Journal of Computations & Modelling, vol.3, no.3, 2013, 75-86
ISSN: 1792-7625 (print), 1792-8850 (online)
Scienpress Ltd, 2013

Econometric Computing Issues with Logit Regression Models: The Case of Observation-Specific and Group Dummy Variables

Steven B. Caudill[1], Franklin G. Mixon, Jr.[2] and Kamal P. Upadhyaya[3]

Abstract

Our study extends work on econometric computing issues in logit regression models by focusing on observation-specific and group dummy variables, wherein all or nearly all of the members of the group are associated with the same value for *y*, rather than the case of continuous regressors. To make our case, we employ a small data set from a previously published study. Lastly, we explore, using various econometric software packages, several prescriptions for dealing with these issues.

Mathematics Subject Classification: 62H12; 62P20

Keywords: limited dependent variables; dummy variables; logit models

[1] Department of Economics, Rhodes College, USA.
[2] Center for Economic Education, Columbus State University, USA.
[3] Department of Economics University of New Haven, USA.

Article Info: *Received* : July 3, 2013. *Revised* : August 8, 2013
Published online : September 15, 2013

1 Introduction

Econometric computing issues associated with maximum likelihood estimation of logit and probit models that include observation-specific and/or group dummy variables have been the subject of econometric research dating back 25 years. The early entries in this genre, which include Oskanen (1986), Anderson (1987) and Caudill (1987 and 1988), indicate that an entire class of dichotomous choice models, including logit, encounter estimation difficulty in the presence of an observation-specific dummy variable. Additionally, inclusion of a group dummy variable, wherein all or nearly all of the members of the group are associated with the same value for the dependent variable, may also present complications for econometric computing using traditional statistical packages (Caudill, 1987 and 1988).

More recently, the Institute for Digital Research and Education (IDRE, 2013) at the University of California – Los Angeles provided a guideline for econometricians in dealing with the issue of complete and quasi-complete separation. The former occurs when the outcome variable, *y*, separates a predictor variable, *x*, completely (IDRE, 2013).[4] Although the IDRE (2013) exposition is based largely on the case where the regressors are continuous, the separation and quasi-complete separation estimation problems can also occur with binary regressors. In the context of binary regressors, complete separation may occur in the presence of an observation-specific dummy variable or when a group dummy variable is included on the right-hand side of the model. In terms of an example concerning an observation-specific dummy variable, complete separation results when (1) for the observations where *x* is equal to 1, *y* is also equal to 1, and (2) for all other observations both *x* and *y* are equal to 0. The other issue, quasi-complete separation, occurs when the outcome variable, *y*, separates a

[4] As Albert and Anderson (1984) point out, complete separation occurs when a vector, α, correctly allocates all observations to their group (see also IDRE, 2013).

predictor variable, *x*, to a certain degree (IDRE, 2013). In the context of binary regressors, quasi-complete separation may also occur in the presence of an observation-specific dummy variable, or when a group dummy variable is included on the right-hand side of the model, and, for example, either (1) nearly all of the members of the category represented by the group dummy make the same choice (i.e., where *x* is equal to 1, *y* is nearly always equal to 1), or (2) all of the members of the category represented by the group dummy make the same choice (i.e., where *x* is equal to 1, *y* is equal to 1), and yet there are other observations where *y* is equal to 1 and *x* is equal to 0.

Although they are presented here for illustrative purposes only, the example data sets in Appendix 1 provide a depiction of the types of data sets leading to the quasi-complete separation scenarios discussed above for both observation-specific and group dummy variables. With only 10 observations each, the example data sets in Appendix 1 also highlight the indication in IDRE (2013) that quasi-complete separation problems are more likely to occur with the use of small data sets. Even given the expansive presentation of the separation problems in IDRE (2013), there is still room for further econometric computing analysis. As stated above, our study extends IDRE (2013) by focusing on observation-specific and group dummy variables wherein members of the group are associated with the same value for *y*, rather than continuous regressors. In doing so, we also employ a small data set from a previously published study (in the field of sports economics). Lastly, we explore, using various econometrics packages, several of the prescriptions described in IDRE (2013), but that are not provided by that same resource.

2 Addressing the Problem

In order to address the logit estimation problems associated with

observation-specific and group dummy variables, we re-examine the econometric model in Caudill and Mixon (2007). Their study models the probability of a University of Alabama (hereafter Alabama) victory in its annual college football game against rival Auburn University (hereafter Auburn), known nationwide as the Iron Bowl, in an effort to draw wider conclusions about the importance of home field advantage in college football. In modeling this probability, Caudill and Mixon (2007) examine the role of four regressors – two continuous variables and two dummy variables – on the outcome of 32 previous Iron Bowl contests. Their econometric model is shown below in equation (1),

$$\text{ALWIN} = \alpha + \beta_1 ln\text{FANS} + \beta_2\text{RECDIF} + \beta_3\text{ALNEED} + \beta_4\text{AUNEED} + \varepsilon, \qquad (1)$$

where ALWIN is a dichotomous variable equal to 1 for Iron Bowl games won by Alabama, and 0 otherwise (i.e., Iron Bowl games won by Auburn). In terms of regressors, *ln*FANS is equal to the log of the ratio of Auburn fans to Alabama fans in attendance during a given Iron Bowl game. Next, RECDIF is equal to the difference between Alabama's record, in ratio form, heading into the Iron Bowl minus Auburn's record at that same point. ALNEED and AUNEED are both binary variables, equal to 1 if Alabama and Auburn, respectively, need an Iron Bowl victory to avoid a non-winning season, and 0 otherwise. Over the period examined by Caudill and Mixon (2007), ALNEED is equal to 1 on a single occasion, thus constituting an observation-specific dummy variable. On the other hand, AUNEED is equal to 1 for multiple observations, and in each case ALWIN is also equal to 1. This represents the type of group dummy variable that has the potential to result in a separation problem once the logit model in equation (1) is estimated.

Although not germane to this particular study, the expected values of the second and third parameter estimates from equation (1) above are, as explained in Caudill and Mixon (2007), positive, while those for the first and fourth parameter estimates are negative. It is also worth noting that the econometric model in Caudill and Mixon (2007) is based on a conceptual (statistical and graphical)

model in an earlier study by Caudill and Mixon (1996) that specifies a linear relationship between the probability of an Alabama victory in a given Iron Bowl and the log of the relative number of Auburn fans in attendance. As such, a linear probability model (LPM) is explored in Caudill and Mixon (2007), which is one of the prescriptions for dealing with separation problems resulting from maximum likelihood estimation of the logit model that is provided by Caudill (1987 and 1988). The results of that LPM, with *t*-values in parentheses, are presented in equation (2) below.

$$\underset{\quad (5.61)\quad (-1.94)\qquad (3.13)\qquad (2.47)\qquad (0.99)}{\text{ALWIN} = 0.442 - 0.098\textit{ln}\text{FANS} + 0.973\text{RECDIF} + 1.076\text{ALNEED} + 0.227\text{AUNEED}} \qquad (2)$$

The results above indicate that all but the final regressor retains its expected sign, and that four of the five LPM parameter estimates are statistically significant. These results should provide a benchmark for the newer estimates using the same data from Caudill and Mixon (2007) that we present below.

The Caudill and Mixon (2007) data are used to re-estimate equation (1) above, which includes the aforementioned observation-specific dummy (ALNEED) and group dummy (AUNEED), by maximum likelihood/logit. The econometric packages chosen for comparison purposes include EViews, R, SAS, SPSS and Stata. The first conventional logit approach employed EViews and Stata. These packages, however, failed to provide estimates for either ALNEED or AUNEED, given the separation issues that are the focus of this study. More specifically, EViews terminated, noting quasi-complete separation involving both ALNEED and AUNEED, while Stata dropped the two dummy variables, ALNEED and AUNEED, and, unlike EViews, provided estimates for the remaining regressors, *ln*FANS and RECDIF.[5] These types of failures are common,

[5] Given the lack of results for ALNEED and AUNEED, the logit estimates provided by Stata are not presented in this study.

at least historically, with various statistical packages (IDRE, 2013), and in the case of Stata, there are some alternative estimation procedures that are discussed below. In the case of EViews, however, there are few solutions. With one solution, the researcher moves forward by estimating a model with only two regressors – *ln*FANS and RECDIF (IDRE, 2013). This result is unsatisfying in that estimates are not obtained for the observation-specific and group dummy variables. Another solution is LPM estimation, as discussed above (Caudill, 1987 and 1988).

Given EViews' time series focus or specialization, researchers who work with limited dependent variables models likely have access to other statistical packages. Two of these are SAS and SPSS. Conventional logit estimates of the parameters in equation (1) using each of these packages are presented in Table 1. The results using either SAS or SPSS for *ln*FANS and RECDIF are much like those of their LPM counterparts in equation (2) above. Unlike conventional logit estimation using either EViews or Stata, these packages provide estimates for the dummy variables of interest. However, the parameter estimates for both ALNEED and AUNEED are relatively large and are accompanied by extremely large standard errors, which is indicative of quasi-complete separation issues. In fact, both packages provide users with the warning that the maximum likelihood estimate may not exist, and that the software package terminated after a number of iterations. For SPSS, 20 iterations were completed, while SAS terminated after an unspecified number of iterations (see Table 1), although it is believed that the default value for SAS is 25 iterations.[6] These additional few iterations contribute to the differences between the estimates when comparing the two sets of results.

[6] The SAS package provided a warning of quasi-complete separation, while SPSS did not provide a similar warning.

Table 1: Conventional Logit Results

Variables	SAS Logit	SPSS Logit
constant	−0.399 (0.514)	−0.399 (0.514)
	[p = .438]	[p = .438]
*ln*FANS	−0.745 (0.407)	−0.744 (0.407)
	[p = .068]	[p = .068]
RECDIF	5.831 (2.488)	5.831 (2.488)
	[p = .019]	[p = .019]
ALNEED	16.038 (451.8)	25.015 (40,193)
	[p = .972]	[p = 1.000]
AUNEED	11.179 (185.0)	20.159 (16,490)
	[p = .952]	[p = .999]
Software Comments	The maximum likelihood estimate may not exist. Results are based on the last maximum likelihood iteration. Validity of the model fit is questionable.	Estimation terminated at iteration number 20 because maximum iterations has been reached. Final solution cannot be found.

Note: In addition to parameter estimates, the cells above also provide standard errors in parentheses and p-values in brackets.

Both Stata and SAS provide alternatives to conventional logit that are not apparently available in either EViews or SPSS.[7] These are the Firth logit in Stata and the Firth Bias-Correction logit in SAS. The SAS Institute's brief exposition of Firth's method, based largely on Firth (1993), Heinze and Schemper (2002) and Heinze (2006), is provided in Appendix 2. Results from this approach, one using

[7] Given SPSS' failure to provide a quasi-complete separation warning, or to offer additional tests to address this issue, use of SPSS in circumstances such as those described in this study is problematic. On the other hand, EViews' provision of a quasi-complete separation warning provides researchers using this package with enough information to, at the very least, employ an LPM approach (Caudill, 1987 and 1988).

Stata and a second using SAS, are presented in Table 2. Both estimations represent dramatic differences from those in Table 1. The Firth logit model available in Stata provides estimates that are also different from those of its SAS counterpart, as indicated in Table 2, particularly with regard to ALNEED. In fact, the Firth Bias-Correction estimation procedure employed by SAS suffered from quasi-complete separation issues, as noted in the SAS warning statement that is reproduced at the bottom on Table 2.[8]

Table 2: Bias-Reduced Logit Results

Variables	Stata Firth Logit	SAS Firth Bias-Correction Logit	R Bias-Reduced Logit
constant	−0.318 (0.466)	−0.318 (0.480)	−0.318 (0.480)
	[p = .495]	[p = .508]	[p = .514]
*ln*FANS	−0.605 (0.358)	−0.605 (0.359)	−0.605 (0.359)
	[p = .091]	[p = .092]	[p = .103]
RECDIF	4.720 (2.136)	4.720 (2.175)	4.720 (2.175)
	[p = .027]	[p = .030]	[p = .039]
ALNEED	4.183 (2.105)	4.310 (2.731)	4.183 (2.663)
	[p = .047]	[p = .115]	[p = .128]
AUNEED	1.604 (1.776)	1.605 (1.929)	1.604 (1.929)
	[p = .366]	[p = .406]	[p = .413]
Software Comments		Convergence was not attained in 25 Iterations. Results shown are based on the last maximum likelihood iteration. Validity of the model fit is questionable.	

Note: In addition to parameter estimates, the cells above also provide standard errors in parentheses and p-values in brackets.

[8] SAS notes that it encounters difficulty providing convergence after an iteration count of

Many researchers are now using the open source econometrics package referred to as R. This package provides a bias-reduced logit estimation procedure (Wessa, 2009) that is based on work by Firth (1992 and 1993), Heinze and Schemper (2002), Zorn (2005), Bewick, Cheek and Ball (2005) and Macdonald (2006). Estimation of equation (1) above using R provides the results presented in the final column of Table 2. These results are generally quite similar to those from Firth Bias-Correction logit estimation using SAS.

3 Conclusion

This study extends research on econometric computing issues associated with maximum likelihood estimation of logit and probit models by focusing on observation-specific and group dummy variables, wherein members of the group are associated with the same value for *y*, rather than continuous regressors. In doing so, we also employ a small data set and various econometric packages, including SAS and R, which is an open source software engine. Although these packages offer bias-reducing estimation procedures, our explorations indicate that researchers must still be concerned with maximum likelihood estimates in these situations.

ACKNOWLEDGEMENTS. The authors thank two anonymous referees for helpful comments. The usual caveat applies.

References

[1] A. Albert and J.A. Anderson, On the existence of maximum likelihood estimates in logistic regression models, *Biometrika*, **71**(1), (1984), 1-10.

[2] G.J. Anderson, Prediction tests in limited dependent variable models, *Journal of Econometrics*, **34**(1-2), (1987), 253-261.

[3] V. Bewick, L. Cheek and J. Ball, Statistics review 14: Logistic regression, *Critical Care*, **9**(1), (2005), 112-118.

[4] S.B. Caudill, Dichotomous choice models and dummy variables, *The Statistician*, **36**(4), (1987), 381-383.

[5] S.B. Caudill, An advantage of the linear probability model over logit or probit, *Oxford Bulletin of Economics and Statistics*, **50**(4), (1988), 425-427.

[6] S.B. Caudill and F.G. Mixon Jr., Stadium size, ticket allotments and home field advantage in college football, *Social Science Journal*, **44**(4), (2007), 751-759.

[7] S.B. Caudill and F.G. Mixon Jr., Winning and ticket allotments in college football, *Social Science Journal*, **33**(4), (1996), 451-457.

[8] D. Firth, Bias reduction, the Jeffreys prior and GLIM, in *Advances in GLIM and Statistical Modelling*, L. Fahrmeir, B.J. Francis, R. Gilchrist and G. Tutz [eds.], New York: Springer, (1992), 91-100.

[9] D. Firth, Bias reduction of maximum likelihood estimates, *Biometrika*, **80**(1), (1993), 27-38.

[10] G. Heinze, A comparative investigation of methods for logistic regression with separated or nearly separated data, *Statistics in Medicine*, **25**(24), (2006), 4,216-4,226.

[11] G. Heinze and M. Schemper, A solution to the problem of separation in logistic regression, *Statistics in Medicine*, **21**(16), (2002), 2,409-2,419.

[12] Institute for Digital Research and Education [IDRE], What is complete or quasi-complete separation in logistic/probit regression and how do we deal with them, (2013), www.ats.ucla.edu/stat/mult_pkg/faq/general/complete_separation_logit_models.htm (accessed on 13 May 2013).

[13] P.D.M. Macdonald, R functions for ROC curves and the Hosmer-Lemeshow Test, (2006), www.math.mcmaster.ca/peter/s4f03/s4f03_0607/rochl.html, (accessed on 13 May 2013).

[14] E.H. Oskanen, A note on observation-specific dummies and logit analysis, *The Statistician*, **35**(4), (1986), 413-416.

[15] P. Wessa, Bias Reduced Logistic Regression (v1.0.4) in *Free Statistics Software (v1.1.23-r7)*, Office for Research Development and Education, (2009), www.wessa.net/rwasp_logisticregression.wasp/ (accessed on 13 May 2013).

[16] C. Zorn, A solution to separation in binary response models, *Political Analysis*, **13**(2), (2005), 157-170.

Appendix 1: Example Data Sets

Quasi-Complete Separation

Observation-Specific Dummy		Group Dummy	
Y	*X*	*Y*	*X*
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
0	0	0	0
1	1	1	1
1	0	1	1
1	0	1	1
1	0	1	1
1	0	1	0

Appendix 2: Firth Bias-Reducing Penalized Likelihood

Following the SAS Institute's exposition, Firth's method replaces the usual score (gradient) equation,

$$g(\beta_j) = \sum_{i=1}^{n} (y_i - \pi_i) x_{ij} = 0 \qquad (j=1, \ldots p), \tag{1}$$

where p is the number of parameters in the model, with the modified score equation,

$$g(\beta_j)^* = \sum_{i=1}^{n} \{y_i - \pi_i + h_i(0.5 - \pi_i)\} x_{ij} = 0 \qquad (j=1, \ldots p), \tag{2}$$

where the h_is are the ith diagonal elements of the hat matrix $\mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}$ and $\mathbf{W}$=diag$\{\pi_i(1-\pi_i)\}$.