Bias in Fixed-Effects Cox Regression with Dummy Variables

Paul D. Allison*

Department of Sociology
University of Pennsylvania

January 2002

Abstract

One approach to doing fixed-effects regression analysis is simply to include dummy variables in the model for all the individuals (less one). Greene (2001) has recently introduced algorithms that make this computationally feasible even for nonlinear models with thousands of dummy variables. The dummy variable approach works well for linear regression and Poisson regression, but may suffer severe "incidental parameters bias" for logistic regression. The performance of the dummy variable method for Cox regression with repeated event data is unknown. I show by simulation that incidental parameters bias for Cox regression may be nearly as severe as that with logistic regression. Fortunately, as is well known, fixed-effects analysis of repeated event data is conveniently done by Cox regression combined with stratification on individuals, thereby eliminating the nuisance parameters.

*3718 Locust Walk, Philadelphia, PA 19104-6299, voice: 215-898-6717, fax: 215-573-2081, e-mail: allison@ssc.upenn.edu, web: www.ssc.upenn.edu/~allison.

A major attraction of longitudinal data is the ability to control for all stable covariates, without actually including them in a regression equation. In general, this is accomplished by using only within-individual variation to estimate the parameters, and then averaging the estimates over individuals. Regression models for accomplishing this are often called fixed-effects models. Fixed-effects models have been developed for a variety of different data types and models, including linear models for quantitative data (Mundlak 1978), logistic regression models for categorical data (Chamberlain 1980), and Poisson or negative binomial regression models for count data (Palmgren 1981).

One approach to fixed-effects regression is simply to include dummy variables for all the individuals in the sample (less one). For linear models, Poisson regression models (Cameron and Trivedi 1998) and negative binomial regression models (Allison and Waterman 2002), this method works very well. For logistic regression, on the other hand, the inclusion of dummy variables for many individuals can lead to severe “incidental parameters” bias (Kalbfleisch and Sprott 1970). The usual asymptotic justification for maximum likelihood estimation depends on the presumption that the number of parameters remains constant as the sample gets larger. For longitudinal data, that works just fine if the number of individuals remains constant but the number of observations per individual gets larger. But if the number of individuals is getting larger while the number of time points remains constant, then the number of parameters in a fixed-effects model (including coefficients of the dummy variables) is increasing at the same rate as the sample size. This tends to produce an inflation of the coefficient magnitudes. When there are exactly two observations for each individual, logistic regression coefficients will be twice as large as they should be (Abrevaya 1997).

The solution to the incidental parameters problem for logistic regression is to do conditional maximum likelihood, conditioning on the number of 1's and 0's for each individual (Chamberlain 1980). This removes the dummy variable coefficients from the likelihood function and yields coefficient estimates that are consistent.

In this paper, I investigate the use of the dummy variable method applied to Cox regression models for repeated event data. To my knowledge, there is no literature addressing the presence or absence of incidental parameters bias for Cox regression. Previous attempts at fixed-effects analysis for Cox regression have used stratification on individuals to remove the dummy variable coefficients from the partial likelihood function (Chamberlain 1985, Yamaguchi 1986), an approach quite similar to conditional maximum likelihood for logistic regression. Elsewhere I have shown by simulation (Allison 1996) that this method produces approximately unbiased estimates under a wide variety of conditions. However, Chamberlain (1979) raised some questions about the validity of the likelihood function, and my own simulations suggest that there may be bias when the regression model includes variables describing the previous event history.

Given the possible limitations of the stratification method, it's reasonable to consider the dummy variable method as an alternative. One possible disadvantage of the dummy variable method is the sheer computational difficulty of estimating models with hundreds, perhaps thousands of dummy variables. Standard Cox regression programs require repeated inversion of of a $K \times K$ matrix, where K is the number of covariates. When K is large, this becomes very computationally intensive. Greene (2001), however, has recently shown that for a large class of nonlinear fixed-effects models, the coefficients for the dummy variables may be estimated without the necessity of inverting a large matrix. Although Greene's method would require modification of conventional software, it does raise the possibility that the dummy variable method could be a convenient way to estimate the Cox regression model in a wide variety of situations. However, before we take the trouble to rewrite existing Cox regression programs, it is essential to determine whether these models are subject to incidental parameters bias. In the simulations that follow, I show that such bias is quite severe when the number of events per individual is small.

Repeated event data generally take two forms, ordered and unordered. With ordered data, the more common form, we begin observing an individual at some time t=0, and then observe a sequence of events that occur at times t_1, t_2, …. Usually, the focus of the modeling is on the gap times:

$$\begin{aligned} y_1 &= t_1 - 0 \\ y_2 &= t_2 - t_1 \\ y_3 &= t_3 - t_2 \\ &\vdots \end{aligned}$$

Observation is terminated at some point τ, so the last gap time is censored.

With unordered data, for each individual we simply observe one or more event times, each of which may be censored or uncensored, and there is no sequence to these event times. This kind of data would arise if the "individuals" were families and the goal was to model death times for all siblings in the family.

For either ordered or unordered data, we begin with the following proportional hazards model

$$\log h_{ik}(t) = \alpha(t) + \beta x_{ik} + \delta_i \qquad (1)$$

where $h_{ik}(t)$ is the hazard of the k'th event for the i'th individual at time t, x_{ik} is a column vector of covariates for the k'th interval for the i'th individual, β is a row vector of coefficients, $\alpha(t)$ is an unspecified function of time, and δ_i is a set of fixed-effects.

For generating simulated data, I used a special case of this model

$$\log h_{ik}(t) = \alpha \log t + \beta x_{ik} + \delta_i \qquad (2)$$

which gives rise to a Weibull distribution for event times, conditional on x and δ. In the simulations, I specified that x and δ were drawn from a bivariate normal distribution, with 0 means for both variables, standard deviations of 1 for x and σ for δ, and a correlation ρ.

The first set of simulations is for ordered data. For each scenario, I constructed 100 samples, each with 100 individuals. For each individual, I generated a sequence of intervals between events until the sum of those intervals exceeded a censoring time c. Hence, the last interval for each individual was censored, while all the earlier intervals were uncensored. The baseline scenario set α=1, β=1, σ=1, ρ=0, and c=2. I then estimated Cox regressions for each sample using the SAS procedure PHREG with dummy variables to estimate δ_i. The results for the 100 samples are shown in line 1 of Table 1. For the baseline scenario, with approximately three intervals per individual and 33 percent of the intervals censored, the mean coefficient for x was 1.27, that is, 27 percent above the true value.

I then proceeded to vary the parameter values one at a time, with results shown in later lines of Table 1. In all cases but one (when β=0), the coefficients are biased away from 0, with percentages varying between 11 and 73. It appears that the percentage increase is primarily a function of the mean number of intervals contributed by each individual, which would be consistent with incidental parameters bias in logistic regression. However, in this data set up, the average number of intervals is completely confounded with the percent censored (the latter is just the reciprocal of the former) because every individual has exactly one censored interval. So, it's possible that the degree of parameter inflation is essentially a function of the percentage censored.

Table 1. Estimates for Cox Regression Model with Dummy Variables, Ordered Data

Model	*Mean Intervals per Individual*	*%Censored*	*Mean Coefficient*	*Standard Error*
1. Baseline	3.06	33	1.270	.017
2. α=1.5	2.91	34	1.358	.018
3. α=0.5	3.29	30	1.226	.015
4. α=0.0	3.71	27	1.157	.013
5. α=-0.5	4.73	21	1.107	.009
6. σ=1.5	3.41	29	1.230	.013
7. σ=0.5	2.85	35	1.347	.020
8. σ=0.0	2.82	35	1.371	.022
9. β=1.5	2.96	34	1.932	.019
10. β=0.5	3.13	32	0.611	.014
11. β=0.0	3.16	32	-0.018	.014
12. ρ=.50	3.46	29	1.183	.014
13. ρ=-.50	2.87	35	1.342	.021
14. *c*=4	5.33	19	1.145	.009
15. *c*=1	1.95	51	1.511	.041
16. *c*=.75	1.64	61	1.727	.061

The influence of these two factors can be distinguished by considering simulations for unordered data. In this setup, each individual has a fixed number of intervals. Each interval may be censored or uncensored, depending on whether it exceeds a censoring time *c*. As before, I generated data based on equation (2) with baseline parameter values α=1, β=1, σ=1, and ρ=0. Table 2 presents results for five different numbers of intervals per individual (all with 24 percent censoring) and five different levels of censoring (all with three interval per individual). The inflation due to number intervals ranges from 7 percent (for 10 intervals) to 80 percent (for two intervals). The inflation due to censoring ranges from 35 percent (with no censoring) to 57 percent (with 88 percent of the cases censored). Clearly, both factors play a role, but number of intervals seems to have the larger effect.

Table 2. Estimates for Cox Regression Model with Dummy Variables, Unordered Data

Censoring Time	*Number Intervals per Individual*	*%Censored*	*Mean Coefficient*	*Standard Error*
1. *c*=1.5	10	24	1.073	.006
2. *c*=1.5	5	24	1.185	.010
3. *c*=1.5	4	24	1.237	.010
4. *c*=1.5	3	23	1.378	.018
5. *c*=1.5	2	23	1.796	.031
6. *c*=∞	3	0	1.347	.014
7. *c*=2	3	15	1.373	.014
8. *c*=1	3	40	1.387	.019
9. *c*=0.5	3	68	1.505	.027
10. *c*=0.25	3	88	1.570	.058

These simulation results demonstrate that fixed-effects Cox regression with dummy variables is prone to serious inflation of parameter estimates when the number of intervals per individual is low and the percentage of censored cases is high. Both conditions are likely to occur with ordered event data when the observation period is short. Fortunately, the implementation of fixed-effects Cox regression via stratification on individuals is readily available and easily employed alternative. Elsewhere (Allison 1996) I have shown that this method shows very little bias in simulations that are very similar to those used here.

References

Abrevaya, Jason (1997) “The Equivalence of Two Estimators of the Fixed-Effects Logit Model.” *Economics Letters* 55: 41-43.

Allison, Paul D. 1996. “Fixed Effects Partial Likelihood for Repeated Events.” *Sociological Methods & Research* 25: 207-222.

Allison, Paul D. and Richard P. Waterman (2002) “Fixed-Effects Negative Binomial Regression Models.” Forthcoming in *Sociological Methodology 2002*.

Cameron, A. Colin and Pravin K. Trivedi. 1998. *Regression Analysis of Count Data.* Cambridge, UK: Cambridge University Press.

Chamberlain, Gary A. 1980. “Analysis of Covariance with Qualitative Data.” *Review of Economic Studies* 47: 225-238.

Chamberlain, Gary A. (1985) “Heterogeneity, omitted variable bias, and duration dependence.” Pp. 3-38 in *Longitudinal Analysis of Labor Market Data*, edited by J. J. Heckman and B. Singer. Cambridge: Cambridge University Press.

Greene, William. 2001. "Estimating Econometric Models with Fixed Effects." Unpublished paper available at http://www.stern.nyu.edu/~wgreene.

Kalbfleisch, John D. And David A. Sprott. 1970. "Applications of Likelihood Methods to Models Involving Large Numbers of Parameters" (with discussion). *Journal of the Royal Statistical Society* Series B, 32: 175-208.

Mundlak, Yair (1978) "On the Pooling of Time Series and Cross Section Data." *Econometrica* 46: 69-85.

Palmgren, Juni. 1981. "The Fisher Information Matrix for Log-Linear Models Arguing Conditionally in the Observed Explanatory Variables." *Biometrika* 68: 563-566.

Yamaguchi, Kazuo. 1986. "Alternative Approaches to Unobserved Heterogeneity in the Analysis of Repeated Events." Pp. 213-49 in *Sociological Methodology 1986*, edited by Nancy Brandon Tuma. Washington, DC: American Sociological Association.