Regression Models for Categorical Outcomes

J. Scott Long and Simon Cheng
Indiana University

April 28, 2001

Forthcoming in Melissa Hardy and Alan Bryman (editors),
Handbook of Data Analysis, Sage Publications

1 Introduction[1]

The linear regression model (LRM) is the most commonly used statistical method in the social sciences. A key advantage of the LRM is that the results have a simple interpretation: for a unit change in a given independent variable, the expected value of the outcome changes by a fixed amount, holding all other variables constant. Unfortunately, the application of the LRM is limited to cases in which the dependent variable is continuous and uncensored.[2] If the LRM is used when the dependent variables is categorical, censored, or truncated, the estimates are likely to be inconsistent, inefficient, or simply nonsensical. When the dependent variable is continuous and censored or truncated, models for limited dependent variables such as tobit need to be used. These are discussed in Chapter ___ of the current volume. Of particular concern for our paper are models for binary, ordinal, or nominal outcomes.

There is a wide and increasing variety of models that can be used for categorical outcomes. These include binary logit and probit, ordinal logit and probit, multinomial and conditional logit, and multinomial probit. Within the last 15 years, computational problems for estimating these models by maximum likelihood have been solved and the models can be easily estimated with readily available software. But, since these models are nonlinear, interpretation is much more difficult than for the LRM. Proper interpretation involves *post-estimation analysis* that transforms the estimated parameters into more substantively useful information.

The focus of our paper is on the most basic models for categorical outcomes. These models are extremely useful in and of themselves and also serve as the foundation for a vast and increasing number of alternative models that are available for

[1] This paper draws on the more detailed presentation in Long (1997). Examples of software to estimate the models considered in this paper can be found at `www.indiana.edu/~jsl650/rm4cldv.htm` We would like to thank Melissa Hardy and an anonymous reviewer for their valuable comments.

[2] The use of the LRM with a binary dependent variables leads to the linear probability model. However, nonlinear models for binary outcomes, discussed in this paper, have key advantages over the linear probability model, as discussed below.

categorical outcomes. Our review emphasizes the similarities among the models, noting that models for ordinal and nominal outcomes can be developed as generalizations of models for binary outcomes. Methods of interpretation are also shared by these models. Accordingly, we begin with a general discussion of issues of interpretation for nonlinear models. This is followed in Section 3 with a review of general issues related to estimating, testing and assessing fit of these models. The remaining sections consider models for binary, ordinal, and nominal outcomes.

2 Nonlinearity and Interpretation

Models for categorical outcomes are nonlinear and understanding the implications of nonlinearity is fundamental to the proper interpretation of these models. Unfortunately, data analysts often limited their "interpretation" to a table of coefficients accompanied by a brief description of the signs and significance levels of the coefficients. This unnecessary limitation can be avoided if the implications of nonlinearity are fully understood. Accordingly, in this section we focus heuristically on the idea of nonlinearity and the implications of nonlinearity for the proper interpretation of these models. Specific details as they apply to particular models are given later in the chapter.

Figure **??** shows a simple, linear regression model, where y is the dependent variables, x is a continuous independent variable, and d is a binary independent variable. The model being estimated is

$$y = \alpha + \beta x + \delta d \ ,$$

where for simplicity we have assumed that there is no error term. The solid line plots y as x changes for $d = 0$: that is, $y = \alpha + \beta x$. The dashed line plots y as x changes when $d = 1$, which simply changes the intercept: $y = \alpha + \beta x + \delta 1 = (\alpha + \delta) + \beta x$.

The effect of x on y can be computed as the partial derivative or slope of the line relating x to y, often called the *marginal change*:

$$\frac{\partial y}{\partial x} = \frac{\partial \left(\alpha + \beta x + \delta d\right)}{\partial x} = \beta \ .$$

This equation is the ratio of the change in y to the change in x, where the change in x is infinitely small, holding d constant. In a linear model, the marginal change is the same at *all* values of x and d. Consequently, when x increases by one unit, y increases by β units regardless of the current values for x and d. This is shown by the four small triangles with bases of length 1 and heights of β.

The effect of d cannot be computed with a partial derivative since d is discrete. Instead, we measure the *discrete change* in y as d changes from 0 to 1, holding x

constant:

$$\frac{\Delta y}{\Delta d} = (\alpha + \beta x + \delta\, 1) - (\alpha + \beta x + \delta\, 0) = \delta\ .$$

When d changes from 0 to 1, y changes by δ units regardless of the level of x. This is shown by the two arrows marking the distance between the solid and dashed lines.

The distinguishing feature of interpretation in the LRM is that the effect of a given change in an independent variable is the same regardless of the value of that variable at the start of its change and regardless of the level of the other variables in the model. Accordingly, interpretation only needs to specify which variable is changing, by how much, and that all other variables are being held constant. Another simplification due to the linearity of the model is that a discrete change of one unit equals the marginal change. This will not be true, however, for nonlinear models, as we now show.

Figure **??** plots a logit model where $y = 1$ if some event occurred, say if a person is in the labor force, else $y = 0$. The curves are from the logit equation[3]

$$\Pr(y = 1) = \frac{\exp(\alpha + \beta x + \delta d)}{1 + \exp(\alpha + \beta x + \delta d)}\ . \tag{1}$$

Once again, x is continuous and d is binary.

The nonlinearity of the model makes it more difficult to interpret the effects of x and d on the probability of an event occurring since neither the marginal nor discrete change with respect to x or d are constant. This is illustrated by the triangles. Since the solid curve for $d = 0$ and the dashed curve for $d = 1$ are not parallel, $\Delta_1 \neq \Delta_4$. And, the effect of a unit change in x differs according to the level of both d and x: $\Delta_2 \neq \Delta_3 \neq \Delta_5 \neq \Delta_6$. *In nonlinear models the effect of a change in a variable depends on the values of all variables in the model and is no longer simply equal to a parameters in the model.*

There are several general approaches for interpreting nonlinear models:

1. Marginal and discrete change coefficients can be computed at a representative value of the independent variables, such as when all variables equal their means. Or, the discrete and marginal changes can be computed for all values in the sample and then averaged.

2. Predicted values can be computed at values of interest and presented in tables or plots.

3. The nonlinear model can be transformed to a model that is linear or multiplicative in some other outcome. For example, the logit model in Equation 1

[3] The α, β, and δ parameters in this equation are unrelated to those in Figure 1.

can be written as

$$\ln\left(\frac{\Pr(y=1)}{1-\Pr(y=1)}\right) = \alpha + \beta x + \delta d \ ,$$

which can then be interpreted with methods for linear model. Or, the model can be expressed as a multiplicative model in terms of odds:

$$\frac{\Pr(y=1)}{1-\Pr(y=1)} = \exp(\alpha + \beta x + \delta d) \ .$$

Note, however, that here the difficulty is in the meaning of the transformed dependent variable.

Each of these approaches is discussed in Section 4.

3 Estimation, Testing, and Fit

While the focus of our review is on the form and interpretation of models for categorical outcomes, it is important to begin with general comments on estimation, testing, and measuring fit.

3.1 Estimation

Each of the model that we consider can be estimated by maximum likelihood (ML).[4] Under the usual assumptions, the ML estimator is consistent, efficient, and asymptotically normal. These desirable properties hold as the sample size approaches infinity. While ML estimators are not necessarily bad estimators in small samples, the small sample behavior of ML estimators for the models we consider is largely unknown. With the exception of the binary logit model, which can be estimated with exact permutation methods using LogXact (Cytel Software Corporation, 2000), alternative estimators with known small sample properties are not available. Based on both his experience with these methods and discussion with other researchers, Long (1997:53-54) proposed the following guidelines for the use of ML in small samples:

> It is risky to use ML with samples smaller than 100, while samples over 500 seem adequate. These values should be raised depending on characteristics of the model and the data. First, if there are may parameters,

[4] A full discussion of ML estimation is beyond the scope of this paper. For further information, see Long (1997) for a general overview, Eliason (1993) for a more detailed introduction, and Cramer (1986) for a more advanced discussion.

more observations are needed. A rule of at least ten observations per parameter seems reasonable (which does not imply that less than 100 is not needed if you have only two parameters). Second, if the data are ill-conditioned (e.g., independent variables are highly collinear) or if there is little variation in the dependent variable (e.g., nearly all of the outcomes are 1), a larger sample is required. Third, some models seem to require more observations, such as the ordinal regression model.

Numerical methods are used to compute the ML estimates. These methods work extremely well when the data are clean, variables are properly constructed, and the model is correctly specified. In some cases, problems with convergence occur if the ratio of the largest standard deviation to the smallest standard deviation among independent variables is large. For example, if income is measured in units of \$1, recoding income to units of $\$1,000$ may resolve problems with convergence. Overall, numerical methods for ML estimation work well when your model is appropriate for your data. In using these models, Cramer's (1986:10) advice should be taken very seriously: "Check the data, check their transfer into the computer, check the actual computations (preferably by repeating at least a sample by a rival program), and always remain suspicious of the results, regardless of the appeal."

3.2 Statistical Tests

Coefficients estimated by ML can be easily tested with standard Wald and likelihood ratio (LR) tests. Even though the LR and Wald tests are asymptotically equivalent, in finite samples they give different answers, particularly for small samples. In general, it is unclear which test is to be preferred. In practice, the choice of which test to use is often determined by convenience, although many statisticians (including us) prefer the LR test. While the LR test requires the estimation of two models, the computation of the test only involves subtraction. The Wald test only requires estimation of a single model, but the computation of the test involves matrix manipulations. Which test is more convenient depends on the software being used. Regarding significance levels for tests based on small samples, Allison (1995:80) suggests that, contrary to standard advice of using larger p-values in small samples, given that the degree to which ML estimates are normally distributed is unknown in small samples, it is reasonable to require smaller p-values in small samples.

3.3 Measures of Fit

Residuals and Outliers When assessing a model it is useful to consider how well the model fits each case and how much influence each case has on the estimates

of the parameters. Pregibon (1981) extended methods of residual and outlier analysis from the LRM to the case of binary logit and probit. See also Cook and Weisberg (1999: Part IV). Similar methods for ordinal and nominal outcomes are not available. However, models for ordinal and nominal outcomes can often be expressed as a series of binary models (as shown below). Methods developed for binary models can be applied to each of these models, providing potentiality useful information about the fit of the model.

Scalar Measures of Fit In addition to assessing the fit of each observation, a single number to summarize the overall goodness of fit of a model would be useful in comparing competing models and ultimately in selecting a final model. While the desirability of a scalar measure of fit is clear, in practice their use is problematic. Selecting a model that maximizes the value of a given measure of fit does not necessarily lead to a model that is optimal in any sense other than the model having a larger value of that measure. While measures of fit provide some information, it is partial information that must be assessed within the context of the theory motivating the analysis, past research, and the estimated parameters of the model being considered. For details on the many measures that have been proposed, see Long (1997: Chapter 4).

4 Models for Binary Outcomes

The binary logit and probit models, referred to jointly as the binary regression model (BRM), can be derived in three ways. First, an unobserved or latent variable can be hypothesized along with a measurement model relating the latent variable to the observed, binary outcome. Second, the model can be constructed as a probability model. And finally, the model can be generated as a random utility or discrete choice model. While we focus on the first two approaches, the third approach is used to explain the multinomial probit model.

4.1 A Latent Variable Model

Assume a *latent* variable y^* ranging from $-\infty$ to ∞ that is related to the observed independent variables by the structural equation

$$y_i^* = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i \ ,$$

where i indicates the observation and ε is a random error. The form of this equation is identical to that of the LRM with the important difference that the dependent variable is unobserved.

The observed binary variable y is related to y^* by a simple measurement equation:

$$y_i = \begin{cases} 1 & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases} .$$

Cases with positive values of y^* are observed as $y=1$, while cases with negative or zero values of y^* are observed as $y=0$. For example, let $y = 1$ if a women is in the labor force and $y = 0$ if she is not. The independent variables might include number of children, education, and expected wages. Not all women in the labor force are there with the same certainty. One woman might be close to leaving the labor force, while another woman could be firm in her decision to work. In both cases, we observe $y=1$. The idea of a latent y^* is that an underlying *propensity to work* generates the observed state. While we cannot directly observe the propensity, at some point a change in y^* results in a change in what we observe, namely, whether a woman is in the labor force.

The latent variable model for binary outcomes is illustrated in Figure **??** for a single independent variable, where we use the simpler notation $y^* = \alpha + \beta x + \varepsilon$. For a given value of x, illustrated in the figure for $x = 5$:

$$\Pr(y = 1 \mid x) = \Pr(y^* > 0 \mid x) .$$

Substituting the structural model and rearranging terms:

$$\begin{aligned} \Pr(y = 1 \mid x) &= \Pr(\alpha + \beta x + \varepsilon > 0 \mid x) \\ &= \Pr(\varepsilon > -[\alpha + \beta x] \mid x) \\ &= \Pr(\varepsilon \leq \alpha + \beta x \mid x) . \end{aligned} \tag{2}$$

This equation shows that the probability depends on the distribution of the error. Two distributions are commonly assumed. First, ε is distributed normally with $E(\varepsilon) = 0$ and $Var(\varepsilon) = 1$, which leads to the binary probit model. Specifically, Equation 2 becomes

$$\Pr(y = 1 \mid x) = \int_{-\infty}^{\alpha+\beta x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) dt .$$

Alternatively, ε is assumed to have a logistic distribution with $E(\varepsilon) = 0$ and $Var(\varepsilon) = \frac{\pi^2}{3}$, leading to the binary logit model:

$$\Pr(y = 1 \mid x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} . \tag{3}$$

The peculiar value assumed for $Var(\varepsilon)$ in the logit model illustrates a basic point regarding the identification of models with latent outcomes. In the LRM, $Var(\varepsilon)$ can

be estimated since y is observed. For the BRM, $Var(\varepsilon)$ must be assumed since the dependent variable is unobserved. The model is unidentified unless an assumption is made about the variance of the errors. For probit, we assume $Var(\varepsilon) = 1$ since this leads to a simple equation for the model. In the logit model, the variance is set to $\pi^2/3$ since this leads to the simple form of Equation 3. While the value assumed for $Var(\varepsilon)$ is arbitrary, the value does *not* affect the computed value of the probability (see Long 1997:49-50 for a simple proof). In effect, changing the assumed variance affects the spread of the distribution, but not the proportion of the distribution above or below the threshold. If a different value is assumed, the values of the structural coefficients are changed in a uniform way. This is illustrated in Figure **??**.

Overall, the probability of the event occurring is the cumulative density function of the error term evaluated at given values of the independent variables:

$$\Pr(y = 1 \mid \mathbf{x}) = F(\mathbf{x}\boldsymbol{\beta}) \ , \tag{4}$$

where F is the normal cdf Φ for the probit model and the logistic cdf Λ for the logit model. The relationship between the linear latent variable model and the resulting S-shaped probability model is shown in Figure **??** for a model with a single independent variable. Panel A shows the error distribution for nine values of x. The area where $y^* > 0$ corresponds to $\Pr(y=1 \mid x)$ and has been shaded. Panel B plots $\Pr(y=1 \mid x)$ corresponding to the shaded regions in Panel A. As we move from $x = 1$ to 2 only a portion of the thin tail crosses the threshold in Panel A, resulting in a small change in $\Pr(y=1 \mid x)$ in Panel B. As we move from $x = 2$ to 3 to 4, thicker regions of the error distribution slide over the threshold and the increase in $\Pr(y=1 \mid x)$ becomes larger. The resulting curve is the well known S-curve associated with the BRM.

4.2 A Nonlinear Probability Model

The BRM can also be derived without appealing to a latent variable. This is done by specifying a nonlinear model relating the x's to the probability of an event. Following Thiel (1970) the logit model can be derived by constructing a model in which $\Pr(y = 1 \mid \mathbf{x})$ is forced to be within the range 0 to 1. For example, in the linear probability model $\Pr(y = 1 \mid \mathbf{x}) = \mathbf{x}\boldsymbol{\beta} + \varepsilon$, the probabilities can be greater than 1 and less than 0. To constrain the range of possible values, first transform the probability into the *odds*:

$$\Omega(\mathbf{x}) = \frac{\Pr(y = 1 \mid \mathbf{x})}{\Pr(y = 0 \mid \mathbf{x})} = \frac{\Pr(y = 1 \mid \mathbf{x})}{1 - \Pr(y = 1 \mid \mathbf{x})} \ .$$

The odds indicate how often something happens (e.g., $y = 1$) relative to how often it does not happen (e.g., $y = 0$). The odds vary from 0 when $\Pr(y = 1 \mid \mathbf{x}) = 0$ to

Table 1: Descriptive Statistics for the Labor Frce Participation Example.

Name	Mean	Std Dev	Min	Max	Description
K5	0.24	0.52	0.00	3.00	# of children ages 5 and younger.
K618	1.35	1.32	0.00	8.00	# of children ages 6 to 18.
Age	42.54	8.07	30.00	60.00	Wife's age in years.
WC	0.28	0.45	0.00	1.00	1 if wife attended college; else 0.
HC	0.39	0.49	0.00	1.00	1 if husband attended college; else 0.
Lwg	1.10	0.59	-2.05	3.22	Log of wife's estimated wage rate.
Income	20.13	11.63	0.00	96.00	Family income excluding wife's wages.

Note: *N*=753.

∞ when $\Pr(y = 1 \mid \mathbf{x}) = 1$. The log of the odds, or *logit,* ranges from $-\infty$ to ∞. This suggests a model that is linear in the logit:

$$\ln \Omega(\mathbf{x}) = \mathbf{x}\boldsymbol{\beta} .$$

This equation is equivalent to our earlier definition of the logit model in Equation 3.

Other binary regression models are created by choosing functions of $\mathbf{x}\boldsymbol{\beta}$ that range from 0 to 1. Cumulative distribution functions have this property and readily provide a number of examples. For example, the cdf for the standard normal distribution results in the probit model.

4.3 Interpretation

To make our discussion concrete, we use an example from Mroz (1987) on the labor force participation of women using data from the 1976 Panel Study of Income Dynamics.[5] The sample is 753 white, married women between the ages of 30 and 60. The dependent variable *LFP=1* if a woman is employed (57%) and else 0. The independent variables are listed in Table 1.

Based on the specification

$$\begin{aligned} \Pr(LFP = 1) &= F(\beta_0 + \beta_1 K5 + \beta_2 K618 + \beta_3 Age \\ &\quad + \beta_4 WC + \beta_5 HC + \beta_6 Lwg + \beta_7 Income) , \end{aligned}$$

a binary logit and probit were estimated, with results given in Table 2. The column "Ratio" shows that the logit coefficients are about 1.7 times larger than those for

[5] These data were generously made available by Thomas Mroz.

Table 2: Logit and Probit Analyses of Labor Force Participation.

	Logit		**Probit**		**Ratio**	
Variable	β	z	β	z	β	z
Constant	3.182	4.94	1.918	5.04	1.66	0.98
K5	-1.463	-7.43	-0.875	-7.70	1.67	0.96
K618	-0.065	-0.95	-0.039	-0.95	1.67	1.00
Age	-0.063	-4.92	-0.038	-4.97	1.66	0.99
WC	0.807	3.51	0.488	3.60	1.65	0.98
HC	0.112	0.54	0.057	0.46	1.95	1.17
Lwg	0.605	4.01	0.366	4.17	1.65	0.96
Income	-0.034	-4.20	-0.021	-4.30	1.68	0.98
$-2 \ln L$	905.27		905.39			1.00

Note: N=753. β is an unstandardized coefficient; z is the z-test for β. "Ratio" is the ratio of a logit to a probit coefficient.

probit, with the exception of the coefficient for *HC* which is the least statistically significant parameter. This illustrates how the magnitudes of the coefficients are affected by the assumed $Var(\varepsilon)$. The significance tests are quite similar since they are not affected by $Var(\varepsilon)$.

4.3.1 Predicted Probabilities

In general, the estimated parameters from the BRM provide only information about the sign and statistical significance of the relationship between an independent variable and the outcome. More substantively meaningful interpretations are based on the predicted probabilities and functions of those probabilities (e.g., ratios, differences). For example, Figure **??** plots the probit model with two independent variables:

$$\Pr(y = 1 \mid x, z) = \Phi(1 + 1\,x + .75\,z) \tag{5}$$

and illustrates the basic issues involved in interpretation. Each point on the surface corresponds to the predicted probability for given values of x and z. For example, the point in the northwest corner corresponds to $\Pr(y = 1 \mid x = -4, z = 8)$. Interpretation can proceed by presenting a table of predicted probabilities at substantively interesting values of the independent variables, by plotting the predicted probability holding all but one variable constant, or by computing how much the predicted probability changes when one independent variable changes holding the others constant.

Table 3: The probability of employment by college attendence and the number of young children.

Number of Young Children	Predicted Probability		
	Did Not Attend	Attended College	Difference
0	0.61	0.78	0.17
1	0.27	0.45	0.18
2	0.07	0.16	0.09
3	0.01	0.03	0.02

While Figure **??** is for two independent variables, the idea extends to more variables. For example, consider the effects of age and income from our example of labor force participation. First, set all variables but *Age* and *Income* to their means. Holding *Age* at 30, compute the predicted probability of labor force participation as *Income* ranges from 0 to 100 using the equation

$$\widehat{\Pr}\left(LFP = 1 \mid \mathbf{x}^*\right) = \Phi\left(\mathbf{x}^*\widehat{\boldsymbol{\beta}}\right) \ ,$$

where $\mathbf{x}^*$ contains the assumed values of each variable. These predictions are plotted with the line marked with circles in Figure **??**. This process is repeated holding age at 40, 50, and 60. The nonlinearities in the effects are apparent, with the effect of income decreasing with age. When relationships are nonlinear, plots are often useful for uncovering relationships.

In other cases, a table is a more useful way to summarize results. For example, holding all variables at their means except for the wife's education and the number of young children, Table 3 clearly shows the effect of eduction and family on labor force participation.

4.3.2 Marginal and Discrete Change

Another useful method of interpretation is to compute the change in the probability of the outcome event as one variable changes, holding all other variables constant. In economics, the most commonly used measure of change is the marginal change, shown by the tangent to the probability curve in Figure **??**:

$$\text{Marginal Change} = \frac{\partial \Pr(y = 1 \mid \mathbf{x})}{\partial x_k} \ .$$

This value is often computed with all variables held at their means or by computing the marginal change for each observation in the sample and then averaging across all observations.

Alternatively, the discrete change in the predicted probabilities for a given change in an independent variables can be used. Let $\Pr(y = 1 \mid \mathbf{x}, x_k)$ be the probability of an event given $\mathbf{x}$, noting in particular the value of x_k. Thus, $\Pr(y = 1 \mid \mathbf{x}, x_k + \delta)$ is the probability with x_k increased by δ, all other variables held constant at specified values. The *discrete change* for a change of δ in x_k equals

$$\frac{\Delta \Pr(y = 1 \mid \mathbf{x})}{\Delta x_k} = \Pr(y = 1 \mid \mathbf{x}, x_k + \delta) - \Pr(y = 1 \mid \mathbf{x}, x_k)$$

which can be interpreted as:

> For a change in variable x_k from x_k to $x_k + \delta$, the predicted probability of an event changes by $\frac{\Delta \Pr(y=1|\mathbf{x})}{\Delta x_k}$, holding all other variables constant.

As shown in Figure **??**, in general the marginal change and discrete change will not be equal:

$$\frac{\partial \Pr(y = 1 \mid \mathbf{x})}{\partial x_k} \neq \frac{\Delta \Pr(y = 1 \mid \mathbf{x})}{\Delta x_k} \ .$$

The two measures of change differ since the model is nonlinear and the rate of change is constantly changing. The discrete change measures the actual amount of change over a finite change in an independent variable while the marginal measures the instantaneous rate of change. The two measures will be similar when the discrete change occurs over a region of the probability curve that is roughly linear. In practice, we prefer the discrete change since it measures the actual change occurring, regardless of the approximate linearity of the model in that area of the curve.

While measures of change are straightforward in the LRM, there is an important problem in nonlinear models: *the magnitude of the change in the probability for a given change in an independent variable depends both on the level of the independent variables and on the start value of the variable that is changing.* This is illustrated in Figure **??**. Consider the effect of a unit change in x, which corresponds to a change along a line running southwest to northeast. For example, consider the change in probability when x changes from -4 to -3, with $z = 8$. This change is quite small since the predicted probability is already near 1 at $x = -4$ and $z = 8$. Now, consider the same change in x when $z = 4$. The change is now much larger. Clearly, the amount of change caused by a unit change in x depends on the level of z and also on the start value for x. The key problem in using measures of change in nonlinear models is to decide on the level of each control variable and the value at which you want to start the change for a given variable.

Figure **??** also illustrates a subtle, but important point about computing discrete change that was raised by Kaufman (1996). Consider the point on the curve at $\Pr(y = 1 \mid x)$. When x increases by 1, $\Pr(y = 1)$ increases by some amount. When x decreases by 1, $\Pr(y = 1)$ decreases by an amount that is smaller than the change caused by the 1 unit increase. Because of this asymmetry around a given point on the curve, it is useful to center the change around a given value of x. For example, rather than examining the quantity

$$\frac{\Delta \Pr(y = 1 \mid \mathbf{x})}{\Delta \, x_k} = \Pr(y = 1 \mid \mathbf{x}, x_k + 1) - \Pr(y = 1 \mid \mathbf{x}, x_k) \ ,$$

Table 4: Discrete change in the probability of employment.

Variable	Centered Unit Change	Centered Standard Deviation Change	Change From 0 to 1
K5	-0.33	-0.18	- - -
K618	-0.02	-0.02	- - -
Age	-0.01	-0.12	- - -
WC	- - -	- - -	0.18
HC	- - -	- - -	0.02
Lwg	0.14	0.08	- - -
Income	-0.01	-0.09	- - -

Note: Changes are computed with other variables held at their means.

the *centered discrete change* can be used:

$$\frac{\Delta \Pr(y = 1 \mid \mathbf{x})}{\Delta x_k} = \Pr\left(y = 1 \mid \mathbf{x}, x_k + \tfrac{1}{2}\right) - \Pr\left(y = 1 \mid \mathbf{x}, x_k - \tfrac{1}{2}\right) \ .$$

This is the measure that we report in our examples.

Table 4 contains measures of discrete change for the probit model of women's labor force participation. For example, the effects can be interpreted as:

> For a woman who is average on all characteristics, an additional young child decreases the probability of employment by .33.
>
> A standard deviation change in age centered around the mean will decrease the probability of working by .12, holding other variables constant.
>
> If a woman attends college, her probability of being in the labor force is .18 greater than a woman who does not attend college, holding other variables at their means.

4.3.3 Odds Ratios

Recall that the logit model, but not the probit model, can be written as linear in the log of the odds of the event occurring:

$$\ln \Omega \left(\mathbf{x}\right) = \mathbf{x}\boldsymbol{\beta} \ .$$

Taking the exponential,

$$\begin{aligned} \Omega \left(\mathbf{x}, x_k\right) &= \exp \left(\mathbf{x}\boldsymbol{\beta}\right) \\ &= e^{\beta_0} e^{\beta_1 x_1} \cdots e^{\beta_k x_k} \cdots e^{\beta_K x_K} \ , \end{aligned}$$

Table 5: Factor change coefficients for labor frce participation.

Variable	Logit Coef	Factor Change	Std Factor Change	z-value
Constant	3.182	- - -	- - -	4.94
K5	-1.463	0.232	0.465	-7.43
K618	-0.065	0.937	0.918	-0.95
Age	-0.063	0.939	0.602	-4.92
WC	0.807	2.242	- - -	3.51
HC	0.112	1.118	- - -	0.54
Lwg	0.605	1.831	1.427	4.01
Income	-0.035	0.966	0.670	-4.20

where $\Omega(\mathbf{x}, x_k)$ makes explicit the value of variable x_k. To assess the effect of x_k, we want to see how the odds change when x_k changes by some quantity δ, which is often set to 1 or the standard deviation of x_k. If we change x_k by δ, the odds become

$$\Omega(\mathbf{x}, x_k + \delta) = e^{\beta_0} e^{\beta_1 x_1} \cdots e^{\beta_k x_k} e^{\beta_k \delta} \cdots e^{\beta_K x_K} \ .$$

The *odds ratio* is simply

$$\frac{\Omega(\mathbf{x}, x_k + \delta)}{\Omega(\mathbf{x}, x_k)} = e^{\beta_k \delta} \ ,$$

which can be interpreted as:

> For a change of δ in x_k, the odds are expected to change by a factor of $\exp(\beta_k \times \delta)$, holding all other variables constant.

Importantly, the effect of a change in x_k does *not* depend on the level of x_k or on the level of any other variable.

The factor change and standardized factor change coefficients for the logit model analyzing labor force participation are presented in Table 5. Here is how some of the coefficients can be interpreted:

> For each additional young child, the odds of being employed are decreased by a factor of .23, holding all other variables constant.
>
> For a standard deviation increase in wages, the odds of being employed are 1.43 times greater, holding all other variables constant.
>
> Being ten years older decreases the odds by a factor of .52 ($=e^{-.063\times 10}$), holding all other variables constant.

Since the odds ratio is a multiplicative coefficient, "positive" effects are greater than one, while "negative" effects are between zero and one. Therefore, *positive and negative effects should be compared by taking the inverse of the negative effect (or vice versa).* For example, a positive factor change of 2 has the same magnitude as a negative factor change of $.5 = 1/2$. Second, *a constant factor change in the odds does not correspond to a constant change or constant factor change in the probability.* For example, if the odds are 2:1 and are doubled to 4:1, the probability changes from .667 to .800, a change of .130. If the odds are 10:1 and double to 20:1, the change is the probability is only: .909-.952=.043. While the odds change by a constant factor of two, the probabilities do not change by a constant amount. Consequently, when interpreting a factor change in the odds, it is *essential* to know what the current level of the odds or probability is.

4.3.4 Summary Regarding Interpretation

For nonlinear models, no single approach to interpretation can fully describe the relationship between a variable and the outcome probability. The data analyst should search for an elegant and concise way to summarize the results that does justice to the complexities of the nonlinear model. To do this, it is often necessary to try each method of interpretation before a final approach is determined.

5 Models for Ordinal Outcomes

While there are several models for ordinal outcomes, we focus on the ordered logit and ordered probit models, which are the most commonly used models for ordinal outcomes in the social sciences. These models, referred to jointly as the ordered regression model (ORM), were introduced by McKelvey and Zavoina (1975) in terms of an underlying latent variable. At about the same time, the model was developed in biostatistics (McCullagh 1980), where it is referred to as the *proportional odds model*,

the *parallel regression model*, or the *grouped continuous model*. After presenting the ORM, we consider several less common models for ordinal outcomes.

5.1 A Latent Variable Model

The close relationship between the BRM and the ORM is easily shown in the latent variable formulation of the model. Using the same structural model

$$y_i^* = \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i \ ,$$

we simply expand the measurement model to divide y^* into J ordinal categories:

$$y_i = m \quad \text{if } \tau_{m-1} \leq y_i^* < \tau_m \quad \text{for } m = 1 \text{ to } J \ .$$

The *cutpoints* or *thresholds* τ_1 through τ_{J-1} are estimated and, for reasons that are explained below, we assume $\tau_0 = -\infty$ and $\tau_J = \infty$. For example, people respond to the statement: "A working mother can establish just as warm and secure of a relationship with her child as a mother who does not work," with the ordinal categories: "Strongly Disagree" (SD), "Disagree" (D), "Agree" (A), and "Strongly Agree" (SA). The latent variable is the propensity to agree that working mothers can be good mothers, leading to the measurement model:

$$y_i = \begin{cases} 1 \Rightarrow \text{SD} & \text{if } \ \tau_0 = -\infty \leq y_i^* < \tau_1 \\ 2 \Rightarrow \text{D} & \text{if } \ \tau_1 \leq y_i^* < \tau_2 \\ 3 \Rightarrow \text{A} & \text{if } \ \tau_2 \leq y_i^* < \tau_3 \\ 4 \Rightarrow \text{SA} & \text{if } \ \tau_3 \leq y_i^* < \tau_4 = \infty \end{cases} \ .$$

For a single independent variable, this ORM is shown in Figure **??**. The predicted probability of an outcome is the area under the curve between a pair of cutpoints at a given level of the independent variables. For example, we observe $y = 2$ when y^* falls between τ_1 and τ_2:

$$\Pr\left(y = 2 \mid \mathbf{x}\right) = \Pr\left(\tau_1 \leq y^* < \tau_2 \mid \mathbf{x}\right) \ .$$

Substituting $y^* = \mathbf{x}\boldsymbol{\beta} + \varepsilon$ and using some algebra, the predicated probability is the difference

$$\Pr\left(y = 2 \mid \mathbf{x}\right) = F\left(\tau_2 - \mathbf{x}\boldsymbol{\beta}\right) - F\left(\tau_1 - \mathbf{x}\boldsymbol{\beta}\right) \ ,$$

where F is the cdf for the assumed distribution of the errors. As with the BRM, if F is normal with $Var(\varepsilon) = 1$, we have the ordinal *probit* model; if F is logistic with $Var(\varepsilon) = \pi^2/3$, we have the ordinal *logit* model. In general, for each outcome m:

$$\Pr\left(y = m \mid \mathbf{x}\right) = F\left(\tau_m - \mathbf{x}\boldsymbol{\beta}\right) - F\left(\tau_{m-1} - \mathbf{x}\boldsymbol{\beta}\right) \ . \tag{6}$$

For $y = 1$, the second term drops out since $F\left(-\infty - \mathbf{x}\boldsymbol{\beta}\right) = 0$; for $y = J$, the first term equals $F\left(\infty - \mathbf{x}\boldsymbol{\beta}\right) = 1$. Thus, with two outcome categories, the model is identical the binary regression model (see Equation 4).

5.1.1 Parameterization

In the BRM, we assumed that $\tau = 0$ in order to identify the model. The ORM is commonly identified in either of two ways. First, some software assumes that $\tau_1 = 0$ and estimates the intercept β_0, while others programs assume that $\beta_0 = 0$ and estimate the threshold τ_1. The choice of parameterization does not affect estimates of the slopes, but does affect the estimates of β_0 and the τ's. Importantly, the parameterization does not affect the predicted probabilities.

5.2 The Parallel Regression Assumption

To understand and interpret the ORM, it is helpful to reformulate the model in terms of the *cumulative probability* that an outcome is less than or equal to m:

$$\Pr(y \leq m \mid \mathbf{x}) = \sum_{j=1}^{m} \Pr(y = j \mid \mathbf{x}) \ .$$

Expanding $\Pr(y = j \mid \mathbf{x})$ with Equation 6 and canceling terms leads to the simple expression

$$\Pr(y \leq m \mid \mathbf{x}) = F(\tau_m - \mathbf{x}\boldsymbol{\beta}) \ . \tag{7}$$

This equation both shows the link between the BRM and the ORM and makes explicit a fundamental assumption of the ORM.

Consider the case with a single independent variable:

$$\begin{aligned} \Pr(y \leq m \mid \mathbf{x}) &= F(\tau_m - [\alpha + \beta x]) \\ &= F([\tau_m - \alpha] + \beta x) \\ &= F(\alpha_m^* + \beta x) \ . \end{aligned} \tag{8}$$

The new notation makes it clear that the cumulative probability equation is identical to a binary regression. That is, the ORM is equivalent to simultaneously estimating $J-1$ binary regressions. For example, with four outcomes we would simultaneously estimate three equations:

$$\begin{aligned} \Pr(y \leq 1 \mid \mathbf{x}) &= F(\alpha_1^* + \beta x) \\ \Pr(y \leq 2 \mid \mathbf{x}) &= F(\alpha_2^* + \beta x) \\ \Pr(y \leq 3 \mid \mathbf{x}) &= F(\alpha_3^* + \beta x) \ . \end{aligned} \tag{9}$$

While the α_m^*'s differ across equations, the β's do not. This is reflected by the lack of subscripts for β. The $J-1$ binary regressions are assumed to have *exactly* the same value across all equations. This assumption is referred to either as the *parallel*

Table 6: Descriptive statistics for the attitudes toward working mothers example.

Name	Mean	Std Dev	Min	Max	Description
Yr89	0.40	0.49	0.00	1.00	Survey Year: 1=1989; 0=1977.
Male	0.47	0.50	0.00	1.00	1=male; 0=female.
White	0.88	0.33	0.00	1.00	1=white; 0=non-white.
Age	44.94	16.78	18.00	89.00	Age in years.
Ed	12.22	3.16	0.00	29.00	Years of education.
Prst	39.59	14.49	12.00	82.00	Occupational prestige.
Note: N=2293.					

regression assumption (since plotting Equation 8 for $m = 1$ to $J - 1$ results in parallel curves) or, for the ordinal logit model, as the *proportional odds assumption*.

While the constraint of parallel regressions or proportional odds is implicit in the ordinal regression model, in our experience the assumption is violated in many applications. This is illustrated with an example.

5.2.1 Attitudes toward Working Mothers

The 1977 and 1989 General Social Survey asked respondents to evaluate the statement: "A working mother can establish just as warm and secure a relationship with her child as a mother who does not work." Responses were coded as: 1=Strongly Disagree (SD); 2=Disagree (D); 3=Agree (A); and 4=Strongly Agree (SA). The variables used in our analysis are described in Table 6. Estimates from ordered logit are given in the first column of Table 7, with estimates from three binary logits on cumulative probabilities in the last three columns. The parallel regression assumption requires that the β's in the last three equations are equal. While some of the estimates are similar across equations (e.g., *White*), others are quite different (e.g., *Male*).

Formal tests are also available. A score test is included in SAS's LOGISTIC (SAS Institute 1990:1090). An approximate LR test is in Stata's *omodel* command (Wolf and Gould, 1998). These are omnibus tests that do not allow you to tell if the problem only exists for some of the independent variables. Brant's (1990) Wald test allows both an overall test that all $\boldsymbol{\beta}_m$'s are equal, and tests of the equality of coefficients for individual variables. This test is available in Stata through the *brant* command (Long and Freese forthcoming). In our example, the value of the Brant test for the hypothesis that all coefficients are parallel is 49.18 with 12 degrees of freedom, providing strong evidence that the assumption of parallel regressions is violated. Looking at the tests for individual variables, we find that the evidence

Table 7: Ordered logit and cumulative logit regressions.

Variable		Ordered Logit	Cumulative Logits $m \leq 1$	$m \leq 2$	$m \leq 3$
Yr89	β	0.524	0.965	0.565	0.319
	z	6.33	6.26	6.09	2.80
Male	β	-0.733	-0.305	-0.691	-1.084
	z	-9.23	-2.36	-7.68	-8.88
White	β	-0.391	-0.553	-0.314	-0.393
	z	-3.27	-2.40	-2.24	-2.49
Age	β	-0.022	-0.017	-0.025	-0.019
	z	-8.52	-4.06	-8.84	-4.94
Ed	β	0.067	0.105	0.053	0.058
	z	4.20	4.14	2.86	2.27
Prst	β	0.006	-0.001	0.010	0.006
	z	1.84	-0.25	2.50	1.14

Note: β is an unstandardized coefficient; z is a z-test of β.

against parallel regressions is strongest for the variables *Yr89* ($X^2 = 13.01$, df= 2, $p < .01$) and *Male* ($X^2 = 22.24$, df= 2, $p < .01$).

5.3 Interpretation

If the idea of a latent variable makes substantive sense (and assuming that the assumption of parallel regressions is not violated), simple interpretations are possible by rescaling y^* and computing standardized coefficients. When concern is with the observed categories, the methods illustrated for the BRM can be extended to multiple outcomes. Since the ORM is nonlinear in the outcome probabilities, no single approach can fully describe the relationship between a variable and the outcome probabilities. Consequently, you should consider each of these methods before deciding which approach is most effective in a given application.

5.3.1 Partial Change in y^*

In the ORM, $y^* = \mathbf{x}\boldsymbol{\beta} + \varepsilon$ and the marginal change in y^* with respect to x_k is

$$\frac{\partial y^*}{\partial x_k} = \beta_k \ .$$

Since y^* is latent, the marginal cannot be interpreted without standardizing y^*. The variance of y^* can be estimated by the quadratic form

$$\widehat{\sigma}^2_{y^*} = \widehat{\boldsymbol{\beta}}' \widehat{Var}\left(\mathbf{x}\right) \widehat{\boldsymbol{\beta}} + Var\left(\varepsilon\right) \ , \tag{10}$$

where $\widehat{Var}\left(\mathbf{x}\right)$ is the covariance matrix for the observed x's; $\widehat{\boldsymbol{\beta}}$ contains ML estimates; and $Var(\varepsilon) = 1$ for the probit model and $\pi^2/3$ for the logit model. Then, the y^*-*standardized* coefficient for x_k is

$$\beta_k^{Sy^*} = \frac{\beta_k}{\sigma_{y^*}} \ ,$$

which can be interpreted as:

> For a unit increase in x_k, y^* is expected to increase by $\beta_k^{Sy^*}$ standard deviations, holding all other variables constant.

With σ_k equal to the standard deviation for x_k, the *fully standardized coefficient* is

$$\beta_k^S = \frac{\sigma_k \beta_k}{\sigma_{y^*}} = \sigma_k \beta_k^{Sy^*} \ ,$$

which can be interpreted as:

> For a standard deviation increase in x_k, y^* is expected to increase by β_k^S standard deviations, holding all other variables constant.

For our example using the ordinal logit model, $\widehat{\sigma}^2_{y^*} = 3.77$ and the standardized coefficients can be interpreted as follows:

> In 1989, support was .27 standard deviations higher than in 1977, holding all other variables constant.
>
> Each standard deviation increase in education increases support by .11 standard deviations, holding all other variables constant.

5.3.2 Predicted Probabilities

The predicted probabilities and cumulative probabilities can be estimated as:

$$\begin{aligned} \widehat{\Pr}\left(y = m \mid \mathbf{x}\right) &= F\left(\widehat{\tau}_m - \mathbf{x}\widehat{\boldsymbol{\beta}}\right) - F\left(\widehat{\tau}_{m-1} - \mathbf{x}\widehat{\boldsymbol{\beta}}\right) \\ \widehat{\Pr}\left(y \leq m \mid \mathbf{x}\right) &= F\left(\tau_m - \mathbf{x}\widehat{\boldsymbol{\beta}}\right) \ . \end{aligned} \tag{11}$$

Table 8: Predicted probabilities by gender and year.

1977	SD	D	A	SA
Men	0.19	0.40	0.32	0.10
Women	0.10	0.31	0.41	0.18
Men-Women	0.09	0.09	-0.09	-0.08
1989	SD	D	A	SA
Men	0.12	0.34	0.39	0.15
Women	0.06	0.23	0.44	0.27
Men-Women	0.06	0.11	-0.05	-0.12
Change from 1977 to 1989	SD	D	A	SA
Men	-0.07	-0.06	0.07	0.05
Women	-0.04	-0.08	0.03	0.09

Either set of probabilities can be plotted. For example, in Panel A of Figure **??**, the predicted probabilities for each outcome are plotted. The probability of strongly agreeing, indicated with circles, shows that at age 20 the probability is .39. As age increases the probability decreases to .25 at age 50 and is .15 at age 80. The probability of disagreeing, indicated by triangles, is nearly the mirror image. It begins at .16 at age 20 and ends at .34 at age 80. There is a smaller change in the probability of strongly disagreeing, indicated by diamonds, that starts at .04 and ends at .12. The probability of agreeing, shown by squares, illustrates an unusual characteristic of the ORM, which also occurs with nominal models. The effect of age on agreeing is initially positive and is then negative. As age increases from 20, more cases from category SA move into category A than move from A into D, which increase the probability of A. With increasing age, more cases leave A for D than enter A from SA, resulting in a smaller probability.

Cumulative probabilities can be plotted as shown in Panel B. The cumulative probabilities "stack" the corresponding probabilities from the top panel and show the overall increase with age in negative attitudes toward the statement that working women can have a warm relationship with their children. The information in the graph can be viewed in two ways. First, height within a category (e.g., the height of the trapezoid labeled "Strongly Agree") corresponds to the predicted probability for that category. Second, the height from the x-axis to the top of a category is the probability for all categories at or below that level.

Tables can also be used to present probabilities. Table 8 contains the predicted

probabilities for men and women by the year of the survey, along with differences by gender in the probabilities within year and across years. The first thing to notice is that men are more likely than women to disagree and strongly disagree that working women can have as warm of relationships with their children. Second, between 1977 and 1989 there was a movement for both men and women towards more positive attitudes.

Both partial and discrete change can also be used. The partial derivative of Equation 6:

$$\begin{aligned}\frac{\partial \Pr(y=m \mid \mathbf{x})}{\partial x_k} &= \frac{\partial F(\tau_m - \mathbf{x}\boldsymbol{\beta})}{\partial x_k} - \frac{\partial F(\tau_{m-1} - \mathbf{x}\boldsymbol{\beta})}{\partial x_k} \\ &= \beta_k \left[f(\tau_{m-1} - \mathbf{x}\boldsymbol{\beta}) - f(\tau_m - \mathbf{x}\boldsymbol{\beta}) \right] ,\end{aligned}$$

is the slope of the curve relating x_k to $\Pr(y=m \mid \mathbf{x})$, holding all other variables constant. *The sign of the marginal is not necessarily the same as the sign of* β_k, since $f(\tau_{m-1} - \mathbf{x}\boldsymbol{\beta}) - f(\tau_m - \mathbf{x}\boldsymbol{\beta})$ can be negative. Accordingly, the sign of the estimated β's in the ORM should not be used as a quick indication of the direction of a variables effect on any of the outcome categories (as illustrated by the curve for agreeing in Panel A of Figure **??**).[6] Since the marginal effect depends on the levels of all variables, we must decide on which values of the variables to use when computing the effect. As with the BRM, the marginal can be averaged over all observations or computed at the mean of all variables. Keep in mind that the marginal change does *not* indicate the change in the probability that would be observed for a unit change in x_k, unless an independent variable is varying over a region of the probability curve that is nearly linear. When the curve is approximately linear, the marginal effect can be used to summarize the effect of a unit change in the variable on the probability of an outcome.

Since interpretation using marginal effects can be misleading when the probability curve is changing rapidly or when an independent variable is a dummy variable, we prefer using discrete change. The discrete change in the predicted probability for a change in x_k from the start value x_{S} to the end value x_{E} (e.g., a change from $x_k = 0$ to $x_k = 1$) is

$$\frac{\Delta \Pr(y=m \mid \mathbf{x})}{\Delta x_k} = \Pr(y=m \mid \mathbf{x}, x_k = x_{\mathrm{E}}) - \Pr(y=m \mid \mathbf{x}, x_k = x_{\mathrm{S}}) ,$$

where $\Pr(y=m \mid \mathbf{x}, x_k)$ is the probability that $y = m$ given $\mathbf{x}$, noting a specific value for x_k. The change is interpreted as:

[6] Note, however, that since Equation 7 is a binary logit for any given m, the sign of an estimated β indicates the direction of the effect of a variable on the probability of being less than or equal to some category.

Table 9: Discrete change in the probability of attitudes about working mothers for the ordered logit model.

Variable	**Change**	$\overline{\Delta}$	**SD**	**D**	**A**	**SA**
Overall Probability		- - -	0.11	0.33	0.40	0.16
Yr89	$0 \rightarrow 1$	0.06	-0.05	-0.08	0.05	0.07
Male	$0 \rightarrow 1$	0.09	0.08	0.11	-0.08	-0.10
White	$0 \rightarrow 1$	0.05	0.04	0.06	-0.04	-0.06
Age	$\Delta 1$	0.00	0.00	0.00	-0.00	-0.00
	$\Delta\sigma$	0.04	0.04	0.05	-0.04	-0.05
	ΔRange	0.18	0.18	0.19	-0.18	-0.19
Ed	$\Delta 1$	0.01	-0.01	-0.01	0.01	0.01
	$\Delta\sigma$	0.03	-0.02	-0.03	0.02	0.03
	ΔRange	0.16	-0.15	-0.17	0.16	0.17
Prst	$\Delta 1$	0.00	-0.00	-0.00	0.00	0.00
	$\Delta\sigma$	0.01	-0.01	-0.01	0.01	0.01
	ΔRange	0.05	-0.04	-0.06	0.04	0.06

Note: $0 \rightarrow 1$ is change from 0 to 1; $\Delta 1$ is centered change of one around the mean; $\Delta\sigma$ is centered change of one standard deviation around the mean; ΔRange is change from the minimum to the maximum. $\overline{\Delta}$ is the average absolute discrete change.

> When x_k changes from x_{S} to x_{E}, the predicted probability of outcome m changes by $\frac{\Delta \Pr(y=m|\mathbf{x})}{\Delta x_k}$, holding all other variables at $\mathbf{x}$.

As with the BRM, the value of the discrete change depends on: (1) the value at which x_k starts; (2) the amount of change in x_k; and (3) the values of all other variables. Most frequently each continuous variable except x_k is held at its mean. For dummy independent variables, the change might be computed for both values of the variable. For example, we could compute the discrete change for age separately for men and women.

Table 9 contains measures of discrete change for our example using the ordered logit model.

> The probability of strongly disagreeing is .08 higher for men than women, holding all other variables at their means.

For variables that are not binary, the discrete change can be interpreted for a unit change centered around the mean, for a standard deviation change centered around

the mean, and when the variable goes from its minimum to its maximum value. For example,

> For each additional year of education, the probability of strongly agreeing increases by .01, holding other variables constant at their means.
>
> For a standard deviation increase in age, the probability of disagreeing increases by .05, holding other variables at their means.
>
> Moving from the minimum prestige to the maximum prestige changes the predicted probability of strongly agreeing by .06, holding all other variables at their means.

The J discrete change coefficients for a variable can be summarized by computing the average of the absolute values of the changes across all of the outcome categories:

$$\overline{\Delta} = \frac{1}{J} \sum_{j=1}^{J} \left| \frac{\Delta \Pr\left(y = j \mid \overline{\mathbf{x}}\right)}{\Delta x_k} \right| .$$

The absolute value is taken since the sum of the changes without taking the absolute value is necessarily zero. The average absolute discrete change in the table clearly shows that the respondent's gender, education, and age have the strongest effects on attitudes about working mothers.

5.3.3 Odds Ratios

To illustrate the use of odds ratios, consider the coefficient for gender from Table 7: $\beta_{Male} = -.73$, so that $\exp(-\beta_{Male}) = 2.1$. This can be interpreted as:

> The odds of SD versus the combined outcomes D, A, and SA are 2.1 times greater for men than women, holding other variables constant. Similarly, the odds of SD and D versus A and SA are 2.1 times greater from men than women; and the odds of SD, D, and A versus SA are 2.1 times greater.

The coefficient for age is $\beta_{Age} = -.02$ with a standard deviation $s_{Age} = 16.8$. Thus, $100\left[\exp\left(-s_{Age} \times \beta_{Age}\right) - 1\right] = .44$, which can be interpreted as:

> For a standard deviation increase in age, the odds of SD versus D, A, and SA are increased by 44 percent, holding other variables constant. Similarly, the odds of SD and D versus A and SA are 44 percent greater for every standard deviation increase in age; and the odds of SD, D, and A versus SA are 44 percent greater.

5.3.4 Summary

The ordered regression model is the most frequently used model for ordinal outcomes. However, as our discussion has shown, this model imposes the strong assumption of parallel regressions or proportional odds. We recommend that you always test this assumption, ideally in a way that allows you to assess which variables are violating the assumption (which can suggest problems in the specification of the model). In our experience, outcomes that are considered ordinal often contain complexities that are "assumed away" by the ORM. For example, a variable could be ordered differently with respect to different independent variables. Or, it might be ordered on more than one dimension or be only partially ordered. Accordingly, we suggest that if your outcome is ordinal, you also consider the ordinal models discussed in the next section as well as models for nominal outcomes.

6 Less Common Models for Ordinal Outcomes

In this section we consider briefly several less commonly used models for ordinal outcomes. While we do not consider methods of interpretation, the same approaches discussed above can be used after making the appropriate change to the formula for computing predicted probabilities. The first two models that we consider, the generalized ordered logit model and the stereotype model, relax the assumption of equal $\boldsymbol{\beta}$'s over outcome categories that is found in the ORM. The last two models, the adjacent categories model and the continuation ratio model, propose alternative comparisons for the ordinal categories.

6.1 Generalized Ordered Logit Model

The parallel regression assumption results from assuming the same coefficient vector $\boldsymbol{\beta}$ for all comparisons in the $J-1$ equations:

$$\ln \Omega_{y \leq m}(\mathbf{x}) = \tau_m - \mathbf{x}\boldsymbol{\beta} \ ,$$

where $\Omega_{y \leq m}(\mathbf{x}) = \frac{\Pr(y \leq m|\mathbf{x})}{\Pr(y > m|\mathbf{x})}$. The generalized ordered logit model (GOLM) removes the restriction of parallel regressions by allowing $\boldsymbol{\beta}$ to differ for each of the $J-1$ comparisons. That is,

$$\ln \Omega_{y \leq m}(\mathbf{x}) = \tau_m - \mathbf{x}\boldsymbol{\beta}_m \quad \text{for } j = 1, J-1 \ .$$

Or, in terms of odds

$$\Omega_{y \leq m}(\mathbf{x}) = \exp(\tau_m - \mathbf{x}\boldsymbol{\beta}_m) \quad \text{for } j = 1, J-1 \ .$$

Predicted probabilities are computed by solving these equations, resulting in:

$$\begin{aligned}
\Pr(y = 1 \mid \mathbf{x}) &= \frac{\exp(\tau_1 - \mathbf{x}\boldsymbol{\beta}_1)}{1 + \exp(\tau_1 - \mathbf{x}\boldsymbol{\beta}_1)} \\
\Pr(y = j \mid \mathbf{x}) &= \frac{\exp\left(\tau_j - \mathbf{x}\boldsymbol{\beta}_j\right)}{1 + \exp\left(\tau_j - \mathbf{x}\boldsymbol{\beta}_j\right)} - \frac{\exp\left(\tau_{j-1} - \mathbf{x}\boldsymbol{\beta}_{j-1}\right)}{1 + \exp\left(\tau_{j-1} - \mathbf{x}\boldsymbol{\beta}_{j-1}\right)} \text{ for } j = 2, J-1 \\
\Pr(y = J \mid \mathbf{x}) &= 1 - \frac{\exp\left(\tau_{J-1} - \mathbf{x}\boldsymbol{\beta}_{J-1}\right)}{1 + \exp\left(\tau_{J-1} - \mathbf{x}\boldsymbol{\beta}_{J-1}\right)} \, .
\end{aligned}$$

To insure that the $\Pr(y = j \mid \mathbf{x})$ is between 0 and 1, it must be the case that $\left(\tau_j - \mathbf{x}\boldsymbol{\beta}_j\right) \geq \left(\tau_{j-1} - \mathbf{x}\boldsymbol{\beta}_{j-1}\right)$. Since this constraint is not imposed during estimation, it is possible that predicted probabilities can be negative or greater than 1. Once predicted probabilities are computed, all of the approaches used to interpret the ORM results can be readily applied.

While we have not seen social science applications of this model, it has been discussed by Clogg and Shihadeh (1994:146-147), Fahrmeir and Tutz (1994:91), and McCullagh and Nelder (1989:155). Applications may become more common since this model has recently been programmed for Stata by Fu (1998).

6.2 The Stereotype Model[7]

The *stereotype ordered regression model* (SORM) was proposed by Anderson (1984) in response to the restrictive assumption of parallel regressions in the ordered regression model. The SORM is a compromise between allowing the coefficients for each independent variable to vary by outcome category and restricting them to be identical across all outcomes. The SORM is defined as[8]

$$\ln \frac{\Pr(y = q)}{\Pr(y = r)} = (\alpha_q - \alpha_r)\,\beta_0 + \left(\phi_q - \phi_r\right)(\mathbf{x}\boldsymbol{\beta}) \, , \tag{12}$$

where β_0 is the intercept and $\boldsymbol{\beta}$ is a vector of coefficients associated with the independent variables; since β_0 is included in the equation, it is not included in $\boldsymbol{\beta}$. The α's and ϕ's are scale factors associated with the outcome categories. To see how

[7] The name of this model appears to come from a line in Anderson's (1984) article in which he discusses how cases might allocated to ordinal outcomes: "One possibility is that the judge has loose stereotypes for each category and that a new case for categorization is fitted into the most appropriate category."

[8] The sterotype model can be set up in several different ways. For example, in some presentations, it is assumed that $\beta_0 = 0$ and fewer constraints are imposed on the α's. Here we parameterize the model to highlight its links to other models that we consider.

these work, consider a model with two independent variables and three outcomes:

$$\begin{aligned}
\ln \frac{\Pr(y=1)}{\Pr(y=2)} &= (\alpha_1 - \alpha_2)\beta_0 + (\phi_1 - \phi_2)\beta_1 x_1 + (\phi_1 - \phi_2)\beta_2 x_2 \\
\ln \frac{\Pr(y=1)}{\Pr(y=3)} &= (\alpha_1 - \alpha_3)\beta_0 + (\phi_1 - \phi_3)\beta_1 x_1 + (\phi_1 - \phi_3)\beta_2 x_2 \\
\ln \frac{\Pr(y=2)}{\Pr(y=3)} &= (\alpha_2 - \alpha_3)\beta_0 + (\phi_2 - \phi_3)\beta_1 x_1 + (\phi_2 - \phi_3)\beta_2 x_2 \ .
\end{aligned}$$

The model allows the coefficients associated with each independent variable to differ by a scalar factor that depends on the pair of outcomes on the left-hand-side of the equation. For example, in the equation comparing outcomes 1 and 2, the coefficient β_1 for x_1 is rescaled by the factor $\phi_1 - \phi_2$; for outcomes 1 and 3, by the factor $\phi_1 - \phi_3$; and for 2 and 3, by the factor $\phi_2 - \phi_3$. The same factors are also used for the coefficient for x_2. Similarly, the α's allow different intercepts for each pair of outcomes.

As the model stands, it is over-parameterized (i.e., there are too many unconstrained α's and ϕ's to allow the parameters to be uniquely determined) and constraints must be imposed to identify the model. The model can be identified in a variety of ways. For example, we can assume $\phi_1 = 1$, $\phi_J = 0$, $\alpha_1 = 1$, and $\alpha_J = 0$. Or, using the approach from loglinear models for ordinal outcomes, the model is identified by the constraints $\sum_{j=1}^{J} \phi_j = 0$ and $\sum_{j=1}^{J} \phi_j^2 = 1$. See DiPrete (1990) for further discussion.

The model we have presented above, which does *not* include any order restrictions, is commonly referred to as the stereotype model. However, Anderson (1984) referred to the model without ordering constraints as the "ordered regression model." The stereotype model includes additional constraints that insures the ordinality of the outcomes: $\phi_1 = 1 > \phi_2 > \cdots > \phi_{J-1} > \phi_J = 0$.

Equation 12 can be used to compute the predicted probabilities:

$$\Pr(y = m \mid \mathbf{x}) = \frac{\exp(\alpha_m \beta_0 + \phi_m \mathbf{x}\boldsymbol{\beta})}{\sum_{j=1}^{J} \exp\left(\alpha_j \beta_0 + \phi_j \mathbf{x}\boldsymbol{\beta}\right)} \ .$$

This formula can be used for interpreting the model using methods discussed above. The model can also be interpreted in terms of the effect of a change in x_k on the odds of outcome q versus r. After rewriting Equation 12 in terms of odds:

$$\Omega_{q|r}(\mathbf{x}, x_k) = \frac{\Pr(y=q)}{\Pr(y=r)} = \exp\left[(\alpha_q - \alpha_r)\beta_0 + \left(\phi_q - \phi_r\right)(\mathbf{x}\boldsymbol{\beta})\right] \ ,$$

it is easy to show that

$$\frac{\Omega_{q|r}(\mathbf{x}, x_k + 1)}{\Omega_{q|r}(\mathbf{x}, x_k)} = e^{(\phi_q - \phi_r)\beta_k} = \left(\frac{e^{\phi_q}}{e^{\phi_r}}\right)^{\beta_k} \ .$$

Thus the effect of x_k on the odds of q versus r differ across outcome comparisons according to the scaling coefficients ϕ.

DiPrete (1990) used a general ML program in GAUSS to estimate this model. Recently, Hendrickx's (2000) *mclest* program in Stata can also be used to estimate the model. Note that these programs do not impose the ordinality constraint $\phi_1 = 1 > \phi_2 > \cdots > \phi_{J-1} > \phi_J = 0$. Since the SORM is closely related to the multinomial logit model (MNLM), discussed below, the model can be informally assessed by examining the parameters from the MNLM to see if the structure of the stereotype model is approximated. This approach was taken by Greenwood and Farewell (1988).

6.3 Adjacent Categories Model

The *adjacent categories model* (Agresti 1990: 318; Clogg and Shihadeh 1994:149-154) is a special case of the multinomial logit model considered in the next section. The model is specified as

$$\ln\left[\frac{\Pr(y=m \mid \mathbf{x})}{\Pr(y=m+1 \mid \mathbf{x})}\right] = \tau_m - \mathbf{x}\boldsymbol{\beta}\ ,$$

where the outcome is the log of the odds of category m versus category $m+1$. Note that the vector $\boldsymbol{\beta}$ is the same for all values of m. Taking exponentials,

$$\begin{aligned}\Omega_{m|m+1}(\mathbf{x}) &= \frac{\Pr(y=m \mid \mathbf{x})}{\Pr(y=m+1 \mid \mathbf{x})} \\ &= \exp(\tau_m - \mathbf{x}\boldsymbol{\beta})\ .\end{aligned}$$

From this it follows readily that for a unit increases in x_k, $\Omega_{m|m+1}$ changes by a factor of $\exp(-\beta_k)$, holding all other variables constant.

Using simple but tedious algebra, these equations can be solved for the predicted probabilities:

$$\begin{aligned}\Pr(y=m \mid \mathbf{x}) &= \frac{\exp\left(\sum_{r=m}^{J-1}[\tau_r - \mathbf{x}\boldsymbol{\beta}]\right)}{1+\sum_{q=1}^{J-1}\left[\exp\left(\sum_{r=q}^{J-1}[\tau_r - \mathbf{x}\boldsymbol{\beta}]\right)\right]} \quad \text{for } m = 1, J-1 \\ \Pr(y=J \mid \mathbf{x}) &= 1 - \sum_{q=1}^{J-1} p_q\end{aligned}$$

These probabilities can be used in the methods of interpretation that were discussed for the ORM.

6.4 The Continuation Ratio Model

The *continuation ratio model* was proposed by Fienberg (1980:110) and designed for ordinal outcomes in which the categories represent the progression of events or stages in some process though which an individual can advance.[9] For example, the outcome could be faculty rank, where the stages are assistant professor, associate professor, and full professor. A key characteristic of the process is that an individual must pass through each stage. For example, to become an associate professor you must be an assistant professor; to be a full professor, an associate professor. While there are versions of this model based on other binary models (e.g., probit), here we consider the logit version.

If $\Pr(y = m \mid \mathbf{x})$ is the probability of being in stage m given $\mathbf{x}$, then the probability of being in stage m or later is:

$$\Pr(y \geq m \mid \mathbf{x}) = \sum_{j=m}^{J} \Pr(y = j \mid \mathbf{x}) \ .$$

The *conditional* probability of being in stage m given that you are in stage m or later (e.g., the probability of being an associate professor given that you have progressed from the rank of assistant professor) is:

$$\Pr(y = m \mid y \geq m, \mathbf{x}) = \frac{\Pr(y = m \mid \mathbf{x})}{\Pr(y \geq m \mid \mathbf{x})} \ .$$

And accordingly the probability of begin beyond stage m is:

$$\begin{aligned} \Pr(y > m \mid y \geq m, \mathbf{x}) &= 1 - \Pr(y = m \mid y \geq m, \mathbf{x}) \\ &= \frac{\Pr(y > m \mid \mathbf{x})}{\Pr(y \geq m \mid \mathbf{x})} \ . \end{aligned}$$

Using these probabilities, we can compute the odds of being in stage m compared to being past stage m given that a respondent is in stage m or later:

$$\frac{\Pr(y = m \mid y \geq m, \mathbf{x})}{\Pr(y > m \mid y \geq m, \mathbf{x})} = \frac{\Pr(y = m \mid \mathbf{x})}{\Pr(y > m \mid \mathbf{x})} \ .$$

We can then construct a model for the log odds::

$$\ln\left[\frac{\Pr(y = m \mid \mathbf{x})}{\Pr(y > m \mid \mathbf{x})}\right] = \tau_m - \mathbf{x}\boldsymbol{\beta} \quad \text{for } m = 1 \text{ to } J - 1$$

[9] For a discussion of the links between this model and survival analysis, see Allison (1995).

where the $\boldsymbol{\beta}$'s are constrained to be equal across outcome categories, while the constant term τ_m differs by stage. As with other logit models, we can also express the model in terms of the odds:

$$\frac{\Pr(y = m \mid \mathbf{x})}{\Pr(y > m \mid \mathbf{x})} = \exp(\tau_m - \mathbf{x}\boldsymbol{\beta}) \ .$$

Accordingly, $\exp(-\beta_k)$ can be interpreted as the effect of a unit increase in x_k on the odds of being in m compared to being in a higher category given that an individual is in category m or higher, holding all other variables constant.

The formula for the predicted probabilities highlights the structure of the model. The probability of $y = 1$ is computed from $\frac{\Pr(y=1|\mathbf{x})}{\Pr(y>1|\mathbf{x})} = \exp(\tau_1 - \mathbf{x}\boldsymbol{\beta})$ just as in the model binary logit:

$$\Pr(y = 1 \mid \mathbf{x}) = \frac{\exp(\tau_1 - \mathbf{x}\boldsymbol{\beta})}{1 + \exp(\tau_1 - \mathbf{x}\boldsymbol{\beta})} \ .$$

The probability of $y = 2$ equals:

$$\Pr(y = 2 \mid \mathbf{x}) = \frac{\exp(\tau_2 - \mathbf{x}\boldsymbol{\beta})}{[1 + \exp(\tau_2 - \mathbf{x}\boldsymbol{\beta})] \times [1 + \exp(\tau_1 - \mathbf{x}\boldsymbol{\beta})]} \ .$$

In general,

$$\begin{aligned} \Pr(y = m \mid \mathbf{x}) &= \frac{\exp(\tau_m - \mathbf{x}\boldsymbol{\beta})}{\prod_{j=1}^{m} [1 + \exp(\tau_j - \mathbf{x}\boldsymbol{\beta})]} \text{ for } m = 1 \text{ to } J - 1 \\ \Pr(y = J \mid \mathbf{x}) &= 1 - \sum_{j=1}^{J-1} \Pr(y = j \mid \mathbf{x}) \ . \end{aligned}$$

These predicted probabilities can be used for interpreting the model.

7 Models for Nominal Outcomes

For ordinal outcomes, we also recommend using models for nominal outcomes, which are now discussed. If a dependent variable is ordinal and a nominal model is used, there is a loss of efficiency since information is being ignored. On the other hand, when an ordinal model is applied to a nominal dependent variable, the resulting estimates are biased or nonsensical. Overall, if there are concerns about the ordinality of the dependent variable, the potential loss of efficiency in using models for nominal outcomes is outweighed by avoiding potential bias. Of course, these models must also be used when the dependent variable is nominal. We consider three closely related models: the multinomial logit model, the conditional logit model, and finally the multinomial probit model.

7.1 Multinomial Logit

The MNLM can be thought of as simultaneously estimating binary logits for all comparisons among the outcome categories. Indeed, Begg and Gray (1984) show that estimates from binary logits are consistent estimates of the parameters of the MNLM. For example, let y be a nominal outcome with categories A, B, and C and a single independent variables x. We can estimate the effect of x on y by running three binary logits:

$$\begin{aligned}
\ln\left[\frac{\Pr(A \mid \mathbf{x})}{\Pr(B \mid \mathbf{x})}\right] &= \beta_{0,A|B} + \beta_{1,A|B}x \\
\ln\left[\frac{\Pr(B \mid \mathbf{x})}{\Pr(C \mid \mathbf{x})}\right] &= \beta_{0,B|C} + \beta_{1,B|C}x \\
\ln\left[\frac{\Pr(A \mid \mathbf{x})}{\Pr(C \mid \mathbf{x})}\right] &= \beta_{0,A|C} + \beta_{1,A|C}x
\end{aligned}$$

where the subscripts to the β's indicate which comparison is being made. The three binary logits include redundant information in the sense that the following equality must hold:

$$\ln\left[\frac{\Pr(A \mid \mathbf{x})}{\Pr(B \mid \mathbf{x})}\right] + \ln\left[\frac{\Pr(B \mid \mathbf{x})}{\Pr(C \mid \mathbf{x})}\right] = \ln\left[\frac{\Pr(A \mid \mathbf{x})}{\Pr(C \mid \mathbf{x})}\right] .$$

This implies that

$$\begin{aligned}
\beta_{0,A|B} + \beta_{0,B|C} &= \beta_{0,A|C} \\
\beta_{1,A|B} + \beta_{1,B|C} &= \beta_{1,A|C} .
\end{aligned} \tag{13}$$

Accordingly, if there are 3 outcomes, only 2 binary logits are needed since the remaining comparison can be derived. In general, with J outcomes, only $J-1$ binary logits are needed.

The problem with estimating the MNLM by a series of binary logits is that each binary logit is based on a different sample since only cases from two outcomes are used. Consequently the equalities in Equation 13 will not hold exactly. Programs for the MNLM *simultaneously* estimate the $J-1$ binary logits, thus insuring that the implied equalities hold. Specific packages differ in which comparisons are estimated. For example, one program might estimate $A|C$ and $B|C$, while another might estimate $A|B$ and $C|B$.

Formally, the MNLM can be written as

$$\Pr(y = m \mid \mathbf{x}_i) = \frac{\exp\left(\mathbf{x}_i\boldsymbol{\beta}_{m|r}\right)}{\sum_{j=1}^{J}\exp\left(\mathbf{x}_i\boldsymbol{\beta}_{j|r}\right)} , \tag{14}$$

where r is the reference category used by the software estimating the model. Regardless of the reference category used, the predicted probability for a given outcome is identical. Alternatively, the model can be written in terms of logits

$$\ln \Omega_{m|r}\left(\mathbf{x}_i\right) = \mathbf{x}_i \boldsymbol{\beta}_{m|r} \tag{15}$$

or in terms of odds

$$\Omega_{m|r}\left(\mathbf{x}_i\right) = \exp\left(\mathbf{x}_i \boldsymbol{\beta}_{m|r}\right) \ . \tag{16}$$

Note that with J dependent categories, there are $J-1$ non-redundant, coefficients associated with each independent variable x_k. In our simple example, the coefficients $\beta_{1,A|C}$ and $\beta_{1,B|C}$ completely describe the effects of x on the three outcome categories. Accordingly, to test that a variable has no effect, you need to test that $J-1$ coefficients are simultaneously equal to zero. In our example, to test the effect of x the hypothesis is:

$$H_0\text{: } \beta_{1,A|C} = \beta_{1,B|C} = 0 \ .$$

Or, more generally, the hypothesis that x_k does not affect the dependent variable can be written as:

$$H_0\text{: } \beta_{k,1|b} = \cdots = \beta_{k,J|b} = 0$$

where b is the base category. Since $\beta_{k,b|b}$ is necessarily 0, the hypothesis imposes constraints on $J-1$ parameters. This hypothesis can be tested with either a Wald or a LR test using standard procedures available with most packages that estimate the MNLM. Both types of test are distributed as chi-square with $J-1$ degrees of freedom.

7.1.1 Interpretation of the MNLM

While the MNLM is mathematically a simple extension of the binary model, interpretation is difficult due to the large number of possible comparisons. For example, with 3 outcomes you can compare 1|2, 1|3, and 2|3. With four outcomes: 1|4, 2|4, 3|4, 2|3, 2|4, and 3|4. And so on. To illustrate how to interpret the model, we consider the effects of race, education, and work experience on occupation.

Example of Occupation The 1982 General Social Survey asked respondents their occupation. These occupations were recoded into five broad categories: menial jobs, blue collar jobs, craft jobs, white collar jobs, and professional jobs. This outcome is one that many would argue is ordered. However, as illustrated by Miller and Volker (1985), different orderings lead to different outcomes. Accordingly, a nominal model is appropriate. Three independent variables are considered, which

Table 10: Descriptive statistics for the occupational attainment example.

Name	Mean	Std Dev	Min	Max	Description
White	0.92	0.28	0.0	1.0	Race: 1= white; 0= nonwhites.
Ed	13.10	2.95	3.0	20.0	Education: Number of years of formal education.
Exp	20.50	13.96	2.0	66.0	Possible years of work experience: Age minus years of education minus 5.

Note: $N = 337$.

are described in Table 10. The estimated coefficients in Table 11 are the standard output from a program that estimates the MNLM and correspond to the equations:

$$\begin{aligned}
\ln \Omega_{B|M}(\mathbf{x}_i) &= \beta_{0,B|M} + \beta_{1,B|M}\, White + \beta_{2,B|M} Ed + \beta_{3,B|M} Exp \\
\ln \Omega_{C|M}(\mathbf{x}_i) &= \beta_{0,C|M} + \beta_{1,C|M}\, White + \beta_{2,C|M} Ed + \beta_{3,C|M} Exp \\
\ln \Omega_{W|M}(\mathbf{x}_i) &= \beta_{0,W|M} + \beta_{1,W|M}\, White + \beta_{2,W|M} Ed + \beta_{3,W|M} Exp \\
\ln \Omega_{P|M}(\mathbf{x}_i) &= \beta_{0,P|M} + \beta_{1,P|M}\, White + \beta_{2,P|M} Ed + \beta_{3,P|M} Exp\ .
\end{aligned}$$

Predicted Probabilities The estimated coefficients can be plugged into Equation 14 to compute predicted probabilities that can be used in the same way as shown for ordinal outcomes.

Marginal and Discrete Change Marginal and discrete change can be used in the same way as in models for ordinal outcomes. Marginal change is defined as

$$\frac{\partial \Pr(y = m \mid \mathbf{x})}{\partial x_k} = \Pr(y = m \mid \mathbf{x}) \left[\beta_{k,m|J} - \sum_{j=1}^{J} \beta_{k,j|J} \Pr(y = j \mid \mathbf{x}) \right] .$$

Since this equation combines all of the $\beta_{k,j|J}$'s, the marginal effect of x_k on $\Pr(y = m \mid \mathbf{x})$ need not have the same sign as the corresponding coefficient $\beta_{k,m|J}$ (keep in mind that the $\beta_{k,j|J}$'s are from equations for the odds of outcomes, not probabilities of being in various outcomes). Discrete change is defined as

$$\frac{\Delta \Pr(y = m \mid \mathbf{x})}{\Delta x_k} = \Pr(y = m \mid \mathbf{x}, x_k = x_E) - \Pr(y = m \mid \mathbf{x}, x_k = x_S) .$$

Table 11: Logit coefficients for a MNLMModel of occupational attainment.

Comparison		Logit Coefficient for *Constant*	*White*	*Ed*	*Exp*
$B \mid M$	β	0.741	1.237	-0.099	0.0047
	z	0.49	1.71	-0.97	0.27
$C \mid M$	β	-1.091	0.472	0.094	0.0277
	z	-0.75	0.78	0.96	1.66
$W \mid M$	β	-6.239	1.571	0.353	0.0346
	z	-3.29	1.74	3.01	1.84
$P \mid M$	β	-11.518	1.774	0.779	0.0357
	z	-6.23	2.35	6.79	1.98

Note: N=337. β is a logit coefficient for the indicated comparison; z is a z-value. Job types: M=menial; B=blue collar; C=craft; W=white collar; P=professional.

One difficulty with nominal outcomes is the large number of coefficients that need to be considered: one for each variable times the number of outcome categories. A plot, such as Figure **??**, can help you see the pattern in the effects. In this case it is easy to see that the effects of education are largest and those of experience are smallest. Or, each coefficient can be interpreted individually, such as:

> The effects of a standard deviation change in education are largest, with an increase in the probability of over .35 for professional occupations.
>
> The effects of race are also substantial, with average blacks being less likely to enter blue collar, white collar, or professional jobs.
>
> The expected changes due to a standard deviation change in experience are much smaller and show that experience increases the probabilities of more highly skilled occupations.

While discrete change is useful, it is essential to remember that different values are obtained at different levels of the variables. Further, discrete change does not indicate the dynamics among the dependent outcomes. For example, a decrease in education increases the probability of both blue collar and craft jobs, but how does it affect the odds of a person choosing a craft job relative to a blue collar job? To deal with these issues, the odds ratios can be used.

Table 12: Odds ratios for the effects of race on occupational attainment.

Factor Change in the Odds of m vs n			**Outcome n**				
			M	*B*	*C*	*W*	*P*
Outcome	*M*	Menial	- - -	0.29	0.62	0.21	0.17
m	*B*	Blue Collar	3.44	- - -	2.15	0.72	0.58
	C	Craft	1.60	0.47	- - -	0.33	0.27
	W	White Collar	4.81	1.40	3.00	- - -	0.82
	P	Professional	5.90	1.71	3.68	1.23	- - -

Note: The coefficients in the table are $\exp\left(\widehat{\beta}_{1,m|n}\right)$.

Odds Ratios As with the binary model, the factor change in the odds of one outcome compared to another is a simple transformation of the estimated coefficients:

$$\frac{\Omega_{m|n}\left(\mathbf{x}, x_k + \delta\right)}{\Omega_{m|n}\left(\mathbf{x}, x_k\right)} = e^{\beta_{k,m|n}\delta} .$$

The odds ratio can be interpreted as:

> For a unit change in x_k, the odds are expected to change by a factor of $\exp(\beta_{k,m|n})$, holding all other variables constant.
>
> For a standard deviation change in x_k, the odds are expected to change by a factor of $\exp(\beta_{k,m|n} \times s_k)$, holding all other variables constant.

To illustrate how to interpret the odds ratios for the MNLM, consider the coefficients for the effect of race on occupational attainment. These are shown in Table 12. The odds ratio for the effect of race on having a professional versus a menial job is 5.90, which can be interpreted as:

> The odds of having a professional occupation relative to a menial occupation are 5.9 times greater for whites than for blacks, holding education and experience constant.

To fully understand the effects of race, the coefficients for comparisons among all pairs of outcomes should be considered, even though they provide redundant information. However, to consider all of the coefficients for even a single variable with only five dependent categories is complicated. Consequently, we recommend that these coefficients be plotted (see Long 1997: Chapter 6 for full details), as illustrated in Figure **??**.

In this plot, the outcome categories are indicated by letters. Distances between letters for a given variable indicate the magnitude of the corresponding $\beta_{k,m|n}$ (i.e., the coefficient for independent variable x_k for outcome m versus n) when measured using the logit coefficient scale at the bottom. The scale at the top indicates the factor change $\exp\left(\beta_{k,m|n}\right)$. The size of the letters is proportional to the square of the discrete change that was plotted in Figure **??**. The square is used so that the area of the letter corresponds to the size of the discrete change. The graph shows that race orders occupations from menial to craft to blue collar to white collar to professional. The dotted lines show that none of the adjacent categories is significantly differentiated by race. Being white increases the odds of being a craft worker relative to having a menial job, but the effect is not significant. However, being white significantly increases the odds of being a blue collar worker, a white collar worker, or a professional relative to having a menial job. The effects of *Ed* and *Exp* can be interpreted similarly.

7.2 The Conditional Logit Model

The conditional logit model (CLM) is closely related to the MNLM.[10] The key difference is that in the CLM each independent variable is measured for each outcome category. For example, in modeling which mode of transportation people use for commuting, we might consider three modes of travel: train, car, and bus. The amount of time it takes to get to work depends on the mode of transportation and specific characteristics of an individual. For example, if you live next to a bus stop, your time by bus will be less than someone who has a 30 minute walk to the bus stop. Thus, each independent variable is defined for each outcome category. This information is entered into the CLM as follows:

$$\Pr\left(y_i = m \mid \mathbf{z}_i\right) = \frac{\exp\left(\mathbf{z}_{im}\boldsymbol{\gamma}\right)}{\sum_{j=1}^{J}\exp\left(\mathbf{z}_{ij}\boldsymbol{\gamma}\right)} \,. \tag{17}$$

This equation can be compared to the MNLM:

$$\Pr\left(y_i = m \mid \mathbf{x}_i\right) = \frac{\exp\left(\mathbf{x}_i\boldsymbol{\beta}_{m|J}\right)}{\sum_{j=1}^{J}\exp\left(\mathbf{x}_i\boldsymbol{\beta}_{j|J}\right)} \,. \tag{18}$$

In Equation 18 there are $J-1$ parameters $\beta_{k,m|r}$ for each x_k, but only a single value of x_k for each individual. In Equation 17 there is a single γ_k for each variable z_k, but there are J values of the variable for each individual.

[10] Indeed, the CLM can be used to estimate the MNLM. This is often useful since programs for CLM allow adding constraints that are not easy to impose in programs for the MNLM.

Applications of the CLM are relatively rare outside of the area of travel demand analysis (where the method was initially developed) since appropriate data are not readily available. Hoffman and Duncan (1988), however, provide a useful comparison on multinomial and conditional logit models applied to outcomes of marriage and welfare status. They also consider a mixed model that includes elements of both models.

7.3 Independence of Irrelevant Alternatives

In both the MNLM and the CLM, there is an implicit assumption referred to as the *Independence of Irrelevant Alternatives* (IIA). To understand this assumption, note that the odds in these models do not depend on other outcomes that might be available:

$$\begin{aligned} \frac{\Pr\left(y = m \mid \mathbf{x}\right)}{\Pr\left(y = n \mid \mathbf{x}\right)} &= \exp\left(\mathbf{x}\left[\boldsymbol{\beta}_{m|J} - \boldsymbol{\beta}_{n|J}\right]\right) \\ \frac{\Pr\left(y = m \mid \mathbf{z}\right)}{\Pr\left(y = n \mid \mathbf{z}\right)} &= \exp\left(\left[\mathbf{z}_m - \mathbf{z}_n\right]\boldsymbol{\gamma}\right) \ . \end{aligned}$$

This implies that adding or deleting outcomes does not affect the odds among the remaining outcomes. This point is often made with the red bus/blue bus example. Suppose that you have the choice of a *red* bus or a car to get to work and that the odds of taking a red bus compared to a car are 1:1. IIA implies that the odds will remain 1:1 between these two alternatives if a new bus company comes to town that is identical to the red bus company except for the color of the bus. Thus, the probability of driving a car can be made arbitrarily small by adding enough different colors of buses! More reasonably, we would expect that the odds of a red bus compared to a car would be reduced to 1:2 since half of those riding the red bus would be expected to ride the blue bus.

There are two tests of the IIA assumption. Hausman and McFadden (1984) proposed a Hausman-type test and McFadden, Tye, and Train (1976) proposed an approximate likelihood ratio test that was improved by Small and Hsiao (1985). Details on computing these tests are found in Zhang and Hoffman (1993) or Long (1997:Chapter 6). Our experience with these tests is that they often give inconsistent results and in practice provide little guidance to violations of the IIA assumption. Unfortunately, there do not appear to be simulation studies that examine their small sample properties. Perhaps as a result of the practical limitations of these tests, McFadden (1973) suggested that IIA implies that the multinomial and conditional logit models should only be used in cases where the outcome categories "can plausibly be assumed to be distinct and weighed independently in the eyes of each

decision maker." Similarly, Amemiya (1981:1517) suggests that the MNLM works well when the alternatives are dissimilar. Care in specifying the model to involve distinct outcomes that are not substitutes for one another seems to be reasonable, even if unfortunately ambiguous, advice.

7.4 Multinomial Probit

While logit and probit models for binary and ordinal outcomes are essentially equivalent, the multinomial probit model (MNPM), initially proposed by Aitchison and Bennett (1970), has important features that are not found in the MNLM. In particular, the MNPM does not require the IIA assumption. However, until recently, the computations necessary for estimating the MNPM made the model impractical for all but the simplest applications. Recent work by McFadden (1989) has made progress in solving the computational problems and there are at least two programs that can estimate the MNPM: Limdep (Greene 1995) and GAUSSX (Breslaw 1999). Now that estimation is computational feasible for modestly sized models, the focus has turned to issues of identification (Keane 1992), which are discussed below.

The MNPM is generally developed as a discrete choice model, which we will do using three outcome categories. See Pudney (1989) for a detailed discussion or Long (1997, Chapter 6) for an introduction. Let u_j be the utility associated with choice j. Then,

$$\begin{aligned} u_1 &= \mathbf{x}\boldsymbol{\beta}_1 + \varepsilon_1 \\ u_2 &= \mathbf{x}\boldsymbol{\beta}_2 + \varepsilon_2 \\ u_3 &= \mathbf{x}\boldsymbol{\beta}_3 + \varepsilon_3 \ , \end{aligned}$$

where $\mathbf{x}$ is a vector of independent variables and ε_j is the error for outcome j. In the MNPM, the ε's are assumed to have a multivariate normal distribution with mean zero and

$$Cov \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 \end{pmatrix} \ .$$

The correlated errors avoid the restriction of IIA. For example, in the red bus/blue bus example we would expect the errors to be negatively correlated to reflect that these options are close substitutes. If the ε's have an extreme value distribution, the resulting model is the MNLM, where the IIA assumption is necessary since $Cov(\boldsymbol{\varepsilon})$ must be diagonal for the extreme value distribution.

The outcome chosen by an individual is based on a comparison of the utilities associated with the choices. An individual chooses outcome j over outcome k if

$u_j > u_k$. For example,

$$\begin{aligned}
\Pr(1 \text{ over } 2 \mid \mathbf{x}) &= \Pr([\mathbf{x}\boldsymbol{\beta}_1 + \varepsilon_1] > [\mathbf{x}\boldsymbol{\beta}_2 + \varepsilon_2]) = \Pr([\varepsilon_1 - \varepsilon_2] > [\mathbf{x}\boldsymbol{\beta}_2 - \mathbf{x}\boldsymbol{\beta}_1]) \\
\Pr(1 \text{ over } 3 \mid \mathbf{x}) &= \Pr([\mathbf{x}\boldsymbol{\beta}_1 + \varepsilon_1] > [\mathbf{x}\boldsymbol{\beta}_3 + \varepsilon_3]) = \Pr([\varepsilon_1 - \varepsilon_3] > [\mathbf{x}\boldsymbol{\beta}_3 - \mathbf{x}\boldsymbol{\beta}_1]) \ .
\end{aligned}$$

Accordingly, the probability of choosing outcome 1 would be

$$\Pr(1 \mid \mathbf{x}) = \Pr([1 \text{ over } 2] \ \& \ [1 \text{ over } 3] \mid \mathbf{x}) \ .$$

The last quantity involves multiple integrals which leads to the computation difficulties in estimating the MNPM. In our experience, for a multinomial logit model that can be estimated in a minute, the corresponding probit model might take hours to estimate.

The model is not identified unless restrictions are placed on $Cov(\boldsymbol{\varepsilon})$, an issue that is discussed by Keane (1992). Without restrictions, a proportional change in all elements of $Cov(\boldsymbol{\varepsilon})$ and the β's does not affect the probabilities. And, adding a constant to each β_0 leaves the probabilities unchanged since it is only the difference in utilities that determines the choice. Standard identification conditions involve normalizing the variance of one alternative and restricting the utility function to 0. Formally, the model can be identified by setting $u_3 = 0$ and $\sigma_1 = 0$, as follows:

$$\begin{aligned}
u_1 &= \mathbf{x}\boldsymbol{\beta}_1 + \varepsilon_1 \\
u_2 &= \mathbf{x}\boldsymbol{\beta}_2 + \varepsilon_2 \\
u_3 &= 0
\end{aligned}$$

with covariance matrix:

$$Cov \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} = \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \ .$$

While these conditions *formally identify* the model, Keane finds that this identification is fragile. That is, additional restrictions beyond those required for formal identification are necessary in order to avoid the substantial risk of obtaining unreliable results. Our experience in experiment with the MNPM leads us to fully endorse 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." Thus, while the MNPM appears to offer substantial advantages over the MNLM in avoiding the IIA assumption, in practice this model remains difficult to use.

8 Conclusions

In this chapter we have considered what we believe are the most basic and useful models for the analysis of categorical dependent variables. However, in the limited space available, it is impossible to consider all of the important issues related to these models. Topics that have not been discussed include: robust, exact, and nonparametric methods of estimation, specification tests (Davidson and MacKinnon 1993:522-528; Greene 2000:827-831), complex sampling, multiple equation systems (see Browne and Arminger 1995 for a review), and hierarchical models (Longford 1995:551-556).

There are several sources that we recommend for obtaining further information. Maddala (1983) considers dozens of models for categorical and limited dependent variables. McCullagh and Nelder (1989) discuss some of the same models from the standpoint of the generalized linear model. King (1989) presents many of these models with particular application to political science. Agresti (1990) is particularly useful if all of your variables are nominal or ordinal. Powers and Xie (2000) consider both regression models and loglinear models. Greene's (1995) Limdep will estimate many of the models discussed here as well as many others; the manual provides a wealth of information. Finally, our review has not considered the many useful models for count outcomes (e.g., how many times did a person go to a doctor; how many patents did a company receive). Fortunately, Cameron and Trivedi (1998) provide an extensive review of these models; Long (1997: Chapter 8) provides a more elementary introduction.

Until recently, the models considered in this chapter required specialized software. At this time, however, each of these models can easily be estimated on typical, desktop computer (with the exception of the MNPM). A detailed discussion of using SAS for these models is found in Allison (1999). Long and Freese (forthcoming) provide information on estimating these models in Stata and provide a series of commands that facilitate the types of interpretation that we recommend. Since many packages do not make it simple to compute many of the quantities that we find useful for interpretation, Cheng and Long (2000) have written a series of Excel files that facilitate post-estimation interpretation.

9 References

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.

Aitchison, J., and Bennett, J. (1970). Polychotomous Quantal Response by Maximum Indicant. *Biometrika*, 57, 253-262.

Allison, P.D. (1995). *Survival Analysis Using the SAS System.* Cary, NC: SAS Institute, Inc.

Allison, P.D. (1999). *Logistic Regression Using the SAS System: Theory and Application.* Cary, NC, SAS Institute Inc.

Amemiya, T. (1981). Qualitative Response Models: A Survey. *Journal of Economic Literature*, 19, 1483-1536.

Anderson, J.A. (1984). Regression and Ordered Categorical Variables (with Discussion). *Journal of the Royal Statistical Society Series B*, 46, 1-30.

Begg, C.B., and Gray, R. (1984). Calculation of Polychotomous Logistic Regression Parameters Using Individualized Regressions. *Biometrika*, 71, 11-18.

Brant, R. (1990). Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression. *Biometrics*, 46, 1171-1178.

Breslaw (1999). *GAUSSX.* Westmount, CANADA, Econotron Software.

Browne, M.W., and Arminger, G. (1995). Specification and Estimation of Mean- and Covariance-Structures Models. In G. Arminger, C. C. Clogg, and M. E. Sobel (Eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences* (pp. 185-249). New York: Plenum Press.

Cameron, A.C. and P.K. Trivedi (1998). *Regression Analysis of Count Data.* New York, Cambridge.

Cheng. S. and Long, J.S. (2000) *EPost: Excel Programs for the Post-estimation Analysis of Models for Categorical Outcomes.* `www.indiana.edu/~jsl650`.

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

Cook, R.D. and S. Weisberg (1999). *Applied Regression Including Computing and Graphics.* New York, Wiley.

Cramer, J.S. (1986). *Econometric Applications of Maximum Likelihood Methods.* Cambridge: Cambridge University Press.

Cytel Software Corporation (2000). *LogXact Version 4.* Cambridge, MA, Cytel Software Corporation.

Davidson, R., and MacKinnon, J.G. (1993). *Estimation and Inference in Econometrics.* New York, NY: Oxford University Press.

DiPrete, T.A. (1990). Adding Covariates to Loglinear Models for the Study of Social Mobility. *American Sociological Review*, 55, 757-773.

Eliason, S. (1993). *Maximum Likelihood Estimation*. Newbury Park, CA, Sage.

Fahrmeir, L. and G. Tutz. 1994. *Multivariate Statistical Modeling Based on Generalized Linear Models*. Springer Series in Statistics. New York: Springer-Verlag.

Fienberg, S.E. (1980). *The Analysis of Cross-Classified Categorical Data*. (2nd ed.). Cambridge, MA: MIT Press.

Fu, V.K. (1998). sg88: Estimating generalized ordered logit models. *Stata Technical Bulletin* 44: 27-30.

Greene, W.H. (2000). *Econometric Analysis*, 4th Ed. New York, Prentice Hall.

Greene, W.H. (1995). *LIMDEP Version 7.0*. Bellport, NY: Econometric Software.

Greenwood, C., and Farewell, V. (1988). A Comparison of Regression Models for Ordinal Data in an Analysis of Transplanted-Kidney Function. *Canadian Journal of Statistics*, 16, 325-335.

Hausman, J.A., and McFadden, D. (1984). Specification Tests for the Multinomial Logit Model. *Econometrica*, 52, 1219-1240.

Hendrickx, J. (2000). sbe37: Special Restrictions in Multinomial Logistic Regression. *Stata Technical Bulletin* 56: 18-26.

Hoffman, S.D. and G.J. Duncan (1988). Multinomial and Conditional Logit Discrete-Choice Models in Demography. *Demography* 25: 415-427.

Kaufman, R.L. (1996). Comparing Effects in Dichotomous Logistic Regression: A Variety of Standardized Coefficients. *Social Science Quarterly* 77: 90-109.

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

King, G. (1989). *Unifying Political Methodology: The Likelihood Theory of Statistical Inference*. Cambridge, Cambridge University Press.

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

Long, J.S. and J. Freese (forthcoming). *Regression Models for Categorical Outcomes Using Stata*. College Station, TX, Stata Press.

Longford, N. T. (1995). Random Coefficient Models. *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. G. Arminger, C.C. Clogg and M.E. Sobel. New York, Plenum Press: 519-577.

Maddala, G.S. (1983). *Limited-dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

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

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*. (2nd ed.). New York: Chapman and Hall.

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

McFadden, D. (1989). A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration. *Econometrica*, 57, 995-1026.

McFadden, D., Tye, W., and Train, K. (1976). An Application of Diagnostic Tests for the Independence from Irrelevant Alternatives Property of the Multinomial Logit Model. *Transportation Research Board Record*, 637, 39-45.

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

Miller, P. W., & Volker, P. A. (1985). On the Determination of Occupational Attainment and Mobility. *The Journal of Human Resources*, 20, 197-213.

Mroz, T. A. (1987). The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions. *Econometrica*, 55, 765-799.

Powers, D.A. and Y. Xie (2000). *Statistical Methods for Categorical Data Analysis*. San Diego, CA, Academic Press.

Pregibon, D. (1981). Logistic Regression Diagnostics. *The Annals of Statistics*, 9, 705-724.

Pudney, S. (1989). *Modelling Individual Choice: The Econometrics of Corners, Kinks and Holes*. Oxford: Basil Blackwell.

SAS Institute, Inc. (1990). *SAS/STAT User's Guide. Version 6.* (4rd ed.). Cary, NC: SAS Institute.

Small, K. A., and Hsiao, C. (1985). Multinomial Logit Specification Tests. *International Economic Review*, 26, 619-627.

Theil, H. (1970). On the Estimation of Relationships Involving Qualitative Variables. American Journal of Sociology 76: 103-154.

Wolfe, R. and Gould, W. (1998). sg76: An Approximate Likelihood Ratio Test for Ordinal Regression Models. *Stata Technical Bulletin* 42.

Zhang, J., and Hoffman, S.D. (1993). Discrete-Choice Logit Models: Testing the IIA Property. *Sociological Methods and Research*, 22, 193-213.