

# An Alternative Way of Estimating a Cumulative Logistic Model with Complex Survey Data

Phillip S. Kott and Peter Frechtel

RTI International

6110 Executive Blvd.; Rockville, MD 20852

## Introduction: Fitting a Regression Model with Complex Survey Data

The standard “design-based” framework for fitting a regression model to survey data was introduced by Fuller (1975) for linear regression and by Binder (1983) more generally. This framework treats the finite population as a realization of independent trials from a conceptual population. A maximum likelihood regression estimator could, in principle, be estimated from the finite-population values. The goal in the Fuller/Binder framework is to estimate the conceptual maximum-likelihood estimator, or its limit as the population grows arbitrarily large, from survey data. Skinner (1989) refers to this as the “pseudo-maximum-likelihood” approach.

Kott (2018) describes an alternative design-sensitive (robust model-based) approach to estimating regression models with complex survey data. Following Kott (2007), the *standard model* is defined in this approach in the following manner:

$$y_k = f(\mathbf{x}_k^T \boldsymbol{\beta}) + \varepsilon_k, \text{ where } E(\varepsilon_k | \mathbf{x}_k) = 0. \quad (1)$$

Although apparently very general, there is key restriction imposed by the standard model in equation (1):  $E(\varepsilon_k) = 0$  no matter the value of  $\mathbf{x}_k$ . This assumption can fail and the standard model not be appropriate in the population being analyzed. For example, suppose  $y_k = x_k^2$  in the population. The linear model  $y_k = \alpha + \beta x_k + \varepsilon_k$  when fit to the population fails as a standard model because  $E(\varepsilon_k | \mathbf{x}_k) \neq 0$ .

In the *extended model*,  $E(\varepsilon_k | \mathbf{x}_k) = 0$  in equation (1) is replaced by  $E(\mathbf{x}_k \varepsilon_k) = \mathbf{0}$ . Unlike the standard model, the more general extended model rarely fails. Indeed, in the above example,  $\beta = \text{Cov}(x_k^2, x_k) / \text{Var}(x_k)$  and  $\alpha = E(x_k^2) - \beta E(x_k)$  so long as  $x_k$  the first three central moments of  $x_k$  are finite.

With an independent identically distributed (*iid*) population  $U$  of  $N$  elements, it is easy to see that

$$p \lim \left\{ N^{-1} \sum_U \left[ y_k - f(\mathbf{x}_k^T \boldsymbol{\beta}) \right] \mathbf{x}_k \right\} = \mathbf{0}$$

under the extended model. Given a complex sample  $S$  with weights  $\{w_k\}$ , each (nearly) equal to the inverse of the corresponding element’s selection probability,

$$p \lim \left\{ N^{-1} \sum_S w_k \left[ y_k - f(\mathbf{x}_k^T \boldsymbol{\beta}) \right] \mathbf{x}_k \right\} = \mathbf{0} \quad (2)$$

under mild conditions on the sampling design. The parenthetical “nearly” needs to be added when the weights include adjustments for unit nonresponse or coverage errors in the frame which the analysts assumes have been accounted for in an asymptotically unbiased manner. Calibration weight adjustments for statistical efficiency are another reason to add “nearly.”

Whether the standard or extended model is assumed to hold in the population, solving for  $\mathbf{b}$  in the *weighted estimating equation* (Godambe and Thompson 1974)

$$\sum_S w_k [y_k - f(\mathbf{x}_k^T \mathbf{b})] \mathbf{x}_k = \mathbf{0} \quad (3)$$

provides a consistent estimator for  $\boldsymbol{\beta}$  under mild conditions.

The pseudo-maximum-likelihood estimating equation in Binder is

$$\sum_S w_k \frac{f'(\mathbf{x}_k^T \mathbf{b})}{v_k} [y_k - f(\mathbf{x}_k^T \mathbf{b})] \mathbf{x}_k = \mathbf{0}.$$

For logistic, Poisson, and ordinary least squares (OLS) linear regression,  $f'(\mathbf{x}_k^T \mathbf{b})/v_k = 1$ . This equality may not hold for general least squares (GLS) linear regression, however even when the elements are uncorrelated. It also need not hold for a cumulative logistic regression model.

The cumulative logistic model is a multinomial logistic regression model for  $L$  categories with a natural ordering (e.g., always, frequently, sometimes, never). Being in the first category is assumed to fit a logistic model. Being in either the first or second category is assumed to fit a logistic model. Being in the first, second, or third category is assumed to fit a logistic model, and so forth.

The *general cumulative logistic model* is (splitting out the intercept from the rest of the covariates)

$$E(y_{\ell k} | \mathbf{x}_k) = \frac{\exp(\alpha_\ell + \mathbf{x}_k^T \boldsymbol{\beta}_\ell)}{1 + \exp(\alpha_\ell + \mathbf{x}_k^T \boldsymbol{\beta}_\ell)} \quad \text{for } \ell = 1, \dots, L-1,$$

where  $y_{\ell k} = 1$  when  $k$  is in one of the first  $\ell$  categories, 0 otherwise. The *parallel-lines assumption* is that  $\boldsymbol{\beta}_\ell = \boldsymbol{\beta}$  for all values of  $\ell$  less than  $L$  with each such value having its own intercept ( $\alpha_\ell$ ). The cumulative logistic model under the parallel-lines assumption is often called a *proportional-odds model*. We will call it the “simple cumulative logistic model,” although it is more commonly referred to as *the* cumulative logistic model.

Finding the  $a_\ell$  and  $\mathbf{b}_\ell$  that satisfy the estimating equation:

$$\sum_{k \in S} w_k \left[ y_{\ell k} - \frac{\exp(a_\ell + \mathbf{x}_k^T \mathbf{b}_\ell)}{1 + \exp(a_\ell + \mathbf{x}_k^T \mathbf{b}_\ell)} \right] \begin{bmatrix} 1 \\ \mathbf{x}_k \end{bmatrix} = \mathbf{0} \quad \text{for } \ell = 1, \dots, L-1 \quad (4)$$

can be used for estimating the general cumulative logistic model. This is *not* the pseudo-maximum-likelihood estimating equation in the *surveylogistic* routine in SAS/STAT 14.1 (SAS Institute Inc. 2015; An (2002), p. 7 has the multivariate pseudo-maximum-likelihood estimating equation), the *logistic* routine in SUDAAN 11 (Research Triangle Institute 2012) or the *gologit2* routine in STATA (Williams 2005) for the simple cumulative logistic model. Only the STATA routine allows the  $\mathbf{b}_\ell$  to vary.

Given  $L$  nominal categories and complex survey data, SAS and SUDAAN *can* fit the *general logistic regression*,

$$E(y_{\ell k} | \mathbf{x}_k) = \frac{\exp(\alpha_\ell + \mathbf{x}_k^T \boldsymbol{\beta}_\ell)}{1 + \sum_{j=1}^{L-1} \exp(\alpha_j + \mathbf{x}_k^T \boldsymbol{\beta}_j)} \quad \text{for } \ell = 1, \dots, L-1,$$

with  $y_{\ell k} = 1$  when  $k$  in  $\ell^{\text{th}}$  category, 0 otherwise, but this is not the same thing as the general *cumulative* logistic model, which these programs cannot estimate with complex survey data.

In what follows, we introduce a modest example of a simple cumulative logistic model. Given complex survey data, we fit the model both with the pseudo-maximum-likelihood technique and with equation (4). The latter is accomplished by repeating each observation  $L - 1$  times with an iteration for each version of  $y_{\ell k}$  except the last (all  $L - 1$  iterations are in the same probability sampling unit). We will call this fitting method the “design-sensitive” technique, even though, strictly speaking, the pseudo-maximum-likelihood approach is also design sensitive. We then go on to test the parallel-lines assumption and provide a discussion.

### A Simple Example

The National Survey on Drug Use and Health (NSDUH) is an annual survey of the civilian, noninstitutionalized population aged 12 or older living in the United States. Using NSDUH data from 2006 to 2010, we focus on a survey question given to adolescents (12-17) who received depression treatment in the past year:

During the past 12 months, how much has treatment or counseling helped you?

The viable responses were: Not at all (1); A little (2); Some (3); A lot (4); or Extremely (5).

We discarded missing and invalid responses both to this question and to the question of whether the respondent received depression treatment in the past year. We will return to this practice in a later section.

Using SAS, we estimated the following simple cumulative logistic model:

$$E(y_{\ell k} | \mathbf{x}_k) = \frac{\exp(\alpha_{\ell} + meds_k \beta)}{1 + \exp(\alpha_{\ell} + meds_k \beta)} \quad \text{for } \ell = 1, \dots, L-1, \quad (5)$$

where  $meds = 1$  when respondent  $k$  was taking medication for depression (0 otherwise), with both pseudo-maximum-likelihood and the design-sensitive technique. For pseudo-maximum-likelihood estimation, we reversed the order of the responses with  $y_{1k} = 1$  when  $k$  responded that treatment (or counseling) helped extremely,  $y_{2k} = 1$  when  $k$  responded that treatment helped a lot,  $y_{3k} = 1$  when  $k$  responded that treatment helped more than a little, and  $y_{4k} = 1$  when  $k$  responded that treatment helped at least a little. Finally,  $y_{5k} = 1$  when  $k$  responded that treatment did not help at all. In SAS, the mean dependent variable  $Y$  was set equal to 1 when treatment helped extremely, to 2 when treatment helped a lot, ..., and to 5 when treatment didn't help at all.

For the design-sensitive technique, we created four iterations of  $k$  in a new data set. In the  $i^{th}$  iteration labeled  $C = i$  in SAS, a class (categorical) variable added to the model statement, we created a dependent variable (D) equal to  $y_{ik}$  in equation (5). We needed to add `EVENT = "1"` after D in the model statement because we were modeling when  $D = 1$ .

SAS code for both estimation techniques are in the appendix. The NSDUH data set we used had 60 variance strata with two variance primary sampling units (PSUs) in each and analysis weights based on the probabilities of selection and unit response.

The parameter estimates from our pseudo-maximum-likelihood and design-sensitive SAS runs are displayed in Tables 1 and 2, respectively. In Table 1,  $Intercept=i$  is the estimate of  $\alpha_{ik}$  in equation (5). The sum of the  $Intercept$   $C=i$  in Table 2 also estimates  $\alpha_{ik}$  for  $i = 1, 2, \text{ or } 3$ , while  $\alpha_{4k}$  is estimated by the Intercept in Table 2 minus  $(C=1 + C=2 + C=3)$ . More obviously,  $meds$  in both tables estimates  $\beta$ .

In all cases, estimates of the same parameter from the two tables are close. The percent increase in every level of satisfaction with treatment due to having taken drugs for depression (the estimate for  $\beta$ ) is roughly 45% (in our discussion of the results of the logistic regressions, we treat differences of the log odds as equal to percent differences in the odds, even though this is only approximately true). That near equality suggests that the parallel-lines assumption is not violated by our NSDUH data.

**Table 1. Pseudo-Maximum-Likelihood Estimates for the Simple Cumulative Logistic Model**

<i>Parameter</i>		<b>Estimate</b>	<b>Standard Error</b>	<b>t Value</b>	<b>Pr &gt;  t </b>
<i>Intercept</i>	1	-2.2917	0.0913	-25.10	<.0001
<i>Intercept</i>	2	-0.7617	0.0685	-11.11	<.0001
<i>Intercept</i>	3	0.2511	0.0624	4.02	0.0002
<i>Intercept</i>	4	1.3695	0.0739	18.53	<.0001
<i>meds</i>		0.4516	0.0965	4.68	<.0001

NOTE: The degrees of freedom for the t tests is 60.

**Table 2. Design-Sensitive Estimates for the Simple Cumulative Logistic Model**

<i>Parameter</i>		<b>Estimate</b>	<b>Standard Error</b>	<b>t Value</b>	<b>Pr &gt;  t </b>
<i>Intercept</i>		-0.3591	0.0583	-6.16	<.0001
<i>C</i>	1	-1.9329	0.0592	-32.63	<.0001
<i>C</i>	2	-0.4039	0.0356	-11.33	<.0001
<i>C</i>	3	0.6087	0.0392	15.52	<.0001
<i>meds</i>		0.4498	0.0955	4.71	<.0001

NOTE: The degrees of freedom for the t tests is 60.

The parallel-lines assumption can be tested directly by adding a class variable M to the design-sensitive data set with M = 1 when C = 1 and *meds* = 1, M = 2 when C = 2 and *meds* = 1, M = 3 when C = 3 and *meds* = 1, and M = 4 otherwise.

When added to the model statement in SAS, the class variable M captures the differing impacts of taking medication for depression in the previous year had on the levels of satisfaction with treatment. For example, the estimated percent increase in the odds of being extremely pleased by treatment due to having taken drugs for depression during the year is, according to Table 3, .3816 (from *meds*) plus .717 (from M = 1) or 45.33%. The other percent increases are lower, but none are significantly different from the others. We see that from the extremely low F value for M in Table 4. In addition, none of the t-values for an M is Table 3 is significant at even the .5 level.

**Table 3. Estimating the General Cumulative Logistic Model**

<b>Parameter</b>		<b>Estimate</b>	<b>Standard Error</b>	<b>t Value</b>	<b>Pr &gt;  t </b>
<i>Intercept</i>		-0.2919	0.1270	-2.30	0.0251
<i>C</i>	1	-1.9636	0.0806	-24.37	<.0001
<i>C</i>	2	-0.4104	0.0440	-9.33	<.0001
<i>C</i>	3	0.6202	0.0490	12.66	<.0001
<i>Meds</i>		0.3816	0.1452	2.63	0.0109
<i>M</i>	1	0.0717	0.1273	0.56	0.5754
<i>M</i>	2	0.0234	0.0652	0.36	0.7215
<i>M</i>	3	-0.0236	0.0719	-0.33	0.7439

NOTE: The degrees of freedom for the t tests is 60.

**Table 4. F tests for the General Cumulative Logistic Model**

Effect	F Value	Num DF	Den DF	Pr > F
C	280.39	3	58	<.0001
<i>Meds</i>	6.91	1	60	0.0109
M	0.16	3	58	0.9239

### Discussion

When there is more than one explanatory variable in the cumulative logistic model then each one needs to be tested like *meds* was in the previous section by adding an analogous class variable for each. A general F test can be used for testing whether every class variable is not significant (say at the .05 level). A better approach with complex survey data may be to follow Korn and Graubard (1990) and use the simple Bonferroni-adjusted *t*-test. For significance at the .05 level, one would compute the *t*-values for every tested component of each added class variable (there are three such in Table 3), then compare the *p*-value of the smallest of these to .05/the number of components tested.

An advantage of the design-sensitive approach to fitting a simple cumulative logistic model to the pseudo-maximum-likelihood approach is not apparent with our NSDUH data. When the parallel-lines assumption doesn't hold and an extended model is being fit, satisfying the first "equation" in (4) assures us that

$$\sum_{k \in S} w_k y_{\ell k} = \sum_{k \in S} w_k \frac{\exp(a_{\ell} + \mathbf{x}_k \mathbf{b})}{1 + \sum_{j=1}^{L-1} \exp(a_j + \mathbf{x}_k \mathbf{b})} \quad \text{for } \ell = 1, \dots, L-1. \quad (6)$$

When  $\mathbf{x}_k$  is produced from a single categorical variable, equation (6) assures that the weighted mean of  $y_{\ell k}$  for each  $\mathbf{x}$ -category and cumulative level  $\ell$  equals its predicted value described by

$$\hat{y}_{\ell k} = \exp(a_{\ell} + \mathbf{x}_k^T \mathbf{b}) / [1 + \sum_{j=1}^{L-1} \exp(a_j + \mathbf{x}_k^T \mathbf{b})],$$

which is a reasonable property. Equation (4) is simply an extension of the property to more general  $\mathbf{x}_k$ .

In our NSDUH example, although not generally, using the design sensitive approach was more efficient than using the PSEUDO-MAXIMUM-LIKELIHOOD approach. This can be seen by comparing the *t*-values of *meds* (the inverses of their respective estimated coefficients of variation) in Tables 1 and 2. When we to ignore the analysis weights, the strata, and the clustering (by setting the weights and strata to 1, and treating each respondent as a primary sampling unit), this result reverses as expected. The point here is that pseudo-maximum likelihood with complex survey data is indeed "pseudo."

Finally, the data set we created dropped responding observations with missing values of the dependent and *meds* variables. When fitting the *extended* model, this is only valid (i.e., resulting estimates are asymptotically unbiased) when an in-scope respondent – an adolescent who had treatment for depression in the previous year – being dropped occurred completely at random. When fitting the *standard* model, the probability of being dropped can be a function only of whether an in-scope adolescent has taken medication for depression in the previous year but nothing else. This suggests it may have been prudent to add variables to the model that are never missing even when they are not significant. If we add class variables for age, sex, race/ethnicity, urbanicity, and family income (all of which have values imputed for them when missing in the NSDUH) to our simple cumulative logistic model, none are significant at the .05 level. The major results do not change meaningfully (the estimate for  $\beta$  increases from roughly .45 to 50), although that the *t*-value for *meds* using the design-sensitive approach ( $b_{meds} = .4948$ ;  $t_{meds} = 5.49$ ) is slightly smaller than that from using the pseudo-maximum-likelihood approach ( $b_{meds} \approx .4987$ ;  $t_{meds} = 5.52$ ).

More importantly, if the *standard* model for the simple cumulative logistic model,

$$y_{\ell k} = \frac{\exp(\alpha_{\ell} + \mathbf{x}_k \boldsymbol{\beta})}{1 + \sum_{j=1}^{L-1} \exp(\alpha_j + \mathbf{x}_k \boldsymbol{\beta})} + \varepsilon_{\ell k} \quad \text{with } E(\varepsilon_{\ell k} | \mathbf{x}_k) = 0 \quad \text{for } \ell = 1, \dots, L-1,$$

is correct (and we have no reason to think otherwise), then dropping the observations as we did does not cause a bias when the probability of an in-scope adolescent being dropped from the analysis is a function of all the explanatory

variables (the components of  $\mathbf{x}_k$ ) in the model including the non-significant ones but not the dependent variable ( $y_{\ell k}$ ). This is because the analysis weight for the data set after deletions, call it  $R$ , is  $w_k / g(\mathbf{x}_k)$ , where  $g(\mathbf{x}_k)$  is the probability that an in-scope sample respondent remained in the data set after the deletions. When the standard model holds, the analysis weight in the estimating equation:

$$\sum_{k \in R} [w_k / g(\mathbf{x}_k)] \left[ y_{\ell k} - \frac{\exp(a_\ell + \mathbf{x}_k \mathbf{b})}{1 + \sum_{j=1}^{L-1} \exp(a_j + \mathbf{x}_k \mathbf{b})} \right] \begin{bmatrix} 1 \\ \mathbf{x}_k \end{bmatrix} = \mathbf{0} \quad \text{for } \ell = 1, \dots, L-1,$$

can be multiplied by any function of  $\mathbf{x}_k$ , such as  $g(\mathbf{x}_k)$  (noting that  $[w_k / g(\mathbf{x}_k)]g(\mathbf{x}_k)$  is  $w_k$ ), and the  $a_\ell$  and  $\mathbf{b}$  that solves the revised estimating equation will remain a consistent estimator for  $\alpha_\ell$  and  $\boldsymbol{\beta}$ . This is because  $E[g(\mathbf{x}_k)\varepsilon_k] = 0$  and  $E[g(\mathbf{x}_k)\varepsilon_k\mathbf{x}_k] = \mathbf{0}$  under the standard model.

### Appendix

/\* PML is a data set of adolescents NSDUH respondents in the 2006 to 2010 survey years who reported having treatment for depression and whether they had taken drugs for depression. Variables include:

Y = 1 treatment was extremely helpful; Y = 2 treatment helped a lot; y = 3 some; Y = 4 a little; Y = 5 not at all  
 meds = 1 had taken drugs for depression, 0 otherwise  
 VESTR variance stratum  
 VEPSU variance primary sampling unit  
 IDNUM respondent identification number  
 ANALWT the analysis weight

This set is used for pseudo-maximum-likelihood estimation of the simple cumulative logistic model and to create the DS\_SIMPLE data set, which is used for design-sensitive estimation of the simple cumulative logistic model and DS\_GENERAL data set, which is used for design-sensitive estimation of the general cumulative logistic model. \*/  
 DATA DS\_SIMPLE; SET PML; BY VESTR VEPSU IDNUM;

D = 0;  
 C = 1; IF Y < 2 THEN D = 1; OUTPUT;  
 C = 2; IF Y < 3 THEN D = 1; OUTPUT;  
 C = 3; IF Y < 4 THEN D = 1; OUTPUT;  
 C = 4; IF Y < 5 THEN D = 1; OUTPUT;

DATA DS\_GENERAL; SET DS\_SIMPLE;  
 M = 4;  
 IF C = 1 AND MEDS = 1 THEN M = 1;  
 IF C = 2 AND MEDS = 1 THEN M = 2;  
 IF C = 3 AND MEDS = 1 THEN M = 3;

/\*The PROC below is used to produce Table 1 \*/  
 PROC SURVEYLOGISTIC DATA = PML; CLUSTER VEPSU;  
 MODEL Y = MEDS;  
 RUN;

STRATA VESTR; WEIGHT ANALWT;

/\*The PROC below is used to produce Table 2 \*/  
 PROC SURVEYLOGISTIC DATA = DS\_SIMPLE; CLASS C;  
 CLUSTER VEPSU;  
 MODEL D(EVENT = '1') = C MEDS;  
 ANALWT; RUN;

STRATA VESTR; WEIGHT

```
/*The PROC below is used to produce Tables 3 and 4*/  
PROC SURVEYLOGISTIC DATA =DS_GENERAL; CLASS M C;  
CLUSTER VEPSU ;  
MODEL D(EVENT = '1') = C MEDS M ;  
STRATA VESTR; WEIGHT ANALWT; RUN;
```

## References

- An, A. (2002). Performing logistic regression on survey data with the new SURVEYLOGISTIC procedure. In *Proceedings of the Twenty-Seventh Annual SAS® Users Group International Conference*, Cary, NC: SAS Institute Inc.  
(<http://www2.sas.com/proceedings/sugi27/p258-27.pdf>)
- Binder, D. A. (1983). On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review*, 51, 279-292.
- Fuller, W. A. (1975). Regression analysis for sample survey. *Sankhya-The Indian Journal of Statistics*, 37(Series C), 117-132.
- Godambe, V.P. & and Thompson, M.E. (1974). Estimating equations in the presence of a nuisance parameter. *Annals of Statistics*, 2, 568-571.
- Kott, P.S. (2018). A design-sensitive approach to fitting regression models with complex survey data. *Statistics Surveys*, 12, 1-17.
- Kott, P.S. (2007). Clarifying some issues in the regression analysis of survey data. *Survey Research Methods*. 1, 11–18.
- Korn, E. L. & Graubard, B. I. (1990). Simultaneous testing of regression coefficients with complex survey data: Use of Bonferroni *t* statistics. *American Statistician*, 44, 270-276.
- Research Triangle Institute (2012). *SUDAAN Language Manual*, Volumes 1 and 2, Release 11. Research Triangle Park, NC: Research Triangle Institute.
- SAS Institute Inc. (2015). *SAS/STAT® 14.1 User's Guide*. Cary, NC: SAS Institute Inc.
- Skinner, C. J. (1989). Domain means, regression and multivariate analysis. In Skinner, C. J., Holt, D. and Smith, T. M. F. eds. *Analysis of Complex Surveys*. Chichester: Wiley, 59-87.
- Williams, R. (2005). *Gologit2: A Program for Generalized Logistic Regression/ Partial Proportional Odds Models for Ordinal Variables*. Retrieved January 3, 2016 (<http://www.nd.edu/~rwilliam/stata/gologit2.pdf>).