

Fall 10-20-2010

# Ordinal Regression Analysis: Fitting the Continuation Ratio Model to Educational Data Using Stata

Xing Liu

*Eastern Connecticut State University, liux@easternct.edu*

Follow this and additional works at: [http://digitalcommons.uconn.edu/nera\\_2010](http://digitalcommons.uconn.edu/nera_2010)

 Part of the [Education Commons](#)

---

## Recommended Citation

Liu, Xing, "Ordinal Regression Analysis: Fitting the Continuation Ratio Model to Educational Data Using Stata" (2010). *NERA Conference Proceedings 2010*. 35.  
[http://digitalcommons.uconn.edu/nera\\_2010/35](http://digitalcommons.uconn.edu/nera_2010/35)

# Ordinal Regression Analysis: Fitting the Continuation Ratio Model to Educational Data Using Stata

---

Xing Liu

Eastern Connecticut State University

## Abstract

Ordinal data are widely available to educational researchers. One of the most commonly used models to analyze ordinal data is the proportional odds (PO) model, which is also known as the cumulative odds model. However, when the research interest is focused on a particular category rather than at or below that category, given that an individual must pass through a lower category before achieving a higher level, the continuation ratio model (Fienberg, 1980; Hardin & Hilbe, 2007; Long & Freese, 2006) is a more appropriate choice than the proportional odds model. The purpose of this paper was to demonstrate the use of the continuation ratio (CR) model to analyze ordinal data in education using Stata, and compare the results of the CR model with the PO model. Ordinal regression analyses are based on a subset of data from the ELS (Educational Longitudinal Study): 2002, in which the ordinal outcome of students' mathematics proficiency was predicted from a set of students' classroom practices.

*Keywords:* Continuation Ratio Models, Proportional Odds Models, Mathematics Proficiency, Stata, Comparison.

# Ordinal Regression Analysis: Fitting the Continuation Ratio Model to Educational Data Using Stata

## Introduction

Ordinal data are widely available to educational researchers. For example, it is common to see student's SES ordered from low to high, a response to a survey item scaled from strongly disagree to strongly agree, children's proficiency in early reading scored from level 0 to 5, and students' educational proficiency level in a state test ranging from fail, pass to proficient. To model these ordinal data, one of the most commonly used models is the proportional odds (PO) model, which is also known as the cumulative odds model (Agresti, 1996, 2002, 2007; Armstrong & Sloan, 1989; Long, 1997, Long & Freese, 2006; McCullagh, 1980; McCullagh & Nelder, 1989; O'Connell, 2000, 2006; Powers & Xie, 2000). It is an extension of binary logistic regression models when the response variable has more than two ordinal categories. The proportional odds model is used to estimate the cumulative probability of being at or below a particular level of the response variable, or its complementary, the probability of being beyond a particular level. However, when the interest of research is focused on a particular category rather than at or below that category, given that an individual must pass through a lower category before achieving a higher level, the continuation ratio (CR) model (Fienberg, 1980; Hardin & Hilbe, 2007; Long & Freese, 2006) is a more appropriate choice than the proportional odds model. In particular, the CR model is more appealing than other models when analyzing educational attainment data (Allison, 1999). Under the No Child Left Behind Act (NCLB), when students' academic proficiency level is measured

annually or frequently using a mastery test, this model is extraordinarily useful in analyzing this type of data.

In the CR model, the ordinal categories represent successive stages or proficiency levels through which an individual can advance; for example, faculty ranks from assistant professor, to associate professor, and to full professor, or educational attainment from a high school diploma, a Bachelor's degree, and a Master's degree to a doctorate degree. Individuals must pass through lower stages or levels in order to reach a higher stage or level. For instance, a faculty member needs to be granted the assistant professor title before he/she achieves the full professor position. Similarly, a person needs to be awarded the lower level degree before he/she received the doctorate degree. This model estimates the odds of being in a certain category relative to the odds of being in that category or beyond. In terms of probability, this model estimates the probability of being in a category, given that an individual has been in that category or beyond. In addition, it estimates the conditional probability of being beyond a category given that person has attained that particular category, since these two conditional probabilities are complementary.

Although the PO model is commonly used, the CR model seems to be overlooked. The purpose of this article was to demonstrate the use of the continuation ratio (CR) model to predict mathematics proficiency of high school students using Stata, and compare the results of fitting the continuation ratio model and the proportional odds model. Ordinal regression analyses were based on the data from the ELS (Educational Longitudinal Study): 2002, in which the ordinal outcome of students' mathematics proficiency was predicted from a set of students' classroom activities, such as, reviewing

work from the previous day in math class, listening to teachers' lectures, copying notes from board, using books besides textbooks, doing problem solving in class, using general and graphing calculators, using computers, explaining work orally, and participating in student-led discussions.

## Theoretical Framework

### *General Logistic Regression Model and the Proportional Odds Model*

In a binary logistic regression model, the outcome variable has two levels, with 1 = experiencing the events, and 0 = not experiencing the events. This model estimates the log odds of the outcome, and thus the probability of success on a set of predictors. The logistic regression model can be defined as:

$$\ln(Y') = \text{logit} [\pi(\underline{x})] = \ln \left( \frac{\pi(\underline{x})}{1 - \pi(\underline{x})} \right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p. \quad (1)$$

In an ordinal logistic regression model, the outcome variable has more than two levels. It estimates the probability of an observation being at or beyond a specific outcome level, given a collection of explanatory variables. The ordinal logistic regression model can be expressed as a latent variable model (Agresti, 2002; Greene, 2003; Long, 1997, Long & Freese, 2006; Powers & Xie, 2000; Wooldridge & Jeffrey, 2001).

Assuming a latent variable,  $Y^*$  exists, we can define  $Y^*$  as a function of a set of predictor variables and a random error. Let  $Y^*$  be divided by some cut points (thresholds):  $\alpha_1, \alpha_2, \alpha_3 \dots \alpha_j$ , and  $\alpha_1 < \alpha_2 < \alpha_3 \dots < \alpha_j$ . The values of the observed ordinal variable,  $Y$ , fall within the regions divided by these cut points (thresholds). For example,  $Y = 0$ , if  $Y^* \leq \alpha_1$ .

Considering the observed mathematics proficiency level is the ordinal outcome,  $y$ , ranging from 0 to 5, we define:

$$y = \begin{cases} 0 & \text{if } y^* \leq \alpha_1 \\ 1 & \text{if } \alpha_1 < y^* \leq \alpha_2 \\ 2 & \text{if } \alpha_2 < y^* \leq \alpha_3 \\ 3 & \text{if } \alpha_3 < y^* \leq \alpha_4 \\ 4 & \text{if } \alpha_4 < y^* \leq \alpha_5 \\ 5 & \text{if } \alpha_5 < y^* \leq \infty \end{cases} \quad (2)$$

Therefore, we can predict the probability of a student achieving each proficiency level, and the cumulative probabilities as well.  $P(Y \leq j) = F(\alpha_j - \mathbf{x}\boldsymbol{\beta})$ , where  $j = 1, 2, \dots, J-1$ .

Since different software packages employ different parameterizations in estimating logit coefficients, the ordinal logistic regression model can be expressed in different forms (Liu, in press). In Stata, the ordinal logistic regression model assumes that the outcome variable is a latent variable. It is expressed in logit form as follows:

$$\ln(Y_j') = \text{logit} [\pi(x)] = \ln \left( \frac{\pi_j(x)}{1 - \pi_j(x)} \right) = \alpha_j + (-\beta_1 X_1 - \beta_2 X_2 - \dots - \beta_p X_p), \quad (3)$$

where  $\pi_j(\underline{x}) = \pi(Y \leq j | x_1, x_2, \dots, x_p)$ , which is the probability of being at or below category  $j$ , given a set of predictors.  $j = 1, 2, \dots, J-1$ .  $\alpha_j$  are the cut points, and  $\beta_1, \beta_2, \dots, \beta_p$  are logit coefficients. This is the form of the proportional odds (PO) model because the odds ratio of any predictor is assumed to be constant across all categories. This equal logit slope assumption can be assessed by the Brant test (Brant, 1990). Similar to logistic regression, in the proportional odds model we work with the logit, or the natural log of the odds. To estimate the  $\ln(\text{odds})$  of being at or below the  $j$ th category, the PO model can be rewritten as:

$$\text{logit} [\pi(Y \leq j | x_1, x_2, \dots, x_p)] = \ln \left( \frac{\pi(Y \leq j | x_1, x_2, \dots, x_p)}{\pi(Y > j | x_1, x_2, \dots, x_p)} \right) = \alpha_j + (-\beta_1 X_1 - \beta_2 X_2 - \dots - \beta_p X_p) \quad (4)$$

Thus, this model predicts cumulative logits across J -1 response categories. By transforming the cumulative logits, we can obtain the estimated cumulative odds as well as the cumulative probabilities being at or below the  $j^{\text{th}}$  category. Just as Stata, the ordinal logit model is also based on the latent continuous outcome variable for SPSS PLUM, it takes the same form.

However, SAS uses a different ordinal logit model for estimating the parameters from Stata. For SAS PROC LOGISTIC (the ascending option), the ordinal logit model has the following form:

$$\text{logit} [\pi(Y \leq j | x_1, x_2, \dots, x_p)] = \ln \left( \frac{\pi(Y \leq j | x_1, x_2, \dots, x_p)}{\pi(Y > j | x_1, x_2, \dots, x_p)} \right) = \alpha_j + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p; \quad (5)$$

Using SAS with the descending option, the ordinal logit model can be expressed as:

$$\text{logit} [\pi(Y \geq j | x_1, x_2, \dots, x_p)] = \ln \left( \frac{\pi(Y \geq j | x_1, x_2, \dots, x_p)}{\pi(Y < j | x_1, x_2, \dots, x_p)} \right) = \alpha_j + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p, \quad (6)$$

where in both equations  $\alpha_j$  are the intercepts, and  $\beta_1, \beta_2 \dots \beta_p$  are logit coefficients.

#### *The Continuation Ratio Model*

When estimating the conditional probability of being beyond a category given that individual has attained that particular category, i.e.,  $\pi(Y > j | Y \geq j)$ , the CO model can be expressed in this form (Agresti, 2007; Allison, 1999; O'Connell, 2006):

$$\ln \left( \frac{\pi(Y \geq j | x_1, x_2, \dots, x_p)}{\pi(Y = j | x_1, x_2, \dots, x_p)} \right) = \alpha_j + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \quad (7)$$



where  $\pi(Y \geq j | x_1, x_2, \dots, x_p)$  is the conditional probability of being beyond a category  $j$ , conditional on being in that category, given a set of predictors.  $j = 1, 2, \dots, J - 1$ .  $\alpha_j$  are the cut points, and  $\beta_1, \beta_2, \dots, \beta_p$  are logit coefficients. SAS follows this form in estimating the continuation ratio model with the PROC LOGISTIC command. Before the model is fitted, the data set needs to be restructured following a series of steps (Allison, 1999; O'Connell, 2006). First, separate sub-data sets need to be constructed with the binary outcome variable being beyond a category coded as 1 and 0 otherwise. Individuals who have not advanced to a particular proficiency level are dropped at each stage. If the ordinal dependent variable has  $j$  categories,  $J - 1$  sub-data sets should be created. Then, these data sets are combined into one data set with a new binary outcome variable with 1 = beyond a particular category. Finally, the CR model is fitted using the SAS PROC LOGISTICS with the descending option.

The CR models also estimates the odds of being in a particular category  $j$  relative to being that category or beyond. In this situation, the CR model can be formulated as (Ananth & Kleinbaum, 1997; Armstrong & Sloan, 1989; Fienberg, 1980; Long & Freese, 2006):

$$\ln \left( \frac{\pi(Y = j | x_1, x_2, \dots, x_p)}{\pi(Y \geq j | x_1, x_2, \dots, x_p)} \right) = \alpha_j + (-\beta_1 X_1 - \beta_2 X_2 - \dots - \beta_p X_p) \quad (8)$$

where  $\pi(Y = j | x_1, x_2, \dots, x_p)$  is the conditional probability of being in category  $j$ , conditional on being that category or beyond, given a set of predictors.  $j = 1, 2, \dots, J - 1$ .  $\alpha_j$  are the cut points, and  $\beta_1, \beta_2, \dots, \beta_p$  are logit coefficients. Different from SAS, Stata follows this form to fit the CR model, which is known as the forward CR model (Bender & Bender, 2000). Another distinctive difference is that Stata does not require data

restructuring before model fitting, which makes data analysis of the CR model much easier. The following analyses demonstrate how to fit the CR model using Stata.

## Methodology

### Sample

The data were from the Educational Longitudinal Study (ELS: 2002). The ELS: 2002 study, conducted by the National Center for Educational Statistics (NCES), was designed to provide longitudinal data regarding the transitions of 2002 high school sophomores to postsecondary school education and their future careers. In the 2002 base year of the study, more than 15,000 high school sophomores, from a national sample of 752 public and private high schools, participated in the study by taking cognitive tests and responding to surveys.

The outcome variable of interest was students' mathematics proficiency levels in high schools, which was an ordinal categorical variable with five levels (1 = students can do simple arithmetical operations on whole numbers; 2 = students can do simple operations with decimals, fractions, powers, and root; 3 = students can do simple problem solving; 4 = students can understand intermediate-level mathematical concepts and/or find multi-step solutions to word problems; and 5 = students can solve complex multiple-step word problems and/or understand advanced mathematical material) (Ingels, Pratt, Roger, Siegel, & Stutts, 2004, 2005). These five proficiency domains were hierarchically structured: mastery of higher proficiency level indicated mastery of all previous levels. Students needed to pass through the first four levels of proficiency before achieving the final fifth level. Those students who failed to pass through level 1 were assigned to level 0. Table 1 provides the frequency of six mathematics proficiency levels.

## Data Analysis

We began by fitting the continuation ratio model with a single explanatory variable using Stata OCRATIO command (Wolfe, 1998) with the link functions of *logit* and *cloglog*. Then, we fitted a proportional odds (PO) model. Finally, a full-model with all 11 explanatory variables was fitted. The *eform* option was used to estimate the odds ratios and corresponding standard errors and the confidence intervals. Stata OLOGIT command was used to fit the proportional odds models. The results from both the continuation ratio models and the proportional odds model were compared.

## Results

### *Continuation Ratio Model with a Single Explanatory Variable*

A continuation ratio model with a single predictor, gender was fitted first. Stata OCRATIO command with the logit function as default was used. Figure 1 displays the Stata output for the single predictor continuation ratio model. The log likelihood ratio Chi-Square test with 1 degree of freedom, , LR  $\chi^2_{(1)} = 38.90$ ,  $p < .001$ , indicating that the logit regression coefficient of the predictor, gender was statistically different from 0, so the full model with one predictor provides a better fit than the null model with no independent variables in predicting conditional probabilities for mathematics proficiency level. The Pseudo  $R^2 = .0008$ , which is the likelihood ratio  $R^2_L$ , suggested that the relationship between the response variable, mathematics proficiency, and the predictor, gender was small.

The estimated logit regression coefficient,  $\beta = .1416$ ,  $z = 6.23$ ,  $p < .001$ , indicating that gender had a significant effect on mathematics proficiency. Substituting the value of the coefficient into the formula (8),  $\text{logit} [\pi(Y = j \mid Y \geq j, \text{gender})] = \alpha_j + (-\beta_1 X_1)$ , we

calculated logit  $[\pi(Y=j | Y \geq j, \text{gender})] = \alpha_j - .1416 (\text{gender})$ . OR =  $e^{(-.1416)} = .8680$ , indicating that male students were .8680 times the odds for female students of being in any category compared to being in higher categories, i.e., female students were more likely than male students to stop out in a particular category, because males are coded as 1 and females are coded as 0.

To estimate the conditional probability of being beyond a category of mathematics proficiency, which is the complement of the conditional probability of being at a category, we just need to change the signs before the cutpoints and the estimated logits in the equation (8) and then calculate logit  $[\pi(Y>j | Y \geq j, \text{gender})] = -\alpha_j + .1416 (\text{gender})$ . By exponentiating .1416, we got the OR of 1.152, which indicated that male students were 1.152 times more likely to be beyond a particular mathematics proficiency level than female students.

The CR model could also be fitted using the complementary log-log link (clog-log) with the cumulative option within the Stata OCRATIO command. The CR model with the complementary log-log link is actually the discrete-time proportional hazards model for the event history analysis or survival analysis (Allison, 1999; O'Connell, 2006). It estimates the hazard ratio (HR) rather than the odds ratio (OR) of being in a particular category relative to advancing to a higher category. Figure 2 displays the Stata output for the clog-log continuation model.

The log likelihood ratio Chi-Square test with 1 degree of freedom, LR  $\chi^2_{(1)} = 51.38$ ,  $p < .001$ , indicating that the full model with one predictor provides a better fit than the null model with no independent variables. The Pseudo  $R^2 = .0011$ , suggested that the

relationship between the response variable, mathematics proficiency, and the predictor, gender was small.

The estimated clog-log coefficient,  $\beta = .1257$ ,  $z = 7.17$ ,  $p < .001$ , indicating that gender had a significant effect on mathematics proficiency. Since Clog-log  $[\pi(Y=j | Y \geq j, \text{gender})] = \log(-\log(1 - \pi)) = \alpha_j + (-\beta_1 X_1)$ , we calculated  $\log(-\log(1 - \pi)) = \alpha_j - .1257$  (gender). By exponentiating  $-.1257$ , we got the hazard ratio,  $HR = e^{(-.1257)} = .8819$ , indicating that the hazard of being in a particular proficiency level rather than beyond for male students was .8819 times the hazard for female students, i.e., the hazard for female students of stopping out in a particular category was 1.134 times as great as that for male students.

*Proportional Odds Model with a Single Explanatory Variable*

Next, for comparison purpose, a proportional odds model analysis with the same single predictor, gender was conducted using the Stata OLOGIT procedure. Figure 3 displays the Stata output for the one-predictor proportional odds model.  $LR \chi^2_{(1)} = 28.13$ ,  $p < .001$ , indicating that the one-predictor PO model provided a better fit than the null model with no independent variables in predicting cumulative probabilities for mathematics proficiency level. The Pseudo  $R^2 = .0006$ , which was as small as that in the continuation ratio model.

The estimated logit regression coefficient,  $\beta = .1527$ ,  $z = 5.30$ ,  $p < .001$ . Since the PO model estimates the cumulative odds and cumulative probabilities of being at or below a particular category of the ordinal response outcome, we calculated  $\text{logit} [\pi(Y \leq j | \text{gender})] = \alpha_j - .1527$  (gender). By exponentiating the logit,  $-.1527$ , we obtained the odds ratio (OR),  $e^{(-.1527)} = .8584$ , indicating that the odds of being at or below a mathematics

proficiency level were .8584 times as great for male students as they were for female students, i.e., female students were more likely than male students to be at or below a particular proficiency level.

The PO model can estimate  $J-1$  cumulative probabilities of being at or below a category of the ordinal response variable with  $j$  levels. When the ordinal response variable, mathematics proficiency, has six levels from 0 to 5, the proportional odds model estimates five cumulative probabilities, which include  $P(Y \leq 0)$ ,  $P(Y \leq 1)$ ,  $P(Y \leq 2)$ ,  $P(Y \leq 3)$ , and  $P(Y \leq 4)$ . The cumulative probabilities of being beyond a category can also be estimated since they are the complementary probabilities of the being at or below a particular category.

Different from the cumulative probabilities in the PO model, the logit CR model estimates the conditional probabilities. In the gender-only CR model, it estimates the conditional probability of being in category  $j$ , conditional on being at or beyond that category, i.e.,  $P(Y = j | Y \geq j, \text{gender})$ . This CR model can also estimate the conditional probability of being beyond a category given that individual has achieved that particular category, since  $P(Y > j | Y \geq j, \text{gender})$  is the complementary form of  $P(Y = j | Y \geq j, \text{gender})$ .

Another difference between the CR model and the PO model is the change in the sample size. In the gender-only PO model, the sample size was 15,325. However, the number of observations has been increased to 51,353 in the CR model due to different comparisons between proficiency levels, which included level 0 versus levels 1, 2, 3, 4, and 5; level 1 versus levels 2, 3, 4, and 5; level 2 versus 3, 4, and 5; level 3 versus 4 and 5; and level 4 versus level 5 (Table 2 provides the comparisons between six proficiency

levels). Fitting the CR model using SAS requires a restructured data set from the J-1 concatenated sub-data sets from the comparisons between proficiency levels (Allison, 1999; O'Connell, 2006), while Stata can fit the CR model directly without the data restructuring procedure.

#### *Continuation Ratio Model with 11 Explanatory Variables*

Next, we fit a CR model with 11 explanatory variables, which was referred to as the Full Model. Figure 4 and Table 3 display the results for the fitting of the full model with 11 explanatory variables.

The log likelihood ratio Chi-Square test,  $LR \chi^2_{(11)} = 3069.32$ ,  $p < .001$ , indicating that the full model with 11 predictor provides a better fit than the null model with no independent variables in predicting conditional probability for mathematics proficiency. The likelihood ratio  $R^2_L = .0777$ , much larger than that of the gender-only model, but still small, suggesting that the relationship between the response variable, mathematics proficiency, and 11 predictors, was still small. AIC Goodness-of-fit statistics were used for model comparisons. Compared with the gender-only model (.9224), the AIC statistic indicated that the full-model fit the data much better (.8483).

Using the eform option, we could obtain odds ratios for all the predictors. Overall, these predictors, such as, being male students (bygender), reviewing work from the previous day in math class (bys29a), listening to teachers' lectures (bys29b), doing problem solving in class (bys29e), using general calculators (bys29f), using graphing calculators (bys29g), and explaining work orally (bys29i), were positively associated with the odds of being beyond a particular mathematics proficiency level. Copying notes from board in class (bys29c), using books besides textbooks (bys29d), using computers

(bys29h), and participating in student-led discussions (bys29j) were less likely to advance to a higher proficiency level, i.e., they were more likely to stop out in a particular proficiency level.

In terms of odds ratios, male students were 1.359 the odds for female students to be beyond a given proficiency level (OR = 1.359), after controlling for the effects of other predictors in the full model. The odds of being beyond a particular proficiency level relative to being in that level were 1.166 times greater with one unit increase in the frequency of reviewing work from the previous day (OR = 1.166). Similarly, listening to teachers' lectures (OR = 1.192), doing problem solving in class (OR = 1.077), using general calculators (OR = 1.179), using graphing calculators (OR = 1.173), and explaining work orally (OR = 1.066) were more likely to be in a higher proficiency level. In the other hand, for every one unit increase in copying notes from board in class, the odds of being beyond a particular category decreased by a factor of .96 (OR = .96). In other words, the more the students copied notes from board, the more likely they would stop out in a mathematics proficiency level. Similarly, the odds decreased by a factor of .785 (OR = .785), for a unit increase in using textbooks besides the mathematics textbook; they decreased by a factor of .833 for a unit increase in using computers in math classes; and they decreased by a factor of .892 in participating in student-led discussions, holding the effects of the other variables constant.

Table 3 also provides the results of the multiple regression (MR) analysis. Although the results of MR analysis looked similar to those estimated by the CR model, they were different in nature: the former estimates the linear effects the classroom practices on mathematics proficiency level, while the latter estimates the conditional probability of being in a proficiency level relative to being beyond, or its complement,



the probability of advancing to a higher proficiency level rather than being in that particular level. The MR analysis could be used as a preliminary analysis before the CR model fitting.

## Conclusions

In this article, the use of continuation ratio models was illustrated to estimate high school students' mathematics proficiency from a set of predictors of classroom practices. Modeling fitting started from a single-variable CR with both logit and clog-log links, and then the PO model, and finally the full CR logit model with 11 predictor variables. Results between the single-variable CR logit model and the PO model, and between the single-variable and the full CR logit model were interpreted and compared.

Results from the CR models suggested that some classroom practices, such as reviewing work from the previous day in math class, listening to teachers' lectures, doing problem solving in class, using general calculators, using graphing calculators, and explaining work orally, had positive effects on the odds of being beyond a particular mathematics proficiency level relative to being in that level; while other classroom practices, such as, copying notes from board, using books besides textbooks, using computers in class, and participating in student-lead discussions were associated with odds of stopping out in a particular proficiency level rather than advancing to a higher proficiency level.

In education research, the use of ordinal categorical data becomes abundant, and it is crucial for researchers to understand different statistical methods while analyzing ordinal response variables. With the availability of statistical software packages, it is

hoped that this article would help researchers be familiar with continuation ratio models and utilize them correctly in their research.

## References

- Agresti, A. (2002). *Categorical data analysis* (2<sup>nd</sup> ed.). New York: John Wiley & Sons.
- Agresti, A. (2007). *An introduction to categorical data analysis* (2<sup>nd</sup> ed.). New York: John Wiley & Sons.
- Agresti, A. (1996). *An introduction to categorical data analysis*. New York: John Wiley & Sons.
- Allison, P.D. (1999). *Logistic regression using the SAS system: Theory and application*. Cary, NC: SAS Institute, Inc.
- Ananth, C. V. and Kleinbaum, D. G. (1997). Regression models for ordinal responses: A review of methods and applications. *International Journal of Epidemiology* 26, 1323-1333.
- Armstrong, B. B. & Sloan, M. (1989). Ordinal regression models for epidemiological data. *American Journal of Epidemiology*, 129(1), 191-204.
- Bender, R. & Benner, A. (2000). Calculating ordinal regression models in SAS and S-Plus. *Biometrical Journal*, 42(6), 677-699.
- Bender, R. & Grouven, U. (1998). Using binary logistic regression models for ordinal data with non-proportional odds. *Journal of Clinical Epidemiology*, 51(10), 809-816.
- Brant (1990). Assessing proportionality in the proportional odds model for ordinal logistic regression. *Biometrics*, 46, 1171-1178.
- Clogg, C.C., Shihadeh, E.S. (1994). *Statistical models for ordinal variables*. Thousand Oaks, CA: Sage.
- Fienberg, S. E. (1980). *The analysis of cross-classified categorical data*. Cambridge, MA: The MIT Press.
- Greene, William H. (2003). *Econometric analysis* (5<sup>th</sup> ed.). Upper Saddle River, NJ: Prentice Hall.
- Hardin, J. W., & Hilbe, J. M. (2007). *Generalized linear models and extensions* (2<sup>nd</sup> ed.). Texas: Stata Press.
- Hosmer, D. W. & Lemeshow, S. (2000). *Applied logistic regression* (2<sup>nd</sup> ed.). New York: John Wiley & Sons.
- Ingels, S. J., Pratt, D. J., Roger, J., Siegel, P. H., & Stutts, E. (2004). *ELS: 2002 Base Year Data File User's Manual*. Washington, DC: NCES (NCES 2004-405).

- Ingels, S. J., Pratt, D. J., Roger, J., Siegel, P. H., & Stutts, E. (2005). Education Longitudinal Study: 2002/04 Public Use Base-Year to First Follow-up Data Files and Electronic Codebook System. Washington DC: NCES (NCES 2006-346).
- Liu, X. (in press). Ordinal regression analysis: Fitting the proportional odds model using Stata, SAS and SPSS. *Journal of Modern Applied Statistical Methods*.
- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Thousand Oaks, CA: Sage.
- Long, J. S. & Freese, J. (2006). *Regression models for categorical dependent variables using Stata* (2<sup>nd</sup> ed.). Texas: Stata Press.
- McCullagh, P. (1980). Regression models for ordinal data (with discussion). *Journal of the Royal Statistical Society Ser. B*, 42, 109-142.
- McCullagh, P. & Nelder, J. A. (1989). *Generalized linear models* (2<sup>nd</sup> ed.). London: Chapman and Hall.
- Menard, S. (1995). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage.
- O'Connell, A.A., (2000). Methods for modeling ordinal outcome variables. *Measurement and Evaluation in Counseling and Development*, 33(3), 170-193.
- O'Connell, A. A. (2006). *Logistic regression models for ordinal response variables*. Thousand Oaks: SAGE.
- Powers D. A., & Xie, Y. (2000). *Statistical models for categorical data analysis*. San Diego, CA: Academic Press.
- Wooldridge, Jeffrey M. (2001). *Econometric analysis of cross section and panel data*. Cambridge, MA: The MIT Press.
- Wolfe, R. (1998). Continuation-ratio models for ordinal response data. *Stata Technical Bulletin*, 44, 18-21.

Table 1: *Proficiency Categories and Frequencies (Proportions) for the Study Sample, ELS:2002 (N = 15,976)*

<i>Proficiency</i>		
<i>Category</i>	<i>Description</i>	<i>Frequency</i>
0	Did not pass level 1	842 (5.27%)
1	Can do simple arithmetical operations on whole numbers	3882 (24.30%)
2	Can do simple operations with decimals, fractions, powers, and root	3422 (21.42%)
3	Can do simple problem solving	4521 (28.30%)
4	Can understand intermediate-level mathematical concepts and/or find multi-step solutions to word problems	3196 (20.01%)
5	Can solve complex multiple-step word problems and/or understand advanced mathematical material	113 (0.71%)

Table 2: Category Comparisons for the Continuation Odds Model with Six Mathematics Proficiency Levels ( $j = 0, 1, 2, \dots, 5$ ).

Category	Conditional Probability $P(Y=j   Y \geq j)$	Odds Ratio	Probability Comparisons
Proficiency 0	$P(Y=0   Y \geq 0)$	$\frac{P(Y=0)}{P(Y>0)}$	Category 0 versus all categories above
Proficiency 1	$P(Y=1   Y \geq 1)$	$\frac{P(Y=1)}{P(Y>1)}$	Category 1 versus Categories 2 through 5
Proficiency 2	$P(Y=2   Y \geq 2)$	$\frac{P(Y=2)}{P(Y>2)}$	Category 2 versus Categories 3 through 5
Proficiency 3	$P(Y=3   Y \geq 3)$	$\frac{P(Y=3)}{P(Y>3)}$	Category 3 versus Categories 4 and 5
Proficiency 4	$P(Y=4   Y \geq 4)$	$\frac{P(Y=4)}{P(Y>5)}$	Category 4 versus 5

Table 3: Results of the Continuation Ratio Model and the OLS Regression Model (Full Model),  $n = 42,992$

Variable	Continuation Ratio Model (logit)		OLS Model
	b (se(b))	OR	
$\alpha_1$	-1.50 (.08)		1.15 (.06)
$\alpha_2$	.49 (.08)		
$\alpha_3$	.89 (.08)		
$\alpha_4$	2.27 (.08)		
$\alpha_5$	5.64 (.13)		
Gender <sup><math>\delta</math></sup>	.31 (.03)**	1.36	.21 (.02)**
Review	.15 (.01)**	1.17	.12 (.01)**
Listen	.18 (.01)**	1.19	.13 (.01)**
Copynote	-.04 (.01)**	.96	-.02 (.01)*
Usebooks	-.24 (.01)**	.79	-.18 (.01)**
Probsolv	.07 (.01)**	1.08	.05 (.01)**
Usecalcu	.16 (.01)**	1.18	.12 (.01)**
Usegraph	.16 (.01)**	1.17	.11 (.01)**
Usecompu	-.18 (.01)**	.83	-.14 (.01)**
Explain	.06 (.01)**	1.06	.05 (.01)**
Participate	-.11 (.01)**	.89	-.09 (.01)**
$R^2$	$R^2_L = .078$		$R^2 = .221$
Model Fit <sup>a</sup>	$\chi^2_{11} = 3039.32$ (p < .0001)		$F(11, 12768) = 329.24^{**}$

<sup>$\delta$</sup>  gender: male=1

<sup>a</sup> Likelihood ratio test

\*Significant at  $p < .05$ ; \*\*  $p < .01$

Figure 1: Stata Continuation Ratio Model with Logit Link: Single Predictor, Gender

```
. ocratio Profmath BYGENDER, link (logit)
```

Continuation-ratio logit Estimates

Log Likelihood = -23683.4

Number of obs = 51353  
 chi2(1) = 38.90  
 Prob > chi2 = 0.0000  
 Pseudo R2 = 0.0008

---

Profmath	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
BYGENDER	.1416361	.0227235	6.23	0.000	.0970989 .1861732

---

(Ancillary parameters)

_cut1	-2.790613	.0372137			
_cut2	-.9961043	.0219305			
_cut3	-.7736138	.0238228			
_cut4	.368887	.026111			
cut5	3.392331	.0966743			

---

```
. ocratio Profmath BYGENDER, link (logit) eform
```

Continuation-ratio logit Estimates

Log Likelihood = -23683.4

Number of obs = 51353  
 chi2(1) = 38.90  
 Prob > chi2 = 0.0000  
 Pseudo R2 = 0.0008

---

Profmath	Odds ratio	Std. Err.	z	P> z	[95% Conf. Interval]
BYGENDER	1.152157	.026181	6.23	0.000	1.101969 1.204631

---

(Ancillary parameters)

_cut1	-2.790613	.0372137			
_cut2	-.9961043	.0219305			
_cut3	-.7736138	.0238228			
_cut4	.368887	.026111			
_cut5	3.392331	.0966743			

---

```
. aic
```

AIC Statistic = .9224153

Figure 2: Stata Continuation Ratio Model with Clog-log Link: Single Predictor, Gender

```
. ocratio Profmath BYGENDER, link (cloglog) cumulative
```

Ordered cloglog Estimates

Log Likelihood = -23677.16

Number of obs = 51353  
 chi2(1) = 51.38  
 Prob > chi2 = 0.0000  
 Pseudo R2 = 0.0011

---

Profmath	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
BYGENDER	.1256615	.0175265	7.17	0.000	.0913103 .1600128

---

(Ancillary parameters)

_cut1	-2.826367	.0356499			
_cut2	-.9834265	.022463			



```

_cut3 | -.2817271 .0217445
_cut4 | .5087509 .0202158
_cut5 | 1.663668 .0274349
-----
. aic
AIC Statistic = .9221723

```

```

. ocratio Profmath BYGENDER, link (cloglog) eform cumulative

Ordered cloglog Estimates                                Number of obs = 51353
                                                         chi2(1)         = 51.38
                                                         Prob > chi2     = 0.0000
Log Likelihood = -23677.16                               Pseudo R2      = 0.0011
-----
Profmath | Haz. ratio   Std. Err.      z    P>|z|    [95% Conf. Interval]
-----+-----
BYGENDER | 1.133898     .0198732    7.17  0.000    1.095609    1.173526
-----+-----
(Ancillary parameters)
_cut1 | -2.826367   .0356499
_cut2 | -.9834265   .022463
_cut3 | -.2817271   .0217445
_cut4 | .5087509    .0202158
_cut5 | 1.663668    .0274349
-----

```

Figure 3: Stata Proportional Odds Model: Single Predictor, Gender

```

ologit Profmath BYGENDER

Iteration 0:  log likelihood = -23702.845
Iteration 1:  log likelihood = -23688.779
Iteration 2:  log likelihood = -23688.778

Ordered logistic regression                                Number of obs   = 15325
                                                         LR chi2(1)     = 28.13
                                                         Prob > chi2    = 0.0000
Log likelihood = -23688.778                               Pseudo R2      = 0.0006
-----
Profmath |      Coef.   Std. Err.      z    P>|z|    [95% Conf. Interval]
-----+-----
BYGENDER | .1527419    .0288057    5.30  0.000    .0962839    .2092
-----+-----
/cut1 | -2.785918   .0381689    -2.86 0.006    -2.860728   -2.711108
/cut2 | -.7893203   .0224898    -3.51 0.000    -.8333995   -.7452411
/cut3 | .1072826    .0214844    4.99 0.000    .065174     .1493911
/cut4 | 1.402499    .0246227    57.00 0.000    1.354239    1.450758
/cut5 | 4.981085    .095611     52.10 0.000    4.793691    5.168479
-----

```

Figure 4: Stata Continuation Ratio Model with Logit Link: Full Model

```

. ocratio Profmath BYGENDER BYS29Arec BYS29Brec BYS29Crec BYS29Drec BYS29Er
> ec BYS29Frec BYS29Grec BYS29Hrec BYS29Irec BYS29Jrec, link (logit)

Continuation-ratio logit Estimates                                Number of obs = 42992
                                                         chi2(11)      = 3069.32

```

```

Log Likelihood = -18223.82
Prob > chi2 = 0.0000
Pseudo R2 = 0.0777

```

---

Profmath	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
BYGENDER	.3068651	.0267692	11.46	0.000	.2543984 .3593318
BYS29Arec	.1533858	.0099134	15.47	0.000	.1339559 .1728157
BYS29Brec	.1758252	.0127855	13.75	0.000	.150766 .2008843
BYS29Crec	-.0407896	.0123618	-3.30	0.001	-.0650184 -.0165608
BYS29Drec	-.2416916	.0093455	-25.86	0.000	-.2600084 -.2233748
BYS29Erec	.073735	.0112281	6.57	0.000	.0517284 .0957416
BYS29Frec	.1644396	.0113946	14.43	0.000	.1421066 .1867726
BYS29Grec	.1594446	.0087424	18.24	0.000	.1423098 .1765794
BYS29Hrec	-.1828204	.0114234	-16.00	0.000	-.2052098 -.160431
BYS29Irec	.0638591	.0107391	5.95	0.000	.0428109 .0849074
BYS29Jrec	-.1145253	.0108834	-10.52	0.000	-.1358564 -.0931942

---

```

_ cut1 | -1.504545 .0834409 (Ancillary parameters)
_ cut2 | .4853076 .0788576
_ cut3 | .8863932 .0812187
_ cut4 | 2.266622 .0846272
_ cut5 | 5.643517 .1346226

```

---

```

. aic
AIC Statistic = .8482891

```

```

. ocratio Profmath BYGENDER BYS29Arec BYS29Brec BYS29Crec BYS29Drec BYS29Er
> ec BYS29Frec BYS29Grec BYS29Hrec BYS29Irec BYS29Jrec, link (logit) eform

```

Continuation-ratio logit Estimates

```

Number of obs = 42992
chi2(11) = 3069.32
Prob > chi2 = 0.0000
Pseudo R2 = 0.0777

```

Log Likelihood = -18223.82

---

Profmath	Odds ratio	Std. Err.	z	P> z	[95% Conf. Interval]
BYGENDER	1.359158	.0363836	11.46	0.000	1.289686 1.432372
BYS29Arec	1.165775	.0115568	15.47	0.000	1.143342 1.188647
BYS29Brec	1.19223	.0152433	13.75	0.000	1.162725 1.222483
BYS29Crec	.9600311	.0118678	-3.30	0.001	.9370503 .9835756
BYS29Drec	.7852983	.007339	-25.86	0.000	.7710451 .799815
BYS29Erec	1.076522	.0120873	6.57	0.000	1.05309 1.100475
BYS29Frec	1.178732	.0134312	14.43	0.000	1.1527 1.205353
BYS29Grec	1.172859	.0102536	18.24	0.000	1.152934 1.193129
BYS29Hrec	.8329177	.0095147	-16.00	0.000	.8144764 .8517766
BYS29Irec	1.065942	.0114472	5.95	0.000	1.043741 1.088616
BYS29Jrec	.8917894	.0097057	-10.52	0.000	.872968 .9110165

---

```

_ cut1 | -1.504545 .0834409 (Ancillary parameters)
_ cut2 | .4853076 .0788576
_ cut3 | .8863932 .0812187
_ cut4 | 2.266622 .0846272
_ cut5 | 5.643517 .1346226

```

---