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

Phillip S. Kott

pkott@rti.org

with Peter Frechtel

Outline

What Does Fitting a Regression Model to Survey Data Mean?

- The Fuller-Binder Approach
- An Alternative Design-Sensitive Approach

The (General) Cumulative Logistic Model for Ordered Categories

Fitting the Cumulative Logistic Model

Testing the “Parallel lines” Assumption

A Simple Example in SAS

Discussion

What Does Fitting a Regression Model to Survey Data Mean?

Fuller (*Sankhya*,1975) for linear regression and then Binder (ISR, 1983) more generally treat the finite population as a realization of independent trials from a conceptual population.

They treat the maximum likelihood estimator that, in principle, could be estimated from the finite-population as the *target*, and then try to estimate that target with the complex sample drawn from the finite population.

That is not what most analysts think they are estimating.

The Design-Sensitive (Robust Model-based) Approach

(Kott, *Statistics Surveys* 2017) points out that *methods* developed by Fuller and Binder –

- fitting weighted estimating equations,
- robust "sandwich" variance estimators with an adjustment for stratified samples –

could be used with an analyst's *Standard Model*:

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

Or, if that failed, than an *Extended Model* where only $E(\varepsilon_k\mathbf{x}_k) = \mathbf{0}$.

The Estimating Equation

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

$$N^{-1}\sum_S w_k \left[y_k - f(\mathbf{x}_k{}^T\boldsymbol{\beta}) \right]\mathbf{x}_k \to \mathbf{0}.$$

We insert the w_k in case $\mathrm{E}(\varepsilon_k \,|\mathbf{x}_k, w_k) \neq 0$
(the weights are not ignorable)

Solving for $\mathbf{b}$ in $\sum_S w_k[y_k - f(\mathbf{x}_k{}^T\boldsymbol{\beta})]\mathbf{x}_k = \mathbf{0}$

provides a consistent estimator for $\boldsymbol{\beta}$

even if only $\mathrm{E}(\varepsilon_k\mathbf{x}_k) = \mathbf{0}$!!

"Design-based" variance estimates work as well.

A Distinction Without a Difference?

The *pseudo-maximum-likelihood* (*PML*) estimating equation of Binder is

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

For logistic, Poisson, and OLS regression, $f'(\mathbf{x}_k{}^T\boldsymbol{\beta})/v_k = 1$

Problems can arise for multi-equation systems.

The Cumulative Logistic Model

This is a multinomial logistic regression model for ordered data,

where there are 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 …

The General CLM

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

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

and $y_{\ell k} = 1$ if k is in one of the first ℓ categories, 0 otherwise.

When $\boldsymbol{\beta}_\ell = \boldsymbol{\beta}$ for all $L - 1$ categories, but each has its own intercept, the (simple) *cumulative logistic model* is often called the *proportional-odds model*.

The assumption that $\boldsymbol{\beta}_\ell = \boldsymbol{\beta}$ is called *the parallel-lines assumption*.

The General Logistic Model

The general logistic model for L *nominal* categories is

$$E(y_{\ell k} \mid \mathbf{x}_k) = \frac{\exp(\alpha_\ell + \mathbf{x}_k \boldsymbol{\beta}_\ell)}{1 + \sum_j^{L-1} \exp(\alpha_j + \mathbf{x}_k \boldsymbol{\beta}_j)} \quad \text{for} \quad \ell = 1, \ldots, L-1,$$

and $y_{\ell k} = 1$ if k is in category ℓ, 0 otherwise.

It is NOT the general CLM.

Unlike the general CLM, SAS and SUDAAN can be used to fit the general logistic model with complex survey data.

The Extended-Model Estimating Equations

To estimate parameters with this alternative estimating equation (and to test equality of the $\mathbf{b}_\ell$) using design-based software:

Repeat each record $L-1$ times,
and use the binary logistic regression routine,
creating a new dependent variable
and adding dummies for the equation-specific intercepts,
perhaps with a class variable.

Treating the $L-1$ equations from the same record as if they were from the same PSU captures any correlation between the equations.

A Simple Example

A NSDUH survey question to adolescents who received depression treatment in the past year (2006-2010):

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

The response set:

Not at all (ℓ =5)

A little (ℓ =4)

Some (ℓ =3)

A lot (ℓ =2)

Extremely (ℓ =1)

The Cumulative Logistic Model

$$E(y_{\ell k} \mid \mathbf{x}_k) = \frac{\exp(\alpha_\ell + meds_k \beta_\ell)}{1 + \exp(\alpha_\ell + meds_k \beta_\ell)} \quad \text{for} \quad \ell = 1, \ldots, L-1,$$

meds = 1 if taking medication for depression (0 otherwise).

With the design-sensitive approach for the proportional odds model:

$$\sum_{k \in S} w_k y_{\ell k} = \sum_{k \in S} w_k \frac{\exp(\hat{\alpha}_\ell + meds_k \hat{\beta})}{1 + \exp(\hat{\alpha}_\ell + meds_k \hat{\beta})} \quad \text{for} \quad \ell = 1, \ldots, 4;$$

The predicted fraction who benefited from meds at a level equals the estimated fraction who benefited.

The above equation may not hold with PML estimation.

SAS CODE

```
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;
```

SAS CODE

```
PROC SURVEYLOGISTIC DATA = PML; CLUSTER VEPSU ;
MODEL Y =  MEDS;
STRATA VESTR; WEIGHT ANALWT; RUN;

PROC SURVEYLOGISTIC DATA = DS_SIMPLE ; CLASS  C;
CLUSTER VEPSU ;
MODEL D(EVENT = '1') = C MEDS;
STRATA VESTR; WEIGHT ANALWT; RUN;

 PROC SURVEYLOGISTIC DATA =DS_GENERAL; CLASS M C;
CLUSTER VEPSU ;
MODEL D(EVENT = '1') = C  MEDS  M ;
STRATA VESTR; WEIGHT ANALWT; RUN;
```

Results: Simple CLM

Pseudo Maximum Likelihood

Parameter		Estimate	t Value	Pr > |t|
Intercept	1	-2.292	-25.10	<.001
Intercept	2	-0.762	11.11	<.001
Intercept	3	0.251	4.02	<.001
Intercept	4	1.370	18.53	<.001
meds		0.452	4.68	<.001

Design-Sensitive

Parameter		Estimate	t Value	Pr > |t|
Intercept		-0.359	-6.16	0.001
C	1	-1.933	-32.63	<.001
C	2	-0.404	-11.33	<.001
C	3	0.609	15.52	<.001
meds		0.450	4.71	<.001

The Intercepts

Design Sensitive		*PML*
Intercept + C1	≈	Intercept1
Intercept + C2	≈	Intercept2
Intercept + C3	≈	Intercept3
Intercept – C1 – C2 – C3	≈	Intercept4

Result: General CLM

Effect	F Value	Pr > F
C	280.39	<.0001
meds	11.84	0.0011
M	0.16	0.9239

Parameter		Estimate	t Value	Pr > |t|
Intercept		-0.4350	-3.19	0.002
⋮				
meds		0.5247	3.44	0.001
M	1	-0.0715	-0.65	0.516
M	2	-0.0236	-0.33	0.744
M	3	0.0234	0.36	0.722

Discussion

PML did not lead to a more efficient estimate than the design-sensitive approach under NSDUH's complex sampling design, while the latter is a more natural extension of simple weighted means.

Korn and Graubard (*Am. Stat.* 1990) show that using a Bonferroni-adjusted *t*-test may be better than an *F*-test with complex survey data. (Either way, M was not significant.)

Listwise deletion is justified under the *standard* model when whether an observation is deleted is a function of the covariates in the model.

That is a reason to add covariates that are not significant like sex, race/ethnicity, age, urbanicity, and family income.