SAGE researchmethods

Ordinal Regression Models

Foundation Entries

?

SAGE Research Methods Foundations

By: Richard A. Williams & Christopher Quiroz
Length: 10,000 Words
DOI: http://dx.doi.org/10.4135/9781526421036
Methods: Ordinal Regression Models
Online ISBN: 9781526421036
Disciplines: Anthropology, Business and Management, Criminology and Criminal Justice, Communication and Media Studies, Counseling and Psychotherapy, Economics, Education, Geography, Health, History, Marketing, Nursing, Political Science and International Relations, Psychology, Social Policy and Public Policy, Social Work, Sociology, Science, Technology, Computer Science, Engineering, Mathematics, Medicine
Access Date: October 5, 2020
Publishing Company: SAGE Publications Ltd
City: London

Abstract

Researchers often encounter ordinal measures that they wish to examine as dependent variables in their research—variables where the categories are ordered (running from high to low or low to high), but the distance between the categories is unknown. For example, respondents might be asked if they strongly disagree, disagree, agree, or strongly agree with a statement. Or, rather than give an exact value for their years of education, respondents might be asked whether they had no education, some grade school, grade school graduate, some high school, and so on. While it might be tempting to treat ordinal dependent variables as though they were continuous and use techniques like ordinary least squares regression, this can result in misleading estimates of independent variable effects and incorrect tests of statistical significance. Ordinal regression models are therefore preferred under these circumstances—but there are many ordinal models to choose from. This entry begins with a detailed discussion of perhaps the most popular choice, the ordered logit model (also called the proportional odds model). The discussion will cover when the model might be appropriate, the major assumptions of the model (and how they can be violated), and how to interpret model results. However, in many cases, other ordinal models and methods will be more powerful or appropriate. This entry therefore also discusses the ordered probit model, the generalized ordered logit model, interval regression, scoring methods, heterogeneous choice/location scale models, stereotype models, stage models, and the rank-ordered logit model—as well as briefly explains when and why each might be preferred.

Introduction

Researchers often estimate models where a continuous dependent variable, *Y*, is regressed on an independent variable, *X*. But suppose the observed *Y* is not continuous. For example, Income might be coded in categories such as <$1,000 = 1, $1,001–$10,000 = 2, $10,001–$30,000 = 3, $30,001–$60,000 = 4, $60,001, or higher = 5. Respondents may also be asked, Do you approve or disapprove of the president's health care plan? The options being 1 = *strongly disapprove*, 2 = *disapprove*, 3 = *approve*, 4 = *strongly approve*. Presumably, values of approval are not limited to four possible values. For example, respondents could express their approval on a 100-point scale if it were to be constructed. However, when the possible responses for approval have been condensed into four choices, the respondents must decide which of the few available options best reflects their feelings.

For such variables, which are also known as limited dependent variables (Long, 1997), we know the categories are ordered (running from high to low or low to high), but the distance between the categories is unknown (Long & Freese, 2014). That is, unlike continuous variables, the distance between values 1 and 2 need not be the same as the distance between values 2 and 3.

The choice of an appropriate statistical technique is heavily affected by the measurement of the dependent variable. When the dependent variable is continuous (e.g., income, or age), ordinary least squares (OLS)

regression is often appropriate. Another instance in which the choice of statistical technique is more clear occurs when the dependent variable is binary (e.g., yes/no, employed/unemployed). In this type of model, the assumptions of OLS regression are violated, and methods like logistic or probit regression are often preferred.

When the dependent variable is ordinal however, the choice of statistical strategies may not be so clear. As Scott Menard (2002) notes, some researchers will treat the variable as continuous and use OLS regression. This may be fine if the variable has several categories but can be problematic if the variable has few categories and/or if the spacing between categories is clearly not consistent. As Richard D. McKelvey and William Zavoina (1975; see also Winship & Mare, 1984) point out, when OLS regression is used with ordinal dependent variables, parameter estimates can be biased and misleading and tests of statistical significance can be inaccurate. Others will use techniques like multinomial logistic regression (MLOGIT) which totally ignore that the categories are ordered. Such techniques tend to be less parsimonious and harder to interpret because they ignore the useful information that might be contained in the ordering of the categories.

Ordinal regression models are therefore often preferred. They avoid the problems of treating ordinal variables as though they are continuous while at the same time still taking advantage of the knowledge that the categories are ordered.

There is more than one ordinal regression model. This entry first focuses on one of the most popular models, commonly called by such names as the ordered logit model (OLOGIT), the proportional odds model, the cumulative logit model, the parallel lines model, or the parallel regressions model. The discussion will cover when the model might be appropriate (and when it might not be), the major assumptions of the model, and how to interpret model results.

This entry then discusses a few of the most popular alternatives: the ordered probit model, the generalized OLOGIT (GOLOGIT), and interval regression. These models will sometimes be preferable to the ordered logit. Still other less common but potentially useful alternatives—scoring methods, heterogeneous choice/location scale models, stereotype models, stage models, and the rank-ordered logit—are also briefly mentioned.

While there will be some key equations in this entry, overall the approach will be relatively nonmathematical and intuitive (for a more technical presentation, see Long, 1997; Long & Freese, 2014; Powers & Xie, 2008; or Hardin & Hilbe, 2012).

The OLOGIT or Proportional Odds Model

Model Basics

Ordinal regression techniques, in particular the OLOGIT, can be motivated in various ways. One approach is to say that there is an observed ordinal variable, Y. Y, in turn, is a function of another continuous, unmeasured latent variable Y^*, whose values determine what the observed ordinal variable Y equals. J. Scott Long and Jeremy Freese (2014) write the model as

Note that there are no intercept terms in the model. Instead, the continuous latent variable *Y** has various threshold points, which are represented using the Greek letter kappa (κ). *Our value on the observed variable* Y *depends on whether or not we have crossed a particular threshold.* For example, when there are three possible responses for the observed variable *Y*,

Put another way, we can think of *Y* as being a collapsed version of *Y**, for example, *Y** can take on an infinite range of values which might then be collapsed into three categories of *Y*.

As Long (1997) points out, we can also motivate the OLOGIT by thinking of it as a nonlinear probability model, that is, we predict the probability of the observed *Y* being a 1, a 2, and so on based on the values of the observed *X* variables. We do not have to rely on the notion of an underlying *Y**, and some prefer not to. The statistical procedures are the same either way.

With observed variables, the metric has to be set in some way, for example, we can measure income in dollars or in thousands of dollars. This is just as true with unobserved variables like *Y**. Typically, this is done by assuming that the error term has a standard logistic distribution and hence has a residual variance of $\pi^2/3$ (about 3.29). This choice may seem peculiar, and there are other ways to set the metric of the latent variable, but this approach tends to work well in practice and has nice mathematical properties that make it easy to compute other quantities of interest.

The K βs and the M-1 κs are parameters that need to be estimated. Once we have done so, using the corresponding sample estimates for each case we compute

Z is our best estimate of the underlying *Y**, given the *X* values that were observed. It is similar to the y-hat

that can be estimated using OLS regression. Because Y^* also has an error term, this estimate may be too high or too low for any particular case.

Because the error term has a standard logistic distribution, we can use the estimated $M - 1$ cutoff terms to estimate the probability that the observed Y will take on a particular value. For the OLOGIT, the formulas are

In the case of $M = 3$, these equations simplify to

As these formulas make clear, using the estimated value of the underlying latent variable and the assumed logistic distribution of the disturbance term, the OLOGIT can be used to estimate the probability that the unobserved variable Y^* falls within the various threshold limits.

Interpreting Results

Here, two empirical examples are presented to illustrate the key features of the OLOGIT. The first uses a very small data set, only 25 cases, making it possible to examine the interpretation of the OLOGIT for individual cases. The second data set is much larger, over 2,000 cases, and will show how more complicated hypotheses and model assumptions can be tested. All analyses were done with Stata 16, but several other statistical packages could have been used as well.

On January 28, 1986, the space shuttle *Challenger* exploded only 73 s after launch, killing all seven crew

members. There was intense public interest in the flight because the crew included Christa McAuliffe, who had been chosen from out of more than 11,000 applicants to become the first teacher in space. In *Statistics With Stata (Updated for Version 7)*, Lawrence C. Hamilton (2001) presents a fascinating example that shows that the disaster might have been averted had NASA officials heeded the warning signs. His analysis is replicated and extended here. Data covers the first 25 flights of the U.S. space shuttle. Table 1 lists the variables that were measured for each flight.

Table 1. Shuttle data variables.

DISTRESS	The number of "thermal distress incidents" in which hot gas damaged the joint seals of a flight's booster rockets. Damage to the joint seals helped lead to the *Challenger* disaster. This is the dependent variable and can be thought of as the observed indicator of the underlying riskiness of the flight. It is coded 1 = *None*, 2 = *1 or 2*, and 3 = *3 or more*.
TEMP	The calculated joint temperature at launch time. Temperature depends largely on weather. Colder temperatures cause the rubber O-rings sealing the booster rocket joints to become less flexible and hence more likely to have problems.
DATE	Date measured in days elapsed since January 1, 1960 (an arbitrary starting point). The rationale for this variable is that undesirable changes in the shuttle program, and aging hardware may have caused launches to become more risky across time.

Table 2 contains excerpts from the data. For each selected flight the observed variable values are given, as well at the estimates of the underlying latent variable and the predicted probability that the flight would experience three or more distress incidents.

Table 2. Shuttle data, selected flights.

flight	**DISTRESS**	**TEMP**	**DATE**	***Z* (Computed)**	**P (*Y* = 3)**
1	none	66	7,772	**14.09598**	**1.753**
2	1 or 2	70	7,986	**14.10568**	**1.770**
3	none	69	8,116	**14.70623**	**3.180**
13	none	78	9,044	**16.19526**	**12.708**
23	1 or 2	76	9,461	**17.91227**	**44.769**
24	3 plus	58	9,508	**21.18747**	**95.543**

flight	DISTRESS	TEMP	DATE	*Z* (Computed)	P (*Y* = 3)
25 (Challenger)	**MISSING**	**31**	**9,524**	**25.92117**	**99.959**

Table 3 shows the results when the OLOGIT is used and DISTRESS is regressed on DATE and TEMP. *Challenger* and one flight with missing data are excluded, yielding an *N* of 23 cases.

Table 3. Ordered logit model for shuttle data.

A tabular representation of ordered logistic regression.
Here is how to interpret the results. The likelihood-ratio χ^2 for the model (sometimes referred to as L^2) is like the global F statistic in an OLS regression. It tests the hypothesis that at least one of the independent variables in the model has a nonzero effect. When the null hypothesis is true (all independent variables have zero effects), the model χ^2 statistic has a χ^2 distribution with degrees of freedom (*df*) equal to the number of variable coefficients that are estimated. In this case, its value is 12.32 with 2 *df*. This is highly significant; the *p* value of .0021 tells us that, if the null hypothesis of all effects equaling zero is true, if we drew 10,000 samples this same way, we would expect only 21 of them to have coefficient estimates that differed this much from 0. This tells us that DATE and/or TEMP has a statistically significant effect on the number of thermal distress incidents.

Ordinary least squares regression offers an R^2 statistic that measures the strength of the association between the dependent and independent variables. Several pseudo-R^2 statistics have been proposed for logit and OLOGITs. One of the most popular is McFadden's R^2, which is used here. The pseudo R^2 ranges between 0 and 1; the bigger the value, the stronger the relationship is. For this model, its value is .247. (Formulas for various pseudo R^2 can be found in Long & Freese, 2014; Allison (2013) discusses the pros and cons of various measures.)

How do we interpret the coefficients themselves? As is the case with OLS regression and many other methods, the signs and statistical significance of the coefficients provide a basic way of interpreting results. The positive coefficient for DATE means that the likelihood of having more distress incidents did increase with time. Similarly, the negative coefficient for TEMP implies that colder temperatures increased the likelihood of having more distress incidents. The standard errors, *z* values, *p* values, and confidence intervals indicate that coefficients of this magnitude were unlikely to arise because of chance factors alone (e.g., because of drawing an atypical sample).

The cut points or threshold parameters may seem unusual to those not familiar with ordinal models. Indeed, they have little intuitive appeal and researchers usually spend little or no time on them when discussing results. But, they do provide important information for computing other quantities of interest; in this case, we will soon see how they can be used to estimate the riskiness of each flight. Here, the estimated values of 16.4281 and 18.1223 tell us the following. Since there are three possible values for *Y* (i.e., M = 3), the values for observed *Y* are estimated to be

As already noted, looking at the signs and statistical significance of coefficients provides us means for interpreting them: Are effects positive or negative and do they significantly differ from zero? But, by themselves, signs and statistical significance give little feel for the practical significance of the findings. Lower temperatures increased the riskiness of a flight, but by how much? In practice, what exactly does a −.173 coefficient for TEMP mean?

There are several ways to make the results more tangible. Many find it useful to think in terms of the odds of an event occurring, and how such odds are expected to change as *X* changes. When *Y* has only two possible values—for example, 1 = *success*, 0 = *failure*—the odds can be expressed as

---

For example, if the probability of success is .6 for a case, the probability of failure is .4, and the odds of success are 1.5. Conversely, if the probability of success is .25, the probability of failure is .75, and the odds of success are .25/.75 = 1/3.

Ordinal variables can have more than two outcomes, so multiple measures of odds can be computed. If the categories are 1, 2, 3,…, *M*, the odds formulas can be written as

---

For example, if there is a .2 probability that y = 1, there is a .80 probability that y will be greater than 1, so the odds of getting a value greater than 1 are .8/.2 = 4. Or, if there is a .70 probability that y will equal 1, 2, or 3, the probability of getting a value greater than 3 is .3, and the odds of getting a value greater than 3 are .3/.7 = .429.

Now, suppose a one-unit increase in *X* multiplies the odds of achieving a higher valued outcome by 1.25,

that is, makes the odds 25% greater. If the odds were 1 before the increase, they would be 1.25 after, which means that the probability of having a higher value would now be 1.25/(1 + 1.25) = 55.6%. If the odds of getting a higher value before *X* increased were 2, after the increase they would be 2.5 and the probability of a higher valued outcome would be 2.5/(1 + 2.5) = 71.4%.

In the case of ordinal regression, we are interested in how the odds of getting higher values on the ordinal variable *Y* are affected by changes in the *X*s. One of the reasons the OLOGIT is so popular is because the effect of *X* on the odds is easily calculated. To estimate the effects of increases in *X* on the odds of having a higher value on *Y*, we exponentiate the coefficients, that is, compute e^{bx}. The resulting quantities are then referred to as odds ratios because they are the odds of having a higher value on *Y after* the increase in *X* divided by the odds *before* the increase. As in the previous examples, this ratio will be the same regardless of what the odds were before *X* increased. Equivalently, if we want to see the percentage change in the odds after a one unit change in *X*, we can use the formula (Long & Freese, 2014)

Table 4 shows the results when the coefficients are exponentiated.

Table 4. Odds ratios for shuttle data.

A tabular representation of ordered logistic regression with coefficients exponentiated.
For TEMP, the odds ratio is exp(−.173) = .841. This means that, with each 1 degree increase in temperature, the odds of obtaining a higher valued outcome on the observed y get multiplied by .841; or, equivalently, decline by 100*exp(−.173) – 1) = 15.9%. For DATE, each additional day increases the odds of a higher outcome by 0.3% (which may not seem like much, but remember, the launches took place over several hundred days).

While many like odds ratios—they are somewhat more tangible and intuitive than the original coefficients—other approaches can provide more or alternative insights. It often helps to plug in some hypothetical or real data values to get a better feel for the coefficients' meaning. For example, with shuttle Flight 13, the temperature was 78 °F on launch date and date equaled 9044. Hence, for Flight 13, the estimated value of the underlying latent variable is

Note that this value is slightly less than the lowest threshold estimate of 16.4281. But, this is just an estimate of the underlying value of *Y**, and the true value of the unobserved variable may be higher or lower. This uncertainly is reflected in the formulas that predict the probability for each of the observed values of *Y* given the observed values of the *X*s. For Flight 13, we can therefore next compute

Hence, for Flight 13, which occurred more than a year earlier than *Challenger* under much warmer conditions, the most likely outcome (55.79%) was that there would be no damage to the booster joints. However, there was still more than a 40% chance Flight 13 would experience one or more distress incidents. In fact, Flight 13 did not have any problems.

Now, consider shuttle Flight 25, *Challenger*. Remember, *Challenger's* own data were not used when calculating these parameters. Hence, it would have been possible for a NASA official to use these numbers on launch day to predict the likelihood of a problem. On *Challenger*'s launch date, DATE equaled 9,524, and the temperature at launch time was 31° F (the previous coldest launch had been at 53° F). Hence, for *Challenger*,

This value is much higher than the upper threshold estimate of 18.1223 presented by the OLOGIT. Using the formulas presented earlier and the threshold estimates, we can now compute the probabilities of *Challenger* falling into each of the three different distress categories:

Hence, based on the experience from the previous 24 flights, there was virtually no chance that *Challenger* would experience no damage to its joint seals. Indeed, it was a virtual certainty (99.96%) that *Challenger* would experience 3 or more distress incidents.

Admittedly, ordinal regression models may not have been widely known back in 1986. But, if we run OLS regression instead, the predicted value for *Challenger* is 4.63, which is not a legitimate value for *Y* as it is

currently coded, but is consistent with the finding that launching on that day was very risky. Indeed, one engineer working on *Challenger*, Bob Ebeling, did try to stop the launch—but his warnings were not heeded (Berkes, 2016).

In this example, we computed the expected probabilities for two of the actual flights. But, we could just as easily have used hypothetical values—for example, we could have calculated the risk of *Challenger* if it had waited 10 more days and the temperature was 75° F (the estimated probability of 3 or more distress incidents would have only been 53.76% in that case). Stata and many other statistical packages make such calculations easy. Long and Freese (2014) and Williams (2012) give many examples of how predicting the likelihood of events using real or hypothetical values can make the substantive meaning and practical significance of results clearer.

Model Comparisons

The model χ^2, also called L^2, tests whether all the variables in the model have zero effects. However, researchers are often interested in testing hypotheses about subsets of variables. For example, a researcher might want to compare a model that has X1, X2, and X3 included, with a model that has the same three variables plus X4 and X5. We refer to the first model as the *constrained* model because, by not including X4 and X5, we in effect constrain their effects to equal 0. For example, X1, X2, and X3 might be demographic variables, and we might want to see whether attitudinal measures X4 and X5 tell us anything more than the demographic variables do. The *unconstrained* model is the model that allows X4 and X5 to have nonzero effects.

In logistic regression, we can do this via χ^2 contrasts. The simplest formula is

When the null hypothesis is true—the effects of the additional set of variables are all zero—the difference between the model χ^2s of the constrained and unconstrained models has a χ^2 distribution with *df* equal to the number of constraints. A large L^2 value suggests that at least one of the added variables has an effect that differs from zero.

The shuttle data are too small and have too few variables to demonstrate these principles, so here a new example is presented to show how this works in practice. The European Social Survey (ESS) is a cross-national study that has been conducted every 2 years across Europe since 2001. For this example, the 2012 ESS survey for Great Britain (ESS Round 6: European Social Survey Round 6 Data (2012)) is used. The study has 2,286 respondents, of which 2,159 (94.4%) have complete data for the variables used in this analysis. Although cases have unequal probabilities of selection, weighting had little effect on the results so to simplify the presentation they were not used. Williams (2016) shows how models can be estimated when data are weighted.

Respondents were asked the extent to which they agreed or disagreed with the following statement: “Gay men and lesbians should be free to live their own life as they wish.” The possible responses were 1 = *strongly disagree*, 2 = *disagree*, 3 = *neither agree nor disagree*, 4 = *agree*, and 5 = *strongly agree*. We use this as our response variable, HMSFREE. Thus, the higher the reported value, the more supportive the person is of gay and lesbian rights.

The explanatory variables are the responses to the following questions. In some cases, the original coding has been modified or reversed to make interpretation easier.

- AGEDECADE—Age of respondent (in decades, e.g., a value of 3.4 means 34 years old)
- FEMALE—Gender of respondent (1 = Female, 0 = Male)
- LIFEWORSE—“For most people in this country life is getting worse rather than better” (coded 1 = *strongly disagree* to 5 = *strongly agree*)
- HINCFEL—“Which of the descriptions on this card comes closest to how you feel about your household’s income nowadays?” (1 = *living comfortably on present income*, 2 = *coping on present income*, 3 = *finding it difficult on present income*, 4 = *finding it very difficult on present income*)
- FEELECON—“On the whole how satisfied are you with the present state of the economy in this country?” (11 point scale where 0 = *extremely satisfied* and 10 = *extremely dissatisfied*).

Age and gender are demographic variables. We might reasonably expect that age is related to attitudes about gays and lesbians because of changing attitudes across time. Gender might also be related whether women tend to be more tolerant than men of different lifestyles. HINCFEL and LIFEWORSE measure different aspects of satisfaction with one’s life. The theoretical argument for including them as explanatory variables is less clear. Dissatisfaction with one’s life could lead to less tolerance toward others, but not necessarily. FEELECON measures satisfaction with the entire economy but not necessarily satisfaction with one’s own life.

Therefore, we are interested in estimating two models. The first model includes only the demographic variables, while the second model adds the attitudinal variables. The two models are presented in Tables 5 and 6, along with an explanation of how they can be compared.

Table 5. Constrained model.

A tabular representation of ordered logistic regression in the constrained model.
Not surprisingly, older people are more likely to disagree that gays should be able to live their lives the way they want. Conversely, women are more supportive. The model χ^2 and the z values for the two demographic variables are all highly significant.

Table 6. Unconstrained model.

A tabular representation of ordered logistic regression in the unconstrained model.
The difference between the two model χ^2s is 164.71 – 137.25 = 27.46 with 3 *df*. This value is highly significant,

indicating that at least one of the attitudinal measures has a statistically significant effect. For LIFEWORSE and HINCFEL, less satisfaction is negatively associated with support for gay and lesbian rights, but neither effect is significant at the .05 level. For FEELECON, higher dissatisfaction levels are positively related to support for gay rights.

However, the use of χ^2 statistics as goodness of fit measures has sometimes been criticized. When sample sizes are large, it is much easier to accept (or at least harder to reject) more complex models because the χ^2 test statistics are designed to detect any departure between a model and observed data. That is, adding more terms to a model will always improve the fit, but with a large sample, it becomes harder to distinguish a "real" improvement in fit from a substantively trivial one. Likelihood-ratio tests therefore often lead to the rejection of acceptable models, and models become less parsimonious than they need to be.

Therefore, as many have noted, including J. Scott Long (1997) and Adrian E. Raftery (1995), information measures—in particular the Bayesian information criterion (BIC) and the Akaike information criterion (AIC)—have become increasingly popular. The BIC and AIC statistics are appropriate for many types of statistical methods (e.g., OLS regression), and are not just limited to logistic regression. The basic idea is to compare the relative plausibility of two models rather than to find the absolute deviation of observed data from a particular model. Unlike many measures, the information measures have penalties for including variables that do little to improve fit. Particularly with large samples, the information measures can lead to more parsimonious but adequate models. There are different formulas for these measures (see Long & Freese, 2014, for a discussion). It really does not matter which we use, so long as we are consistent when making comparisons between models.

When comparing two models, the model with the *smaller* BIC value is preferred. The same is true with AIC. How much one model is preferred over the other depends on the magnitude of the difference. For BIC, Raftery (1995) proposed the guidelines presented in Table 7:

Table 7. BIC guidelines.

Absolute difference	Evidence
0–2	Weak
2–6	Positive
6–10	Strong
>10	Very strong

Returning to our earlier constrained and unconstrained models, the BIC and AIC statistics, along with the χ^2 contrast presented earlier, are shown in Table 8.

Table 8. BIC, AIC, and likelihood-ratio tests.

A tabular representation of the likelihood-ratio test.
The BIC statistic does favor the unconstrained model, but, by Raftery's criteria, the difference between the model BIC values (4,855.055 − 4,850.623 = 4.432) provides only positive support for the unconstrained model, not strong or very strong. The AIC statistic also supports the unconstrained model.

Often, as in this case, the Likelihood-ratio χ^2 contrast and the BIC and AIC statistics all support the same model. The measures do not always agree though, and researchers will then have to decide which model they think is most defensible. Further, even though in this case all three measures preferred the unconstrained model that does not mean that all the variables that were added in the unconstrained model should have been added. The z values for the individual variables suggest that FEELECON should be in the model, but the statistical case for including LIFEWORSE and HINCFEL is shakier.

Testing Model Assumptions

Regardless of how reasonable any of the models presented so far may be, none of them should be accepted without further testing. The OLOGIT or proportional odds model makes certain assumptions about the data. If these assumptions are not met, the use of the model may be inappropriate. The assumptions of the model are explained in this section, and then ways for testing the assumptions are presented.

Williams (2016) presents hypothetical examples that illustrate what the proportional odds assumption is and when the assumptions are and are not violated. The first example presents an ideal (and probably never realized) situation, whereas the second example is typical of what is often encountered in practice.

Table 9. Hypothetical example—Perfect proportional odds/parallel lines.

A tabular representation of Perfect Proportional Odds or Parallel Lines.
In Table 9, looking at the column labeled 1 versus 2, 3, 4, we see that men are 3 times as likely to be in one of the higher categories as they are to be in the lowest category, so the odds for men are 3 (i.e., 750/250). Women, on the other hand, are 9 times as likely to be in one of the higher categories, so the odds for women are 9 (i.e., 900/100). Ergo, the ratio of the odds for women to men (i.e., the odds ratio) is 9/3 = 3.

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

Finally, the odds ratio is again 3 for the 1, 2, 3 versus 4 column. The Brant test (to be explained shortly) says the data meet the proportional odds assumption perfectly.

Table 10. Hypothetical example—Proportional odds violated.

A tabular representation of proportional odds violated.

Table 10 presents a second example. In this case, women are again clearly more likely to agree than men are; and yet, the assumptions of the OLOGIT are not met. Gender has its greatest effect at the lowest levels of attitudes—that is, as the odds ratio of 3 indicates, women are much less likely to strongly disagree than men are. But other differences are smaller—that is, in the 1 and 2 versus 3 and 4 cumulative logit, the odds ratio is only 1.5, and in the last cumulative logit, 1, 2, 3 versus 4, the odds ratio is only 1.28. The odds for women being in a higher category are consistently greater than the odds for men (i.e., women are more likely to agree than men are). But, because the odds ratios are not the same across the different regressions, the Brant test is highly significant (40.29 with 2 *df*). Thus, even though women do hold more favorable attitudes than do men, the assumptions of the proportional odds model are not met. There is an ordinal relationship between gender and attitudes, but it is not the kind of ordinal relationship that meets the assumptions of the proportional odds model.

There are several ways to test the proportional odds or parallel lines assumption of the OLOGIT. The most common is the Brant test (Brant, 1990; see Long, 1997, for details on how the test is computed). In Stata, the Brant test can be calculated using Long and Freese's (2014) brant command, which is part of their SPost suite of commands. For the unconstrained model with the ESS, the Brant test (Table 11) suggests that the model assumptions are violated:

Table 11. Brant test of parallel regression/proportional odds assumption.

A tabular representation of brant test of parallel regression.

The Brant test for all the variables, 43.90 with 15 *df*, is highly significant. At least one variable in the model violates the assumptions of the proportional odds model. However, Brant tests can also be done on individual variables. The tests reveal that only HINCFEL clearly violates the model assumptions. While the Brant test for AGEDECADE is significant at the .05 level, Williams (2006) argues that, since multiple variables are being tested, more stringent α levels should be used (e.g., .01), before deciding that any given variable violates proportional odds.

The detail option (see Table 12) for Long and Freese's brant command clarifies why the OLOGIT is also sometimes called the parallel lines model, the parallel regressions model, or the cumulative logit model.

Table 12. Estimated coefficients from cumulative logits.

A tabular representation of coefficients from cumulative logits.

In the cumulative logits, the ordinal variable is dichotomized. First it is Category 1 versus all higher categories, then Categories 1 and 2 versus all higher categories, and so on. If the assumptions of the OLOGIT are met, the coefficients (other than the constants) should be the same for each logistic regression—that is, the regression lines will be parallel, differing only in their intercepts. To the extent that the coefficients are not identical, the assumptions of the OLOGIT are violated. Because of sampling variability we never expect the proportional odds assumption to hold perfectly in a data set, but the Brant test tells us whether the violations

of assumptions are too large just to attribute to chance factors alone. Visual inspection suggests that the coefficients for HINCFEL vary greatly across models.

What are the implications of these violations of assumptions? Some would suggest that, when assumptions are violated, a multinomial logit model (MLOGIT) should be used instead. Table 13 shows what happens when an MLOGIT model is estimated:

Table 13. MLOGIT.

A tabular representation of multinomial logistic regression.
The MLOGIT model is much less parsimonious. The OLOGIT estimated five coefficients for five variables; MLOGIT estimates 20. Interpretation is certainly possible, but it is far more complicated than it is with the OLOGIT. Does the OLOGIT model really need to be totally abandoned, even when as few as one or two variables are problematic? The GOLOGIT model, discussed shortly, suggests that a less extreme approach may be possible.

Alternative Ordinal Models

The OLOGIT or proportional odds model may be the most popular ordinal regression model. But, it is not the only one. Depending the circumstances, some models may work just as well or better. Some of the most common alternatives are discussed next.

Ordered Probit

The ordered probit model is very similar to the ordered logit, except the error term is assumed to have a normal distribution with mean 0 and variance 1, that is, N(0, 1). In practice, the ordered logit and ordered probit models almost always lead to the same substantive conclusions. Some prefer the ordered logit because they like using the exponentiated coefficients or odds ratios to discuss effects, and the ordered probit model does not have that nice mathematical property. In most cases, though, the choice between logit and probit is based on whatever the most common practice is within a discipline.

GOLOGIT

When the assumptions of the OLOGIT are violated, some authors recommend that the MLOGIT model be used instead. The MLOGIT model makes no assumptions about the ordering of a variable; indeed, categories could be randomly renumbered and the MLOGIT model would give the same results. However, since the MLOGIT model ignores all the information about the ordering of categories, it estimates many more parameters, making it less parsimonious and more difficult to interpret. Williams (2006, 2016) suggests that another alternative also be considered: the GOLOGIT model. (GOPROBIT models can also be estimated if a researcher prefers them; conclusions are usually the same either way.) The GOLOGIT model can relax the proportional odds assumption for those variables that violate it, while keeping the constraints on those

variables that do not violate it. It therefore avoids the use of a model whose assumptions are violated (OLOGIT) while also avoiding the use of a model that is much less parsimonious (MLOGIT) than it needs to be.

The GOLOGIT model (Williams, 2006) can be written as

---

where *M* is the number of categories of the ordinal dependent variable.

When *M* = 2, the GOLOGIT model is equivalent to the binary logistic regression model. The proportional odds or parallel lines model is also a special case of the GOLOGIT model. The parallel lines model can be written as

---

The formulas for the parallel lines model and GOLOGIT model are the same, except that in the parallel lines model, the betas (but not the alphas) are the same for all values of *j*.

As noted previously, a key problem with the parallel lines model is that its assumptions are often violated, while common solutions like MLOGIT often go too far in the other direction, estimating far more parameters than is really necessary. Another special case of the GOLOGIT model overcomes these limitations. In the *partial proportional odds model*, some of the beta coefficients can be the same for all values of j, while others can differ. For example, in the following, the betas for X1 and X2 are the same for all values of j, but the betas for X3 are free to differ.

Consider again the example from the ESS. The assumptions of the OLOGIT were violated, but only one variable, HINCFEL, was clearly problematic. Table 14 shows the parameter estimates when a GOLOGIT or partial proportional odds model is used instead.

Table 14. GOLOGIT model for ESS data.

A tabular representation of generalized ordered logit estimates.
Like MLOGIT, there are *M* − 1 panels. But, the interpretation is very different. GOLOGIT is like running a series of logistic regressions, where the ordinal variable has been collapsed into a dichotomy. In the first

category, it is Category 1 versus Categories 2, 3, 4, and 5. In the second panel, it is Categories 1 and 2 versus 3, 4, and 5; then 1, 2, and 3 versus 4 and 5; and finally 1, 2, 3, and 4 versus 5. In each panel a positive coefficient means that increases in *X* make it more likely that a respondent will have one of the *higher* values for *Y*, while negative coefficients for *X* mean that increases in *X* make it more likely the subject will be in the *current category of* Y *or a lower one*.

At first glance, the GOLOGIT model might not appear to be very parsimonious compared to MLOGIT; but, other than HINCFEL, the coefficients are the same for each variable across panels. Hence, only 8 unique beta coefficients need to be examined, just 3 more than OLOGIT, and 12 less than the 20 coefficients that would be produced by MLOGIT.

Because the repetition of identical parameters is potentially confusing, Williams (2016) suggests alternate, more parsimonious ways of presenting GOLOGIT results. In Table 15, only one set of coefficients is presented for explanatory variables that meet the proportional odds assumption, while $M - 1$ coefficients are presented for those that do not. The overall *p* value is based on a test of the joint significance of all coefficients for the variable that are in the model.

Table 15. OLOGIT and partial proportional odds models for gays and lesbians should be free to live their lives as they want.

	Model 1: Proportional odds		**Model 2: Partial proportional odds**				
Explanatory variables	***P* value**	**Coef.**	**Overall *p* value**	***SD* vs D, N, A, SA**	***SD*, D vs N, A, SA**	***SD*, D, N vs A, SA**	***SD*, D, N, A vs SA**
Feelings about household income	.064	−.100	.000	−.957	−.387	−.217	−.037
Life is getting worse	.369	−.046	.3633	−.047			
Age (in decades)	.000	−.248	.000	−.251			
Gender (1= *female*, 0 = *male*)	.000	.388	.000	.387			
Satisfaction with state of economy	.000	.112	.000	.113			

This model is only slightly more difficult to interpret than the earlier parallel lines model, and it provides insights that were previously obscured. Effects of the constrained variables can be interpreted much as before. Older people are less supportive of gay rights, females are more supportive, and those expressing

dissatisfaction in different areas of their personal life also tend to be less supportive.

For HINCFEL, the differences from before are largely a matter of degree. All the coefficients are negative, but they get smaller in magnitude across each panel. Those with high levels of dissatisfaction are less supportive of gay rights, and are especially likely to express strong disapproval. Furthermore, the effect of HINCFEL is highly significant in the GOLOGIT model whereas it was not in the OLOGIT model. Hence, if we had only estimated an OLOGIT, not only would we have misestimated the effects of HINCFEL, we might have even erroneously concluded that it did not have any effect at all.

To sum up, with the GOLOGIT or partial proportional odds model, the effects of the variables that meet the parallel lines assumption are easily interpretable (we interpret them the same way as we do in OLOGIT). For other variables, an examination of the pattern of coefficients reveals insights that would be obscured or distorted if a parallel lines model were estimated instead. An MLOGIT analysis might lead to similar conclusions as GOLOGIT, but there would be many more parameters to look at, and the increased number of parameters could cause some estimated effects to become statistically insignificant.

Williams (2006) outlines procedures besides Brant that are more flexible for identifying which variables violate the proportional odds assumption. Williams (2006) also suggests different criteria for when a GOLOGIT model should be used. When relatively few variables violate the proportional odds assumption, a partial proportional odds model can avoid violating the assumptions of the OLOGIT while at the same maintaining most of OLOGIT's advantages with regard to ease of interpretation. If several variables violate proportional odds, however, a GOLOGIT model provides little parsimony and researchers may prefer to use the better known MLOGIT model or some other ordinal alternative. Williams (2016) also suggests several ways that the patterns of coefficients can yield substantive insights that might be missed by an OLOGIT model. In this case, the coefficients for HINCFEL differed in both their magnitude across panels, and also in their statistical significance. In other cases, the signs for a variable may actually switch from being positive to negative. Such a pattern might suggest, for example, that women take less extreme positions, high or low, than do men. Important relationships might be missed or obscured if only an OLOGIT is used.

Interval Regression

We earlier gave the example where Income might be coded in categories like ≤$1,000 = 1, $1,001–$10,000 = 2, $10,001–$30,000 = 3, $30,001–$60,000 = 4, $60,001, or higher = 5. Or, rather than give an exact value for their years of education, respondents might be asked if they had no education, some grade school, grade school graduate, some high school, and so on. Such variables are common in research. Rather than give the exact value of their income (or education, or years employed), respondents are asked to tell what interval they fall into. For example, someone whose income was $13,782 would code themselves as a 3. Note that the lower and upper bounds (≤ $1,000 and > $60,001) are not given. These can be treated as negative infinity and positive infinity, although the real values will usually fall into a much smaller range. Interval regression programs (e.g., intreg in Stata) typically ask the user to specify what the lower and upper bounds are for the interval a respondent falls into.

Here, two examples, one real and one hypothetical, are used to illustrate how interval regression works. StataCorp (2019) provides the first example. Women were asked via a questionnaire to indicate a category for their yearly income from employment. The categories were less than $5,000, $5,001–$10,000, … , $25,001–$30,000, $30,001–$40,000, $40,001–$50,000, and more than $50,000. To use Stata's intreg, the user must create two variables, wage1 and wage2, containing the lower and upper endpoints of the wage categories. The dependent variables can be thought of as measuring income in thousands of dollars, but instead of having the exact value for income only the interval in which it falls is known. Other variables in the model include NEV_MAR (0 = *has been married*, 1 = *never married*), RURAL (0 = *urban resident*, 1 = *rural resident*, SCHOOL (years of schooling), TENURE (job tenure, in years), AGE (age in years), and AGESQUARED (AGE * AGE). Table 16 shows the results.

Table 16. Interval regression model for women's yearly income.

A tabular representation of interval regression for women's annual income.

As Stata Corp (2019, p. 1064) points out, "Because the conditional mean modeled by interval regression is linear, the coefficients are interpreted the same way they are in ordinary least-squares regression." Therefore, the coefficients in interval regression are very easy to interpret. The results indicate that, on average, those who have never been married make $208 less a year (but the effect is not statistically significant). Rural residents average $3,043 a year less than nonrural residents. Each additional year of schooling is worth $1,335 a year more, and each year in the job is worth another $800. The effects of age are less straightforward, because the squared term makes the relation between age and income curvilinear. Effects can easily be computed for specific values of age, however.

While the results are reasonable, it is not clear how accurate they are. Do a few intervals provide a good substitute for exact values? Therefore, a hypothetical data set can be constructed for a second example where we know what the true parameter values are. The data are constructed so that y is a continuous variable that ranges from about −70 to 88. It is normally distributed. All 1,000 cases have a unique value for y. ycat is a collapsed, ordinal version of y. Table 17 shows how the collapsing was done.

Table 17. Hypothetical continuous Y collapsed into 5 intervals.

A tabular representation of Y collapsed into five intervals.

Table 18 illustrates the results of an interval regression, where t is regressed on x1, x2, and x3:

Table 18. Interval regression model with hypothetical collapsed data.

A tabular representation of interval regression.

Again, a nice feature of interval regression, as opposed to other ordinal methods, is that we can interpret parameters the same way we do the parameters from an OLS regression. There is no need to compute odds ratios or predicted probabilities like with other methods. For example, in this case, a one unit increase in x1 is expected to produce, on average, a 1.22 unit increase in y.

Table 19 shows the results when OLS regression is used on the original, noncollapsed y. The closer these are to the interval regression estimates, the better interval regression is working.

Table 19. OLS regression model with hypothetical noncollapsed data.

A tabular representation of OLS regression model.
In this particular example, interval regression does remarkably well. Its coefficients, standard errors, so on are very similar to those produced by OLS regression on the uncollapsed y variable. That is, even though the collapsed y (which only has 5 possible values) loses much of the information contained in the original y (which has 1,000 different values), we still reach very similar conclusions about the effects of the x variables on y.

Of course, if we knew the exact values of y, we would not be using interval regression. It is therefore suggested that the results from the interval regression model (Stata Corp, 2019) be compared with the results of an ordered probit model. In this case (see Table 20), the interval regression model produces almost the exact same model χ^2 and log likelihood as does the ordered probit model and also has very similar z values for the individual coefficients.

Table 20. Ordered probit model with hypothetical data.

A tabular representation of ordered probit regression model.
But, the coefficients from interval regression are much easier to interpret. If, on the other hand, the ordinal probit model had fit much better than the interval regression model did, the researcher might want to modify the interval regression model (e.g., take logs of the interval points) or use some other ordinal method instead.

It is important to remember that this example is “rigged” in interval regression’s favor. Interval regression assumes that variables are normally distributed, and the hypothetical data set was constructed accordingly. You cannot always count on interval regression working this well, and researchers should check whether its results are consistent with those provided by other ordinal methods.

Other Ordinal Regression Models of Interest

Several other ordinal regression models may potentially be of interest to researchers and are briefly discussed here.

Scoring Methods

Daniel A. Powers and Yu Xie (2008) note that various scoring methods are sometimes used to assign values to ordinal variables. For example, if an interval ranged between 5 and 10, the midpoint of 7.5 might be used; if another interval ranged between 15 and 30 the midpoint value assigned would be 22.5. While often done, midpoints can be poor estimates of the true values; for example, Powers and Xie say that for a category like

“less than 12 years of schooling” a value such as 5.5 would likely greatly underestimate years of schooling. There is also the problem of how to score an interval that does not contain an upper bound (e.g., greater than 50). Powers and Xie also discuss more complicated scoring schemes which use normal transformations or require the use of auxiliary information. A good scoring scheme may require a lot of knowledge of the topic and measures and statistical sophistication by the researcher.

Heterogeneous Choice or Location Scale Models

Both Allison (1999) and Williams (2009, 2010) note that OLS regression assumes that error terms are homoscedastic—for example, the error variances for men are equal to the error variances for women. If the assumption is violated—error variances are heteroskedastic—OLS estimates of variable coefficients remain unbiased, but the standard error terms and significance tests will be distorted. However, in ordered logit and probit models, the consequences of heteroscedasticity can be much greater. Coefficient estimates can be biased and cross-group comparisons in particular can be misleading. Allison (1999) gives an example where apparent discrimination against women in the tenure process may be an artifact of failing to control for differences in residual variability between men and women. Williams (2009, 2010) argues that heterogeneous choice models (also called location-scale models) can address the problem. J. S. Long and S. A. Mustillo (2018) suggest a different approach using predictions and marginal effects which they say avoids making what may be questionable assumptions.

Stereotype Logistic Model (SLM)

The SLM, also called the stereotype ordered regression model (Anderson, 1984; Long & Freese, 2014), has also been proposed as a way to deal with violations of the proportional odds assumption. These models can be helpful when the relevance of the ordering of categories is unclear. For example, there might be two or three underlying latent variables that give rise to the observed y. These different dimensions can be estimated and categories can even be reordered if deemed appropriate.

Stage Models

Andrew S. Fullerton and Jun Xu (2016; see also Fullerton, 2009) have outlined several models—cumulative, stage, and adjacent—and shown how they can be modified to relax the parallel regressions assumption. A stage (also called continuation-ratio) model might be appropriate for a dependent variable where respondents go through a series of steps. For example, rather than give their exact number of years of education, respondents might be asked whether they had no education, some grade school, graduated from grade school, some high school, and so on. People can go on to higher stages but cannot go back to lower ones (e.g., one can go on to get more education but cannot lose the education one already has; whereas with attitudes, change is possible in either direction). Depending on the nature of the dependent variable and the goals of the analysis, these may be preferable to the other models that have been discussed.

Rank-Ordered Logit

Sometimes respondents may be asked to do several rankings. For example, the top five job applicants might be ranked from 1 to 5. Or, subjects might be given a list, and asked to indicate which item is most important to them, which is the second most important, then the third, and so on. Rank-ordered logit regressions (Long & Freese, 2014) assess how important different attributes are in determining ratings—for example, how much impact does the education, years of job experience, gender, race, and other characteristics of job applicants have on how they are ranked? Further, rank-ordered logit models can assess how characteristics of the rater affect how they rate—for example, are women less influenced by a candidate's gender than men are?

Conclusion

There are many statistical techniques available when the dependent variable is ordinal. The ordered logit and ordered probit models may be the most popular. But, when their assumptions are violated, other techniques, such as GOLOGITs, may be preferable. In still other cases, such as when options are ranked or when upper and lower boundaries for categories are clearly stated, other ordinal regression techniques may be more powerful or informative. Researchers have many options for analyzing ordinal dependent variables, and they should think carefully about which best meets the specific needs of their data and topic.

References

Allison, P. D. (1999). Comparing logit and probit coefficients across groups. *Sociological Methods & Research*, 28, 186–208.

Allison, P. D. (2013, February13). *What's the best R-squared for logistic regression?* Retrieved August 7, 2019, from https://statisticalhorizons.com/r2logistic

Anderson, J. A. (1984). Regression and ordered categorical variables (with discussion). *Journal of the Royal Statistical Society, Series B*, 46, 1–30.

Berkes, H. (2016, March21). *Challenger engineer who warned of shuttle disaster dies. National Public Radio*. Retrieved August 7, 2019, from https://www.npr.org/sections/thetwo-way/2016/03/21/470870426/challenger-engineer-who-warned-of-shuttle-disaster-dies

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

ESS Round 6: European Social Survey Round 6 Data (2012). *Data file edition 2.1. Norwegian social science data services, Norway—Data archive and distributor of ESS data*. Retrieved May 27, 2015, from http://www.europeansocialsurvey.org/

Fullerton, A. (2009). A conceptual framework for ordered logistic regression models. *Sociological Methods & Research*, 38, 306–347.

Fullerton, A. S., & **Xu, J.** (2016). *Ordered regression models: Parallel, partial, and non-parallel alternatives*. Boca Raton, FL: Chapman & Hall/CRC Press.

Hamilton, L. C. (2001). *Statistics with Stata (Updated for Version 7)*. Belmont, CA: Duxbury Press.

Hardin, J. W., & **Hilbe, J. M.** (2012). *Generalized linear models and extensions* (3rd ed.). College Station, TX: Stata Press.

Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: SAGE.

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

Long, J. S., & **Mustillo, S. A.** (2018). Using predictions and marginal effects to compare groups in regression models for binary outcomes. *Sociological Methods & Research*. doi:10.1177/0049124118799374

McKelvey, R. D., & **Zavoina, W.** (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4, 103–120.

Menard, S. (2002). *Applied logistic regression analysis* (3rd ed.). Thousand Oaks, CA: SAGE.

Powers, D. A., & **Xie, Y.** (2008). *Statistical methods for categorical data analysis* (2nd ed.). London, England: Emerald.

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

StataCorp LLC. (2019). *Stata base reference manual release 16*. College Station, TX: Stata Press. Retrieved August 7, 2019, from https://www.stata.com/manuals/r.pdf

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

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

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

Williams, R. (2012). Using the margins command to estimate and interpret adjusted predictions and marginal effects. *The Stata Journal*, 12, 308–331. Retrieved from http://www.stata-journal.com/article.html?article=st0260

Williams, R. (2016). Understanding and interpreting generalized ordered logit models. *Journal of Mathematical Sociology*, 40, 7–20.

Winship, C., & **Mare, R. D.** (1984). Regression models with ordinal variables. *American Sociological Review*, 49, 512–525.