Fitting Proportional Odds Models for Complex Sample Survey Data with SAS, IBM SPSS, Stata, and R

Xing Liu
Eastern Connecticut State University

An ordinal logistic regression model with complex sampling designs is different from a conventional proportional odds model since the former needs to take weights and design effects in account. While general-purpose statistical packages, such as SAS, IBM SPSS, Stata, and R, are all capable of fitting proportional odds models with complex survey data, they may use different techniques to estimate the models and have different features. The purpose of this article was to illustrate the use of SAS, IBM SPSS, Stata, and R to fit proportional odds models with complex survey data, and to compare the features and results for fitting the models using SAS PROC SURVEYLOGISTIC, IBM SPSS CSORDINAL, Stata `svy: ologit`, and R `survey`. The linearization method was used to estimate the sampling variance.

The proportional odds (PO) model, which is well documented in the literature (Agresti, 2002, 2007, 2010; Ananth & Kleinbaum, 1997; Armstrong & Sloan, 1989; Hardin & Hilbe, 2007; Hilbe, 2009; Liu, 2009, 2016; Long, 1997; Long & Freese, 2014; McCullagh, 1980; McCullagh & Nelder, 1989; O'Connell, 2000, 2006; O'Connell & Liu, 2011; Powers & Xie, 2000), is commonly used for ordinal response variables. An important assumption of this model is that the estimated logit coefficients for each predictor are constant across the ordinal categories. In other words, only one regression coefficient is estimated for each predictor although multiple intercepts or thresholds may be estimated. This is called the PO assumption or the parallel lines assumption.

Although simple random sampling is the ideal method for data collection, complex sampling designs are often used for large-scale national studies, such as the Educational Longitudinal Study of 2002 (ELS: 2002) and the High School Longitudinal Study of 2009 (HSLS: 2009), and international studies, such as the Programme for International Student Assessment (PISA). Complex sampling designs in these studies include the elements such as strata, clusters, and sampling weights. When analyzing such data, researchers need to choose appropriate techniques taking the sampling weights and design variables into account to obtain accurate parameter estimates and variances.

While the general-purpose statistical packages, SAS, IBM SPSS, Stata, and R, are all capable of fitting PO models with complex survey data, they may have different features of setting up the elements of complex sampling designs and use different techniques to estimate these models. Understanding the differences among software packages would help applied researchers and practitioners to clarify the confusion of different parameterizations of PO models and interpret the results correctly. In addition, a broader introduction of statistical software would provide applied researchers and practitioners with more options when they conduct PO models with complex sampling designs.

To fill this gap, the purpose of this article was to illustrate the use of SAS, IBM SPSS, Stata, and R to fit the PO model with complex survey data, and compare the features and results for fitting the models using the SAS PROC SURVEYLOGISTIC, IBM SPSS CSORDINAL, Stata `svy: ologit`, and R `svyolr` commands. For demonstration purposes, ordinal regression analyses were conducted using the same Educational Longitudinal Study (ELS): 2002 data, where the ordinal outcome of students' mathematics proficiency was predicted from student effort, including students can get no bad grades if they decide to, they keep studying even if material is difficult, and they do best to learn.

Theoretical Framework

The Proportional Odds Model

The conventional PO model estimates the cumulative odds of being at or below a specific level of an ordinal response variable, given a set of predictor variables. This model can be expressed in the logit form as follows:

$$\ln(Y_j¢) = \text{logit}\,[p(x)] = \ln\left(\frac{\pi_j(x)}{1-\pi_j(x)}\right) = a_j + (- b_1X_1 - b_2X_2 - \ldots - b_pX_p), \quad (1)$$

where $\pi_j(x) = \pi(Y \leq j \mid x_1, x_2, \ldots, x_p)$, which is the cumulative probability of being at or below category *j* given a set of predictor variables, $j = 1, 2, \ldots J-1$. a_j are the cut points or intercepts, and b_1, $b_{2 \ldots,} b_p$ are the logit coefficients for the corresponding predictor variables. The PO model assumes that the effects of any predictor variable are the same across all categories, so only one logit coefficient is estimated for each predictor variable. The PO model can be also expressed in the form of cumulative odds as follows:

$$\text{logit}\,[p(Y \leq j \mid x_1, x_2, \ldots, x_p)] = \ln\left(\frac{\pi(Y \leq j \mid x_1, x_2, \ldots x_p)}{\pi(Y > j \mid x_1, x_2, \ldots x_p)}\right) = a_j + (-b_1X_1 - b_2X_2 - \ldots - b_pX_p) \qquad (2)$$

In this model, the cumulative odds of being at or below a category equal the ratio of the probability of being at or below a category to the probability of being above that category. The PO model also estimates the cumulative odds of being above a particular category since they are the inversed odds of being at or below that category.

When analyzing complex sampling survey data, researchers (Hahs-Vaughn, 2005; Lee & Forthofer, 2006; Levy & Lemeshow, 2008; Liu, 2016; Liu & Koirala, 2013; Lohr, 2010; Osborne, 2011; Thomas & Heck, 2001) recommended using the appropriate statistical methods which take the weights and design effects into account. Methods for variance estimation in ordinal logistic regression were introduced in Binder (1983) and Heeringa, West, and Berglund (2010). The Taylor series approximation method, which is the default in SAS, IBM SPSS, and Stata, is commonly used for variance estimation for the complex survey data.

Why Compare SAS, IBM SPSS, Stata, and R?

Researchers should be aware that software packages may use different forms to express the PO model and parameterize it differently. Liu (2009) illustrated the use of Stata, SAS, and SPSS to fit PO models and compared the features and the results of the fitted models using these three software packages. This study found that Stata and SPSS both followed the same equation (see Equation 2 above). Compared to both SPSS and Stata, the cut points were the same using SAS with the ascending option. However, the estimated coefficients were the same in magnitude but were reversed in sign. In addition, SAS with ascending and descending options produced the same cut points and coefficients in magnitude with reversed signs. These differences in parameterization may also exist when fitting the model to the data with complex survey sampling, so researchers may feel confused when interpreting the results with different software packages. In addition, these packages may be different in specifying the elements of complex survey samples when fitting the PO model. While Heeringa et al. (2010) introduced some general features of these packages for complex survey data analysis, this articles focuses on the PO model for ordinal response variables only.

Methodology

Sample

The Educational Longitudinal Study of 2002 (ELS: 2002) base-year data was used for the analyses. The ELS: 2002 study, conducted by the National Center for Educational Statistics (NCES), tracked the cohort of 2002 high school sophomores with a longitudinal design regarding their postsecondary school education and future careers. A two-stage sampling design was used (Ingels, Pratt, Roger, Siegel, & Stutts, 2004, 2005). First, a stratified sampling strategy was used to select 1,221 eligible schools from a population of approximately 27,000 schools having 10th grade students. Among these eligible public and private schools, a total of 752 schools agreed to participate in the study. Second, in each of those schools, approximately 25 10th grade students were randomly selected from the enrollment lists.

The ordinal response variable was high-school students' mathematics proficiency levels, including five levels with 1 = students can do simple arithmetical operations on whole numbers and 5 = students can solve complex multiple-step word problems and/or understand advanced mathematical material (Ingels et al., 2004, 2005). In addition, level 0 was assigned to the students who failed to pass through level 1. Table 1 provides a detailed description of these six proficiency levels including level 0 and their frequencies (Liu & Koirala, 2013).

Table 1. *Descript of Mathematics Proficiency Levels and Frequencies (Proportions) for the ELS: 2002 Data (N = 15,976)*

Proficiency Level	*Description*	*Frequency*
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%)

Data Analysis

First, SAS PROC SURVEYLOGISTIC with both the ascending and descending options was used to fit PO models for complex survey samples, taking strata, clusters, and sampling weights into account. Second, the same analysis was conducted using the IBM SPSS CSORDINAL command. Before conducting the complex sample data analysis, steps on how to specify the sample design variables and weights via the analysis preparation wizard using IBM SPSS were discussed. Third, the Stata `svyset` command was used to define the complex sampling design features, and the `svy: ologit` command was used for the ordinal regression analysis. Fourth, the R survey package was used to replicate the analysis. Finally, the similarities and differences of the results across packages were compared and a sample write-up of the results was provided. The linearization method (Taylor series approximation) was used to estimate the sampling variance for the complex survey data.

Results

Proportional Odds Models for Complex Survey Data Using SAS PROC SURVEYLOGISTIC with the Ascending and Descending Options

The SAS PROC SURVEYLOGISTIC procedure, which is the survey analysis procedure for logistic regression models, was used to fit PO models with complex sample survey data. This procedure can be used to estimate binary, ordinal, and nominal response variables. The variables for the strata, clusters, and sampling weights in the data are STRAT_ID, PSU, and BYSTUWT, respectively. Table 2 displays the syntax for the proportional odds models for complex sample survey data with this procedure with the ascending and descending options.

Table 2. Proportional Odds Models for Complex Survey Data Using SAS PROC SURVEYLOGISTIC with the Ascending and Descending Options: Syntax

```
***PO model with the ascending option***
proc surveylogistic data = 'C:\complexdata';
stratum STRAT_ID;
cluster PSU;
model Profmath (order = internal) = BYS89N_REC BYS89O_REC BYS89S_REC;
weight BYSTUWT;
run;

***PO model with the descending option***
proc surveylogistic data = 'C:\complexdata';
stratum STRAT_ID;
cluster PSU;
model Profmath (order = internal descending) = BYS89N_REC BYS89O_REC BYS89S_REC;
weight BYSTUWT;
run;
```

Table 3. Results of the Proportional Odds Model for Complex Survey Data Using SAS (Ascending and Descending), IBM SPSS, Stata, and R: A Comparison

	SAS (Ascending)	SAS (Descending)	SPSS	Stata	R
Model estimates	P(Y≤j)	P(Y>j)	P(Y≤j)	P(Y≤j)	P(Y≤j)
Cut points (Stata)/ Intercepts (SAS)/ Thresholds (SPSS)	α_1 = -.955	α_5 = -7.245	α_1 = -.955	_cut1(α_1) = -.955	α_1 = -.954
	α_2 = 1.153	α_4 = -3.490	α_2 = 1.153	_cut2(α_2) = 1.153	α_2 = 1.153
	α_3 = 2.154	α_3 = -2.154	α_3 = 2.153	_cut2(α_3) = 2.153	α_3 = 2.153
	α_4 = 3.490	α_2 = -1.153	α_4 = 3.490	_cut2(α_4) = 3.490	α_4 = 3.490
	α_5 = 7.245	α_1 = .955	α_5 = 7.245	_cut2(α_5) = 7.245	α_5 = 7.246
decide	-.530**	.530**	.530**	.530**	.530**
	(.033)	(.033)	(.033)	(.034)	(.034)
keeplrn	-.053	.053	.052	.052	.053
	(.033)	(.033)	(.033)	(.033)	(.033)
dobest	-.161**	.161**	.161**	.161**	.161**
	(.039)	(.039)	(.040)	(.040)	(.040)
LR R^2	N/A	N/A	.035	N/A	N/A
Model fit	*LR* $\chi^2_{(3)}$ = 255,015.847**	*LR* $\chi^2_{(3)}$ = 255,015.847**	N/A	*F*(3, 387) = 201.71**	N/A

* p<.05; **p<.01.

In the syntax, the `stratum` statement was used to specify strata and the cluster statement was used to define clusters. In this example, the stratum was STRAT_ID and the cluster was PSU. In addition, the sampling weight, BYSTUWT, was specified using the weight statement. In the model statement, the option, `order = internal`, tells SAS to arrange the values of the ordinal response variable, Promath, from 0 to 5 in the ascending order. The other option, `order = internal descending`, reversed the order of the values of the ordinal outcome so that it is ordered from 5 to 0. The estimated results with both the ascending and descending options are displaced in Table 3.

With the ascending option in SAS, the PO model with complex sample survey data estimates the cumulative odds of being at or below a particular category versus being above that category. The log likelihood ratio chi-square test, *LR* $\chi^2_{(3)}$ = 255,015.847, $p < .001$; the score test and Wald chi-square test were 245,323.426 and 606.768, respectively, and both were significant ($p < .001$). The results indicated that the overall model fit the data better than the null model with no independent variables.

Among the three predictors, two were significantly different from zero. The estimated logit coefficient for *decide* (getting no bad grades if deciding to), β = –.530, Wald χ^2 = 261.357, $p < .001$; the logit coefficient for *keeplrn* (keeping studying if material is difficult), β = -.053, Wald χ^2 = 2.589, $p > .05$; and the coefficient for *dobest* (doing best to learn), β = -.161, Wald χ^2 = 17.495, $p < .001$.

The odds ratios (OR) for these three predictor variables were .589, .949, and .851, respectively, which could be interpreted as follows. The odds of being at or below a particular proficiency level versus being above that level decreased by a factor of .589 with a one unit increase in the value of the predictor variable, getting no bad grades if deciding to (*decide*). In addition, the odds of being at or below a particular proficiency level decreased by a factor of .851 for each one-unit increase in *dobest.* However, the predictor variable, *keeplrn* (keeping studying if material is difficult) did not influence the odds, since the coefficient was not significant ($p > .05$).

With the descending option in SAS, the order of the ordinal response variable is reversed so the PO model estimates the cumulative odds of being above a particular category, which are the inverse of the odds of being at or below that category estimated in the preceding model with the ascending option. With

both options, the results of all model fit statistics are the same except that the signs before the estimated intercepts and logit coefficients are reversed. The odds ratios can be interpreted as the change in the odds of being above a particular category versus being at or below that category for a one-unit change in the predictor variable. The results of the odds ratios for all three predictor variables were interpreted as follows. The odds of being beyond a proficiency level increased by 1.699 with a one-unit increase in the frequency of getting no bad grades if deciding to and increased by 1.175 with a one-unit increase in the frequency of doing best to learn, while keeping studying if material is difficult did not influence the odds since the coefficient was not significant ($p > .05$).

Proportional Odds Model for Complex Survey Data Using IBM SPSS

IBM SPSS has an add-on module, the Complex Samples (CS), for the analysis of complex survey data. The CSORDINAL command can be used to analyze ordinal response variables for complex sampling designs. Two steps need to be followed before we conduct data analysis for complex survey data. First, before conducting the analysis, we need to create an analysis plan via the analysis preparation wizard. Go to **Analysis**, **Complex Samples**, and then create the analysis plan with a name. Figure 1 displays the screenshot for the IBM SPSS analysis preparation wizard.

Second, specify the design variables, such as the strata, clusters, and the sampling weights after the plan file is named. In the dialogue box for design variables, select the design variables and move them to the corresponding boxes on the right. Figure 2 displays the screenshot for specifying complex sampling designs. The three design variables are shown in the corresponding boxes.

Creating the analysis plan file and specifying the design variables can also be done via the command syntax. The ordinal logistic regression for complex survey data was conducted using the CSORDINAL command after the survey designs were specified. Table 4 displays the CSORDINAL command syntax. The estimated results are presented in Table 3.

The same analysis can be conducted using the point-and-click function. Go to **Analysis**, **Complex Samples**, and then **Ordinal Regression**. It brings you to the dialog box for locating the complex sample plan file which has been already created. Next, browse the plan file where you saved and click on **Continue**. Finally, in the dialog box for the **Complex Samples Ordinal Regression**, select the ordinal response variable and predictor variables and then click on the OK button. Table 5 displays the output of the sample design information and the weighted percentages of the ordinal response variable.

The estimated thresholds or intercepts and logit coefficients are displayed in Table 3. The five thresholds were -.955, 1.153, 2.153, 3.490, and 7.245, respectively, which were the intercepts for the underlying binary logistic models for the ordinal response variable. The first threshold α_1 was the intercept for the binary model comparing level 0 versus levels 1 to 5; α_2 was the second threshold for the model comparing levels 0 and 1 with levels 2 to 5; and the final α_5 was the intercept for the model comparing levels from 1 to 4 with level 5.

Table 4. IBM SPSS Syntax for the CSORDINAL Command

```
CSORDINAL  Profmath (ASCENDING) WITH BYS89N_REC BYS89O_REC BYS89S_REC
 /PLAN FILE = 'C:\efficacy.csaplan'
 /LINK FUNCTION=LOGIT
 /MODEL  BYS89N_REC BYS89O_REC BYS89S_REC
 /STATISTICS PARAMETER EXP SE CINTERVAL TTEST
 /NONPARALLEL TEST
 /TEST TYPE=F PADJUST=LSD
 /ODDSRATIOS COVARIATE=[BYS89N_REC(1)]
 /ODDSRATIOS COVARIATE=[BYS89O_REC(1)]
 /ODDSRATIOS COVARIATE=[BYS89S_REC(1)]
 /MISSING CLASSMISSING=EXCLUDE
 /CRITERIA MXITER=100 MXSTEP=5
 PCONVERGE=[1e-006 RELATIVE] LCONVERGE=[0]
 METHOD=NEWTON CHKSEP=20
 CILEVEL=95
 /PRINT SUMMARY VARIABLEINFO SAMPLEINFO.
```

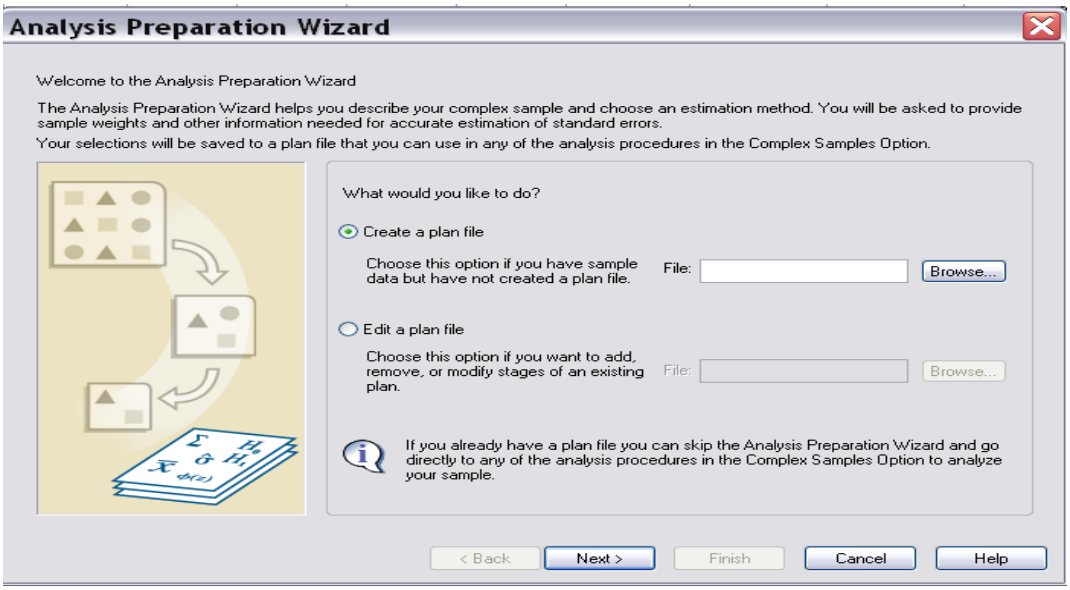

Figure 1. IBM SPSS analysis preparation wizard.

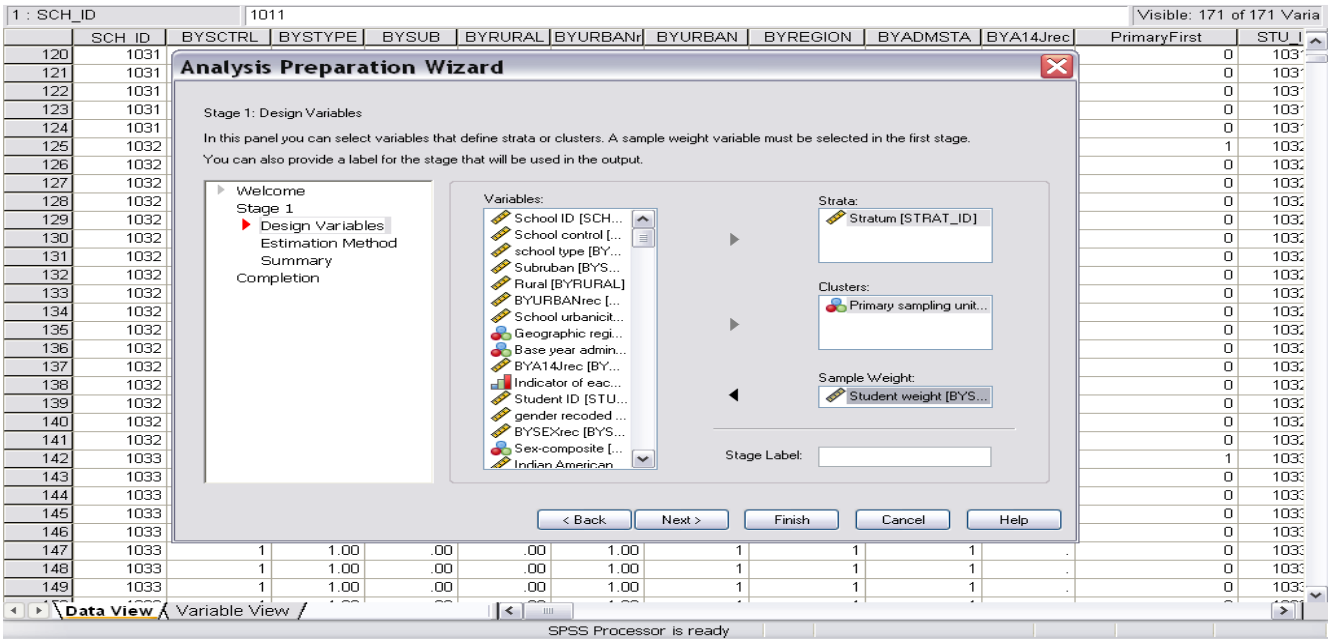

Figure 2. Specifying complex sampling designs in IBM SPSS.

Table 5. Sample Design Information and the Categories of the Ordinal Response Variable

Sample Design Information

		N
Unweighted Cases	Valid	10590
	Invalid	5662
	Total	16252
Population Size		2394546.730
Stage 1: 2	Strata	361
	Units	748
Sampling Design Degrees of Freedom		387

Categorical Variable Information

		Weighted Count	Weighted Percent
Profmath(a)	.00	117931.178	4.9%
	1.00	556275.674	23.2%
	2.00	516478.743	21.6%
	3.00	662330.571	27.7%
	4.00	524056.047	21.9%
	5.00	17474.516	.7%
Population Size		2394546.730	100.0%

a Dependent variable values are sorted in ascending order

Proportional Odds Model for Complex Survey Data Using Stata

While the command for conventional PO models in Stata is `ologit`, the `svy: ologit` command where `svy` is the prefix command is used for the PO model with complex survey data. The `svyset` command needs to be used first to specify the complex sampling design variables and weights before fitting the model. The following Stata command syntax was used for specifying the design features:

```
svyset PSU [pweight = BYSTUWT] , strata (STRAT_ID)
singleunit(certainty)
```

In the syntax, immediately following the `svyset` command were the primary sampling units or clusters, PSU, and the probability weight (`pweight`), `BYSTUWT`, which was the student weight for the based year data. The strata variable, `STRAT_ID`, was included in `strata ()` as an option in the command syntax. The option `singleunit(certainty)` was specified to deal with the singleton stratum. Table 6 presents the result of the specified sampling design information.

Table 6. Sampling Design Variables and Weights Using the `Stata svyset` Command

```
. svyset PSU [pweight = BYSTUWT] , strata (STRAT_ID)
singleunit(certainty)
      pweight: BYSTUWT
          VCE: linearized
  Single unit: certainty
     Strata 1: STRAT_ID
         SU 1: PSU
        FPC 1: <zero>
```

Next, the `svy: ologit` command was used to fit the PO model for complex sampling survey data. Table 7 displays the syntax of `svy: ologit`.

Table 7

PO Model for Complex Survey Data: Stata `svy: ologit` Syntax

```
svy: ologit Profmath  BYS89N_REC BYS89O_REC BYS89S_REC
```

The estimated results of the PO model with complex survey designs using Stata are presented in Table 3. The estimated cut points and logit coefficients were the same as those estimated by IBM SPSS.

Proportional Odds Model for Complex Survey Data Using R

The survey package (Lumley, 2004, 2010, 2014) can be used to analyze ordinal response variables for complex sampling designs in R. This package needs to be installed and loaded first before model fitting since it is an add-on package. The svydesign function in the package is used first to specify the sampling design information, such as the primary sampling units, strata, and weights so that the design object can be created. The following was the R syntax for specifying the design features:

```
svydes<-svydesign(strata=~STRAT_ID, id=~PSU, weights=~BYSTUWT,
data = complexdata, nest=TRUE)
```

In the syntax, `svydes` is the created design object using the `svydesign` function. The strata variable `STRAT_ID`, the primary sampling units PSU, and the probability weight `BYSTUWT`, were specified following the `svydesign` function.

After the design object was created, the `svyolr` function was used to fit the PO model for complex sampling survey data. Table 8 displays the R syntax using the `svyolr` function.

Table 8. PO Model for Complex Survey Data: R Syntax

```
> svymod<-svyolr(factor(Profmath)~BYS89N_REC + BYS89O_REC +
BYS89S_REC, design = svydes)
> summary(svymod)
```

In the syntax, the ordinal outcome variable `Profmath` was estimated by the three predictor variables. The `design = svydes` argument specified the design object. The estimated results displayed by the `summary(svymod)` function are the same as those estimated by Stata and IBM SPSS (see Table 3).

A Comparison of the Results of the PO Model for Complex Survey Data Using SAS, IBM SPSS, Stata, and R

Table 3 provides a comparison of the results of the PO model for complex survey data using SAS, IBM SPSS, Stata, and R. The results estimated by SAS SURVEYLOGISTIC with both the ascending and descending options are displayed. Please note that the different output produced by the four software packages is mainly due to the differences in model parameterizations. With a simple transformation, the final results of the cumulative odds ratios are the same using these packages.

1. When estimating the odds of being at or below a response category, the estimates for the cut points using Stata were the same as the intercepts using SAS with the ascending option in both sign and magnitude. However, the estimated logit coefficients were the same in magnitude but were opposite in sign.
2. Comparing the results of the PO model for complex survey data using Stata and SAS with the descending option, it was found that the estimated logit coefficients were the same in both magnitude and sign. However, the estimated intercepts were opposite in sign.
3. IBM SPSS, Stata, and R produced almost the identical results: the estimated logit coefficients and cutpoints or intercepts were the same.
4. Regarding model fit statistics, IBM SPSS reported the log likelihood ratio R^2 while neither SAS nor Stata reported this statistic. On the other hand, Stata reported *F*-statistic and SAS reported the log likelihood ratio chi-square test statistic for the overall model, while IBM SPSS and R reported neither of them.

Sample Write-Up of the Results

To help researchers to report the results for presentation and/or publication, a sample write-up of the results was provided as follows. The proportional odds model for complex sample survey data was fitted to predict the ordinal outcome variable, mathematics proficiency, from the three predictor variables on student effort. The sampling design information was specified in the model. The Taylor series linearization method was used for variance estimation. The log likelihood ratio test, $LR\ \chi^2_{(3)} = 255{,}015.847$, $p < .001$, which indicated that the overall model fit the data better than the null model with no independent variables.

The logit coefficients of the two predictors, *decide* and *dobest*, were significant. The logit coefficient for *decide* (getting no bad grades if deciding to), $\beta = -.530$, Wald $\chi^2 = 261.357$, $p < .001$; and the coefficient for *dobest* (doing best to learn), $\beta = -.161$, Wald $\chi^2 = 17.495$, $p < .001$. However, the coefficient of *keeplrn* was not significant. The logit coefficient for *keeplrn* (keeping studying if material is difficult), $\beta = -.053$, Wald $\chi^2 = 2.589$, $p > .05$.

In terms of odds ratios (OR), the odds of being beyond a proficiency level increased by 1.699 with a one-unit increase in the frequency of getting no bad grades if deciding to and increased by 1.175 with a one-unit increase in the frequency of doing best to learn. However, keeping studying if material is difficult did not influence the odds (OR = 1.054) since the coefficient was not significant.

Conclusions

This article illustrated the use of SAS, IBM SPSS, Stata, and R to fit the proportional odds models with complex survey sampling for ordinal response variables. Model fitting started from using SAS PROC SURVEYLOGISTIC with both the ascending and descending options, to using IBM SPSS CSORDINAL. The same PO model for complex survey data was fitted using Stata and R. The results using all four statistical software packages were compared. The logit coefficients and corresponding odds ratios for the predictors can be interpreted in the same way as those in the conventional PO model.

In summary, although SAS, IBM SPSS, Stata, and R are all capable of analyzing complex sampling data, they are different in the way of specifying complex design features and sampling weights. Stata, IBM SPSS, and R follow a two-step procedure to analyze complex survey data, while SAS offers survey analysis procedures which include the statements for complex sample designs and sampling weights. Specifically, in the SAS PROC SURVEYLOGISTIC procedure for the ordinal logistic regression analysis, we need to specify statements of survey designs together with other statements. IBM SPSS offers the analysis preparation wizard to create an analysis plan for complex samples (CSPLAN) by specifying elements of complex sample designs. When conducting ordinal logistic regression using the IBM SPSS CSORDINAL command, we need to first open the analysis plan created via the analysis preparation wizard. In Stata, the `svyset` command is used to specify complex survey designs before modeling fitting. Similarly, the `svydesign` function in the R `survey` package is used to specify the sampling design information.

Using SAS with the ascending option and Stata, the estimated logit coefficients are the same in magnitude but are opposite in sign. Using SAS with the descending option and Stata, the estimated logit coefficients are the same in both magnitude and sign. However, the estimated intercepts are opposite in sign. Comparing the results using IBM SPSS, Stata, and R, we found that they produced the same or similar logit coefficients and intercepts. Although these differences may seem trivial, they should not be overlooked, since they do cause confusion and incorrect interpretation of the results if the model parameterization and the software package do not match correctly.

These findings of the PO models for the complex survey data comparing all four packages extended those of the conventional PO models introduced in Liu (2009). Therefore, it is important for researchers to be aware this differences in parameterization of ordinal logistic models for complex survey data when using different statistical packages and apply the sound methodology in their own research.

This article focuses on the linearization method for variance estimation for all four statistical packages. For future research, other variance estimation methods (i.e., replicated methods) in the ordinal regression analysis for complex sampling data will be examined.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). New York: John Wiley & Sons.

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed.). New York: John Wiley & Sons.

Agresti, A. (2010). Analysis of ordinal c*ategorical data* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Ananth, C. V., & 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.

Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.

Hahs-Vaughn, D. L. (2005). A primer for understanding and using weights with national datasets. *Journal of Experimental Education, 73*(3), 221-240.

Hardin, J. W., & Hilbe, J. M. (2007). *Generalized linear models and extensions* (2nd ed.). Texas: Stata Press.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2010). *Applied survey data analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Hilbe, J. M. (2009). *Logistic regression models*. Boca Raton, FL: Chapman & Hall/CRC.

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).

Lee, E. S., & Forthofer, R. N. (2006). *Analyzing complex survey data* (2nd ed.). Thousand Oaks, CA: Sage.

Levy, P. S., & Lemeshow, S. (2008). *Sampling of populations: Methods and application* (4th ed.). New York, NY: John Wiley & Sons.

Liu, X. (2009). Ordinal regression analysis: Fitting the proportional odds model using Stata, SAS and SPSS. *Journal of Modern Applied Statistical Methods, 8(2), 632-645.*

Liu, X. (2016). *Applied ordinal logistic regression using Stata: From single-level to multilevel modeling*. Thousand Oaks, CA: Sage.

Liu, X., & Koirala, H. (2013). Fitting proportional odds models to educational data with complex sampling designs in ordinal logistic regression. *Journal of Modern Applied Statistical Methods, 12*(1), 235-248.

Lohr, S. L. (2010). *Sampling: Design and analysis* (2nd ed.). Boston: Brooks/Cole, Cengage Learning.

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.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software, 9*(1), 1-19.

Lumley, T. (2010). *Complex surveys: A guide to analysis using R*. Hoboken, NJ: John Wiley & Sons.

Lumley, T. (2014). *Survey: Analysis of complex survey samples* (R package version 3.29). Retrieved from http://cran.r-project.org/web/packages/survey/index.html

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* (2nd ed.). London: Chapman and Hall.

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, CA: Sage.

O'Connell, A.A., & Liu, X. (2011). Model diagnostics for proportional and partial proportional odds models. *Journal of Modern Applied Statistical Methods, 10*(1), 139-175.

Osborne, J. W. (2011). Best practices in using large, complex samples: The importance of using appropriate weights and design effect compensation. *Practical Assessment, Research & Evaluation, 16*(12), 1-7.

Powers, D. A., & Xie, Y. (2000). *Statistical models for categorical data analysis*. San Diego, CA: Academic Press.

Thomas, S. L., & Heck, R. H. (2001). Analysis of large-scale secondary data in higher education research: Potential perils associated with complex sampling designs. *Research in Higher Education, 42*(5), 517–540.

Send correspondence to:
Xing Liu
Eastern Connecticut State University
Email: liux@easternct.edu