10 Ordinal Logistic Regression

Contents

Introduction

In this chapter, the standard logistic model is extended to handle outcome variables that have more than two ordered categories. When the categories of the outcome variable have a natural order, ordinal logistic regression may be appropriate.

The mathematical form of one type of ordinal logistic regression model, the proportional odds model, and its interpretation are developed. The formulas for the odds ratio and confidence intervals are derived, and techniques for testing hypotheses and assessing the statistical significance of independent variables are shown.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

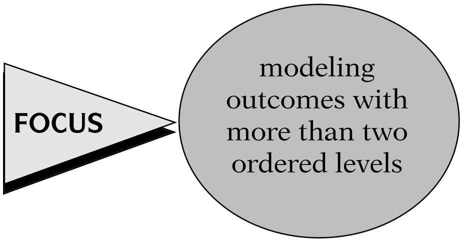
Objectives

Upon completing this chapter, the learner should be able to:

1. State or recognize when the use of ordinal logistic regression may be appropriate.
2. State or recognize the proportional odds assumption.
3. State or recognize the proportional odds model.
4. Given a printout of the results of a proportional odds model:
   a. state the formula and compute the odds ratio;
   b. state the formula and compute a confidence interval for the odds ratio;
   c. test hypotheses about the model parameters using the likelihood ratio test or the Wald test, stating the null hypothesis and the distribution of the test statistic with the corresponding degrees of freedom under the null hypothesis.

Presentation

I. Overview



This presentation and the presentation in Chapter 9 describe approaches for extending the standard logistic regression model to accommodate a disease, or outcome, variable that has more than two categories. The focus of this presentation is on modeling outcomes with more than two *ordered* categories. We describe the **form** and key **characteristics** of one model for such outcome variables: ordinal logistic regression using the proportional odds model.

Ordinal: levels have natural ordering

EXAMPLE

Tumor grade:

- well differentiated
- moderately differentiated
- poorly differentiated

Ordinal variables have a natural ordering among the levels. An example is cancer tumor grade, ranging from well differentiated to moderately differentiated to poorly differentiated tumors.

Ordinal outcome ⇒ Polytomous model or Ordinal model

An ordinal outcome variable with three or more categories can be modeled with a polytomous model, as discussed in Chapter 9, but can also be modeled using ordinal logistic regression, provided that certain assumptions are met.

Ordinal model takes into account order of outcome levels

Ordinal logistic regression, unlike polytomous regression, takes into account any inherent ordering of the levels in the disease or outcome variable, thus making fuller use of the ordinal information.

II. Ordinal Logistic Regression: The Proportional Odds Model

Proportional Odds Model / Cumulative Logit Model

The ordinal logistic model that we shall develop is called the proportional odds or cumulative logit model.

Illustration

0	1	2	3	4

0	1	2	3	4

0	1	2	3	4

0	1	2	3	4

0	1	2	3	4

But, cannot allow

0	4	1	2	3

To illustrate the proportional odds model, assume we have an outcome variable with five categories and consider the four possible ways to divide the five categories into two collapsed categories preserving the natural order.

We could compare category 0 to categories 1 through 4, or categories 0 and 1 to categories 2 through 4, or categories 0 through 2 to categories 3 and 4, or, finally, categories 0 through 3 to category 4. However, we could not combine categories 0 and 4 for comparison with categories 1, 2, and 3, since that would disrupt the natural ordering from 0 through 4.

For G categories $\Rightarrow$ G–1 ways to dichotomize outcome:

$D \geqslant 1$ vs. $D < 1$;
$D \geqslant 2$ vs. $D < 2$, ...,
$D \geqslant G–1$ vs. $D < G–1$

$$\text{odds } (D \geqslant g) = \frac{\mathrm{P}(D \geq g)}{\mathrm{P}(D < g)}$$
where $g = 1, 2, 3, \ldots, G–1$

More generally, if an ordinal outcome variable D has G categories ($D = 0, 1, 2, \ldots, G–1$), then there are G–1 ways to dichotomize the outcome: ($D \geqslant 1$ vs. $D < 1$; $D \geqslant 2$ vs. $D < 2, \ldots, D \geqslant G–1$ vs. $D < G–1$). With this categorization of D, the odds that $D \geqslant g$ is equal to the probability of $D \geqslant g$ divided by the probability of $D < g$, where ($g = 1, 2, 3, \ldots, G–1$).

Proportional odds assumption

EXAMPLE

OR ($D \geqslant 1$) = OR ($D \geqslant 4$)
Comparing two exposure groups
e.g., E=1 vs. E=0
where

$$\mathrm{OR}_{(D\geq 1)} = \frac{\text{odds }[(D \geq 1) \mid E = 1]}{\text{odds }[(D \geq 1) \mid E = 0]}$$

$$\mathrm{OR}_{(D\geq 4)} = \frac{\text{odds }[(D \geq 4) \mid E = 1]}{\text{odds }[(D \geq 4) \mid E = 0]}$$

The proportional odds model makes an important assumption. Under this model, the odds ratio assessing the effect of an exposure variable for any of these comparisons will be the same regardless of where the cut-point is made. Suppose we have an outcome with five levels and one dichotomous exposure (E=1, E=0). Then, under the proportional odds assumption, the odds ratio that compares categories greater than or equal to 1 to less than 1 is the same as the odds ratio that compares categories greater than or equal to 4 to less than 4.

Same odds ratio regardless of where categories are dichotomized

In other words, the odds ratio is **invariant** to where the outcome categories are dichotomized.

Ordinal

Variable	Parameter
Intercept	$\alpha_1, \alpha_2, \ldots, \alpha_{G-1}$
X_1	β_1

This implies that if there are G outcome categories, there is only one parameter (β) for each of the predictors variables (e.g., β_1 for predictor X_1). However, there is still a separate intercept term (α_g) for each of the G–1 comparisons.

Polytomous

Variable	Parameter
Intercept	$\alpha_1, \alpha_2, \ldots, \alpha_{G-1}$
X_1	$\beta_{11}, \beta_{21}, \ldots, \beta_{(G-1)1}$

This contrasts with polytomous logistic regression, where there are G–1 parameters for each predictor variable, as well as a separate intercept for each of the G–1 comparisons.

Odds are *not* invariant

EXAMPLE

$\text{odds}(D \geqslant 1) \neq \text{odds}(D \geqslant 4)$

where, for $E = 0$,

$$\text{odds}(D \geqslant 1) = \frac{P(D \geq 1 \mid E=0)}{P(D < 1 \mid E=0)}$$

$$\text{odds}(D \geqslant 4) = \frac{P(D \geq 4 \mid E=0)}{P(D < 4 \mid E=0)}$$

but

$\text{OR}(D \geqslant 1) = \text{OR}(D \geqslant 4)$

The assumption of the invariance of the odds ratio regardless of cut-point is *not* the same as assuming that the **odds** for a given exposure pattern is invariant. Using our previous example, for a given exposure level E (e.g., E=0), the odds comparing categories greater than or equal to 1 to less than 1 does *not* equal the odds comparing categories greater than or equal to 4 to less than 4.

Proportional odds model:
***G* outcome levels and one predictor (*X*)**

$$P(D \geq g \mid X_1) = \frac{1}{1+\exp[-(\alpha_g + \beta_1 X_1)]}$$

where $g = 1, 2, \ldots, G$–1

We now present the form for the proportional odds model with an outcome (D) with G levels ($D = 0, 1, 2, \ldots, G$–1) and one independent variable (X_1). The probability that the disease outcome is in a category greater than or equal to g, given the exposure, is 1 over 1 plus e to the negative of the quantity α_g plus $\beta_1 X_1$.

$$1 - P(D \geq g \mid \mathbf{X}_1) = 1 - \frac{1}{1+\exp[-(\alpha_g + \beta_1 X_1)]}$$

$$= \frac{\exp[-(\alpha_g + \beta_1 X_1)]}{1+\exp[-(\alpha_g + \beta_1 X_1)]}$$

$$= P(D < g \mid X_1)$$

The probability that the disease outcome is in a category *less* than g is equal to 1 minus the probability that the disease outcome is greater than or equal to category g.

Equivalent model definition

$$\text{odds} = \frac{P(D \geq g \mid X_1)}{1 - P(D \geq g \mid X_1)} = \frac{P(D \geq g \mid X_1)}{P(D < g \mid X_1)}$$

$$= \frac{\dfrac{1}{1+\exp[-(\alpha_g + \beta_1 X_1)]}}{\dfrac{\exp[-(\alpha_g + \beta_1 X_1)]}{1+\exp[-(\alpha_g + \beta_1 X_1)]}} = \exp(\alpha_g + \beta_1 X_1)$$

The model can be defined equivalently in terms of the odds of an inequality. If we substitute the formula $P(D \geqslant g \mid X_1)$ into the expression for the odds and then perform some algebra (as shown on the left), we find that the *odds* is equal to e to the quantity α_g plus $\beta_1 X_1$.

Proportional odds model: $P(D \geqslant g \mid \mathbf{X})$ versus Standard logistic model: $P(D = g \mid \mathbf{X})$

The proportional odds model is written differently from the standard logistic model. The model is formulated as the probability of an inequality, that is, that the outcome D is greater than or equal to g.

Proportional odds model: β_1 versus Polytomous model β_{g1}

no g subscript   *g* subscript

The model also differs from the polytomous model in an important way. The beta is not subscripted by g. This is consistent with the proportional odds assumption that only one parameter is required for each independent variable.

Alternate model formulation:

key differences

$$\text{odds} = \frac{P(D^* \leq g \mid X_1)}{P(D^* > g \mid X_1)} = \exp(\alpha_g^* - \beta_1^* X_1)$$

where $g = 1, 2, 3, \ldots, G\text{–}1$
and $D^* = 1, 2, \ldots, G$

An alternate formulation of the proportional odds model is to define the model as the odds of D^* less than or equal to g given the exposure is equal to e to the quantity $\alpha_g^* - \beta_1^* X_1$, where $g = 1, 2, 3, \ldots, G-1$ and where $D^* = 1, 2, \ldots, G$. The two key differences with this formulation are the direction of the inequality ($D^* \leqslant g$) and the negative sign before the parameter β_1^*. In terms of the beta coefficients, these two key differences "cancel out" so that $\beta_1 = \beta_1^*$. Consequently, if the same data are fit for each formulation of the model, the same parameter estimates of beta would be obtained for each model. However, the intercepts for the two formulations differ as $\alpha_g = -\alpha_g^*$.

Comparing formulations

$$\beta_1 = \beta_1^*$$

$$\text{but } \alpha_g = -\alpha_g^*$$

Formulation affects computer output

- SAS: consistent with first
- SPSS and Stata: consistent with alternative formulation

We have presented two ways of parameterizing the model because different software packages can present slightly different output depending on the way the model is formulated. SAS software presents output consistent with the way we have formulated the model, whereas SPSS and Stata software present output consistent with the alternate formulation (see Appendix).

Advantage of ($D \geqslant g$):
consistent with formulations of standard logistic and polytomous models

$\Downarrow$

For 2-level outcome ($D = 0, 1$), all three reduce to same model.

An advantage to our formulation of the model (i.e., in terms of the odds of $D \geqslant g$) is that it is consistent with the way that the standard logistic model and polytomous logistic model are presented. In fact, for a two-level outcome (i.e., $D = 0, 1$), the standard logistic, polytomous, and ordinal models reduce to the same model. However, the alternative formulation is consistent with the way the model has historically often been presented (McCullagh, 1980). Many models can be parameterized in different ways. This need not be problematic as long as the investigator understands how the model is formulated and how to interpret its parameters.

EXAMPLE

Black/White Cancer Survival Study

$$\mathbf{E} = \text{RACE} \begin{cases} 0 & \text{if white} \\ 1 & \text{if black} \end{cases}$$

$$\mathbf{D} = \text{GRADE} \begin{cases} 0 & \text{if well differentiated} \\ 1 & \text{if moderately differentiated} \\ 2 & \text{if poorly differentiated} \end{cases}$$

Next, we present an example of the proportional odds model using data from the Black/White Cancer Survival Study (Hill et al., 1995). Suppose we are interested in assessing the effect of RACE on tumor grade among women with invasive endometrial cancer. RACE, the exposure variable, is coded 0 for white and 1 for black. The disease variable, tumor grade, is coded 0 for well-differentiated tumors, 1 for moderately differentiated tumors, and 2 for poorly differentiated tumors.

Ordinal: Coding of disease meaningful

Polytomous: Coding of disease arbitrary

Here, the coding of the disease variable reflects the ordinal nature of the outcome. For example, it is necessary that moderately differentiated tumors be coded between poorly differentiated and well-differentiated tumors. This contrasts with polytomous logistic regression, in which the order of the coding is not reflective of an underlying order in the outcome variable.

EXAMPLE *(continued)*

	White (0)	Black (1)
Well differentiated	104	26
Moderately differentiated	72	33
Poorly differentiated	31	22

The 3×2 table of the data is presented on the left.

A simple check of the proportional odds assumption:

In order to examine the proportional odds assumption, the table is collapsed to form two other tables.

	White	Black
Well + moderately differentiated	176	59
Poorly differentiated	31	22

$\widehat{OR} = 2.12$

The first table combines the well-differentiated and moderately differentiated levels. The odds ratio is 2.12.

	White	Black
Well differentiated	104	26
Moderately + poorly differentiated	103	55

$\widehat{OR} = 2.14$

The second table combines the moderately and poorly differentiated levels. The odds ratio for this data is 2.14.

Requirement: Collapsed ORs should be "close"

The odds ratios from the two collapsed tables are similar and thus provide evidence that the proportional odds assumption is not violated. It would be unusual for the collapsed odds ratios to match perfectly. The odds ratios do not have to be exactly equal; as long as they are "close," the proportional odds assumption may be considered reasonable.

	$E=0$	$E=1$
$D=0$	45	30
$D=1$	40	15
$D=2$	50	60

Here is a different 3×2 table. This table will be collapsed in a similar fashion as the previous one.

	$E=0$	$E=1$
$D=0+1$	85	45
$D=2$	50	60

$\widehat{\text{OR}} = 2.27$

	$E=0$	$E=1$
$D=0$	45	30
$D=1+2$	90	75

$\widehat{\text{OR}} = 1.25$

The two collapsed table are presented on the left. The odds ratios are 2.27 and 1.25. In this case, we would question whether the proportional odds assumption is appropriate, since one odds ratio is nearly twice the value of the other.

Statistical test of assumption: **Score test**
Compares ordinal versus polytomous models

Test statistic ~ χ^2 under H_0
with df = number of OR parameters tested

There is also a statistical test—a **Score test**—designed to evaluate whether a model constrained by the proportional odds assumption (i.e., an ordinal model) is significantly different from the corresponding model in which the odds ratio parameters are not constrained by the proportional odds assumption (i.e., a polytomous model). The test statistic is distributed approximately chi-square, with degrees of freedom equal to the number of odds ratio parameters being tested.

Alternate models for ordinal data

- continuation ratio
- partial proportional odds
- stereotype regression

If the proportional odds assumption is inappropriate, there are other ordinal logistic models that may be used that make alternative assumptions about the ordinal nature of the outcome. Examples include a continuation ratio model, a partial proportional odds model, and stereotype regression models. These models are beyond the scope of the current presentation. [See the review by Ananth and Kleinbaum (1997).]

III. Odds Ratios and Confidence Limits

ORs: same method as SLR to compute ORs.

After the proportional odds model is fit and the parameters estimated, the process for computing the odds ratio is the same as in standard logistic regression (SLR).

Special case: one independent variable
$X_1 = 1$ or $X_1 = 0$

$$\text{odds}(D \geq g) = \frac{P(D \geq g \mid X_1)}{P(D < g \mid X_1)} = \exp(\alpha_g + \beta_1 X_1)$$

We will first consider the special case where the exposure is the only independent variable and is coded 1 and 0. Recall that the odds comparing $D \geq g$ versus $D < g$ is e to the α_g plus β_1 times X_1. To assess the effect of the exposure on the outcome, we formulate the ratio of the odds of $D \geq g$ for comparing $X_1=1$ and $X_1=0$ (i.e., the odds ratio for $X_1=1$ vs. $X_1=0$).

$$\text{OR} = \frac{P(D \geq g \mid X_1 = 1) / P(D < g \mid X_1 = 1)}{P(D \geq g \mid X_1 = 0) / P(D < g \mid X_1 = 0)}$$

$$= \frac{\exp[\alpha_g + \beta_1(1)]}{\exp[\alpha_g + \beta_1(0)]} = \frac{\exp(\alpha_g + \beta_1)}{\exp(\alpha_g)}$$

$$= e^{\beta_1}$$

This is calculated, as shown on the left, as the odds that the disease outcome is greater than or equal to g if X_1 equals 1, divided by the odds that the disease outcome is greater than or equal to g if X_1 equals 0.

Substituting the expression for the odds in terms of the regression parameters, the odds ratio for $X_1=1$ versus $X_1=0$ in the comparison of disease levels $\geq g$ to levels $<$ g is then e to the β_1.

General case
(levels X_1^{**} and X_1^{*} of X_1)

$$\text{OR} = \frac{\exp(\alpha_g + \beta_1 X_1^{**})}{\exp(\alpha_g + \beta_1 X_1^{*})}$$

$$= \frac{\exp(\alpha_g)\ \exp(\beta_1 X_1^{**})}{\exp(\alpha_g)\ \exp(\beta_1 X_1^{*})}$$

$$= \exp[\beta_1(X_1^{**} - X_1^{*})]$$

To compare any two levels of the exposure variable, X_1^{**} and X_1^{*}, the odds ratio formula is e to the β_1 times the quantity X_1^{**} minus X_1^{*}.

CIs: same method as SLR to compute CIs

General case (levels X_1^{**} and X_1^{*} of X_1)

$$95\%\ \text{CI: } \exp\left[\hat{\beta}_i(X_1^{**} - X_1^{*}) \pm 1.96(X_1^{**} - X_1^{*})s_{\hat{\beta}_i}\right]$$

Confidence interval estimation is also analogous to standard logistic regression. The general large-sample formula for a 95% confidence interval, for any two levels of the independent variable (X_1^{**} and X_1^{*}), is shown on the left.

EXAMPLE

Black/White Cancer Survival Study

Test of proportional odds assumption:
H_0: assumption holds
Score statistic: $\chi^2 = 0.0008$, df=1, $P = 0.9779$.
Conclusion: fail to reject null

Returning to our tumor-grade example, the results for the model examining tumor grade and RACE are presented next. The results were obtained from running PROC LOGISTIC in SAS (see Appendix).

We first check the proportional odds assumption with a **Score test**. The test statistic, with one degree of freedom for the one odds ratio parameter being tested, was clearly not significant, with a P-value of 0.9779. We therefore fail to reject the null hypothesis (i.e., that the assumption holds) and can proceed to examine the model output.

EXAMPLE *(continued)*

Variable	Estimate	S.E.
Intercept 1	–1.7388	0.1765
Intercept 2	–0.0089	0.1368
RACE	0.7555	0.2466

$\widehat{\text{OR}} = \exp(0.7555) = 2.13$

With this ordinal model, there are two intercepts, one for each comparison, but there is only one estimated beta for the effect of RACE. The odds ratio for RACE is e to β_1. In our example, the odds ratio equals exp(0.7555) or 2.13.

Interpretation of OR

Black versus white women with endometrial cancer over twice as likely to have more severe tumor grade:

Since $\widehat{\text{OR}}\ (D \geq 2) = \widehat{\text{OR}}\ (D \geq 1) = 2.13$

The results indicate that for this sample of women with invasive endometrial cancer, black women were over twice (i.e., 2.13) as likely as white women to have tumors that were categorized as poorly differentiated versus moderately differentiated or well differentiated *and* over twice as likely as white women to have tumors classified as poorly differentiated or moderately differentiated versus well differentiated. To summarize, in this cohort, black women were over twice as likely to have a more severe grade of endometrial cancer compared with white women.

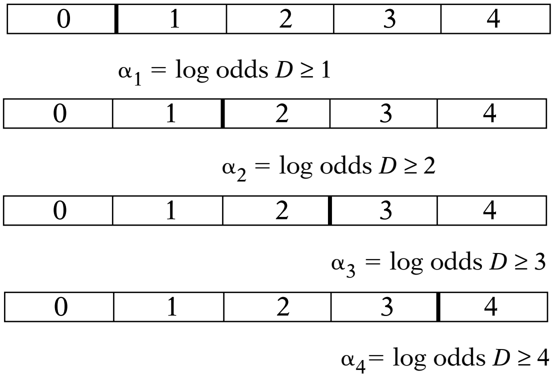
Interpretation of intercepts (α_g)

α_g = log odds of $D \geq g$ where all independent variables equal zero; $g = 1, 2, 3, \ldots, G-1$

$$\alpha_g > \alpha_{g+1}$$
$$\Downarrow$$
$$\alpha_1 > \alpha_2 > \cdots > \alpha_{G-1}$$

What is the interpretation of the intercept? The intercept α_g is the log odds of $D \geq g$ where all the independent variables are equal to zero. This is similar to the interpretation of the intercept for other logistic models except that, with the proportional odds model, we are modeling the log odds of several inequalities. This yields several intercepts, with each intercept corresponding to the log odds of a different inequality (depending on the value of g). Moreover, the log odds of $D \geq g$ is greater than the log odds of $D \geq (g+1)$ (assuming category g is nonzero). This means that $\alpha_1 > \alpha_2 \cdots > \alpha_{G-1}$.

Illustration



As the picture on the left illustrates, with five categories ($D = 0, 1, 2, 3, 4$), the log odds of $D \geq 1$ is greater than the log odds of $D \geq 2$, since for $D \geq 1$, the outcome can be in categories 1, 2, 3, or 4, whereas for $D \geq 2$, the outcome can only be in categories 2, 3, or 4. Thus, there is one more outcome category (category 1) contained in the first inequality. Similarly, the log odds of $D \geq 2$ is greater than the log odds of $D \geq 3$; and the log odds of $D \geq 3$ is greater than the log odds of $D \geq 4$.

EXAMPLE *(continued)*

95% confidence interval for OR

$$95\%\ \text{CI} = \exp[0.7555 \pm 1.96\ (0.2466)] = (1.31, 3.45)$$

Returning to our example, the 95% confidence interval for the OR for AGE is calculated as shown on the left.

Hypothesis testing

Likelihood ratio test or Wald test

H_0: $\beta_1 = 0$

Hypothesis testing about parameter estimates can be done using either the likelihood ratio test or the Wald test. The null hypothesis is that β_1 is equal to 0.

Wald test

$$Z = \frac{0.7555}{0.2466} = 3.06, \quad P = 0.002$$

In the tumor grade example, the P-value for the Wald test of the beta coefficient for RACE is 0.002, indicating that RACE is significantly associated with tumor grade at the 0.05 level.

IV. Extending the Ordinal Model

$$P(D \geq g \mid \mathbf{X}) = \frac{1}{1 + \exp[-(\alpha_g + \sum_{i=1}^{p} \beta_i X_i)]}$$

where $g = 1, 2, 3, \ldots, G-1$

Note: $P(D \geq 0|\mathbf{X}) = 1$

Expanding the model to add more independent variables is straightforward. The model with p independent variables is shown on the left.

$$\text{odds} = \frac{P(D \geq g \mid \mathbf{X})}{P(D < g \mid \mathbf{X})} = \exp(\alpha_g + \sum_{i=1}^{p} \beta_i X_j)$$

The *odds* for the outcome greater than or equal to level g is then e to the quantity α_g plus the summation the X_i for each of the p independent variable times its beta.

OR = $\exp(\beta_i)$, if X_i is coded (0, 1)

The odds ratio is calculated in the usual manner as e to the β_i, if X_i is coded 0 or 1. As in standard logistic regression, the use of multiple independent variables allows for the estimation of an odds ratio for one variable controlling for the effects of the other covariates in the model.

EXAMPLE

$$D = \text{GRADE} = \begin{cases} 0 & \text{if well differentiated} \\ 1 & \text{if moderately differentiated} \\ 2 & \text{if poorly differentiated} \end{cases}$$

$$X_1 = \text{RACE} = \begin{cases} 0 & \text{if white} \\ 1 & \text{if black} \end{cases}$$

$$X_2 = \text{ESTROGEN} = \begin{cases} 0 & \text{if never user} \\ 1 & \text{if ever user} \end{cases}$$

To illustrate, we return to our endometrial tumor grade example. Suppose we wish to consider the effects of estrogen use as well as RACE on GRADE. ESTROGEN is coded as 1 for ever user and 0 for never user.

The model now contains two predictor variables: X_1 = RACE and X_2 = ESTROGEN.

EXAMPLE *(continued)*

$$P(D \geq g \mid \mathbf{X}) = \frac{1}{1+\exp[-(\alpha_g + \beta_1 X_1 + \beta_2 X_2)]}$$

where X_1 = RACE (0, 1)
X_2 = ESTROGEN (0, 1)
g = 1, 2

The odds that the tumor grade is in a category greater than or equal to category 2 (i.e., poorly differentiated) versus in categories less than 2 (i.e., moderately or well differentiated) is e to the quantity α_2 plus the sum of $\beta_1 X_1$ plus $\beta_2 X_2$.

$$\text{odds} = \frac{P(D \geq 2 \mid \mathbf{X})}{P(D < 2 \mid \mathbf{X})} = \exp(\alpha_2 + \beta_1 X_1 + \beta_2 X_2)$$

different α's

same β's

$$\text{odds} = \frac{P(D \geq 1 \mid \mathbf{X})}{P(D < 1 \mid \mathbf{X})} = \exp(\alpha_1 + \beta_1 X_1 + \beta_2 X_2)$$

Similarly, the odds that the tumor grade is in a category greater than or equal to category 1 (i.e., moderately or poorly differentiated) versus in categories less than 1 (i.e., well differentiated) is e to the quantity α_1 plus the sum of $\beta_1 X_1$ plus $\beta_2 X_2$. Although the alphas are different, the betas are the same.

Test of proportional odds assumption

H_0: assumption holds
Score statistic: $\chi^2 = 0.9051$, 2 df, $P = 0.64$
Conclusion: fail to reject null

Before examining the model output, we first check the proportional odds assumption with a Score test. The test statistic has two degrees of freedom because we have two fewer parameters in the ordinal model compared to the corresponding polytomous model. The results are not statistically significant, with a *P*-value of 0.64. We therefore fail to reject the null hypothesis that the assumption holds and can proceed to examine the remainder of the model results.

Variable	Estimate	S.E.	Symbol
Intercept 1	–1.2744	0.2286	$\hat{\alpha}_2$
Intercept 2	0.5107	0.2147	$\hat{\alpha}_1$
RACE	0.4270	0.2720	$\hat{\beta}_1$
ESTROGEN	–0.7763	0.2493	$\hat{\beta}_2$

The output for the analysis is shown on the left. There is only one beta estimate for each of the two predictor variables in the model. Thus, there are a total of four parameters in the model, including the two intercepts.

EXAMPLE *(continued)*

Odds ratio

$\widehat{OR} = \exp \hat{\beta}_1 = \exp(0.4270) = 1.53$

The estimated odds ratio for the effect of RACE, controlling for the effect of ESTROGEN, is e to the $\hat{\beta}_1$, which equals e to the 0.4270 or 1.53.

95% confidence interval

$$95\%\ \text{CI} = \exp[0.4270 \pm 1.96\ (0.2720)] = (0.90,\ 2.61)$$

The 95% confidence interval for the odds ratio is e to the quantity $\hat{\beta}_1$ plus or minus 1.96 times the estimated standard error of the beta coefficient for RACE. In our two-predictor example, the standard error for RACE is 0.2720 and the 95% confidence interval is calculated as 0.90 to 2.61. The confidence interval contains one, the null value.

Wald test

H_0: $\beta_1 = 0$

$$Z = \frac{0.4270}{0.2720} = 1.57, \qquad P = 0.12$$

Conclusion: fail to reject H_0

If we perform the Wald test for the significance of $\hat{\beta}_1$, we find that it is not statistically significant in this two-predictor model (P=0.12). The addition of ESTROGEN to the model has resulted in a decrease in the estimated effect of RACE on tumor grade, suggesting that failure to control for ESTROGEN biases the effect of RACE away from the null.

V. Likelihood Function for Ordinal Model

$$\text{odds} = \frac{P}{1-P}$$

so solving for P,

$$P = \frac{\text{odds}}{\text{odds}+1} = \frac{1}{1+\left(\dfrac{1}{\text{odds}}\right)}$$

Next, we briefly discuss the development of the likelihood function for the proportional odds model. To formulate the likelihood, we need the probability of the observed outcome for each subject. An expression for these probabilities in terms of the model parameters can be obtained from the relationship P = odds/(odds+1), or the equivalent expression P = 1/[1+(1/odds)].

$P(D=g) = [P(D \geq g)] - [P(D \geq g+1)]$

For $g=2$

$P(D=2) = P(D \geq 2) - P(D \geq 3)$

In the proportional odds model, we model the probability of $D \geq g$. To obtain an expression for the probability of $D=g$, we can use the relationship that the probability $(D=g)$ is equal to the probability of $D \geq g$ minus the probability of $D \geq (g+1)$. For example, the probability that D equals 2 is equal to the probability that D is greater than or equal to 2 minus the probability that D is greater than or equal to 3. In this way we can use the model to obtain an expression for the probability an individual is in a specific outcome category for a given pattern of covariates ($\mathbf{X}$).

Use relationship to obtain probability individual is in given outcome category.

L is product of individual contributions.

$$\prod_{j=1}^{n} \prod_{G-0}^{G-1} P(D=g \mid \mathbf{X})^{y_{jg}}$$

where

$$y_{jg} = \begin{cases} 1 & \text{if the } j\text{th subject has } D=g \\ 0 & \text{if otherwise} \end{cases}$$

The likelihood (L) is then calculated in the same manner discussed previously in the section on polytomous regression—that is, by taking the product of the individual contributions.

VI. Ordinal Versus Multiple Standard Logistic Regressions

Proportional odds model: order of outcome considered.

Alternative: several logistic regression models

The proportional odds model takes into account the effect of an exposure on an ordered outcome and yields one odds ratio summarizing that effect across outcome levels. An alternative approach is to conduct a series of logistic regressions with different dichotomized outcome variables. A separate odds ratio for the effect of the exposure can be obtained for each of the logistic models.

Original variable: 0, 1, 2, 3

Recoded:
≥ 1 vs. <1, ≥ 2 vs. <2, and ≥ 3 vs. <3

For example, in a four-level outcome variable, coded as 0, 1, 2, and 3, we can define three new outcomes: greater than or equal to 1 versus less than 1, greater than or equal to 2 versus less than 2, and greater than or equal to 3 versus less than 3.

Three separate logistic regressions

Three sets of parameters

$\alpha_{\geq 1 \text{ vs. } <1}$, $\beta_{\geq 1 \text{ vs. } <1}$

$\alpha_{\geq 2 \text{ vs. } <2}$, $\beta_{\geq 2 \text{ vs. } <2}$

$\alpha_{\geq 3 \text{ vs. } <3}$, $\beta_{\geq 3 \text{ vs. } <3}$

With these three dichotomous outcomes, we can perform three separate logistic regressions. In total, these three regressions would yield three intercepts and three estimated beta coefficients for each independent variable in the model.

Logistic models (three parameters)	Proportional odds model (one parameter)
$\beta_{\geq 1 \text{ vs. } <1}$	
$\beta_{\geq 2 \text{ vs. } <2}$	β
$\beta_{\geq 3 \text{ vs. } <3}$	

If the proportional odds assumption is reasonable, then using the proportional odds model allows us to summarize the relationship between the outcome and each independent variable with one parameter instead of three.

Is the proportional odds assumption met?

- Crude OR's "close"?
  (No control of confounding)

The key question is whether or not he proportional odds assumption is met. There are several approaches to checking the assumption. Calculating and comparing the crude odds ratios is the simplest method, but this does not control for confounding by other variables in the model.

- Beta coefficients in separate logistic models similar?
  (Not a statistical test)

Is $\beta_{\geq 1 \text{ vs. } <1} \cong \beta_{\geq 2 \text{ vs. } <2} \cong \beta_{\geq 3 \text{ vs. } <3}$?

Running the separate (e.g., 3) logistic regressions allows the investigator to compare the corresponding odds ratio parameters for each model and assess the reasonableness of the proportional odds assumption in the presence of possible confounding variables. Comparing odds ratios in this manner is not a substitute for a statistical test, although it does provide the means to compare parameter estimates. For the four-level example, we would check whether the three coefficients for each independent variable are similar to each other.

- Score test provides a test of proportional odds assumption

  H_0: assumption holds

The Score test enables the investigator to perform a statistical test on the proportional odds assumption. With this test, the null hypothesis is that the proportional odds assumption holds. However, failure to reject the null hypothesis does not necessarily mean the proportional odds assumption is reasonable. It could be that there are not enough data to provide the statistical evidence to reject the null.

If assumption not met, may

- use polytomous logistic model
- use different ordinal model
- use separate logistic models

If the assumption does not appear to hold, one option for the researcher would be to use a polytomous logistic model. Another alternative would be to select an ordinal model other than the proportional odds model. A third option would be to use separate logistic models. The approach selected should depend on whether the assumptions underlying the specific model are met and on the type of inferences the investigator wishes to make.

SUMMARY

✓ Chapter 10: Ordinal Logistic Regression

This presentation is now complete. We have described a method of analysis, ordinal regression, for the situation where the outcome variable has more than two ordered categories. The proportional odds model was described in detail. This may be used if the proportional odds assumption is reasonable.

We suggest that you review the material covered here by reading the detailed outline that follows. Then do the practice exercises and test.

Chapter 11: Logistic Regression for Correlated Data: GEE

All of the models presented thus far have assumed that observations are statistically independent, (i.e., are not correlated). In the next chapter (Chapter 11), we consider one approach for dealing with the situation in which study outcomes are not independent.

Detailed Outline

I. Overview (page 304)

A. Focus: modeling outcomes with more than two levels.

B. Ordinal outcome variables.

II. Ordinal logistic regression: The proportional odds model (pages 304–310)

A. Ordinal outcome: variable categories have a natural order.

B. Proportional odds assumption: the odds ratio is invariant to where the outcome categories are dichotomized.

C. The form for the proportional odds model with one independent variable (X_1) for an outcome (D) with G levels ($D = 0, 1, 2, \ldots, G-1$) is

$$P(D \geq g \mid X_1) = \frac{1}{1+\exp[-(\alpha_g + \beta_1 X_1)]} \quad \text{where } g = 1, 2, \ldots, G-1$$

III. Odds ratios and confidence limits (pages 310–313)

A. Computation of the OR in ordinal regression is analogous to standard logistic regression, except that there is a single odds ratio for all comparisons.

B. The general formula for the odds ratio for any two levels of the predictor variable (X_1^{**} and X_1^{*}) is

$$\text{OR} = \exp\left[\beta_1 (X_1^{**} - X_1^{*})\right].$$

C. Confidence interval estimation is analogous to standard logistic regression.

D. The general large-sample formula for a 95% confidence interval for any two levels of the independent variable (X_1^{**} and X_1^{*}), is

$$\exp\left[\hat{\beta}_1 (X_1^{**} - X_1^{*}) \pm 1.96 (X_1^{**} - X_1^{*}) s_{\hat{\beta}_1}\right].$$

E. The likelihood ratio test is used to test hypotheses about the significance of the predictor variable(s).

i. there is one estimated coefficient for each predictor;

ii. the null hypothesis is that the beta coefficient (for a given predictor) is equal to zero;

iii. the test compares the log likelihood of the full model with the predictor(s) to that of the reduced model without the predictor(s).

F. The Wald test is analogous to standard logistic regression.

IV. Extending the ordinal model (pages 314–316)

A. The general form of the proportional odds model for G outcome categories and p independent variables is

$$1 - P(D \geq g \mid \mathbf{X}_1) = 1 - \frac{1}{1+\exp[-(\alpha_g + \beta_1 \mathrm{X}_1)]} \quad \text{where } g = 1, 2, \ldots, G-1$$

B. The calculation of the odds ratio, confidence intervals, and hypothesis testing using the likelihood ratio and Wald tests remain the same.

C. Interaction terms can be added and tested in a manner analogous to standard logistic regression.

V. Likelihood function for ordinal model (pages 316–317)

A. For an outcome variable with G categories, the likelihood function is

$$\prod_{j=1}^{n} \prod_{g=0}^{G-1} \mathrm{P}(D = g \mid \mathbf{X}^{y_{jg}})$$

where

$$y_{jg=} \begin{cases} 1 & \text{if the jth subject has } D=g \\ 0 & \text{if otherwise} \end{cases}$$

where n is the total number of subjects, $g = 0, 1, \ldots, G–1$ and $\mathrm{P}(D = g|\mathbf{X}) = [\mathrm{P}(D \geq g|\mathbf{X})] – [\mathrm{P}(D \geq g+1)|\mathbf{X})]$

VI. Ordinal versus multiple standard logistic regressions (pages 317–319)

A. Proportional odds model: order of outcome considered.

B. Alternative: several logistic regressions models

i. one for each cut-point dichotomizing the outcome categories;

ii. example: for an outcome with four categories (0, 1, 2, 3), we have three possible models.

C. If the proportional odds assumption is met, it allows the use of one parameter estimate for the effect of the predictor, rather than separate estimates from several standard logistic models.

D. To check if the proportional odds assumption is met:

i. evaluate whether the crude odds ratios are “close”;

ii. evaluate whether the odds ratios from the standard logistic models are similar:

a. provides control of confounding but is not a statistical test;

iii. perform a Score test of the proportional odds assumption.

E. If assumption is not met, can use a polytomous model, consider use of a different ordinal model, or use separate logistic regressions.

Practice Exercises

Suppose we are interested in assessing the association between tuberculosis and degree of viral suppression in HIV-infected individuals on antiretroviral therapy, who have been followed for 3 years in a hypothetical cohort study. The outcome, tuberculosis, is coded as none ($D=0$), latent ($D=1$), or active ($D=2$). Degree of viral suppression (VIRUS) is coded as undetectable (VIRUS=0) or detectable (VIRUS=1). Previous literature has shown that it is important to consider whether the individual has progressed to AIDS (no=0, yes=1) and is compliant with therapy (no=1, yes=0). In addition, AGE (continuous) and GENDER (female=0, male=1) are potential confounders. Also there may be interaction between progression to AIDS and COMPLIANCE with therapy (AIDSCOMP=AIDS × COMPLIANCE).

We decide to run a proportional odds logistic regression to analyze these data. Output from the ordinal regression is shown below. (The results are hypothetical.) The descending option was used, so Intercept 1 pertains to the comparison $D \geq 2$ to $D<2$ and Intercept 2 pertains to the comparison $D \geq 1$ to $D<1$.

Variable	Coefficient	S.E.
Intercept 1	–2.98	0.20
Intercept 2	–1.65	0.18
VIRUS	1.13	0.09
AIDS	0.82	0.08
COMPLIANCE	0.38	0.14
AGE	0.04	0.03
GENDER	0.35	0.19
AIDSCOMP	0.31	0.14

1. State the form of the ordinal model in terms of variables and unknown parameters.

2. For the above model, state the fitted model in terms of variables and estimated coefficients.

3. Compute the estimated odds ratio for a 25-year-old noncompliant male with a detectable viral load, who has progressed to AIDS, compared to a similar female. Consider the outcome comparison active or latent tuberculosis versus none ($D \geq 1$ vs. $D < 1$).

4. Compute the estimated odds ratio for a 38-year-old noncompliant male with a detectable viral load, who has progressed to AIDS, compared to a similar female. Consider the outcome comparison active tuberculosis versus latent or none ($D \geq 2$ vs. $D < 2$).

5. Estimate the odds of a compliant 20-year-old female, with an undetectable viral load and who has not progressed to AIDS, of having active tuberculosis ($D \geq 2$).

6. Estimate the odds of a compliant 20-year-old female, with an undetectable viral load and who has not progressed to AIDS, of having latent or active tuberculosis ($D \geq 1$).

7. Estimate the odds of a compliant 20-year-old male, with an undetectable viral load and who has not progressed to AIDS, of having latent or active tuberculosis ($D \geq 1$).

8. Estimate the odds ratio for noncompliance versus compliance. Consider the outcome comparison active tuberculosis versus latent or no tuberculosis ($D \geq 2$ vs. $D < 2$).

Test

True or False (Circle T or F)

T F 1. The disease categories absent, mild, moderate, and severe can be ordinal.

T F 2. In an ordinal logistic regression (using a proportional odds model) in which the outcome variable has five levels, there will be four intercepts.

T F 3. In an ordinal logistic regression in which the outcome variable has five levels, each independent variable will have four estimated coefficients.

T F 4. If the outcome D has seven levels (coded 1, 2, ..., 7), then $P(D{\geq}4)/P(D{<}4)$ is an example of an odds.

T F 5. If the outcome D has seven levels (coded 1, 2, ..., 7), an assumption of the proportional odds model is that $P(D{\geq}3)/P(D{<}3)$ is assumed equal to $P(D{\geq}5)/P(D{<}5)$.

T F 6. If the outcome D has seven levels (coded 1, 2, . . . , 7) and an exposure E has two levels (coded 0 and 1), then an assumption of the proportional odds model is that $[P(D{\geq}3|E{=}1)/P(D{<}3|E{=}1)]/[P(D{\geq}3|E{=}0)/P(D{<}3|E{=}0)]$ is assumed equal to $[P(D{\geq}5|E{=}1)/P(D{<}5|E{=}1)]/[P(D{\geq}5|E{=}0)/P(D{<}5|E{=}0)]$.

T F 7. If the outcome D has four categories coded $D{=}0, 1, 2, 3$, then the log odds of $D{\geq}2$ is greater the log odds of $D{\geq}1$.

T F 8. Suppose a four level outcome D coded $D{=}0, 1, 2, 3$, is recoded $D^{*}{=}$ 1, 2, 7, 29, then the choice of using D or D^{*} as the outcome in a proportional odds model has no effect on the parameter estimates as long as the order in the outcome is preserved.

9. Suppose the following proportional odds model is specified assessing the effects of AGE (continuous), GENDER (female=0, male=1), SMOKE (nonsmoker=0, smoker=1), and hypertension status (HPT) (no=0, yes=1) on four progressive stages of disease ($D{=}0$ for absent, $D{=}1$ for mild, $D{=}2$ for severe, and $D{=}3$ for critical).

$$\ln\frac{P(D \geq g \mid \mathbf{X})}{P(D < g \mid \mathbf{X})} = \alpha_g + \beta_1\text{AGE} + \beta_2\text{GENDER} + \beta_3\text{SMOKE} + \beta_4\text{HPT}$$

where $g = 1, 2, 3$

Use the model to obtain an expression for the odds of a severe or critical outcome ($D \geq 2$) for a 40-year-old male smoker without hypertension.

10. Use the model in Question 9 to obtain the odds ratio for the mild, severe, or critical stage of disease (i.e., $D \geq 1$)] comparing hypertensive smokers versus nonhypertensive nonsmokers, controlling for AGE and GENDER.

11. Use the model in Question 9 to obtain the odds ratio for critical disease only ($D \geq 3$) comparing hypertensive smokers versus nonhypertensive nonsmokers, controlling for AGE and GENDER. Compare this odds ratio to that obtained for Question 10.
12. Use the model in Question 9 to obtain the odds ratio for mild or no disease ($D < 2$) comparing hypertensive smokers versus nonhypertensive nonsmokers, controlling for AGE and GENDER.

Answers to Practice Exercises

1. Ordinal model

$$\ln\left[\frac{P(D \geq g \mid \mathbf{X})}{P(D < g \mid \mathbf{X})}\right] = \alpha_g + \beta_1\text{VIRUS} + \beta_2\text{AIDS} + \beta_3\text{COMPLIANCE} + \beta_4\text{AGE} + \beta_5\text{GENDER} + \beta_6\text{AIDSCOMP}$$

where $g = 1, 2$

2. Ordinal fitted model

$$\hat{\ln}\left[\frac{P(D \geq 2 \mid \mathbf{X})}{P(D < 2 \mid \mathbf{X})}\right] = -2.98 + 1.13\text{VIRUS} + 0.82\text{AIDS} + 0.38\text{COMPLIANCE} + 0.04\text{AGE} + 0.35\text{GENDER} + 0.31\text{AIDSCOMP}$$

$$\hat{\ln}\left[\frac{P(D \geq 1 \mid \mathbf{X})}{P(D < 1 \mid \mathbf{X})}\right] = -1.65 + 1.13\text{VIRUS} + 0.82\text{AIDS} + 0.38\text{COMPLIANCE} + 0.04\text{AGE} + 0.35\text{GENDER} + 0.31\text{AIDSCOMP}$$

3. $\widehat{OR} = \exp(0.35) = 1.42$
4. $\widehat{OR} = \exp(0.35) = 1.42$
5. Estimated odds = $\exp[-2.98 + 20(0.04)] = 0.11$
6. Estimated odds = $\exp[-1.65 + 20(0.04)] = 0.43$
7. Estimated odds = $\exp[-1.65 + 20(0.04) + 0.35] = 0.61$
8. Estimated odds ratios for noncompliant (COMPLIANCE=1) versus compliant (COMPLIANCE=0) subjects:
   for AIDS = 0: $\exp(0.38) = 1.46$
   for AIDS = 1: $\exp(0.38 + 0.31) = 1.99$