7

FIXED-EFFECTS NEGATIVE BINOMIAL REGRESSION MODELS

*Paul D. Allison**
*Richard P. Waterman**

This paper demonstrates that the conditional negative binomial model for panel data, proposed by Hausman, Hall, and Griliches (1984), is not a true fixed-effects method. This method—which has been implemented in both Stata and LIMDEP—does not in fact control for all stable covariates. Three alternative methods are explored. A negative multinomial model yields the same estimator as the conditional Poisson estimator and hence does not provide any additional leverage for dealing with overdispersion. On the other hand, a simulation study yields good results from applying an unconditional negative binomial regression estimator with dummy variables to represent the fixed effects. There is no evidence for any incidental parameters bias in the coefficients, and downward bias in the standard error estimates can be easily and effectively corrected using the deviance statistic. Finally, an approximate conditional method is found to perform at about the same level as the unconditional estimator.

1. INTRODUCTION

A major attraction of panel 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.

*University of Pennsylvania.

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), Cox regression models for event history data (Yamaguchi 1986; Allison 1996), and Poisson regression models for count data (Palmgren 1981).

Here we consider some alternative fixed-effects models for count data. First, we show that the fixed-effects negative binomial model proposed by Hausman, Hall, and Griliches (1984) (hereafter HHG) is not a true fixed-effects method. Next we consider a negative multinomial model, which leads back to the estimator for the fixed-effects Poisson model. We then use simulated data to compare an unconditional negative binomial estimator with the fixed-effects Poisson estimator. The negative binomial estimator does not appear to suffer from any "incidental parameters" bias, and is generally superior to the Poisson estimator. Finally, we investigate an approximate conditional likelihood method for the negative binomial model. Its performance on the simulated data is roughly comparable to that of the unconditional negative binomial estimator.

2. THE FIXED-EFFECTS POISSON MODEL

The fixed-effects Poisson regression model for panel data has been described in detail by Cameron and Trivedi (1998). The dependent variable y_{it} varies over individuals ($i = 1, \dots, n$) and over time ($t = 1, \dots, T_i$). It is assumed to have a Poisson distribution with parameter μ_{it}, which in turn depends on a vector of exogenous variables $\mathbf{x}_{it}$ according to the loglinear function

$$\ln \mu_{it} = \delta_i + \beta \mathbf{x}_{it}, \tag{1}$$

where δ_i is the "fixed effect."

One way to estimate this model is to do conventional Poisson regression by maximum likelihood, including dummy variables for all individuals (less one) to directly estimate the fixed effects. An alternative method is conditional maximum likelihood, conditioning on the count total $\sum_t y_{it}$ for each individual. For the Poisson model, this yields a conditional likelihood that is proportional to

$$\prod_i \prod_t \left(\frac{\exp(\beta \mathbf{x}_{it})}{\sum_s \exp(\beta \mathbf{x}_{is})} \right)^{y_{it}}, \tag{2}$$

which is equivalent to the likelihood function for a multinomial logit model for grouped data. Note that conditioning has eliminated the δ_i parameters from the likelihood function.

For logistic regression models, it is well known that estimation of fixed-effects models by the inclusion of dummy variables yields inconsistent estimates of β (Hsiao 1986) due to the "incidental parameters" problem (Kalbfleisch and Sprott 1970), while conditional estimation does not suffer from this problem. For Poisson regression, on the other hand, these two estimation methods—unconditional maximization of the likelihood and conditional likelihood—always yield identical estimates for β and the associated covariance matrix (Cameron and Trivedi 1998). Hence, the choice of method should be dictated by computational convenience.

The fixed-effects Poisson regression model allows for unrestricted heterogeneity across individuals but, for a given individual, there is still the restriction that the mean of each count must equal its variance:

$$E(y_{it}) = \operatorname{var}(y_{it}) = \mu_{it}. \tag{3}$$

In many data sets, however, there may be additional heterogeneity not accounted for by the model.

As an example, let's consider the patent data analyzed by HHG and reanalyzed by Cameron and Trivedi (1998). The data consist of 346 firms with yearly data on number of patents from 1975 to 1979. Thus, y_{it} is the number of patents for firm i in year t. This variable ranged from 0 to 515 with a mean of 35 and a standard deviation of 71. A little over half of the firm years had patent counts of five or less. The regressor variables include the logarithm of research and development expenditures in the current year and in each of the previous five years. All the fitted models also include four dummy variables corresponding to years 1976 to 1979.

To analyze the data, we created a separate observation for each firm year, for a total of 1730 working observations. We then estimated a fixed-effects Poisson regression model by conventional Poisson regression software,[1] with 345 dummy variables to estimate the fixed effects. Results for the research and development variables are shown in the first two columns of Table 1. These numbers differ somewhat from those in Cameron and Trivedi (1998), but they are identical to the corrected results reported in their website (http://www.econ.ucdavis.edu/faculty/cameron/).

[1]We used the GENMOD procedure in SAS.

TABLE 1
Conditional Regression Models for Number of Patents

	Conditional Poisson		Conditional Negative Binomial			
	Coefficient	Standard Error	Coefficient	Standard Error	Coefficient	Standard Error
LogRD-0	.322	.046	.363	.085	.272	.071
LogRD-1	−.087	.049	.156	.099	−.098	.077
LogRD-2	.079	.045	.174	.090	.032	.071
LogRD-3	.001	.041	.015	.083	−.020	.066
LogRD-4	−.005	.038	.029	.076	.016	.063
LogRD-5	.003	.032	.136	.062	−.010	.053
Log SIZE					.207	.078
SCIENCE					.018	.198
Intercept					1.660	.343

A potential problem with these results is that there is still some evidence of overdispersion in the data. The ratio of the deviance to the degrees of freedom is 2.04 (deviance = 2807 with 1374 d.f.) and the ratio of the Pearson goodness-of-fit chi-square to the degrees of freedom is 1.97 (chi-square = 2709 with 1374 d.f.). For a good-fitting model, these measures should be close to 1. Substantial departures from this ratio may indicate a problem with the model specification, and also suggest that the estimated standard errors may be downwardly biased.

3. THE HHG NEGATIVE BINOMIAL MODEL

HHG deal with the problem of overdispersion by assuming that y_{it} has a negative binomial distribution, which can be regarded as a generalization of the Poisson distribution with an additional parameter allowing the variance to exceed the mean. There are several different ways to parameterize the negative binomial distribution, and the choice can be consequential for regression models. In the HHG model, the negative binomial mass function can be written as

$$f(y_{it} \mid \lambda_{it}, \theta_i) = \frac{\Gamma(\lambda_{it} + y_{it})}{\Gamma(\lambda_{it})\Gamma(y_{it} + 1)} \left(\frac{\theta_i}{1 + \theta_i}\right)^{y_{it}} \left(\frac{1}{1 + \theta_i}\right)^{\lambda_{it}} \tag{4}$$

whereT is the gamma function. The parameter θ_i is assumed to be constant over time for each individual while λ_{it} depends on covariates by the function

$$\ln \lambda_{it} = \beta \mathbf{x}_{it}. \tag{5}$$

The decision to decompose λ_{it} as a function of the covariates is somewhat surprising, since λ is usually regarded as an overdispersion parameter. That's because (4) becomes the Poisson mass function as $\lambda \to \infty$.

The mean and variance of y_{it} are given by

$$\begin{aligned} E(y_{it}) &= \theta_i \lambda_{it} \\ \operatorname{var}(y_{it}) &= (1 + \theta_i)\theta_i \lambda_{it}. \end{aligned} \tag{6}$$

Under this model, the ratio of the variance to the mean is $1 + \theta_i$, which can vary across individuals but, as already noted, is constant over time.

HHG further assume that for a given individual i, the y_{it} are independent over time. These assumptions imply that $\sum_t y_{it}$ also has a negative binomial distribution with parameters θ_i and $\sum_t \lambda_{it}$. Conditioning on these total counts, the likelihood function for a single individual is given by

$$\frac{\Gamma\left(\sum_t y_{it} + 1\right)\Gamma\left(\sum_t \lambda_{it}\right)}{\Gamma\left(\sum_t y_{it} + \sum_t \lambda_{it}\right)} \prod_t \frac{\Gamma(\lambda_{it} + y_{it})}{\Gamma(\lambda_{it})\Gamma(y_{it} + 1)}, \tag{7}$$

thereby eliminating the θ_i parameters. The likelihood for the entire sample is obtained by multiplying together all the individual terms like (7). This likelihood may be maximized with respect to the β parameters using conventional numerical methods. In fact, the method has been implemented in at least two commercial statistical packages, Stata (www.stata.com) and LIMDEP (www.limdep.com).

In the middle two columns of Table 1, we report results of applying this method to the patent data,[2] using the same covariates as Cameron and Trivedi (1998). The numbers reported here are the same as the corrected numbers given in their website. Note that the coefficients are

[2]To estimate the model, we used the NLMIXED procedure in SAS. This required the specification of the log-likelihood for a single individual.

similar in magnitude to those for the conditional Poisson method, but the estimated standard errors are appreciably larger because the model allows for overdispersion.

Unfortunately, this negative binomial model and its conditional likelihood does not really fit the bill as a fixed-effects method. The basic problem is that the θ_i parameters that are conditioned out of the likelihood function do not correspond to different intercepts in the log-linear decomposition of λ_{it}. HHG's rationale is that if we write $\theta_i = \exp(\delta_i)$, equations (5) and (6) imply that

$$E(y_{it}) = \exp(\delta_i + \beta x_{it})$$

$$\mathrm{var}(y_{it}) = (1 + e^{\delta_i})E(y_{it}).$$

Therefore, it appears that this model does allow for an arbitrary intercept δ_i for each individual. The problem with this approach is that the δ_i's play a different role than x_{it}. Specifically, changes in x_{it} affect the mean directly, and affect the variance only indirectly through the mean. But changes in δ_i affect the variance both indirectly, through the mean, and directly. If we regard δ_i as representing the effects of omitted explanatory variables, then there is no compelling reason why these variables should have a different kind of effect from that of x_{it}.

To put it another way, suppose we begin with equations (6) and specify

$$\lambda_{it} = \exp(\delta_i + \beta \mathbf{x}_{it} + \gamma z_i),$$

where δ_i is an individual-specific intercept and z_i is a vector of time-invariant covariates. Then conditioning on the total count for each individual does *not* eliminate δ_i or γz_i from the likelihood function.

Symptomatic of this problem is that using HHG's conditional likelihood in (7), one can estimate regression models with both an intercept and time-invariant covariates, something that is usually not possible with conditional fixed-effects models. The last two columns of Table 1 show results for estimating the conditional negative binomial model with an intercept and two time-invariant covariates.[3] Both the intercept and one of the two covariates are statistically significant at beyond the .01 level.

[3] SIZE is the firm book value in 1972. SCIENCE is an indicator variable equal to 1 if the firm is in the science sector.

4. A NEGATIVE MULTINOMIAL MODEL

We now consider an alternative parameterization of the negative binomial model that is a more natural generalization of the Poisson model. The mass function for a single y_{it} is given by

$$f(y_{it}|\mu_{it},\lambda_i) = \frac{\Gamma(\lambda_i + y_{it})}{\Gamma(\lambda_i)\Gamma(y_{it}+1)}\left(\frac{\mu_{it}}{\mu_{it}+\lambda_i}\right)^{y_{it}}\left(\frac{\lambda_i}{\mu_{it}+\lambda_i}\right)^{\lambda_i}, \tag{8}$$

with mean and variance functions

$$\begin{aligned} E(y_{it}) &= \mu_{it} \\ \text{var}(y_{it}) &= \mu_{it}(1+\mu_{it}/\lambda_i). \end{aligned} \tag{9}$$

Note that the mean is allowed to vary with time, but the overdispersion parameter λ_i is assumed to be constant for each individual. To model dependence on covariates, we let

$$\ln \mu_{it} = \delta_i + \beta \mathbf{x}_{it}. \tag{10}$$

Cameron and Trivedi (1998) refer to this as an NB2 model, to distinguish it from the previous NB1 model.

If we assume (along with HHG) that the event counts are independent across time for each individual, then this model is not tractable for deriving a conditional likelihood. That's because $\sum_t y_{it}$ does not itself have a negative binomial distribution, so it's awkward to condition on it. More technically, under this specification, there is no complete sufficient statistic for the δ_i's that is a function of the data alone.

As an alternative approach, let's assume that the y_{it} have a negative multinomial distribution, a well-known multivariate generalization of the negative binomial distribution (Johnson and Kotz 1969). For a single individual, the joint mass function is given by

$$f(y_{i1},\ldots,y_{iT}|\lambda_i,\mu_{i1},\ldots,\mu_{iT}) = \frac{\Gamma\left(\lambda_i + \sum_t y_{it}\right)}{\Gamma(\lambda_i)y_{i1}!\ldots y_{iT}!}\left(\frac{\lambda_i}{\lambda_i + \sum_t \mu_{it}}\right)^{\lambda_i} \times \prod_t \left(\frac{\mu_{it}}{\lambda_i + \sum_t \mu_{it}}\right)^{y_{it}} \tag{11}$$

with μ_{it} specified as in (10).[4] This multivariate distribution has the property that the marginal distribution of each y_{it} is negative binomial as defined in (8). Furthermore, the sum $\sum_t y_{it}$ has a negative binomial distribution with parameters $\sum_t \mu_{it}$ and λ_i. Unlike the HHG model, this one does *not* assume that event counts in different time intervals are independent for a given individual. In fact, the correlation (Johnson and Kotz 1969) between y_{it} and $y_{is}(s \neq t)$ is

$$\rho(y_{it}, y_{is}) = \sqrt{\left(\frac{\mu_{it}}{\mu_{it} + \lambda_i}\right)\left(\frac{\mu_{is}}{\mu_{is} + \lambda_i}\right)} \tag{12}$$

To derive a fixed effects estimator for β, we can condition the joint mass function on the total $\sum_t y_{it}$, which yields

$$\begin{aligned} f\left(y_{i1}, \ldots, y_{iT} \middle| \sum_t y_{it}\right) &= \frac{y_{i1}! \ldots y_{iT}!}{\Gamma\left(1 + \sum_t y_{it}\right)} \prod_t \left(\frac{\mu_{it}}{\sum_t \mu_{it}}\right)^{y_{it}} \\ &\propto \prod_t \left(\frac{\exp(\beta \mathbf{x}_{it})}{\sum_s \exp(\beta \mathbf{x}_{is})}\right)^{y_{it}} \end{aligned} \tag{13}$$

Thus, conditioning gives us a distribution that doesn't depend on the parameter λ_i but is proportional to the conditional likelihood for the Poisson model in equation (2). In other words, the fixed-effects negative multinomial model leads to the same conditional estimator of β as the fixed-effects Poisson model.[5]

So it seems that the negative multinomial approach doesn't accomplish anything with respect to overdispersion. To understand this, recall that the negative binomial distribution can be generated by compounding a Poisson random variable with a gamma random variable. The negative multinomial can be generated by compounding a set of independent Poisson random variables with a single gamma random variable. Thus, the overdispersion in the negative multinomial can be thought of as arising from a single random variable that is common to all the event counts

[4] This distribution is the same as the one described by Cameron and Trivedi (1998, p. 288) as a Poisson random-effects model with gamma distributed random effects.

[5] See Guo (1996) for an application of the negative multinomial model in a random-effects setting.

for a given individual (which is why the correlation in [12] is not zero). Conditioning on the total count for each individual removes all the unobserved heterogeneity, both that arising from the δ_i fixed-effects and the unobserved heterogeneity that is intrinsic to the negative multinomial distribution.

5. CONVENTIONAL APPROACHES TO OVERDISPERSION

We have seen that the HHG method does not condition out the fixed effects, while the negative multinomial method conditions out too much to be useful. What's left? A relatively simple approach is to estimate the β coefficients under the fixed-effects Poisson model but to adjust the standard errors upward for overdispersion. A commonly used adjustment is to multiply the standard errors by the square root of the ratio of the goodness-of-fit chi-square to the degrees of freedom. (Either Pearson's chi-square or the deviance could be used.) The first two columns of Table 2 show the Poisson coefficients and adjusted standard errors for the patent data. The coefficients are the same as those in Table 1. The standard errors were obtained by multiplying the standard errors in Table 1 by 1.404, the square root of Pearson's chi-square divided by the degrees of freedom.

An alternative approach is to estimate an unconditional negative binomial model. That is, to specify a conventional NB2 regression model, with dummy variables to estimate the fixed effects. Results of doing

TABLE 2
Regression Models with Overdispersion Corrections

	Fixed-Effects Poisson		Unconditional Negative Binomial	
	Coefficient	Adjusted Standard Error	Coefficient	Standard Error
LogRD-0	.322	.064	.356	.093
LogRD-1	−.087	.068	.021	.102
LogRD-2	.079	.063	.007	.096
LogRD-3	.001	.058	.008	.089
LogRD-4	−.005	.053	.117	.081
LogRD-5	.003	.045	.011	.067

that for the patent data are shown in the last two columns of Table 2.[6] The coefficients are similar to those obtained with a Poisson specification, but the negative binomial standard errors are notably larger than the Poisson standard errors, even though the latter are already adjusted for overdispersion.

There are two potential problems with the unconditional negative binomial method. First, since there is a potential incidental parameters problem, it is questionable whether the coefficient estimates are consistent. As yet, there is no proof of this one way or the other. Second, in the case of large sample sizes, it may be computationally impractical to estimate coefficients for large numbers of dummy variables. Greene (2001) has shown that the computational problem can be readily overcome for this and many other nonlinear fixed-effects models, although conventional software would have to be modified to implement his methods.

To investigate the performance of the unconditional negative binomial estimator and the fixed-effects Poisson estimator, we generated simulated data under the following model. For 100 individuals ($i = 1,\ldots,100$) and two time periods ($t = 1, 2$), let y_{it} have a negative binomial distribution with conditional mean μ_{it} and overdispersion parameter λ (constant over individuals and time). Assume that y_{i1} and y_{i2} are independent, conditional on μ_{it}. Restricting the panel to only two time periods produces conditions most likely to yield evidence of bias due to the incidental parameters problem. Using samples of only 100 cases facilitates the use of conventional software to estimate the unconditional models (by including 99 dummies).

The conditional mean is specified as

$$\mu_{it} = \eta \exp(\beta x_{it} + \gamma z_i),$$

where x_{it} and z_i have standard normal distributions with correlation ρ. The variable z_i will be treated as unobserved. It can be interpreted as representing all the stable, unobserved characteristics of individual i that have some effect on y_{it}. Conditional on z_i, the observed variables x_{i1} and x_{i2} are uncorrelated. Unconditionally, their correlation is ρ^2.

As a baseline model, we set $\beta = 1$, $\gamma = 1$, $\lambda = 1$, and $\rho = 0$. For these parameter values, we generated data for 500 samples, each of size 100. (With two observations per case, the working sample size was 200.) For each sample, we estimated β using a conventional negative binomial

[6] Estimates were obtained with SAS PROC GENMOD.

regression program with x as the predictor, along with 99 dummy variables to capture the fixed effects. We then estimated β via a fixed-effects Poisson regression model, with an overdispersion correction for the standard errors. (Standard errors were multiplied by the square root of the ratio of the Pearson chi-square goodness-of-fit statistic to its degrees of freedom.)

This process was replicated over a range of plausible values for each parameter, with other parameters held at their baseline values. For each set of parameter values, Table 3 gives the mean of the coefficient estimates, standard error (standard deviation across the repeated samples), root mean squared error, and proportion of times that the nominal 95 percent confidence intervals contained the true value. For ease of comparison, the baseline model is replicated within each subpanel of Table 3. These baseline

TABLE 3
Simulation Results for Negative Binomial and Poisson Models

	Negative Binomial					Poisson			
Model	β	SE	RMSE	95% CI Coverage	Non-Conv.	β	SE	RMSE	95% CI Coverage
$\lambda = .2$	.978	.326	.327	.854	0	1.045	.458	.460	.724
$\lambda = .5$	.966	.191	.194	.854	3	1.011	.278	.278	.746
$\lambda = 1$ (base)	.982	.145	.146	.826	156	1.018	.202	.203	.778
$\lambda = 10$	.995	.063	.063	.902	500	1.005	.078	.078	.866
$\lambda = 50$	.996	.052	.052	.952	500	1.002	.053	.053	.928
$\gamma = 0$	.966	.124	.129	.846	327	1.003	.144	.144	.900
$\gamma = .5$	.966	.139	.143	.819	281	1.005	.169	.169	.850
$\gamma = 1$ (base)	.974	.138	.140	.838	140	1.016	.202	.203	.782
$\gamma = 1.5$	.967	.142	.146	.860	53	.999	.281	.281	.640
$\beta = 0$	.008	.116	.116	.872	1	.006	.163	.163	.730
$\beta = .5$	.474	.124	.126	.850	25	.490	.164	.164	.794
$\beta = 1$ (base)	.978	.131	.132	.866	144	1.014	.194	.194	.800
$\beta = 1.5$	1.454	.151	.158	.806	261	1.513	.279	.279	.726
$\eta = 1$	.978	.167	.168	.836	452	1.025	.221	.222	.860
$\eta = 2$	.974	.137	.139	.870	353	1.008	.189	.189	.836
$\eta = 4$ (base)	.968	.139	.142	.844	152	1.009	.211	.211	.758
$\eta = 6$	.972	.125	.128	.850	60	1.004	.202	.202	.752
$\eta = 8$	.977	.131	.133	.840	19	1.006	.209	.209	.760
$\rho = 0$ (base)	.959	.140	.146	.822	139	.989	.197	.197	.780
$\rho = .50$	.972	.160	.162	.846	38	1.011	.311	.311	.646
$\rho = .75$	.978	.204	.205	.872	6	1.016	.451	.451	.588

Note: SE is the standard error, RMSE is the root mean squared error, and CI is the confidence interval.

estimates were made from new random draws in each subpanel, which should provide some feel for the sampling variability of these estimates.

One potential problem that occurred with the negative binomial estimator was that, for many of the samples, the estimate for the overdispersion parameter λ did not converge. The number of nonconvergent samples is shown in Table 3. For the baseline model, this happened in about 20 percent of the samples. For other models, the percentage of convergent samples ranged from zero for true $\lambda = 50$ to 100 for true $\lambda = .2$. Nonconvergence for λ did not seem to affect the estimates for β, however. For all models with appreciable numbers of nonconvergent samples, we compared the means and standard errors of β for the convergent and nonconvergent samples. In no case was there a statistically significant difference, so the results in Table 3 are based on all samples combined.

The general conclusions to be drawn from Table 3 are these:

- There is little evidence for incidental parameters bias. Both the negative binomial and Poisson estimates appear to be approximately unbiased under all conditions, although the negative binomial estimates are always a bit too low.
- Root mean squared errors are appreciably lower for the negative binomial estimator, except when $\lambda = 50$ when the negative binomial distribution is very close to a Poisson distribution.
- Both estimators have confidence intervals that are too small, yielding coverage rates that are often considerably lower than the nominal 95 percent level. The Poisson estimator is much worse in this regard, especially for some of the more extreme parameter values. Although not obvious from the table, these reduced coverage rates stem from standard error estimates that are generally too small.

Now for the details. Variation in λ is crucial for comparing the negative binomial with the Poisson because it controls the degree of overdispersion. More specifically, as $\lambda \to \infty$, the negative binomial converges to the Poisson. Interestingly, both estimators do better in both RMSE and CI coverage when λ is large rather than small, although the degradation in performance with decreasing λ is more rapid for the Poisson estimator.

The parameter γ controls the variance of the stable, unobserved heterogeneity. The performance of the negative binomial estimator is hardly affected at all by changes in γ. But for the Poisson, increases in γ produce both substantial increases in RMSE and major decreases in CI coverage. Variations in the true value of β also show little impact on the performance of the negative binomial estimator. For the Poisson estimator, the CI coverage remains fairly stable with variations in β, but there is some evidence for an increase in the RMSE as β gets larger.

The parameter η is a scale factor that affects both the mean and variance of the counts. For these models, $\eta = 1$ produces a mean of about 3.8 while $\eta = 8$ yields a mean of 23. This is potentially important because when the mean is small, large proportions of the sample will have a count of 0 and it becomes increasingly difficult to discriminate between a Poisson distribution and a negative binomial distribution. In Table 3, we see that for $\eta = 1$, the Poisson estimator actually does a little better than the negative binomial estimator in CI coverage, although its RMSE is still about 30 percent larger. As η gets larger, the coverage rate for the Poisson estimator deteriorates, while remaining stable for the negative binomial estimator.

Finally, we examine the impact of ρ, the correlation between the observed variable x and the source of unobserved heterogeneity. When $\rho = 0$, as with all the models examined thus far, we satisfy the assumptions of a random effects model and could, presumably, do better using a random-effects negative binomial or Poisson estimator. When $\rho \neq 0$, random-effects estimators are likely to be biased, while fixed-effects estimators should remove that bias. Table 3 shows that both the negative binomial and Poisson estimators do a good job of avoiding bias in the estimate of β. However, with increasing ρ, the performance of the negative binomial estimator remains stable, while the Poisson estimator deteriorates substantially in both RMSE and CI coverage.

In sum, the message of Table 3 is that, under the specified model, the unconditional fixed-effects negative binomial estimator is virtually always a better choice than the fixed-effects Poisson estimator. But it is still troubling that the negative binomial estimator is accompanied by underestimates of the standard errors, leading to insufficient coverage of confidence intervals. It is natural to ask whether there is some way to adjust the standard errors upward. Table 4 shows the consequences of multiplying the standard errors by the square root of the ratio of the deviance to its degrees of freedom,[7] where the deviance is defined as

$$D = \sum_i \sum_t \{y_{it} \log(y_{it}/\mu_{it}) - (y_{it} + \lambda)\log[(y_{it} + \lambda)/(\mu_{it} + \lambda)]\}. \quad (14)$$

[7]With SAS PROC GENMOD, this correction can be implemented with the DSCALE option on the MODEL statement.

TABLE 4
Confidence Interval Coverage for Negative Binomial Model with Deviance Overdispersion Correction.

Model	95% CI Coverage	Model	95% CI Coverage
$\lambda = .2$	.982	$\beta = 0$	.960
$\lambda = .5$	.972	$\beta = .5$	.972
$\lambda = 1$ (baseline)	.956	$\beta = 1$ (baseline)	.948
$\lambda = 10$	.956	$\beta = 1.5$	.940
$\lambda = 50$	.956	$\eta = 1$	.954
$\gamma = 0$	.966	$\eta = 2$	.964
$\gamma = .5$	.956	$\eta = 4$ (baseline)	.956
$\gamma = 1$ (baseline)	.960	$\eta = 6$	.968
$\gamma = 1.5$	.952	$\eta = 8$	.952
		$\rho = 0$ (baseline)	.962
		$\rho = .50$	.964
		$\rho = .75$	.950

With this correction, confidence intervals have close to their nominal coverage for all parameter values considered in the simulation. Somewhat surprisingly, standard error correction using the Pearson chi-square goodness-of-fit statistic did not produce any noticeable improvement over the conventional model-based standard error estimates (results not shown). Also, use of the deviance-based correction did not improve the confidence interval coverage for the Poisson estimator.

6. AN APPROXIMATE CONDITIONAL ESTIMATOR

Previously we remarked that conditional inference is not feasible for the NB2 model (with event counts independent over time for each individual) because there is no complete, sufficient statistic for the incidental parameters that is a function of the data alone. However, Waterman and Lindsay (1996a), following the work of Small and McLeish (1989), have introduced an approximate method that mimics the beneficial properties of conditional inference, even in situations where a straightforward conditioning approach fails. This methodology is termed the projected score method.

In conventional maximum likelihood estimation, the log-likelihood is differentiated to produce the score function. This function is then set equal to zero, and the solutions to this equation are the MLE's. The projected score method brings the score function itself to the center of attention, and engineers a version of the score function that has properties equivalent to the conditional score function if it existed; the desirable property is that among all estimating functions that are insensitive to the incidental parameters, it provides the maximal information.

Here are some details of the method. Let β be a vector of parameters of interest and let δ be a vector of nuisance (incidental) parameters. Let $U_0(\beta, \delta)$ be the conventional score function—that is, the first derivative of the log-likelihood function with respect to β. Let $U_\infty(\beta, \delta)$ denote the optimal estimating function, which is defined as follows. We restrict attention to all square, integrable functions $g(\beta, \delta)$ that satisfy the strong unbiasedness condition,

$$E\{g(\beta_0, \delta_0)\} = 0,$$

for all true values of δ, and for any values of β_0 and δ_0. This condition implies that the estimating function is insensitive to the values of the nuisance parameters, which is what we desire in a conditional method. Among functions that satisfy this condition, the optimal estimating equation is the one whose solution has lowest asymptotic variance. This function exists whenever certain regularity conditions are satisfied (Waterman and Lindsay 1996a). When a complete sufficient statistic exists, this optimal estimating function is identical to the score function for the conditional likelihood.

It can be shown (Waterman and Lindsay 1996a) that the optimal estimating function U_∞ can be expressed as an infinite series. Consider a single individual i with nuisance parameters δ_i. Define $V_\alpha = f^{(\alpha)}/f$ where f is the density function and $f^{(\alpha)}$ is the αth derivative of f with respect to δ_i. Then, we have for individual i

$$U_\infty(\beta, \delta) = U_0(\beta, \delta) - \sum_{\alpha=1}^{\infty} \rho_\alpha V_\alpha,$$

where the ρ_a are coefficients that depend on the parameters but not the data.

We approximate U by the first r terms of this series:

$$U_r(\beta, \delta) = U_0(\beta, \delta) - \sum_{\alpha=1}^{r} \rho_\alpha V_\alpha$$

Clearly one could construct an entire sequence of approximations to the optimal estimating function, but the hope is that the first approximation, denoted as the U_2 estimating function, is close enough for practical purposes. Waterman and Lindsay (1996b) show a number of examples for which this is the case. The way in which these approximate score functions are engineered to be close to the optimal one is identical to the way a least squares line is engineered to be close to the data. That is, a regression approach is used to obtain estimated values of ρ_a, but here the objects of interest are functions rather than data points. This is achieved by taking a set of derivatives of the score functions and their cross products and then finding expectations, so that the mathematical operations are differentiation and expectation. The effort in accomplishing this is minimized by using symbolic software, such as Mathematica or Maple, which can derive the functions with relatively modest input from the analyst. Once the projected score function has been obtained, the ML solutions can be obtained using standard software packages.

Using the U_2 approximation, we applied the projected score method to the NB2 model, which was the basis for the simulation study of the previous section. (Mathematica and R programs for accomplishing this are available from the authors.) Simulation results are displayed in Table 5.

Comparing the projected score estimates in Table 5 with the unconditional estimates in Tables 3 and 4, we find noticeably less bias in the projected score estimates for every condition. On the other hand, the standard errors for the projected score estimates are somewhat larger than those for the unconditional estimates in every case but one. Combining these results into the root mean squared errors, we find that the unconditional method does better in 13 out of the 21 conditions. With respect to confidence interval coverage, the projected score method is always appreciably better than the unconditional method using the uncorrected standard errors (Table 3). But when the unconditional estimates are corrected by the deviance (Table 4), the resulting confidence interval coverage is always closer to the nominal level than the coverage of the projected score method.

In sum, it does not appear that the projected score method based on the U_2 approximation offers any substantial advantage over the unconditional method with corrected standard errors, at least with respect to estimating β, the regression coefficients. However, the projected score method was much better at estimating λ, the overdispersion parameter. The number of convergence failures was far lower using the projected score method. Furthermore, if we restrict our attention to samples in which the estimate

TABLE 5
Simulation Results for Projected Score Method

	Projected Score Negative Binomial				
Model	β	SE	RMSE	95% CI Coverage	Non-Convg.
$\lambda = .2$	.994	.336	.336	.929	22
$\lambda = .5$	.993	.211	.211	.916	7
$\lambda = 1$ (baseline)	.988	.139	.139	.934	1
$\lambda = 10$	1.002	.069	.069	.919	7
$\lambda = 50$	1.001	.053	.053	.933	65
$\gamma = 0$	.995	.136	.136	.932	0
$\gamma = .5$	.997	.141	.141	.924	0
$\gamma = 1$ (baseline)	.996	.140	.140	.944	0
$\gamma = 1.5$	1.003	.143	.143	.927	4
$\beta = 0$	.005	.125	.125	.940	0
$\beta = .5$	.504	.133	.133	.936	1
$\beta = 1$ (baseline)	1.004	.140	.140	.936	0
$\beta = 1.5$	1.493	.155	.155	.934	2
$\eta = 1$	1.004	.178	.178	.930	0
$\eta = 2$	1.008	.164	.164	.932	1
$\eta = 4$ (baseline)	.989	.141	.142	.920	0
$\eta = 6$	1.000	.136	.136	.914	1
$\eta = 8$	1.002	.132	.132	.934	0
$\rho = 0$ (baseline)	.993	.140	.140	.924	0
$\rho = .50$	.996	.141	.141	.920	0
$\rho = .75$	1.007	.136	.136	.948	0

of λ converged, the unconditional estimates of λ had substantially greater upward bias than the projected score estimates (not shown in the tables). In principle, the projected score method could be improved by using more terms in the approximation.

7. CONCLUSION

The negative binomial model of Hausman, Hall, and Griliches (1984) and its associated conditional likelihood estimator does not accomplish what is usually desired in a fixed-effects method, the control of all stable covariates. That is because the model is based on a regression decomposition of the *overdispersion* parameter rather than the usual regression decomposition of the mean. Symptomatic of the problem is that programs that implement the conditional estimator have no difficulty estimating an intercept or coefficients for time-invariant covariates.

A good alternative is to do conventional negative binomial regression with direct estimation of the fixed effects rather than conditioning them out of the likelihood. Greene (2001) has demonstrated the computational feasibility of this approach, even with large sample sizes. Simulation results strongly suggest that this estimation method does not suffer from incidental parameters bias, and has much better sampling properties than the fixed-effects Poisson estimator. Bias in standard error estimates can be virtually eliminated by using a correction factor based on the deviance.

The approximate conditional score method is another attractive alternative. The approximation used here showed slightly less bias in the coefficient estimates but slightly more sampling variability than the unconditional estimator. This performance could be improved still further by using a higher-order approximation. Furthermore, estimation of the overdispersion parameter was much better with the approximate conditional method than with the unconditional method.

REFERENCES

Allison, Paul D. 1996. "Fixed Effects Partial Likelihood for Repeated Events." *Sociological Methods and Research* 25:207–22.

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

Chamberlain, Gary A. 1980. "Analysis of Covariance with Qualitative Data." *Review of Economic Studies* 47:225–38.

Guo, Guang. 1996. "Negative Multinomial Regression Models for Clustered Event Counts." Pp. 113–32 in *Sociological Methodology 1996*, edited by Adrian E. Raftery. Washington, DC: American Sociological Association.

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

Hausman, Jerry, Bronwyn H. Hall, and Zvi Griliches. 1984. "Econometric Models for Count Data with an Application to the Patents–R&D Relationship." *Econometrica* 52:909–38.

Hsiao, C. 1986. *Analysis of Panel Data.* Cambridge, England: Cambridge University Press.

Johnson, Norman L., and Samuel Kotz. 1969. *Discrete Distributions.* New York: Wiley.

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*, Ser. B, 32:175–208.

Mundlak, Y. 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* 68:563–66.

Small, C. G., and D. L. McLeish. 1989. "Projection as a Method for Increasing Sensitivity and Eliminating Nuisance Parameters." *Biometrika* 76:693–703.

Waterman, Richard P., and Bruce G. Lindsay. 1996a. "Projected Score Methods for Approximating Conditional Scores." *Biometrika* 83:1–13.

———. 1996b. "A Simple and Accurate Method for Approximate Conditional Inference Applied to Exponential Family Models." *Journal of the Royal Statistical Society*, Ser. B, 58:177–88.

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.