where the outcome is divided.

In the GOLM the percentage change for income gradually decreases as the dividing point moves from *SD* to the right, but the differences are not significant ($X^2_3=2.09$, $p=0.554$). The coefficients for age, however, are significantly different ($X^2_3 = 36.1$, $p<.001$). We know from the MNLM that age increase more partisan party affiliation whether on the right or the left. A similar result is found with the GOLM. For age, I find it harder to understand the pattern of *OR*s since they involve comparisons of grouped categories. We know from the MNLM that age increases both *SD* and *SR*, but the GOLM these similar categories are always in different cumulative probabilities. If the combined categories include outcomes that do not change in the same direction with respect to a regressor, it is harder to tell what is going on.

When interpreting odds ratios all I need to say is that the other regressors do not change. The specific values where the controlled regressors are held constant does not matter as long as they do not change. While this simplifies interpretation, it is impossible to understand the magnitude of the OR in terms of probabilities unless you know the value of the odds before it is changed. To know this, you need to know the specific predicted probabilities which depend on the values of all predictors.

3.3 Predicted probabilities

Let $\mathbf{x}^*$ contain specific values of the regressors where I want to compute predictions. For example, in the OLM:

$$\begin{aligned} \widehat{\Pr}(y \leq j \mid \mathbf{x}^*) &= \Lambda\left(\widehat{\tau}_j - \mathbf{x}^*\widehat{\boldsymbol{\beta}}\right) \\ \widehat{\Pr}(y > j \mid \mathbf{x}^*) &= 1 - \Lambda\left(\widehat{\tau}_j - \mathbf{x}^*\widehat{\boldsymbol{\beta}}\right) \end{aligned}$$

where the odds ratio is $\widehat{OR}_k = \exp\left(-\widehat{\beta}_k\right)$. Suppose that $\widehat{OR}_k = 2$. Knowing that the odds double does not tell use how much $\widehat{\Pr}(y \leq j \mid \mathbf{x}^*)$ changes. The magnitude of the change depends on the probability at the start of the change, which in turn depends on the levels of all regressors. Accordingly, it is important to compute the probabilities at values of the regressors that are of substantive interest. You can compute predictions at observed values of the regressors, but you can also compute predictions at hypothetical values. For example, I might want to know the probability of being SR for a black woman with a college education who is 40 and earning \$30,000, even if there is no person in the sample with those specific characteristics. Or I might want to know the probability for someone who is average on all regressors, even though nobody can be average with binary regressors.

To explain the effects of age and income I can plot the probabilities as these variables change. Let $\overline{\mathbf{x}}_{[-age]}$ contain the means of all regressors except age, with $\widehat{\boldsymbol{\beta}}_{[-age]}$ containing the corresponding estimates from the OLM. The predicted cumulative probabilities at a given age with other variables at their mean is:

$$\widehat{\Pr}\left(y \leq j \mid \overline{\mathbf{x}}_{[-age]}, age\right) = \Lambda\left(\widehat{\tau}_j - \overline{\mathbf{x}}_{[-age]}\widehat{\boldsymbol{\beta}}_{[-age]} - \widehat{\beta}_{age} age\right)$$

Probabilities for individual outcomes are:

$$\widehat{\Pr}\left(y = j \mid \overline{\mathbf{x}}_{[-age]}, age\right) = \widehat{\Pr}\left(y \leq j \mid \overline{\mathbf{x}}_{[-age]}, age\right) - \widehat{\Pr}\left(y \leq j-1 \mid \overline{\mathbf{x}}_{[-age]}, age\right)$$

I can do the same for income. These values are plotted in Figures 7 and 6. The solid line with filled circles shows that as income increases from $0 to $100,000 the probability of being SD decreases from .23 to .10 while the probability of SR increases from .08 to .18. As age increases from 20 to 85 the probability of being a SD increases from .15 to .21 while the probability of SR decreases from .12 to .08. Since the OLM is an ordinal model, the probabilities of the highest and lowest categories must change in opposite directions. Plots from the ACLM and SL1M (not shown) are nearly identical to those for the OLM. The maximum absolute difference in probabilities for the ACLM compared to the OLM was less than .02, with most differences less than .005. The SL1M differed by less than .02 for age and .03 for income. Looking at the graphs from the three models would lead you to the same conclusions.

– Figures 6, 7, 8, and 9 here –

Figures 8 and 9 show corresponding predictions from the MNLM. The graph for income is similar to that for the OLM with the average absolute difference in predictions of less than .02. The plot for age, however, is very different. As age increases from 20 to 85 the probability of being a SD increases from .10 to .33, while the probability of being a SR also increases, albeit less strongly, from .07 to .17.[8] Since the MNLM is not ordinal, it does not force the changes in the extreme categories to be in opposite directions as required by ordinal models. This illustrates why

[8]Since the data is cross-sectional the "effect" of age could reflect cohort differences rather than the effect of aging.

when using ordinal models it is prudent to compare the results to those from non-ordinal models such as the MNLM or the GOLM.

Predicted probabilities can also be used in tables to show the effects of key variables. For example, to show the effects of race and gender on party affiliation I compute probabilities by race and gender holding other variables at their means. The results for the GOLM and the OLM are shown in Table 5. Both models show that blacks are far more likely than whites to be a SD or D and less likely to be R or SR. The magnitudes of the differences are similar in both models, although larger differences are found in predictions for Independents. Much smaller differences are found for men and women.

– Table 5 here –

To show the effect of race I use discrete changes also known as first differences. Let $\mathbf{x}^*$ contain values for all regressors except x_k. Let x_k^{Start} be the start value for x_k and x_k^{End} the end value. The discrete change for outcome j is:

$$\frac{\Delta \Pr(y = j \mid \mathbf{x})}{\Delta x_k} = \Pr\left(y = j \mid \mathbf{x}^*,\ x_k = x_k^{End}\right) - \Pr\left(y = j \mid \mathbf{x}^*,\ x_k = x_k^{Start}\right)$$

Here I change the value of only x_k, but I could change multiple variables such as changing from being a white woman to black man holding other variables at the mean. The discrete changes for race by gender from the GOLM and OLM are shown in Table 6. For each discrete change I tested H_0: $\frac{\Delta \Pr(y=j|\mathbf{x})}{\Delta x_k} = 0$ using the delta method with Stata's margins command (StataCorp 2011). The results are similar, although the effects are more significant in the OLM. When presenting these results I might want to comment on whether the race differences are the same for men and women. This requires testing whether the discrete change for men is equal to the discrete change for women, that is, a second difference. Specifically, I want to test:

$$H_0 \quad : \quad [\Pr(y=j \mid \mathbf{x}^*,\ women,\ black) - \Pr(y=j \mid \mathbf{x}^*,\ women,\ white)]$$
$$- [\Pr(y=j \mid \mathbf{x}^*,\ men,\ black) - \Pr(y=j \mid \mathbf{x}^*,\ men,\ white)]$$

The results of these tests are shown in Table 7. Race differences in party preference do not vary by gender.

– Tables 6, 7, 8, and 9 here –

The advantage of odds ratios is that the same odds ratio applies at all values of the regressors. If the OR is 2 for a white, male high school graduates earning \$20,000 at age 30, the OR is 2 for a sixty year old black women with college degrees earning \$50,000. But, the meaning of the odds ratio in terms of changes in probabilities differs for these two individuals. To illustrate this, Table 8 presents discrete changes for race for at two sets of values for the control variables. First, I look at the effects of race by gender for college graduates aged 30 with an income of \$40,000. Second, I consider 60 year old high school graduates earning \$25,000. The patterns are similar to that degree that blacks are more likely to be on the left and whites on the right. The magnitudes of the discrete changes by race, however, differ substantially. For example, in Panel 1 being black increases the probability of being a SD by .197, while the effect is .330 in Panel 2. If we examined the discrete change at other values of the control values, we would obtain different values for the discrete change.

While the changing size of discrete changes may be troubling, they are implicit in the nonlinearity of the model. To see the link between ORs and discrete change, consider Panel 1 where the probability of being SD increases from .083 for white females to .280 for black females, for a discrete change of .197. At the probability .083 the odds is $.083/(1-.083) = .090$ and at .280 it is $.280/(1-.280) = .389$, providing an odds ratio of $.231 = .090/.289$. This matches the OR for race from the GOLM. In Panel 2 the discrete change of .330 as the probability change from .221 for white females to .551 for black females, corresponding an the odds of $.221/(1-.221) = .283$ and $.551/(1-.551) = 1.229$ with an OR of $.231 = .283/1.229$. The different sizes of the discrete change merely reflect the meaning of the OR at two places in the data space. There is no way around the difficulties that are inherent in nonlinear models.

Numerous variations on these methods can be made. Probabilities for ideal types representing characteristics of individuals of particular interest can be presented. Discrete changes can be computed with regressors changing by any amount of interests, such as four years for education, 15 points for IQ, or a standard deviation for a continuous variable. You can use discrete change to examine the differences between individuals that differ on multiple characteristics. Marginal or partial changes in probabilities can be computed.

In our examples we computed predictions at fixed values of the regressors. Since predictions are often made at the mean, this is referred to as the "marginal effect at the mean" or MEM approach. I think of it as the *at mean* approach since it is not limited to marginal effects. An alternative approach computes predictions and quantities based on predictions for each observation and then averages these quantities. This method is sometimes called the "average marginal effect" or AME, but I think of it as the *mean of* approach since it applies to quantities other than the marginal. To explain the mean-of approach, let $\mathbf{x}_i^*$ contain the values for all regressors except x_k for the ith person in the sample. Suppose that x_k is binary and we are interested in the discrete change from $x_k^{Start} = 0$ to $x_k^{End} = 1$ for outcome j for the ith observation:

$$\frac{\Delta \Pr(y = j \mid \mathbf{x}_i^*)}{\Delta x_k} = \Pr\left(y = j \mid \mathbf{x}_i^*, x_k = x_k^{End}\right) - \Pr\left(y = j \mid \mathbf{x}_i^*, x_k = x_k^{Start}\right)$$

The mean of the discrete changes over the sample is:

$$\text{mean}\frac{\Delta \Pr(y = j \mid \mathbf{x}_i^*)}{\Delta x_k} = \frac{1}{N}\sum_{i=1}^{N}\frac{\Delta \Pr(y = j \mid \mathbf{x}_i^*)}{\Delta x_k}$$

This idea can be extended to look only at a particular group, say only men. Following this approach, Table 9 presents discrete changes at the mean and the mean of discrete changes by race for men and women. The findings are very similar with both approaches. In general, differences between discrete changes at the mean and the mean of discrete changes are greatest when the range of predicted probabilities in the sample span a region of the probability or difference in probability curve that is nonlinear.

4 Caveats and frequent errors

Sensitivity analysis Ordinal models impose constraints on how regressors are related to the probabilities of the outcomes. Always compare the predictions from an ordinal model to those from a nominal model such as the MNLM or the GOLM. If non-ordinal models provide predictions that differ noticeably from those from the ordinal model, you should carefully asses whether the ordinal model is appropriate for your application.

Over fitting If such stepwise procedures are used to reduce the complexity of your model, randomly divide your sample into two sub-samples. Use the first sub-sample as the exploration sample where stepwise procedures are used to select your candidate model. After you have selected a model, verify its fit with the second, verification sub-sample. If the results differ markedly, your model could be reflecting peculiarities of the sample rather than the underlying process.

Lack of interpretation Do not limit interpretation to a table of coefficients with a brief discussion of the signs and significance, while ignoring the substantive impact of regressors. Similarly, simply presenting odds ratios without information that allows determination of the magnitude of the effect in terms of changes in probabilities is incomplete. Use predicted probabilities and related measures to show the magnitude of the effect.

Look at all of the parameters implied by your model Standard software for the MNLM shows estimates for the minimal set of parameters. This set might not include comparisons that are of greatest interest. To estimate parameters for all comparisons you can recode the values for your outcome and re-estimate the model (see Long and Freese 2006 for details). If a significance test that a regressor has no effect is not rejected, it is possible that coefficients for specific comparisons are significant. If those comparisons are substantively important, you should test the individual coefficients. If you use the SLM, you should compute the odds ratios from the estimated parameters.

IIA There is no commonly available model for nominal outcomes avoids the IIA property. Tests for IIA do not have good properties and often produce contradictory results, with several tests rejecting the null hypothesis while others accept the hypothesis. The OLM model does avoid IIA but has other limitations.

Test of parallel regressions Tests of the parallel regression assumption in the ORM often reject the hypothesis. Some evidence suggests that tests are sensitive to issues unrelated to the parallel regression assumption. If the null hypothesis is rejected, compare the predictions from the ORM to those from the GOLM or the MNLM to determine if there are substantively meaningful differences in the predictions of the two models. If not, it is reasonable to use the ORM.

Consider nonlinearity on RHS Consider nonlinearities on the right-hand-side of the model. For example, include polynomials of key regressors such as age, age-squared, and age-cubed.

5 Conclusions

In this paper I reviewed the most common regression models for nominal and ordinal outcomes. In practice you will find that most of the applications in the social sciences use either the MNLM or the ORM. With advances in software to estimate models such as the GOLM and the stereotype model their use is increasing. While ordinal models can simplify interpretation and the added information from ordinality allows more efficient estimates, it is critical to assess whether the restrictions implicit in ordinal models are appropriate for your substantive application. Before selecting an ordinal model, compare the results from that model to those from a nominal model.

I have not considered models known as heterogeneous choice models or location-scale models (see Williams 2009 for a review). While these models are theoretically promising, simulations by Keele and Park (2006) suggest that the model is highly sensitive to specification of the variance function. In my experience estimates vary widely with what seem like minor changes to the specification. Another model of potential interest is the continuation ratio model that is appropriate when the outcome reflects stages that individual pass through in sequence, such as the ranks of assistant professor, to associate professor, to full professor, to named professorship. Mixed logit models (see Train 2009) shows promise as models that do not impose the IIA assumption, but requires intensive calculation to estimate and involve more complicated data structures. To date software is not readily available.

6 Further reading

Agresti, A. (2010). The Analysis of Ordinal Categorical Data. Second edition. New York, Wiley.

> This book provides a detailed discussion of regression models for ordinal variables as well as models for the analysis of contingency tables with ordinal variables. The author's web contains data sets and sample programs using SAS and R: www.stat.ufl.edu/~aa/ordinal/ord.html.

Long, J. S. (1997). Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA, Sage Press.

This book provides a more technical discussion of regression models for ordinal and nominal outcomes as well as binary and count variables. Methods of interpretation using predictions are discussed in detail.

Long, J. S. and J. Freese (2006). Regression Models for Categorical Dependent Variables Using Stata. Second Edition. College Station, Texas, Stata Press.

This book focuses on the estimation and interpretation of regression models using Stata, including most of the models discussed in this paper.

O'Connell, A. A. (2006). Logistic regression models for ordinal response variables. Thousand Oaks, Calif., SAGE Publications.

This book focuses on logit models for ordinal outcomes with examples of programs in SAS, SPSS and Stata. Many examples are given.

Train, K. (2009). Discrete choice methods with simulation. Second Edition. Cambridge: New York, Cambridge University Press.

This book includes a detailed discussion of models for discrete choice including new models that can be estimated by simulation. Models with alternative-specific regressors are considered.

7 Bibliography

Agresti, A. (2010). The Analysis of Ordinal Categorical Data. Second edition. New York, Wiley.

American National Election Studies (n.d.) The ANES Guide to Public Opinion and Electoral Behavior. Ann Arbor, MI: University of Michigan, Center for Political Studies [producer and distributor]. www.electionstudies.org.

O'Connell, A. A. (2006). Logistic regression models for ordinal response variables. Thousand Oaks, Calif., SAGE Publications.

Cheng, S. and J. S. Long (2007). "Testing for IIA in the Multinomial Logit Model" Sociological Methods & Research 35(4): 583-600.

Clogg, C. C. and E. S. Shihadeh (1994). Statistical Models for Ordinal Variables. Thousand Oaks, CA, Sage.

Fry, T. R. L. F. and M. N. Harris (1996). "A Monte Carlo study of tests for the independence of irrelevant alternatives property." Transportation Research Part B: Methodological 30(1): 19-30.

Fry, T. R. L. F. and M. N. Harris (1998). "Testing for independence of irrelevant alternatives: some empirical results." Sociological Methods and Research 26(3): 401-423.

Fullerton, A. S. (2009). "A Conceptual Framework for Ordered Logistic Regression Models." Sociological Methods and Research 38: 306-347.

Goodman, L. A. (1983). "The analysis of dependence in cross-classifications having ordered categories, using log-liner models for frequencies and log-linear models for odds." Biometrics 39: 149-160.

Graubard, B. and E. Korn (1999). "Predictive margins with survey data." Biometrics 55(2): 652-659.

Keane, M. P. (1992). "A Note on Identification in the Multinomial Probit Model." Journal of Business and Economic Statistics 10: 193-200.

Keele, Luke and David K. Park. 2006. "Difficult Choices: An Evaluation of Heterogeneous Choice Models." Working Paper, March 3, 2006.

Long, J. S. (1997). Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA, Sage Press.

Long, J. S. and J. Freese (2006). Regression Models for Categorical Dependent Variables Using Stata. Second Edition. College Station, Texas, Stata Press.

McCullagh, P. (1980). "Regression Models for Ordinal Data (with Discussion)." Journal of Royal Statistical Society 42(2): 109-142.

McCullagh, P. and J. A. Nelder (1989). Generalized Linear Models. Second edition. New York, Chapman and Hall.

McFadden, D. (1973). Conditional Logit Analysis of Qualitative Choice Behavior. Frontiers of Econometrics. P. Zarembka. New York, NY, Academic Press: 105-142.

McKelvey, R. D. and W. Zavonia (1975). "A Statistical Model for the Analysis of Ordinal Level Dependent Variables." Journal of Mathematical Sociology 4: 103-120.

StataCorp. (2011). Stata: Release 12. Statistical Software. College Station, TX: StataCorp LP.

Stevens, S. S. (1946). "On the Theory of Scales of Measurement." Science 7: 677-680.

Train, K. (2009). Discrete choice methods with simulation. Second Edition. Cambridge: New York, Cambridge University Press.

Velleman, P. F. and L. Wilkinson (1993). "Nominal, Ordinal, Interval and Ratio Typologies are Misleading." American Statistician 47(1): 65-72.

Williams, R. (2006). "Generalized Ordered Logit/ Partial Proportional Odds Models for Ordinal Dependent Variables." Stata Journal 6(1): 58-82.

Williams, R. (2009). "Using Heterogeneous Choice Models to Compare Logit and Probit Coefficients Across Groups." Sociological Methods & Research 37(4): 531-559.

Williams, R. (2012) gologit2 help file as of 2012-05-02.

Contact information

J. Scott Long Office: Departments of Sociology and Statistics; Ballantine 744; 1020 E. Kirkwood Ave.; Indiana University; Bloomington, IN 47405; jslong@indiana.edu; Phone: 812-855-7617; Fax: 812-855-0781

Biographical sketch Scott Long is Chancellor's Professor of Sociology and Statistics at Indiana University-Bloomington. His research focuses on gender differences in the scientific career, stigma and mental health, aging and labor force participation, human sexuality, and statistical methods. His recent research on the scientific career was published as From Scarcity to Visibility published by the National Academy of Sciences. He is past Editor of Sociological Methods and Research and the recipient of the American Sociological Associations Paul F. Lazarsfeld Memorial Award for Distinguished Contributions in the Field of Sociological Methodology. He is author of Confirmatory Factor Analysis, Covariance Structure Analysis, Regression Models for Categorical and Limited Dependent Variables, and Regression Models for Categorical and Limited Dependent Variables with Stata (with Jeremy Freese) and several edited volumes.

Tables and Figures

Table 1: Nominal and ordinal regression models.

Model	*Is the model ordinal?*
Multinomial logit (MNLM)	No
Adjacent category logit(ACLM)	Yes
Stereotype logit with 1 dimension (SL1M)	Yes
Stereotype logit with 2+ dimensions (SL2M, SL3M...)	No
Ordinal logit (OLM)	Yes
Generalized ordinal logit (GOLM)	No

Table 2: Descriptive statistics for regressors predicting party affiliations (N=1,382).

Variable	*Mean*	*Standard Deviation*	*Minimum*	*Maximum*	*Label*
age	4.59	1.68	1.80	9.10	Age in decades.
income	3.75	2.78	0.15	13.13	Income in $10,000.
black	0.14	—	0	1	1 if black; 0 if not.
female	0.49	—	0	1	1 if female; 0 if male.
highschool#	0.58	—	0	1	High school is the highest degree.
college#	0.26	—	0	1	College is the highest degree.

Note: # - Having less than a high school degree is the excluded category.

Table 3: Odds ratios for income and age in models of party affiliation.

Effects of income				*SL2M*		
on the odds of	*MNLM*	*ACLM*	*SL1M*	Dim 1+2	Dim 1	Dim 2
SD vs. D	0.93*	0.96***	0.94***	0.92***	0.84***	1.10***
D vs. I	0.99	0.96***	1.02	1.01	0.99	1.02
I vs. R	0.93*	0.96***	0.91***	0.91***	0.94	0.97
R vs. SR	0.99	0.96***	0.97	0.98	1.08**	0.91**
Effects of age				*SL2M*		
on the odds of	*MNLM*	*ACLM*	*SL1M*	Dim 1+2	Dim 1	Dim 2
SD vs. D	1.27***	1.02	1.03	1.27***	1.03	1.24***
D vs. I	1.08	1.02	0.99	1.06	1.00	1.06
I vs. R	0.93	1.02	1.05	0.95	1.01	0.95
R vs. SR	0.80***	1.02	1.01	0.80***	0.99	0.80***

Note: Factor changes in the odds for a $10,000 increase in income.

*** indicates significant at the .001; ** at the .05 level; * at the .10 level

for a two-tailed test.

Table 4: Percentage change in the odds for the OLM and the GOLM.

	Income		**Age**	
Odds of	*OLM*	*GOLM*	*OLM*	*GOLM*
SR+R+I+D vs SD	10.1**	13.4***	-6.2**	-20.2***
SR+R+I vs D+SD	10.1**	10.3***	-6.2**	-6.4*
SR+R vs I+D+SD	10.1**	9.7***	-6.2**	-1.6
SR vs R+I+D+SD	10.1**	6.6***	-6.2**	15.0***

Note: Table contains estimates of percent change in the odds for a ten-year increase in age or a $10,000 increase in income. *** indicates significant at the .001 level; ** at the .05 level; * at the .10 level for a two-tailed test.

Table 5: Predicted probabilities of party affiliation by race and gender in the GOLM and OLM.

GOLM	**Outcome**				
Group	*SD*	*D*	*I*	*R*	*SR*
Black women	0.413	0.383	0.142	0.045	0.017
White women	0.140	0.354	0.085	0.310	0.111
Black men	0.402	0.345	0.185	0.046	0.021
White men	0.135	0.290	0.130	0.306	0.139
OLM	*SD*	*D*	*I*	*R*	*SR*
Black women	0.443	0.355	0.068	0.105	0.029
White women	0.154	0.321	0.122	0.287	0.116
Black men	0.405	0.367	0.075	0.119	0.034
White men	0.134	0.301	0.122	0.308	0.134

Note: Other variables are held at their means.

Table 6: Discrete changes in predicted probabilities of party affiliation in the GOLM and OLM.

GOLM	**Outcome**				
Comparison	*SD*	*D*	*I*	*R*	*SR*
Black women-white women	0.273	0.029	0.057	-0.266	-0.094
p-value	<0.001	0.457	0.044	<0.001	<0.001
Black men-white men	0.268	0.055	0.055	-0.260	-0.118
p-value	<0.001	0.185	0.090	<0.001	<0.001
OLM	*SD*	*D*	*I*	*R*	*SR*
Black women-white women	0.289	0.034	-0.054	-0.182	-0.087
p-value	<0.001	0.045	<0.001	<0.001	<0.001
Black men-white men	0.270	0.066	-0.047	-0.189	-0.100
p-value	<0.001	<0.001	<0.001	<0.001	<0.001

Note: Other variables are held at their means. The p-value is for testing that the discrete change is 0.

Table 7: Comparing race differences for men and women in the GOLM and OLM.

GOLM	**Outcome**				
	SD	*D*	*I*	*R*	*SR*
Second difference	0.004	-0.023	0.002	-0.007	0.024
p-value	0.763	0.244	0.910	0.724	0.140
OLM	*SD*	*D*	*I*	*R*	*SR*
Black women-white women	0.005	-0.026	0.002	-0.005	0.024
p-value	0.762	0.273	0.910	0.800	0.139

Note: The table contains the difference between the discrete change for race computed for women and the discrete change for race computed for men. Other variables are held at their means. The p-value is for testing that the second difference is 0.

Table 8: Discrete changes in predicted probabilities of party affiliation for the young college graduates and older high school graduates in the GOLM.

Panel 1	**Outcome**				
Young college graduates	*SD*	*D*	*I*	*R*	*SR*
Black women-white women	0.197	0.122	0.115	-0.311	-0.123
p-value	<0.001	0.007	0.001	<0.001	<0.001
Black men-white men	0.192	0.140	0.119	-0.298	-0.152
p-value	<0.001	0.002	0.002	<0.001	<0.001
Panel 2	**Outcome**				
Older high school graduates	*SD*	*D*	*I*	*R*	*SR*
Black women-white women	0.330	-0.054	0.039	-0.222	-0.092
p-value	<0.001	0.187	0.133	<0.001	<0.001
Black men-white men	0.326	-0.023	0.029	-0.217	-0.115
p-value	<0.001	0.587	0.338	<0.001	<0.001

Note: Other variables are held at their means. The p-value is for testing that the discrete change is 0.

Table 9: Discrete changes at mean and mean of discrete change in party affiliation using the GOLM.

	Outcome				
Discrete change at mean	*SD*	*D*	*I*	*R*	*SR*
Black women-white women	0.273	0.029	0.057	-0.266	-0.094
p-value	<0.001	0.457	0.044	<0.001	<0.001
Black men-white men	0.268	0.055	0.055	-0.260	-0.118
p-value	<0.001	0.185	0.090	<0.001	<0.001
Mean of discrete change					
Black women-white women	0.272	0.020	0.054	-0.250	-0.096
p-value	<0.001	0.597	0.051	<0.001	<0.001
Black men-white men	0.249	0.069	0.067	-0.255	-0.130
p-value	<0.001	0.086	0.043	<0.001	<0.001

Note: Other variables are held at their means. The p-value is for testing that the discrete change is 0.

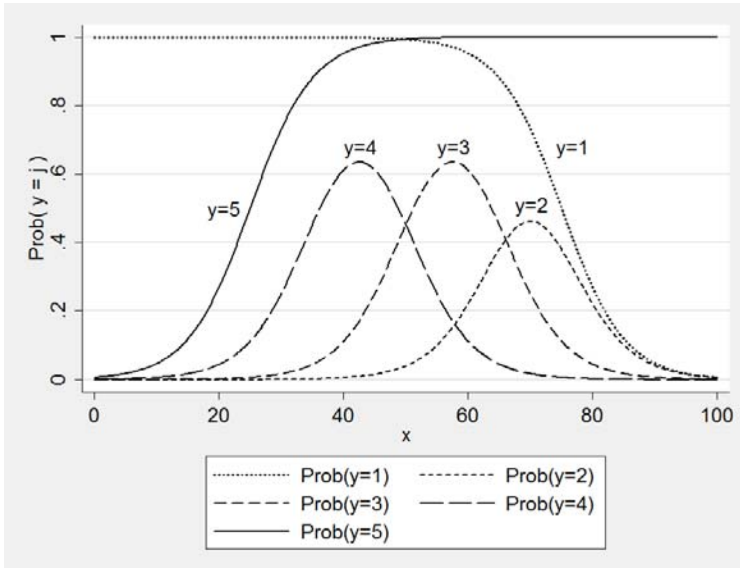

Figure 1: Illustration of Anderson's definition of an ordinal regression model.

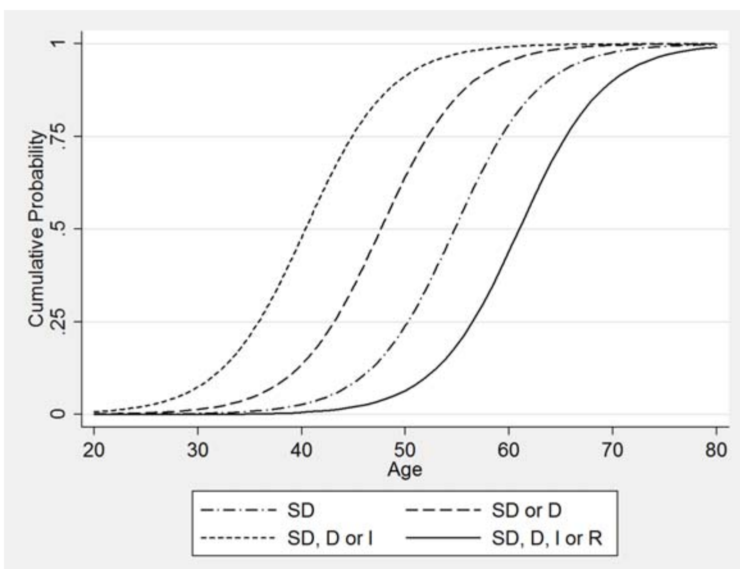

Figure 2: Parallel cumulative probability curves in the ordinal regression model.

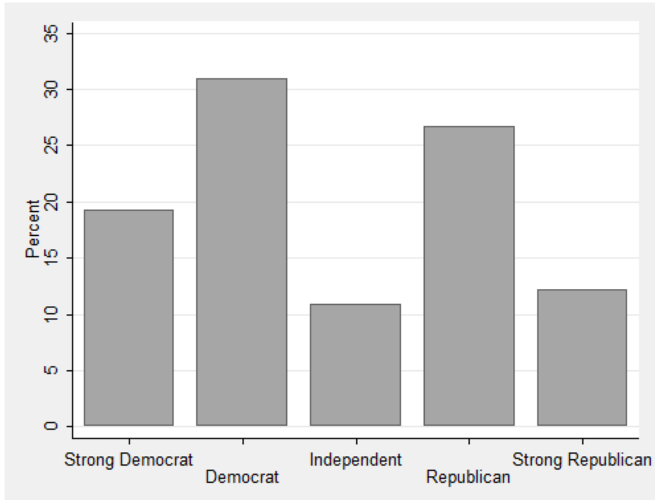

Figure 3: Distribution of party affiliation.

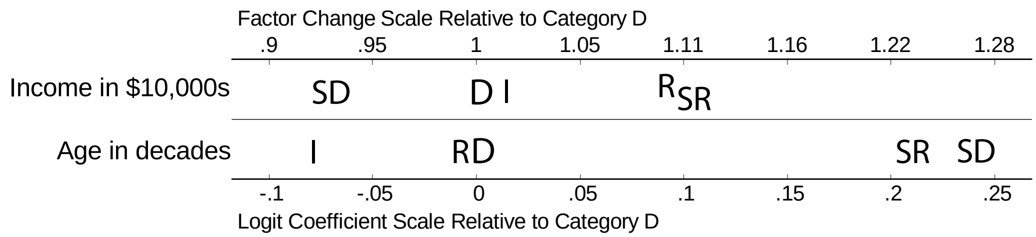

Figure 4: Plot of odds ratios for income and age for MNLM.

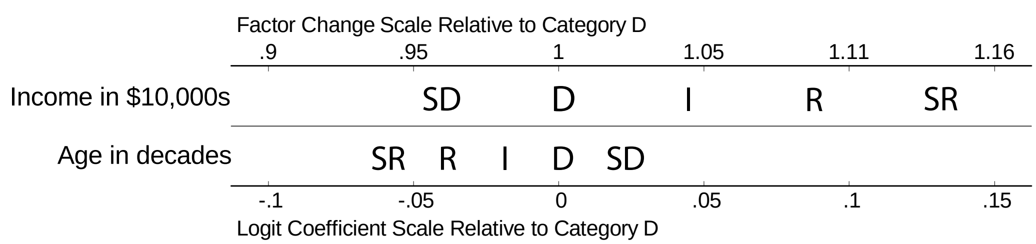

Figure 5: Plot of odds ratios for age and income for ACLM.

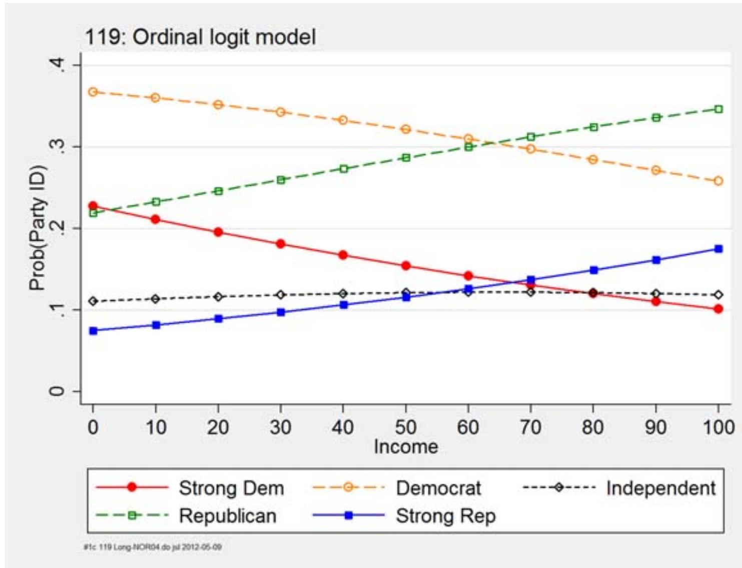

Figure 6: Predicted probabilities of party affiliation by income for the OLM with other variables held at their mean.

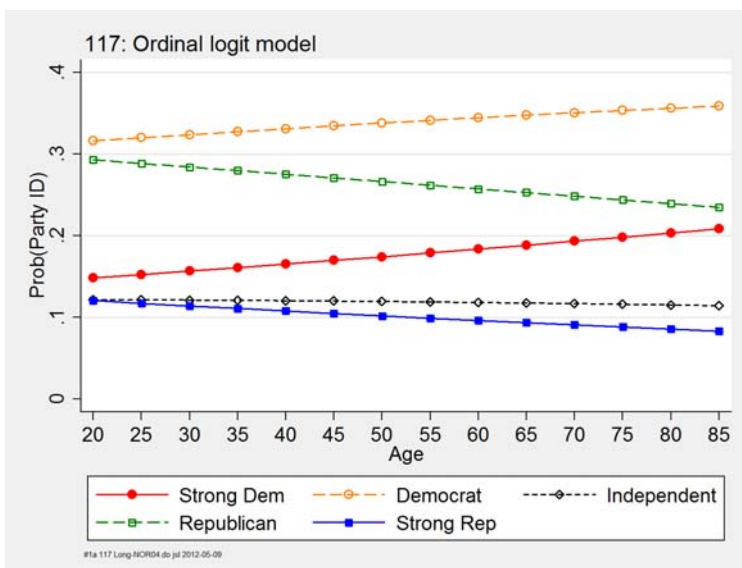

Figure 7: Predicted probabilities of party affiliation by age for the OLM with other variables held at their mean.

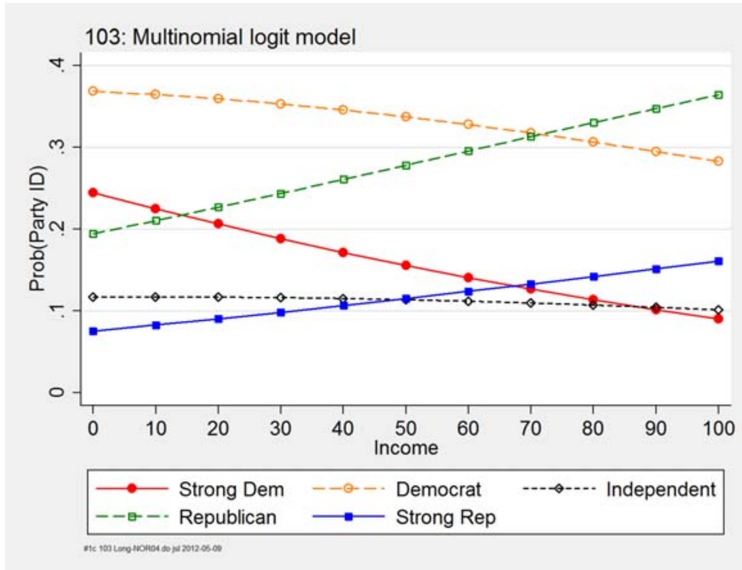

Figure 8: Predicted probabilities of party affiliation by income for the MNLM with other variables held at their mean.

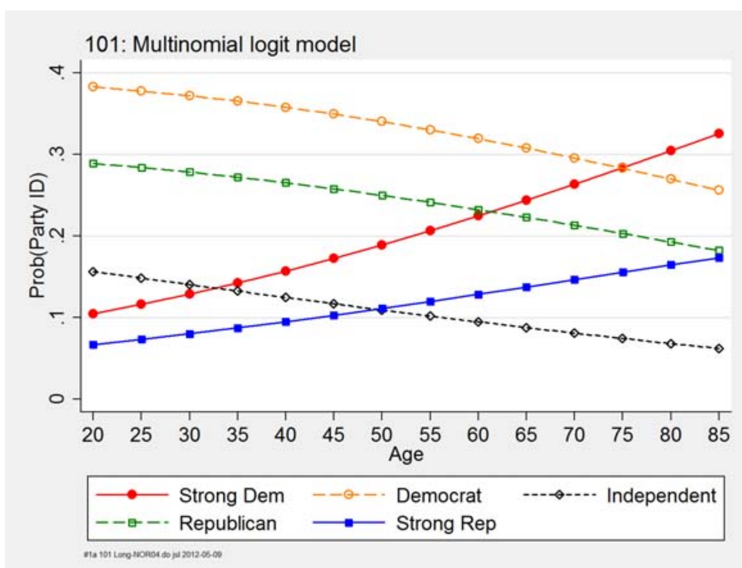

Figure 9: Predicted probabilities of party affiliation by age for the MNLM with other variables held at their mean.