The Stata Journal (2008)
8, Number 2, pp. 299–303

Stata tip 63: Modeling proportions

Christopher F. Baum
Department of Economics
Boston College
Chestnut Hill, MA
baum@bc.edu

You may often want to model a response variable that appears as a proportion or fraction: the share of consumers' spending on food, the fraction of the vote for a candidate, or the fraction of days when air pollution is above acceptable levels in a city. To handle these data properly, you must take account of the bounded nature of the response. Just as a linear probability model on unit record data can generate predictions outside the unit interval, using a proportion in a linear regression model will generally yield nonsensical predictions for extreme values of the regressors.

One way to handle this for response variables' values strictly within the unit interval is the logit transformation

$$y = \frac{1}{1 + \exp(-X\beta)}$$

which yields the transformed response variable y^*

$$y^* = \log\left(\frac{y}{1-y}\right) = X\beta + \epsilon$$

where we have added a stochastic error process ϵ to the model to be fitted. This transformation may be performed with Stata's `logit()` function. We can then use linear regression ([R] **regress**) to model y^*, the *logit transformation* of y, as a linear function of a set of regressors, X. If we then generate predictions for our model ([R] **predict**), we can apply Stata's `invlogit()` function to express the predictions in units of y. For instance,

```
. use http://www.stata-press.com/data/r10/census7
(1980 Census data by state)
. generate adultpop = pop18p/pop
. quietly tabulate region, generate(R)
. generate marrate = marriage/pop
. generate divrate = divorce/pop
. generate ladultpop = logit(adultpop)
```



```
. regress ladultpop marrate divrate R1-R3
      Source |       SS       df       MS              Number of obs =      49
-------------+------------------------------           F(  5,    43) =    4.33
       Model |  .164672377     5  .032934475           Prob > F      =  0.0028
    Residual |  .327373732    43  .007613343           R-squared     =  0.3347
-------------+------------------------------           Adj R-squared =  0.2573
       Total |  .492046109    48  .010250961           Root MSE      =  .08725

------------------------------------------------------------------------------
   ladultpop |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     marrate |  -18.26494   7.754941    -2.36   0.023    -33.90427   -2.625615
     divrate |   8.600844   12.09833     0.71   0.481    -15.79777    32.99946
          R1 |   .1192464   .0482428     2.47   0.017     .0219555    .2165373
          R2 |   .0498657    .042209     1.18   0.244    -.0352569    .1349883
          R3 |   .0582061   .0357729     1.63   0.111    -.0139368     .130349
       _cons |    .999169    .093568    10.68   0.000     .8104712    1.187867
------------------------------------------------------------------------------

. predict double ladultpophat, xb
. generate adultpophat = invlogit(ladultpophat)
. summarize adultpop adultpophat
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    adultpop |        49    .7113268    .0211434   .6303276   .7578948
 adultpophat |        49    .7116068    .0120068   .6855482   .7335103
```

Alternatively, we could use Stata's grouped logistic regression ([R] **glogit**) to fit the model. This command uses the same transformation on the response variable, which must be provided for the number of positive responses and the total number of responses (that is, the numerator and denominator of the proportion). For example,

```
. glogit pop18p pop marrate divrate R1-R3
Weighted LS logistic regression for grouped data
      Source |       SS       df       MS              Number of obs =      49
-------------+------------------------------           F(  5,    43) =    4.24
       Model |  .129077492     5  .025815498           Prob > F      =  0.0032
    Residual |  .261736024    43  .006086884           R-squared     =  0.3303
-------------+------------------------------           Adj R-squared =  0.2524
       Total |  .390813516    48  .008141948           Root MSE      =  .07802

------------------------------------------------------------------------------
      pop18p |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     marrate |  -22.82454   8.476256    -2.69   0.010    -39.91853   -5.730537
     divrate |   18.44877   12.66291     1.46   0.152    -7.088418    43.98596
          R1 |   .0762246   .0458899     1.66   0.104    -.0163212    .1687704
          R2 |  -.0207864   .0362001    -0.57   0.569    -.0937909    .0522181
          R3 |   .0088961   .0354021     0.25   0.803     -.062499    .0802912
       _cons |   1.058316   .0893998    11.84   0.000     .8780241    1.238608
------------------------------------------------------------------------------
```

These results differ from those of standard regression because `glogit` uses weighted least-squares techniques. As explained in [R] **glogit**, the appropriate weights correct for the heteroskedastic nature of ϵ has zero mean but variance equal to

$$\sigma_j^2 = \frac{1}{n_j p_j (1 - p_j)}$$

By generating those weights, where n_j is the number of responses in the jth category and p_j is the predicted value we computed above, we can reproduce the `glogit` results with `regress` by using analytic weights, as verified with the commands:

```
. generate glswt = adultpophat * (1 - adultpophat) * pop
. quietly regress ladultpop marrate divrate R1-R3 [aw=glswt]
```

In the case of these state-level census data, values for the proportion y must lie within the unit interval. But we often consider data for which the limiting values of zero or one are possible. A city may spend 0% of its budget on preschool enrichment programs. A county might have zero miles of active railway within its boundaries. There might have been zero murders in a particular town in each of the last five years. A hospital may have performed zero heart transplants last year. In other cases, we may find values of one for particular proportions of interest. Neither zeros nor ones can be included in the strategy above, as the logit transformation is not defined for those values.

A strategy for handling proportions data in which zeros and ones may appear as well as intermediate values was proposed by Papke and Wooldridge (1996). At the time of their writing, Stata's generalized linear model ([R] **glm**) command could not handle this model, but it has been enhanced to do so. This approach makes use of the logit link function (that is, the logit transformation of the response variable) and the binomial distribution, which may be a good choice of family even if the response is continuous. The variance of the binomial distribution must go to zero as the mean goes to either 0 or 1, as in each case the variable is approaching a constant, and the variance will be maximized for a variable with mean of 0.5.

To illustrate, consider an alternative dataset that contains zeros and ones in its response variable, `meals`: the proportion of students receiving free or subsidized meals at school.

(*Continued on next page*)

```
. use http://www.ats.ucla.edu/stat/stata/faq/proportion, clear
. summarize meals
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       meals |      4421    .5188102    .3107313          0          1

. glm meals yr_rnd parented api99, link(logit) family(binomial) vce(robust) nolog
note: meals has noninteger values

Generalized linear models                         No. of obs      =      4257
Optimization     : ML                             Residual df     =      4253
                                                  Scale parameter =         1
Deviance         =  395.8141242                   (1/df) Deviance =   .093067
Pearson          =  374.7025759                   (1/df) Pearson  =  .0881031

Variance function: V(u) = u*(1-u/1)               [Binomial]
Link function    : g(u) = ln(u/(1-u))             [Logit]

                                                  AIC             =  .7220973
Log pseudolikelihood = -1532.984106               BIC             = -35143.61

------------------------------------------------------------------------------
             |               Robust
       meals |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      yr_rnd |   .0482527   .0321714     1.50   0.134    -.0148021    .1113074
    parented |  -.7662598   .0390715   -19.61   0.000    -.8428386   -.6896811
      api99 |  -.0073046   .0002156   -33.89   0.000    -.0077271   -.0068821
       _cons |    6.75343   .0896767    75.31   0.000     6.577667    6.929193
------------------------------------------------------------------------------
```

The techniques used above can be used to generate predictions from the model and transform them back into the units of the response variable. This approach is preferred to that of dropping the observations with zero or unit values, which would create a truncation problem, or coding them with some arbitrary value ("winsorizing") such as 0.0001 or 0.9999.

Some researchers have considered using censored normal regression techniques such as `tobit` ([R] **tobit**) on proportions data that contain zeros or ones. However, this is not an appropriate strategy, as the observed data in this case are not censored: values outside the $[0, 1]$ interval are not feasible for proportions data.

One concern was voiced about proportions data containing zeros or ones.[1] In the context of the generalized tobit or "heckit" model ([R] **heckman**), we allow for limit observations (for instance, zero values) being generated by a different process than non-censored observations. The same argument may apply here, in the case of proportions data: the managers of a city that spends none of its resources on preschool enrichment programs have made a discrete choice. A hospital with zero heart transplants may be a facility whose managers have chosen not to offer certain advanced services.

In this context, the `glm` approach, while properly handling both zeros and ones, does not allow for an alternative model of behavior generating the limit values. If different factors generate the observations at the limit points, a sample selection issue arises. Li and Nagpurnanand (2007) argue that selection issues arise in numerous variables of interest in corporate finance research. In a forthcoming article, Cook, Kieschnick, and

1. See, for instance, McDowell and Cox (2001).

McCullough (2008) address this issue for proportions of financial variables by developing what they term the "zero-inflated beta" model, which allows for zero values (but not unit values) in the proportion and for separate variables influencing the zero and nonzero values.[2]

References

Cook, D. O., R. Kieschnick, and B. D. McCullough. 2008. Regression analysis of proportions in finance with self selection. *Journal of Empirical Finance*. In press.

Li, K., and R. Nagpurnanand. 2007. Self-selection models in corporate finance. In *Handbook of Corporate Finance: Empirical Corporate Finance*, ed. B. E. Eckbo, chap. 2. Amsterdam: Elsevier.

McDowell, A., and N. J. Cox. 2001. FAQ: How do you fit a model when the dependent variable is a proportion? http://www.stata.com/support/faqs/stat/logit.html.

Papke, L. E., and J. M. Wooldridge. 1996. Econometric methods for fractional response variables with an application to 401(K) plan participation rates. *Journal of Applied Econometrics* 11: 619–632.

2. Their approach generalizes models fit with the beta distribution; user-written programs for that purpose may be located by typing `findit beta distribution`.