Stereotype Ordinal Regression

Mark Lunt, ARC Epidemiology Unit, University of Manchester

June 12, 2001

Introduction

There are a number of reasonable approaches to analysing an ordinal outcome variable. One common approach, known as the Proportional Odds (PO) Model, is implemented in Stata as `ologit`. If the assumptions of the PO model are not satisfied, an alternative is to treat the outcome as categorical, rather than ordinal, and use multinomial logistic regression (`mlogit`) in Stata. This insert describes an alternative form of ordinal regression model, the Stereotype Ordinal Regression (SOR) Model, which can be thought of as imposing ordering constraints on a multinomial model. The multinomial model provides the best possible fit to the data, at the cost of a large number of parameters which can be difficult to interpret. Stereotype regression aims to reduce the number of parameters by imposing constraints, without reducing the adequacy of the fit.

Distinguishability & Dimensionality

In introducing his Stereotype Ordinal Regression Model, Anderson ([1]) justified it in terms of the concepts of distinguishability and dimensionality.

Dimensionality A fundamental assumption of the Grouped Continuous Model is that given a set of predictor variables x_{ij}, where i indexes the subject $1 \ldots n$ and j indexes the variable, $1 \ldots p$, the same combination of variables, $\sum_{j=0}^{p} (x_{ij}\beta_j)$, can be used to distinguish between all levels of the outcome variable. If, however, one combination can distinguish between levels 1 and 2, but a different one is required to distinguish between levels 2 and 3, the relationship is two-dimensional.

Distinguishability Two outcome categories are *indistinguishable* with respect to a variable x_j if x_j is not predictive between the two categories.

The PO model, implemented in `ologit` assumes that the association between the predictor variables x_j and the outcome variable y is one-dimensional. In addition, if there are k possible outcome categories, it assumes that the odds ratio for being in group s or higher, relative to the odds of being in group $s-1$ or lower, associated with each variable x_j is the same for all $s, 2 \leq s \leq k$. If either of these assumptions are untrue, the PO model is inappropriate.

The Stereotype Ordinal Regression Model

A full multinomial model model is of the form

$$\text{pr}(y_i = s | x_{i1} \ldots x_{ip}) = \frac{\exp\left(\beta_{0s} + \sum_{j=1}^{p} x_{ij}\beta_{sj}\right)}{\sum_{t=1}^{k} \exp\left(\beta_{0t} + \sum_{j=1}^{p} x_{ij}\beta_{tj}\right)}$$

As it stands, this model is not identified: adding a fixed constant to every β will give exactly the same predicted probabilities. To identify the model, constraints need to be placed on the parameters. Commonly β_k and β_{0k} are constrained to be 0. If the x_j variables are all categorical, this model is then saturated.

The multinomial model assumes that different linear combinations of the predictor variables are required to discriminate between different pairs of levels of the outcome variable. All of these combinations need to be estimated separately. However, if the outcome is ordinal, rather than categorical, it may be that a single linear combination can discriminate between all levels. If this is the case, the above model can be simplified to

$$\text{pr}(y_i = s | x_{i1} \ldots x_{ip}) = \frac{\exp\left(\beta_{0s} + \phi_s \sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{t=1}^{k} \exp\left(\beta_{0t} + \phi_t \sum_{j=1}^{p} x_{ij}\beta_j\right)}$$

This is a one-dimensional stereotype model. In this model, the β parameters no longer differ between the different levels of the outcome. The combination that best discriminates between the outcome variables is given by $\sum_{j=1}^{p} x_{ij}\beta_j$ and the distance between the outcome levels in terms of this linear predictor is given by the ϕ_j parameters. If the relationship between the predictors and outcome is ordinal, then $\phi_1 \geq \phi_2 \geq \ldots \geq \phi_k$.

Again, constraints are needed to make the model identifiable. Anderson recommended setting $\phi_k = 0$ and $\phi_1 = 1$, but other constraints are possible.

The stereotype model can be considered as a constrained multinomial model, in which the ratios β_{sj}/β_{tj} are constant for all variables x_j. In fact, if there is only one predictor, the stereotype model is simply a re-parameterisation of the multinomial model: the goodness of fit, predicted values etc. are all identical. If there is more than one predictor variable, there are two comparisons to be made when deciding if a stereotype model is adequate: it should be compared to the null (intercept only model) to see if its ability to predict outcome is greater than would be expected by chance. This is equivalent to the χ-squared test supplied by most regression routines in stata[1]. Secondly, it should be compared to the full (multinomial) model, to see if there has been a significant loss of predictive ability when the simplifying constraints of the stereotype model were imposed. This can be thought of as a 'goodness of fit' test, analogous to the Pearson of Hosmer-Lemeshow test after logistic regression (`lfit`) or the poisson goodness of fit test (`poisgof`) after poisson regression.

---

[1] `regress` presents an equivalent F-test, but `logit`, `ologit` and `poisson`, to name a few, present χ-squared tests.

Two Simple Examples

Consider the nausea data presented by Farewell [2] and re-analysed by Anderson. The file nausea.dta contains this data, with a single entry for each of the 219 subjects. Typing

```
soreg nausea treat
```

will produce the following output:

```
. soreg nausea treat
iteration 0: Log Likelihood = -371.4567
iteration 1: Log Likelihood = -371.4567
iteration 2: Log Likelihood = -371.4567

Stereotype Logistic Regression                     Number of obs =        219

Comparison to null model                           LR Chi2(5)    =      18.01
                                                   Prob > chi2   =     0.0029

Comparison to full model                           LR Chi2(0)    =       0.00
                                                   Prob > chi2   =          .

------------------------------------------------------------------------------
         |      Coef.    Std. Err.       z     P>|z|      [95% Conf. Interval]
---------+--------------------------------------------------------------------
   phi11 |         1           .         .       .              .           .
   phi21 |   .9101804    .5096764     1.786    0.074      -.088767    1.909128
   phi31 |   .6789914    .6490182     1.046    0.295      -.593061    1.951044
   phi41 |  -.1123247    .4708013    -0.239    0.811     -1.035078     .810429
   phi51 |  -.6699209    .6525803    -1.027    0.305     -1.948955    .6091131
   phi61 |   (dropped)
  beta11 |  -1.087051    .5215595    -2.084    0.037     -2.109289   -.0648137
------------------------------------------------------------------------------
beta1 = treat
```

The ϕ_i variables are labeled `phi11` to `phi61` (in a two dimensional model, there would be two sets of ϕ_i variables, `phi11` to `phi61` and `phi12` to `phi62`). The predictor variables for the first dimension are labeled `beta11` to `betak1`, for the second dimension `beta12` to `betak2` etc.

There are no standard errors for phi11 or phi61, since they are constrained to be 1 and 0 respectively. Compared to the null (intercept only model), the stereotype model fits significantly better (chi2 = 18.0 on 5 d.f., p = 0.003). However, there is no difference in the fit when compared to the full (multinomial) model, since the SOR model is simply a re-parameterisation of the multinomial model.

A more complex example, also analysed by Anderson, concerns the prognosis for suffers of backpain originally presented by Doran and Newell [3]. In this case, the outcome was scored from 1 ("worse") to 6 ("complete relief"), and there were three predictor variables, x1, x2 and x3. Fitting the stereotype model to this data gives the following output:

```
. soreg pain x1 x2 x3
iteration 0: Log Likelihood = -151.1720
iteration 1: Log Likelihood = -151.4538
iteration 2: Log Likelihood = -151.5506
iteration 3: Log Likelihood = -151.5501
iteration 4: Log Likelihood = -151.5501
iteration 5: Log Likelihood = -151.5501

Stereotype Logistic Regression                    Number of obs =       101

Comparison to null model                          LR Chi2(7)    =     39.96
                                                  Prob > chi2   =    0.0000

Comparison to full model                          LR Chi2(8)    =      4.09
                                                  Prob > chi2   =    0.8494

        -------------------------------------------------------------------------
              |      Coef.   Std. Err.       z     P>|z|      [95% Conf. Interval]
        ---------+---------------------------------------------------------------
        phi11 |         1          .          .        .              .           .
        phi21 |  .3092677   .1294594      2.389    0.017        .0555321    .5630034
        phi31 |  .3465159   .1308123      2.649    0.008        .0901285    .6029034
        phi41 |  .5097036   .1608969      3.168    0.002        .1943514    .8250558
        phi51 |  .1414456   .0933218      1.516    0.130       -.0414618     .324353
        phi61 |  (dropped)
        beta11|  5.374471   2.000401      2.687    0.007        1.453757    9.295185
        beta21|   3.08102   1.055979      2.918    0.004        1.011339      5.1507
        beta31|  2.712501   1.180846      2.297    0.022        .3980843    5.026917
        -------------------------------------------------------------------------

beta1 = x1
beta2 = x2
beta3 = x3
```

In this case, the stereotype model requires 8 fewer parameters than the full multinomial model, although the multinomial model does not fit significantly better ($\chi^2 = 4.1$ on 8 d.f., $p = 0.85$). However, the stereotype model does fit significantly better than the null model ($\chi^2 = 40.0$ on 7 d.f., $p < 0.0001$). Thus, the simplification can be considered successful. However, both of the above models can be simplified further, using constraints.

Constraints

Constraints play two important roles in Anderson's ordinal regression strategy. Firstly, it is necessary to impose constraints on the parameters, in order to make the model identifiable. In addition, questions of distinguishability can be addressed through constraints.

Identifiability

We have already seen that the basic stereotype model is not identified (multiplying all of the β parameters by a constant c and dividing all the ϕ parameters by the same constant

4

would produce exactly the same fitted values). To identify the model, constraints must be applied. Anderson recommended setting ϕ_k to 0 and ϕ_1 to 1, and these constraints are implemented in soreg by default. However, the defaults can be easily overridden by user-defined constraints. For example, consider the analysis by Greenland [4] of the pneumoconiosis data from Ashford [5]. Greenland fitted the constraints $\phi_1 = 0$ and $\phi_2 = 1$, and estimated ϕ_3. The fitting of this model is shown below.

```
. constraint define 1 phi11 = 0

. constraint define 2 phi21 = 1

. soreg pneum lyears, c(1 2)

 ( 1)  phi11 = 0.0
 ( 2)  phi21 = 1.0
iteration 0: Log Likelihood = -204.4344
iteration 1: Log Likelihood = -204.4344
iteration 2: Log Likelihood = -204.4344

Stereotype Logistic Regression                    Number of obs =        371

Comparison to null model                          LR Chi2(2)    =      96.29
                                                  Prob > chi2   =     0.0000

Comparison to full model                          LR Chi2(0)    =       0.00
                                                  Prob > chi2   =          .

-------------------------------------------------------------------------------
         |      Coef.   Std. Err.       z    P>|z|      [95% Conf. Interval]
---------+---------------------------------------------------------------------
   phi11 |   (dropped)
   phi21 |          1          .         .        .            .           .
   phi31 |     1.4166   .3649241     3.882    0.000     .7013616    2.131838
  beta11 |   2.165373   .4574869     4.733    0.000     1.268715    3.062031
-------------------------------------------------------------------------------
beta1 = lyears
```

It can be seen that after setting the constraints, phi11 is fixed at 0 (i.e. dropped from the model) and phi21 is fixed at 1. With these constraints, the estimates for phi31 and beta11 from soreg match those given by Greenland.

A two-dimensional model requires additional constraints: again defaults are available, although they can be overridden. However, for models of greater than 2 dimensions, defaults are not provided, since the appropriate constraints will often depend on the context. If no constraints are defined in such a model, an error message will be printed and no model estimated.

Distinguishability

The ϕ parameters give a measure of the distinguishability of the various categories with respect to the predictors: if the ϕ parameters of two categories are similar, it is likely that the categories are indistinguishable. In this case, the model can be simplified by constraining the corresponding ϕ parameters to be equal. For example, consider the

nausea data analysed above. Looking at the phi parameters and their standard errors, he considered suggested that there were only two distinguishable groups, and that phi11 = phi21 = phi31, and phi41 = phi51 = phi61. We can therefore define the following constraints.

```
. constraint define 1 phi11 = phi21

. constraint define 2 phi11 = phi31

. constraint define 3 phi41 = phi51

. constraint define 4 phi41 = phi61

. constraint define 5 phi11 = 1

. constraint define 6 phi61 = 0
```

Fitting the stereotype model with these constraints applied gives the following output:

```
. soreg nausea treat, c(1/6)

 ( 1)  phi11 - phi21 = 0.0
 ( 2)  phi11 - phi31 = 0.0
 ( 3)  phi41 - phi51 = 0.0
 ( 4)  phi41 - phi61 = 0.0
 ( 5)  phi11 = 1.0
 ( 6)  phi61 = 0.0
iteration 0: Log Likelihood = -371.4567
iteration 1: Log Likelihood = -372.7759
iteration 2: Log Likelihood = -372.7759

Stereotype Logistic Regression                    Number of obs =       219

Comparison to null model                          LR Chi2(1)    =     15.37
                                                  Prob > chi2   =    0.0001

Comparison to full model                          LR Chi2(4)    =      2.64
                                                  Prob > chi2   =    0.6200

------------------------------------------------------------------------------
          |      Coef.   Std. Err.       z     P>|z|     [95% Conf. Interval]
----------+-------------------------------------------------------------------
    phi11 |          1          .         .       .             .           .
    phi21 |          1          .         .       .             .           .
    phi31 |          1          .         .       .             .           .
    phi41 |  (dropped)
    phi51 |  (dropped)
    phi61 |  (dropped)
   beta11 |  -1.244581   .3299873    -3.772   0.000     -1.891344   -.5978177
------------------------------------------------------------------------------
beta1 = treat
```

This model is significantly better than the null model, and not significantly worse than the full model. Thus, it provides a good fit to the data, despite having only one parameter. It says that the odds of being in group 1, 2 or 3 rather than 4, 5 or 6

decrease by a factor of 3.5 ($e^{1.24}$) for those treated with cisplatinum relative to those untreated. The interpretation of this model is therefore extremely simple: explaining the meaning of each of the parameters in the unconstrained model would have been far more difficult.

Note that the default constraints were redefined: this is necessary since the default constraints are only applied if *no* user-defined constraints are given. If insufficient constraints are given, the model may be estimated and the likelihood ratio tests give exactly the same results, but the parameters will have arbitrary values, which makes interpretation difficult.

Constraints can cause some difficulties. Firstly, the first predictor variable has a special role in the model-fitting process. it is therefore not possible to constrain `beta11`, `beta12` etc to be equal to another variable: these variables can, however, be constrained to a constant value. Secondly, constraining parameters to have a fixed value is better, if possible, than constraining them to be equal. For example, the pair of constraints

```
. constraint define 1 phi11 = 1

. constraint define 2 phi21 = phi11
```

should be equivalent to the constraints

```
. constraint define 1 phi11 = 1

. constraint define 2 phi21 = 1
```

However, there can be circumstances in which the model will not converge using the former constraints but will with the latter. This is a weakness of the constraint handling code within `soreg`.

Saved Results

`soreg` saves in `e()`:

Scalars

    e(N)   number of observations

    e(ll_0)   log likelihood, constant only model

    e(ll)   log likelihood, stereotype model

    e(ll_full)   log likelihood, full model

    e(df_m)   model degrees of freedom

    e(df_full)   model degrees of freedom for multinomial model

    e(k_cat)   number of categories

Macros

    e(cmd)   `soreg`

    e(chi2type)   type of model χ^2 test (``LR'')

    e(depvar)   name of dependent variable

Matrices

    e(b)   coefficient vector

    e(V)   variance-covariance matrix of the estimators

    e(cat)   category values

Methods and Formulas

Since the stereotype model is non-linear (it contains the product of parameters), it cannot be fitted by standard generalised linear model methods. The model is fitted using a technique due to Box and Tidwell [6], described by McCullagh & Nelder [7]. Briefly, a linear model containing a non-linear function, say

$$\eta = \alpha + \beta g(x|\theta)$$

can be estimated by iteratively fitting instead

$$\eta = \alpha + \beta u_i + \gamma v_i$$

where

$$
\begin{aligned}
u_i &= g(x|\theta_{i-1}) \\
v_i &= \left[\frac{\partial g}{\partial \theta}\right]_{\theta=\theta_{i-1}}
\end{aligned}
$$

and θ_{i-1} is the value of θ calculated from the $(i-1)^{th}$ iteration. Then $\theta_i = \theta_{i-1} + \hat{\gamma}/\hat{\beta}$, and the process iterates to convergence. A final iteration with $v = \hat{\beta}v$ gives the variance of θ and its covariances with α and β. Unfortunately, this iteration process is not guaranteed to converge. Since the first variable in the list has a special role in this process, it can happen that a model that will not converge can be forced to by changed the order of the predictor variables.

Further Remarks

There already exist two Stata commands, `mclgen` and `mclest` which can be used to estimate one-dimensional stereotype ordinal regression models. However, `soreg` has a number of advantages:

- `soreg` can estimate models with more than 1 dimension, which `mclest` cannot. This was the motivation for writing `soreg`: I needed to fit a two-dimensional model.

- `soreg` allows the use of constraints, which `mclest` does not.

- `mclest` does not produce standard errors for the ϕ_{ij} parameters, which are important when considering distinguishability.

- The standard errors for the β_{ij} parameters produced by `mclest` are conditional on the ϕ_{ij} parameters and hence underestimated. The standard errors produced by `soreg` allow for the uncertainty in estimating the $\phi_i j$ parameters.

It should be pointed out that whilst there are models that can be estimated with `soreg` but not `mclest`, there are also models that can be estimated with `mclest` but not `soreg`.

Care needs to be taken in interpreting the results of significance tests on Stereotype Ordinal Regression models. Unlike generalised linear models, it cannot be shown that the log likelihood follows a χ^2 distribution asymptotically. Hence the likelihood ratio chi-squared tests and corresponding p-values should be treated with circumspection.

In addition, it should be remembered that when considering the distinguishability of k groups, there are implicitly $k-1$ tests being performed. This multiple testing should be borne in mind in interpreting the resultant model. However, this is true in any case in which data-based variable selection takes place.

Finally, the standard errors reported by `soreg` differ, in some cases, from those reported by Anderson in his paper. However, the differences are small. In addition, in some cases the standard errors reported by Anderson are smaller than those reported by `mclest`, despite the fact that those reported by `mclest` are conditional on the estimated values of the phi parameters and therefore should be underestimated. The standard errors reported by `soreg` are always greater than those reported by `mclest` and are similar to those reported by Anderson. Where it is possible to constrain a multinomial model to produce the same parameterisation as a stereotype model, the standard errors are the same by both methods.

However, the significance of an individual parameter reported by the Wald test (i.e. using the standard error) may be quite different from that reported by the likelihood ratio test. This is due to the correlations between the ϕ parameters and the β parameters. The likelihood ratio tests are likely to be a more reliable guide to inference. For example, the overall test of the model in the first nausea example is highly significant, but only one of the individual parameters has a p-value of less than 0.05 (`beta11 = 0.04`).

Acknowledgements

References

[1] J. A. Anderson. Regression and ordered categorical variables. *Journal of the Royal Statistical Society, Series B.*, 46(1):1–30, 1984.

[2] V.T. Farewell. A note on regression analysis of ordinal data with variability of classification. *Biometrika*, 69:533–538, 1982.

[3] D.M.L. Doran and D.J. Newell. Manipulation in treatment of low back pain: A multicentre trial. *British Medical Journal*, 1:161–164, 1975.

[4] Sander Greenland. Alternative models for ordinal logistic regression. *Statistics in Medicine*, 13:1665–1677, 1994.

[5] J.R. Ashford. An approach to the analysis of data for semi-quantal responses in biological assay. *Biometrics*, 26:535–546, 1959.

[6] G.E.P Box and P.W. Tidwell. Transformation of the independent variables. *Technometrics*, 4:531–550, 1962.

[7] P McCullagh and J Nelder. *Generalized Linear Models*. Monographs on Statistics and Applied Probability. Chapman & Hall, 2-6, Boundary Row, London, SE1 8HN, UK., second edition, 1989.