COUNT DATA MODELS WITH VARIANCE OF UNKNOWN FORM: AN APPLICATION TO A HEDONIC MODEL OF WORKER ABSENTEEISM

Miguel A. Delgado and Thomas J. Kniesner*

Abstract—We examine an econometric model of counts of worker absences due to illness in a sluggishly adjusting hedonic labor market. We compare three estimators that parameterize the conditional variance—least squares, Poisson, and negative binomial pseudo maximum likelihood—to generalized least squares (GLS) using nonparametric estimates of the conditional variance. Our data support the hedonic absenteeism model. Semiparametric GLS coefficients are similar in sign, magnitude, and statistical significance to coefficients where the mean and variance of the errors are specified ex ante. In our data, coefficient estimates are sensitive to a regressor list but not to the econometric technique, including correcting for possible heteroskedasticity of unknown form.

I. Introduction

THE NUMBER of days sick, the number of visits to a physician, the number of jobs held, and the number of purchases of a good or service are examples of microeconomic data that are counts of events in an interval of time. We investigate the causes of worker absenteeism via a discrete regression model where the dependent variable measures the number of times a worker is absent from a job in a year. The regression function is linearly exponential, a specification commonly applied to count data to ensure nonnegative conditional expectations. (See Hausman et al. (1984), Cameron and Trivedi (1986), and Cameron et al. (1988).) In the context of examining the microeconometrics of worker absenteeism our research compares the empirical performance of semiparametric generalized least-squares (GLS) estimators with the empirical performance of popular estimators of count data models.

The theoretical model underlying our absence count regressions admits sluggish adjustment to hedonic labor market equilibrium. We examine four models: least squares, Poisson, negative binomial pseudo maximum likelihood, and generalized least squares with heteroskedasticity of unknown form. Regression coefficients and standard errors are generally similar across the econometric models we estimated. In our data the underlying economic model (equilibrium versus sluggish adjustment) is much more important to the parameter estimates than the regression model, including correcting for subtle heteroskedasticity.

Received for publication October 5, 1992. Revision accepted for publication April 18, 1995.

* Universidad Carlos III de Madrid and Indiana University, respectively.

We are grateful for financial support provided by the Spanish Direccion General de Investigacion Cientifica y Tecnica (DGICYT, reference no. PB92-0247), the Economic and Social Research Council of the United Kingdom (ERSC, reference no. R000 231441), both the Department of Economics of the Research School of the Social Sciences and the Department of Statistics of the Faculties at the Australian National University, the Hoover Institution of Stanford University, the Department of Economics of University College London, the CentER for Economic Research of Tilburg University, and the Office of the Vice-President for Research and the University Graduate School of Indiana University. We also thank London Regional Transport for providing data, James P. Ziliak for skillful computing assistance, Elaine Silver for typing tables, and Elisabeth Soller and Deborah Garlow for referencing and proofreading. Finally, thanks go to Steven G. Allen, Colin Cameron, Jose Carbajo, Anthony D. Hall, V. Joseph Hotz, Peter Robinson, Thanasis Stengos, Pravin K. Trivedi, and two anonymous referees for their helpful comments.

II. A Microeconometric Model of Worker Absenteeism

The ideal microeconometric model of worker absenteeism has several distinguishing properties (Avery and Hotz (1984), Barmby et al. (1991)). First, absences depend on both personal (supply) and job (demand) characteristics. Second, work attendance is a dynamic decision with possibly sluggish adjustment in the short run to a changing economic environment. Third, because absences are counts, conditional variances are typically a function of absences' conditional means (Patil (1970)). Least-squares regression produces inefficient estimators when absence counts are the dependent variable, and ignoring the accompanying conditional heteroskedasticity yields inconsistent standard errors and invalidates hypothesis tests. Because it is not obvious how to parametrize the heteroskedasticity in an absence count model, it is desirable to minimize the number of ex ante assumptions. The union of desirable econometric dimensions of a model of worker absenteeism suggests that investigating semiparametric regressions of individual absence counts on the worker's personal and job characteristics and past absenteeism could prove informative. To our knowledge, we present the first microeconometric absenteeism model jointly recognizing supply and demand forces, sluggish adjustment, and allowing heteroskedasticity of unknown form.

The economic structure underlying labor market outcomes involving job attributes, including the regularity of work attendance, is the theory of hedonic labor market outcomes (Rosen (1986)). A matching of workers and firms in the labor market produces a locus of wage–absenteeism pairings that is positioned by the personal traits of workers, the economic and technological characteristics of employers, and the encompassing institutional and legal environment. For some issues, a researcher must uncover the employers' cost functions and workers' utility functions supporting the hedonic locus. Stringent a priori restrictions are needed to identify the complete structure of hedonic equilibrium models (Brown and Rosen (1982), Epple (1987), and Kahn and Lang (1988)). Alternatively, a researcher can numerically simulate hedonic equilibrium over a set of cost and utility function parameters (Kniesner and Leeth (1995)). Our interest is in robustly estimating the market locus of matches of pay and nonwage characteristics of employment.

When absenteeism is an aspect of the employment relationship, hedonic labor market equilibrium is described

algebraically as

$$W_i = f(a_i, C_i; S_i, D_i, E_i, \epsilon_i) \quad (1)$$

where i indexes individuals, W is the wage rate, a is the absence rate, C is the vector of other nonwage characteristics of employment, $f_a < 0$ and $f_C \gtrless 0$, depending on whether the worker views the particular nonwage characteristic as an (un)desirable aspect of employment. The information conditioning the hedonic locus in (1) includes a vector of the worker's personal and economic characteristics (S), a vector of the employing firm's technological and economic traits (D), the surrounding legal and institutional environment (E), and a stochastic error term with unknown distribution (ϵ) to emphasize that the labor market outcomes described in (1) incorporate unpredictable random components.

Because we study absenteeism and workers in our data have only a few pay grades (P), we estimated the inverse hedonic locus

$$a_i = f^{-1}(P_i, C_i; S_i, D_i, E_i, \epsilon_i). \quad (2)$$

Moreover, family and production schedules can be difficult to change quickly. A sluggish adjustment version of the inverse hedonic equilibrium locus (2),

$$a_i = f^{-1}(a_{-ji}, P_i, C_i; S_i, D_i, E_i, \epsilon_i) \quad (3)$$

where $j = 1, \ldots, T$ indexes the time period, acknowledges that absenteeism may be part of a worker's short-run labor supply decision with the adjustment in work attendance due to new health or economic circumstances distributed over time.[1] Semiparametric count data regressions of the lagged adjustment absenteeism equation (3) encompass the desirable characteristics of a microeconometric model of worker absenteeism (Avery and Hotz (1984)).

III. Count Data Models—Econometric Background

Our econometric estimates of the theoretical absenteeism equation (3) have

$$E(a_i|X_i) = \exp(X_i'\beta_0) \quad (4)$$

where X_i' is the vector of explanatory variables $[a_{-ji}, P_i, C_i; S_i, D_i, E_i]$ and β_0 is a vector of unknown parameters. The linear exponential specification of the absence rate in (4) is common in count data models to ensure a positive conditional expectation estimate. (See Gourieroux et al. (1984a, b), Hausman et al. (1984), Cameron and Trivedi (1986), and Gurmu and Trivedi (1993).) Unlike other specifications ensuring positive conditional means, such as a logistic curve, the linear exponential specification (4) emits convenient economic interpretations.

A. *Basic Count Data Specifications*

Given our ex ante choice of a linear exponential regression model of absence counts in (4), we note that for any vector of functions $g(X)$ the moment restriction

$$E[(a_i - \exp(X_i'\beta_0))g(X_i)|X_i] = 0 \quad (5)$$

holds. The moment restriction in (5) is the basis for many estimators.

Because absenteeism is the sum of absences in an interval of given length, an obvious first econometric specification is the Poisson, where $E(a_i|X_i) = \exp(X_i'\beta_0) = \text{Var}(a_i|X_i)$. The choice of $g(X_i) = X_i$ in (5) produces the Poisson pseudo maximum-likelihood estimator[2] (PMLE) of β_0, which solves

$$\sum_i [a_i - \exp(X_i'\beta)]X_i = 0. \quad (6)$$

The asymptotic variance of the Poisson PMLE in (5) and (6) is

$$E[X_iX_i' \exp(X_i'\beta_0)]^{-1}E(X_iX_i'\sigma_i^2)E[X_iX_i' \exp(X_i'\beta_0)]^{-1}. \quad (7)$$

When the underlying model is Poisson and $\sigma_i^2 = \exp(X_i'\beta_0)$ the estimator is fully efficient. Choosing $g(X_i) = X_i \cdot \exp(X_i'\beta_0)W(X_i)$, where $W(X_i)$ are weights depending on the regressors, produces the weighted least squares estimator solving the equation

$$\sum_i [a_i - \exp(X_i'\beta)]X_i \exp(X_i'\beta)W(X_i) = 0 \quad (8)$$

which has asymptotic variance

$$E[X_iX_i' \exp(2X_i'\beta_0)W(X_i)]^{-1}E[X_iX_i' \times \exp(2X_i'\beta_0)\, W(X_i)^2\sigma_i^2]^{-1}E[X_iX_i' \exp(2X_i'\beta_0)W(X_i)]^{-1}. \quad (9)$$

Among the class of weighted least-squares estimators the most efficient uses the weights $W(X_i) = \sigma_i^{-2}$, which is termed infeasible GLS. Notice that the infeasible GLS and the Poisson maximum likelihood (when the underlying distribution is indeed Poisson) are asymptotically equally efficient.

[1] We acknowledge that lagged absences (a_{ji}) may not be uncorrelated with the current errors (ϵ_i), but developing an instrumental-variables count data estimator to confront the possible econometric consequences of a lagged dependent variable in a cross-section context is tangential to our research objectives. As an alternative we present regressions where lagged absences are removed from the regressor list.

[2] Pseudo maximum likelihood refers to the case where an ex ante specified probability distribution may not be the true distribution, but maximum-likelihood estimation is used as though the specified distribution applied. In general, model misspecification leads to an inconsistent estimator. In the special case where the number of absences realized is specified to have a distribution belonging to the linear exponential family the PMLE is consistent if the mean is correctly specified. (See Gourieroux et al. (1984a), Cameron and Trivedi (1986), and McCullagh and Nelder (1989).)

An assumed equality between the mean and the variance is restrictive in economic applications, and this is why researchers have proposed more general count data models (Cameron and Trivedi (1986), Lawless (1987), and Gurmu and Trivedi (1993)). A popular generalization is the compound Poisson, where

$$\Pr\{a_i = \delta_i | X_i\} = \int \frac{\exp[-\exp(X_i'\beta_0 + \epsilon_i)]\exp[\delta_i(X_i'\beta_0 + \epsilon_i)]}{\delta_i!} \times h(\epsilon_i)\, d(\epsilon_i) \quad (10)$$

and $h(\epsilon_i)$ is the probability density function of ϵ_i. For convenience, researchers sometimes assume that $u_i = \exp(\epsilon_i)$ is distributed as a $\Gamma(\varphi, \alpha)$. When there is a constant term in $X_i'\beta$, there is no loss of generality in setting $E[\exp(\epsilon_i)] = 1$, which means that $\alpha = \phi^{-1} = \text{Var}[\exp(\epsilon_i)]$, and the conditional distribution of a_i given δ_i is negative binomial. Specifically,

$$\Pr\{a_i = \delta_i | X_i\} = \frac{\Gamma(\alpha^{-1} + \delta_i)}{\Gamma(\alpha^{-1})\delta_i!} \times [\alpha\exp(X_i'\beta_0)]^{\delta_i}[1 + \alpha\exp(X_i'\beta_0)]^{-(\alpha^{-1}+\delta_i)}. \quad (11)$$

The negative binomial PMLE maximizes the likelihood based on the probability distribution (11).[3] The estimator we are describing can be inconsistent when $\exp(\epsilon_i)$ does not have a gamma distribution (Gourieroux et al. (1984b)).[4] It is then interesting to propose simple estimation methods leading to consistent estimators for all possible distributions $h(\epsilon_i)$ having second moments.

Assuming that $E[\exp(\epsilon_i)] = 1$ and $\text{Var}[\exp(\epsilon_i)] = \eta^2$ in (10),

$$E[a_i|X_i] = \exp(X_i'\beta_0)$$
$$\text{Var}(a_i|X_i) = \exp(X_i'\beta_0)[1 + \eta^2\exp(X_i'\beta_0)]. \quad (12)$$

In the negative binomial case $\eta^2 = \alpha$. Feasible GLS based on the variance specification in (12) is always consistent, but less efficient than negative binomial maximum likelihood when the true conditional distribution is (11) (Gourieroux et al. (1984a, b) and McCullagh and Nelder (1989)).

To cover a wide range of specifications representative of the count data literature, we study linear exponential absenteeism parameter estimates from least squares, Poisson PMLE, negative binomial PMLE, and an optimal GLS that we now explain.

[3] Lawless (1987) provides a computationally convenient expression for the negative binomial likelihood.

[4] Notice that the maximum likelihood solves simultaneously k equations such as (8), corresponding to the derivatives of the log likelihood function with respect to β and an additional equation corresponding to the derivative with respect to α.

B. Variance of Unknown Form

There is no obvious a priori reason to begin with a particular specification for the variance of absenteeism such as (12). An alternative approach that we use is linear exponential absences in (4) plus a nonparametric function for the conditional variances

$$\sigma_i^2 = \text{Var}(a_i|X_i). \quad (13)$$

The conditional mean from (4) and the conditional variance in (13) form the semiparametric count data model we apply to absenteeism data for London bus drivers. We also estimated the hedonic absenteeism locus with semiparametric GLS using as estimated conditional variances

$$\hat{\sigma}_i^2 = \sum_j [a_j - \exp(X_j'\tilde{\beta})]^2 w_{ij} \quad (14)$$

where $\tilde{\beta}$ is a preliminary root-n consistent parameter estimate, and w_{ij} are nonparametric k nearest neighbor (nn) probablistic weights (see appendix A). Specifically, we used so-called uniform k nn weights to estimate the variances in (14) (Robinson (1987a)).[5]

The semiparametric GLS estimator we applied estimates β_0 via the solution to

$$\sum_i [a_i - \exp(X_i'\beta)]X_i \exp(X_i'\beta)\hat{\sigma}_i^{-2} = 0. \quad (15)$$

Under regularity conditions the vector $\hat{\beta}$ that solves the first-order condition (15) has asymptotic variance

$$\text{AsyVar}\{n^{1/2}(\hat{\beta} - \beta_0)\} \equiv I_0^{-1} = E[X_iX_i'\exp(2X_i'\beta)\sigma_i^{-2}]^{-1}. \quad (16)$$

The semiparametric efficiency bound in (16) cannot be bettered under the information set in the model, equation (4) (Chamberlain (1987)).

Regularity conditions needed for asymptotic normality of the solution to semiparametric GLS are similar to the moment conditions needed for asymptotic normality of GLS. Our nearest neighbor weights require that the smoothing parameter, $k \equiv$ number of nearest neighbors, increase with the sample size but at a slower rate (Robinson (1987a)

[5] In general, given observations $\{(y_1, x_1), \ldots, (y_n, x_n)\}$ of the stochastic pair (Y, X), a nonparametric estimate of $E(Y|X = x)$ is a weighted average of Y_j's, where the weights depend on how close the corresponding X_j is to x. Specifically, k nn estimates are a weighted average of the Y_j's such that the corresponding X_j is one of the $k$$X$'s closest to x, according to the scaled Euclidean distance. As the number of observations increases, the number of X_j's close to x also increases, which intuitively explains why the number of terms in the weighted average (the number of nearest neighbors) must increase with the sample size. So, fixed weights in the estimated conditional variances in (14), such that $w_{ij} = 1$ if $i = j$ and $w_{ij} = 0$ otherwise, produce an inconsistent estimator. In appendix A we formally explain the k nn weights. For more discussion see also Härdle (1990) and Delgado and Robinson (1992).

and Delgado (1992)). We examine two different nearest neighbor specifications ($k = n^{1/2}$ and $k = n^{3/5}$) to illustrate how sensitive the procedure is to the choice of the number of nearest neighbors. We also estimate the covariance matrix implied by (16) using both the corresponding sample analog and Eicker–White heteroskedasticity robust procedures, as recommended by Robinson (1987b), to protect against a possibly poor choice of the number of nearest neighbors in a finite sample. Specifically, we also estimated the coefficient (co)variances I^{-1} by

$$\tilde{I}_0^{-1} = \left[n^{-1}\sum_i \tilde{X}_i\tilde{X}_i'\tilde{\sigma}_i^{-2}\right]^{-1}\left[n^{-1}\sum_i \tilde{X}_i\tilde{X}_i'\tilde{u}_i^2\hat{\sigma}_i^{-2}\right] \times \left[n^{-1}\sum_i \tilde{X}_i\tilde{X}_i'\hat{\sigma}_i^{-2}\right]^{-1} \tag{17}$$

where $\tilde{X}_i = X_i \exp(X_i'\tilde{\beta})$ and $\tilde{u}_i = a_i - \exp(X_i'\tilde{\beta})$. For the linear exponential specification (4) we present robust standard errors of the least squares plus Poisson and negative binomial PMLE coefficients (White (1982)).

Recapitulation. In examining count models of worker absences we first estimated least squares, then Poisson and negative binomial PMLE. For maximum generality we provide semiparametric GLS estimates using the initial β's from least squares.[6]

IV. Empirical Results

Our data cover absences by persons working for London buses as conductors, drivers, and single-person operators during January 1, 1981, to December 31, 1985.[7] Information on absences includes the starting and returning dates, reason, and justification. Also our data include personal characteristics, including sex, age, and home address, plus job-related characteristics, including garage name and starting date of work. There are over 200,000 absence histories for 17,720 workers. We restricted the sample to persons employed in all five years (12,549) minus cases for which we could not determine garage location, leaving 5501 workers.[8]

A. *Econometric Strategy*

Because the medical documentation required for a given type of absence changed during the sample years, we focused on 1985 absenteeism. The frequency distribution of absences in our data and the physical and institutional differences among absenteeism spell groups, in particular medical documentation required, made it natural to group spells as 1–7 days, 8–14 days, and 14+ days. Because short-term absences are the absences most subject to individual discretion, and short-term absences have the most interpersonal variation, the dependent variable in our regression is the number of absence spells of one week or less.[9]

We estimate a linear exponential regression of $a_i \equiv$ absence spells of seven days or less in 1985 on the vector $X_i \equiv [a_{-1i}, a_{-2i}, P_i, C_i; S_i, D_i, E_i]$, which is the regression model capturing the outcomes of hedonic labor markets with sluggish adjustment described theoretically by equations (3) and (4). Given the regressor vector contains lagged absenteeism, personal characteristics, and workplace characteristics (X_i), the conditional expectation of absenteeism is $\exp(X_i'\beta_0)$, where β_0 is the unknown vector of parameters to estimate. The workers' personal and workplace characteristics regressors include age, sex, marital status, health, length of service with the bus company, distance of journey to work, and plant size as metered by the number of people working in the bus garage (Jones (1971)).

We present results from four estimators: least squares, Poisson, negative binomial PMLE, and a semiparametric GLS estimator that was iterated until convergence from the least-squares coefficient estimates. To illustrate the sensitivity of the semiparametric estimates of the choice of the number of nearest neighbors (k), we report two different choices, $k = [n^{1/2}]$ and $k = [n^{3/5}]$. For all regression models we report robust and unrobust standard errors (Eicker (1963) and White (1980, 1982)).

Unlike least squares and semiparametric GLS, the Poisson and negative binomial PMLEs are not weighted least-squares procedures. However, the Poisson PMLE can be computed by means of an iterative weighted least-squares procedure. The negative binomial PMLE can also be computed by means of an iterative procedure, where in a first step the objective function is concentrated with respect to α, and then β is estimated by iterative least squares. The resulting $\hat{\beta}$ is substituted into the objective function, which is then optimized with respect to α. The procedure is then repeated until convergence. It is important to recognize that neither Poisson nor negative binomial PMLE can be expressed as solving equations such as (8).

Economic Focus. Before discussing regression results we want to foreshadow our contribution to the economic literature on worker absenteeism. Because our data are for a single employer in a single city, we did not estimate the effects of potentially important absenteeism policies, such as sick leave benefits and work schedule flexibility. We also examine how workplace health hazards affect absenteeism only to the extent that distance from the bus garage to the center of London reflects worker health risks due to pollution or stress. Our emphasis is on whether two core results

[6] Computer programs for estimating semiparametric regressions are described in Delgado (1993).

[7] Norman and Spratling (1956) investigated absences caused by sickness among the personnel of the London Transport Company. Cornwall and Raffle (1961) studied the absenteeism of women bus conductors in London during 1953–1957.

[8] Regression variables are defined in appendix B.

[9] As a point of reference, other studies have typically measured absenteeism as a logistic of either the proportion of time absent during a survey reference period, such as the two weeks prior to the survey, or as whether the person was absent from work on the survey date. See, for example, Allen (1981a, b) and Barmby et al. (1991).

from the microeconometric literature on absenteeism appeared in our count data regressions. Specifically, do we find a substantial negative impact of age on absences coupled with statistically insignificant effects of other demographic characteristics (Allen (1981a, b))?

B. *Coefficient Estimates*

The results in table 1 support a sluggish adjustment hedonic model of worker absenteeism. Pay level and absenteeism vary inversely, ceteris paribus. Because we have two effective pay grades in our data for London buses, accepting the hedonic labor market interpretation requires rejecting the null hypothesis that the coefficient of *Driver* is zero against the alternative that the coefficient of *Driver* is negative. All specifications in table 1 have significantly lower absence rates for the higher wage workers—drivers.[10] The coefficients of the two lagged dependent variables *Abs*84 and *Abs*83, are significantly positive across models, and their sum is in the range of 0.13 to 0.20, so that the estimated long-run effects of regressors on absenteeism are about 15% to 25% larger than the short-run effects of regressors on absenteeism. Satisfied that we can interpret the regressions in table 1 in the spirit of hedonic labor markets with sluggish adjustment to changing economic circumstances, we now turn our attention to how the remaining coefficient estimates square with the existing microeconometric literature on worker absenteeism.

A well-known result in the absenteeism literature is that more mature workers are absent less often. In all regressions in table 1 the short-run elasticity of age is significantly negative, so that a firm whose workers are 10% older than average will have 5% to 9% fewer short-term absence spells. Also consistent with previous research is a haphazard pattern of demographic effects. Although the effects of gender and health status (as captured by long-term absence spells, *LongAbs*) are insignificant, the coefficient of *Family* is generally significant and implies that married workers have about 7% to 10% higher absenteeism in the short run, with the estimated effects of marriage larger in the regressions with variance of unknown form than in their counterparts with the first two error moments specified ex ante. Overall the results in table 1 are consistent with the theoretical model guiding our empirical research, and the coefficient estimates conform to the pattern appearing in previous microeconomic research on worker absenteeism.

C. *Model Selection Results*

Although the estimator with variance of unknown form removes heteroskedasticity, the semiparametric regressions in the last two columns of table 1 can be viewed as slightly less efficient than the Poisson PMLE in the first column of table 1. To elaborate, the robust standard errors of the Poisson PMLE in column 1 tend to be about 10% smaller than the robust standard errors of the coefficients of the regression models in table 1 that permit ex ante unspecified heteroskedasticity. The difference between robust and unrobust standard errors is an indication of the correct specification of the models. Standard error differences are larger for the negative binomial PMLE than for the rest of the estimators. The differences between (un)robust standard errors for the Poisson PMLE are relatively small, which suggests that the Poisson specification is not bad. However, as we will soon note, a test for overdispersion rejects the Poisson specification. Where coefficients' signs, magnitudes, and statistical significance are concerned, it makes little difference in our data whether we used least squares, Poisson or negative binomial PMLE, or semiparametric GLS.

A convenient additional check of the Poisson absenteeism model is a regression-based test for equality of the conditional mean and conditional variance (Cameron and Trivedi (1990)). We tested the equidispersion property of the Poisson absence count regression model in the first column of table 1 by testing the null hypothesis, H_0: $\alpha = 0$, in the artificial regression

$$\frac{[a_i - \exp(X_i'\hat{\beta})]^2}{\exp(X_i'\hat{\beta})} - 1 = \alpha \exp(X_i'\hat{\beta}) + \text{error}. \qquad (18)$$

Rejecting the null hypothesis H_0: $\alpha = 0$ rejects the Poisson specification because the estimated conditional mean and variance are not equal. In the regression-based test of equality of conditional mean and conditional variance in equation (18) $\hat{\alpha} = 0.19$ with a $t = 15.4$ so that our data reject the Poisson specification against the more general alternative, where $\sigma_i^2 = \exp(X_i'\beta)[1 + \alpha \exp(X_i'\beta)]$.[11] However, because $\hat{\alpha}$ is small, the distinction between Poisson and negative binomial PMLE should not be overemphasized.

Whether the Poisson or the negative binomial specifications are (un)convincing, the semiparametric GLS seems a sensible alternative.

D. *Robustness Checks*

We also examined the robustness of our results to an increase in the number of nearest neighbors and to two

[10] We note that the coefficient of the dummy variable for the driver reflects the absence rate effects of the entire vector of attributes of the driver occupation, including higher education and possibly greater job satisfaction. We do not claim that the coefficient of *Driver* reflects only higher pay, but that in order not to reject the hedonic interpretation of our absenteeism count regression the coefficient of *Driver* need be significantly greater than zero.

[11] To elaborate, we rejected the null hypothesis of equidispersion by rejecting the hypothesis H_0: $\alpha = 0$ in the ancillary regression (18). The 95% confidence interval for $\hat{\alpha}$, which is [0.166, 0.214], emphasizes the low level of overdispersion. In our raw data the ratio of the variance of absences to the mean of absences is Var (abs)/mean (abs) = $(2.33)^2/2.17$ = 2.5. When conditioning on the regressors X, the mean scaled variance will fall below 2.0. The point is that there is not much overdispersion in our data.

TABLE 1.—ESTIMATES OF SLUGGISH ADJUSTMENT MODELS OF WORKER ABSENTEEISM[a] (n = 5501)

Regressor {Mean/Std. Dev.}	PMLE (Poisson)[b]	PMLE (NegBin)[c]	NLLS[d]	SEMIPAR[e] ($k = n^{1/2}$)	SEMIPAR ($k = n^{3/5}$)
Intercept	4.1779	4.1374	4.1781	3.8483	4.0758
	(0.6202)[f]	(0.7545)[f]	(0.6579)[f]	(0.7334)[f]	(0.7231)[f]
	[0.7939][f]	[1.2818][f]	[0.9449][f]	[0.8548][f]	[0.8534][f]
*Abs*84	0.1318	0.1492	0.1116	0.1308	0.1249
{2.866/2.710}	(0.0040)[f]	(0.0056)[f]	(0.0036)[f]	(0.0045)[f]	(0.0042)[f]
	[0.0061][f]	[0.0203][f]	[0.0068][f]	[0.0066][f]	[0.0067][f]
*Abs*83	0.0373	0.0502	0.0206	0.0265	0.0271
{2.862/2.944}	(0.0036)[f]	(0.0051)[f]	(0.0032)[f]	(0.0041)[f]	(0.0040)[f]
	[0.0074][f]	[0.0291]	[0.0095][g]	[0.0106][g]	[0.0107][g]
Log (*Age*)	−0.7445	−0.8087	−0.5211	−0.8701	−0.8096
{3.749/0.234}	(0.0503)[f]	(0.0606)[f]	(0.0530)[f]	(0.0596)[f]	(0.588)[f]
	[0.0658][f]	[0.1214][f]	[0.0789][f]	[0.0734][f]	[0.0733][f]
Doctor	−0.3517	−0.3444	−0.3239	−0.3776	−0.3687
{0.551/0.305}	(0.0474)[f]	(0.0593)[f]	(0.0504)[f]	(0.0553)[f]	(0.0548)[f]
	[0.0589][f]	[0.1034][f]	[0.0754][f]	[0.0651][f]	[0.0655][f]
Driver	−0.0069	−0.0653	−0.0737	−0.0939	−0.0889
{0.645/0.475}	(0.0236)[f]	(0.0255)[f]	(0.0209)[f]	(0.0234)[g]	(0.0231)[f]
	[0.0262][g]	[0.0484]	[0.0322][g]	[0.0288][f]	[0.0289][f]
Log (*Employees*)	0.1094	0.0794	0.1920	0.1621	0.1572
{5.531/0.363}	(0.0253)[f]	(0.0323)[g]	(0.0264)[f]	(0.0290)[f]	(0.0288)[f]
	[0.0367][f]	[0.0999]	[0.0551][f]	[0.0455][f]	[0.0475][f]
Family	0.0749	0.0972	0.0156	0.1033	0.0736
{0.774/0.418}	(0.0247)[f]	(0.0308)[f]	(0.0254)	(0.0295)[f]	(0.0289)[g]
	[0.0312][g]	[0.0556]	[0.0371]	[0.0340][f]	[0.0327][g]
Garage	−0.4887	−0.4308	−0.6631	−0.4440	−0.4873
{4.474/0.074}	(0.1294)[f]	(0.1595)[f]	(0.1399)[f]	(0.1518)[f]	(0.1505)[f]
	[0.1638][f]	[0.2995]	[0.2074][f]	[0.1777][g]	[0.1771][f]
Home	0.0689	0.0882	0.0454	0.0772	0.0713
{0.099/0.299}	(0.0292)[g]	(0.0373)[g]	(0.0286)	(0.0333)[g]	(0.0330)[g]
	[0.0463]	[0.1331]	[0.0675]	[0.0592]	[0.619]
LongAbs	−0.0228	−0.0265	−0.0306	−0.0243	−0.0255
{0.616/0.914}	(0.0099)[g]	(0.0126)[g]	(0.0099)[f]	(0.0117)[g]	(0.0117)[g]
	[0.0148]	[0.0389]	[0.0173]	[0.0176]	[0.0179]
*Lost*84	−0.0009	−0.0012	−0.0007	−0.0009	−0.0007
{23.212/40.015}	(0.0003)[f]	(0.0003)[f]	(0.0003)[g]	(0.0003)[f]	(0.0003)[g]
	[0.0003][f]	[0.0009]	[0.0006]	[0.0006]	[0.0006]
*Lost*83	0.0015	0.0016	0.0010	0.0014	0.0012
{21.222/40.784}	(0.0002)[f]	(0.0002)[f]	(0.0002)[f]	(0.0003)[f]	(0.0003)[f]
	[0.0003][f]	[0.0004][f]	[0.0005][g]	[0.0005][f]	[0.0005][g]
Male	−0.0485	−0.0644	−0.0229	−0.0497	−0.0612
{0.930/0.255}	(0.0369)	(0.0479)	(0.0363)	(0.0409)	(0.0407)
	[0.0482]	[0.0944]	[0.0609]	[0.0546]	[0.0515]
Log (*Service*)	0.1523	0.1783	0.0553	0.2819	0.2189
{2.215/1.226}	(0.0161)[f]	(0.0235)[f]	(0.0188)[f]	(0.0175)[f]	(0.0178)[f]
	[0.0167][f]	[0.0411][f]	[0.0207][f]	[0.0167][f]	[0.0164][f]
ShortAbs	0.4021	0.4176	0.3127	0.3937	0.3758
{0.412/0.285}	(0.0524)[f]	(0.0659)[f]	(0.0583)[f]	(0.0612)[f]	(0.0609)[f]
	[0.0629][f]	[0.1156][f]	[0.0790][f]	[0.0704][f]	[0.0709][f]
ESS =	—	—	16,978.8	18,393.9	17,696.0
LL =	−9634.7	−9386.9	—	—	—
R^2 =	0.41	0.41	0.39	0.38	0.38

[a] Absences = exp ($X'\beta + \epsilon_i$). The dependent variable in all regressions is absences in 1985, which has mean 2.172 and standard deviation 2.332. Nonrobust standard errors are in parentheses ( ), robust standard errors are in square brackets [ ].

[b] Poisson pseudo maximum likelihood.

[c] Negative binomial pseudo maximum likelihood.

[d] Nonlinear least squares.

[e] Semiparametric generalized least squares using nonlinear least-squares residuals.

[f] Significanct at the 1% level.

[g] Significant at the 5% level.

changes in the regressor list. Comparing results in the last two columns of table 1 illustrates that as the number of nearest neighbors increases from $n^{1/2}$ to $n^{3/5}$, the coefficient estimates from the model with variance of unknown form move closer to the least-squares coefficient estimates, generally declining in absolute value.[12] The coefficient of (*male* * *family*) was insignificant when we added the interaction between gender and marital status to the regressor list in table 2, which suggests that the greater absenteeism among women, ceteris paribus, is not caused by child care duties. When we ignored sluggishly adjusting work attendance and estimated the regressions in table 2 without the potentially endogenous lagged absence rates *Abs*83 and *Abs*84, the partial effects of the other regressors on absences, particularly sex and age, were magnified as expected. The conclusion to be drawn from our robustness checks is that in our data the choice of the theoretical model to estimate, specifically the regressor list, is much more important to the results than whether or not to use specialized count data regression models, such as Poisson or negative binomial PMLE, or whether to permit a general form of heteroskedasticity.

E. *Goodness of Fit*

We report R^2 values based on Pearson's residuals, as suggested by Cameron and Windmeijer (1996). The Pearson residuals are the raw residuals standardized by the estimated standard deviation. Let us define $\hat{\mu}_i = \exp(X_i'\hat{\beta})$. The R^2 for the Poisson model is then

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2/\hat{\mu}_i}{\sum_{i=1}^{n} (y_i - \bar{y})^2/\bar{y}} \tag{19}$$

where $\bar{y} = (1/n)\,\Sigma_{i=1}^{n} y_i$ is the Poisson PMLE under the restriction that all the coefficients be zero except the intercept. The R^2 for the negative binomial is

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2/(\hat{\mu}_i - \hat{\alpha}\hat{\mu}_i^2)}{\sum_{i=1}^{n} (y_i - \bar{y})^2/(\bar{y} - \hat{\alpha}\bar{y}^2)}. \tag{20}$$

[12] In a linear model with only one regressor, the mean-squared error of the conditional expectation *k*nn estimate is minimized by a k that is proportional to $n^{4/5}$ (Härdle 1990). However, an optimal k is a function of the number of regressors, and $k = n^{4/5}$ is also not necessarily optimal for our count data regression models with heteroskedasticity of unknown form. It is popular to choose $k = n^{1/2}$. To the best of our knowledge there is no evidence concerning the optimal, or data-dependent k in the semiparametric models we estimated and present here.

TABLE 2.—ABSENTEEISM REGRESSIONS OMITTING LAGGED ABSENCES[a] (n = 5501)

Regressor	PMLE (Poisson)[b]	PMLE (NegBin)[c]	NLLS[d]	SEMIPAR[e] ($k = n^{1/2}$)
Intercept	7.0615 (0.6106)[f] [0.9011][f]	7.0901 (0.8944)[f] [1.9711][f]	6.776 (0.8186)[f] [0.9599][f]	6.3134 (0.9205)[f] [1.0077][f]
Log (*Age*)	−1.2638 (0.0496)[f] [0.0704][f]	−1.2943 (0.0948)[f] [0.1699][f]	−1.1637 (0.0686)[f] [0.0762][f]	−1.4997 (0.0745)[f] [0.0769][f]
Doctor	−0.4314 (0.0467)[f] [0.0654][f]	−0.4982 (0.0686)[f] [0.1452][f]	−0.3299 (0.0648)[f] [0.0697][f]	−0.4196 (0.0705)[f] [0.0723][f]
Driver	−0.0689 (0.0204)[f] [0.0301][g]	−0.0805 (0.0298) [0.0678]	−0.0506 (0.0272) [0.0328]	−0.0801 (0.0301)[f] [0.0329][g]
Log (*Employees*)	0.1803 (0.0258)[f] [0.0388][f]	0.1548 (0.0372)[f] [0.0808]	0.2159 (0.0352)[f] [0.0414][f]	0.1833 (0.0379)[f] [0.0427][f]
Family	−0.0336 (0.0684) [0.1026]	−0.0587 (0.1136) [0.3174]	0.0199 (0.0827) [0.1115]	0.0363 (0.0973) [0.1268]
Garage	−0.7362 (0.1270)[f] [0.1775][f]	−0.7325 (0.1920)[f] [0.4146]	−0.7359 (0.1720)[f] [0.1879][f]	−0.4722 (0.1868)[g] [0.1962][g]
Home	0.1525 (0.0291)[f] [0.0432][f]	0.1477 (0.0470)[f] [0.1203]	0.1649 (0.0355)[f] [0.0475][f]	0.2114 (0.0411)[f] [0.0487][f]
LongAbs	0.1161 (0.0094)[f] [0.0142][f]	0.1176 (0.0159)[f] [0.0476][f]	0.1032 (0.0114)[f] [0.0162][f]	0.1281 (0.0129)[f] [0.0161][f]
*Lost*84	0.0021 (0.0002)[f] [0.0003][f]	0.0033 (0.0003)[f] [0.0014][g]	0.0013 (0.0002)[f] [0.0004][f]	0.0021 (0.0003)[f] [0.0004][f]
*Lost*83	0.0024 (0.0002)[f] [0.0003][f]	0.0036 (0.0003)[f] [0.0018][f]	0.0018 (0.0002)[f] [0.0004][f]	0.0022 (0.0002)[f] [0.0004][f]
Male	−0.2251 (0.0488)[f] [0.0662][f]	−0.2186 (0.0714)[f] [0.1741]	−0.2230 (0.0575)[f] [0.0697][f]	−0.2693 (0.0677)[f] [0.0735][f]
Male * *Family*	0.1555 (0.0734) [0.1110]	0.1926 (0.1190) [0.3272]	0.0883 (0.0905) [0.1200]	0.1404 (0.1072) [0.1352]
Log (*Service*)	0.2724 (0.0164)[f] [0.198][f]	0.3234 (0.282)[f] [0.0573][f]	0.1856 (0.252)[f] [0.0250][f]	0.4101 (0.0228)[f] [0.0149][f]
ShortAbs	0.5504 (0.0499)[f] [0.0680][f]	0.6749 (0.0767)[f] [0.1640][f]	0.4173 (0.0694)[f] [0.0716][f]	0.5268 (0.0749)[f] [0.0750][f]
ESS =	—	—	22910.8	23891.3
LL =	−10,947	−10,140	—	—
R^2 =	0.20	0.20	0.16	0.20

[a] Absences = exp ($X'\beta + \epsilon_i$). The dependent variable in all regressions in absences in 1985, which has mean 2.172 and standard deviation 2.332. Nonrobust standard errors are in parentheses ( ), robust standard errors are in square brackets [ ].
[b] Poisson pseudo maximum likelihood.
[c] Negative binomial pseudo maximum likelihood.
[d] Nonlinear least squares.
[e] Semiparametric generalized least squares using nonlinear least-squares residuals.
[f] Significant at the 1% level.
[g] Significant at the 5% level.

The R^2 for the semiparametric GLS is

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2/\hat{\sigma}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y}^*)/\hat{\sigma}_i^2} \tag{21}$$

where $\bar{y}^*$ is the semiparametric GLS estimate of the intercept under the restriction that all the slope coefficients are zero. The R^2 for the semiparametric GLS in (21) was suggested by Buse (1973) in the context of GLS estimation with known variance. Finally, in the goodness-of-fit measure for the (nonlinear) least-squares estimates we use the usual raw residuals so that

$$R^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}.$$

As expected, all R^2 values are similar because the regression models are identical and only the weights differ across estimators. In table 1, R^2 is about 0.4 and in table 2, where the lagged dependent variable is deleted, R^2 is about 0.2—which are commonly appearing values in cross-section regression contexts.

V. Conclusion

How valuable are estimators of count data models with error variance of unknown form when applied to worker absenteeism? We examined the relative benefits of semiparametric estimation where heteroskedasticity of unknown form may be present in the context of a hedonic econometric model of employee absences incorporating sluggish adjustment to changing economic circumstances. Our empirical results support the hedonic theoretical model. Overdispersion tests rejected the Poisson specification. Other parametric estimators used, namely, negative binomial PMLE and least squares, are consistent so that coefficient point estimates are much more sensitive to the economic model estimated (regressor list, in particular) than to the estimation method applied. The semiparametric GLS estimator has the advantage of being asymptotically efficient with known asymptotic covariance matrix. Inferences based on the semiparametric procedure we present are always valid asymptotically and more efficient than the estimators that parametrize the conditional variance incorrectly.

Our application to worker absences showed how semiparametric GLS is a sensible procedure to follow in practice. Estimates are computationally easy to obtain, and the practitioner is always sure that inferences are correct and efficient asymptotically without having to pay attention to the functional form of the conditional variances or any other feature of the data generating process. Our estimated semiparametric GLS coefficients are similar in sign, magnitude, and significance to parallel regression coefficients estimated with ex ante variance specifications.

REFERENCES

Allen, Steven G., "An Empirical Model for Work Attendance," this REVIEW 63 (1981a), 77–87.

——— "Compensation, Safety, and Absenteeism: Evidence from the Paper Industry," *Industrial and Labor Relations Review* 34 (1981b), 207–218.

Avery, Robert B., and V. Joseph Hotz, "Statistical Models for Analyzing Absentee Behavior," in Paul S. Goodman, Robert S. Atkin, and Associates, *Absenteeism* (San Francisco: Jossey-Bass, 1984), 158–193.

Barmby, Timothy A., Christopher D. Orme, and J. G. Treble, "Worker Absenteeism: An Analysis Using Micro Data," *Economic Journal* 101 (1991), 214–229.

Brown, James N., and Harvey S. Rosen, "On the Estimation of Structural Hedonic Price Models," *Econometrica* 50 (1982), 765–768.

Buse, A., "Goodness of Fit in Generalized Least Squares Estimation," *American Statistician* 27 (1973), 106–108.

Cameron, A. Colin, and Pravin K. Trivedi, "Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators and Tests," *Journal of Applied Econometrics* 1 (1986), 29–53.

——— "Regression-Base Tests for Overdispersion in the Poisson model," *Journal of Econometrics* 46 (1990), 347–364.

Cameron, A. Colin, Pravin K. Trivedi, Frank Milne, and John Piggott, "A Microeconomic Model of the Demand for Health Care and Health Insurance in Australia," *Review of Economic Studies* 55 (1988), 85–106.

Cameron, A. Colin, and Frank A. G. Windmeijer, "*R*-Squared Measures for Count Data Regression Models with Applications to Health Care Utilization," *Journal of Business and Economic Statistics* 14 (1996), 209–220.

Chamberlain, Gary, "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics* 34 (1987), 305–334.

Cornwall, C. J., and P. A. Raffle, "Sickness Absence of Women Bus Conductors in London Transport 1953–57," *British Journal of Industrial Medicine* 18 (1961), 197–212.

Delgado, Miguel A., "Semiparametric Generalized Least Squares in the Multivariate Nonlinear Regression Model," *Econometric Theory* 8 (1992), 203–222.

——— "Computing Nonparametric Functional Estimates in Semiparametric Problems," *Econometric Reviews* 12 (1993), 25–128.

Delgado, Miguel A., and Peter A. Robinson, "Nonparametric and Semiparametric Methods for Econometric Research," *Journal of Economic Surveys* 6 (1992), 1–50.

Eicker, F., "Asymptotic Normality and Consistency of the Least Squares Estimators for Families for Linear Regressions," *Annals of Mathematical Statistics* 34 (1963), 447–456.

Epple, Dennis, "Hedonic Prices and Implicit Markets: Estimating Demand and Supply Functions for Differentiated Products," *Journal of Political Economy* 95 (1987), 59–80.

Gourieroux, C., A. Montfort, and A. Trognon, "Pseudo Maximum Likelihood Methods: Theory," *Econometrica* 52 (1984a), 681–700.

——— "Pseudo Maximum Likelihood Methods: Applications to Poisson Models," *Econometrica* 52 (1984b), 701–720.

Gurmu, Shifferaw, and Pravin K. Trivedi, "Recent Developments in Models of Event Counts: A Survey," Department of Economics, Indiana University, Bloomington, Working Paper (May 1993).

Härdle, Wolfgang, *Applied Nonparametric Regression.* (Cambridge, UK: Cambridge University Press, 1990).

Hausman, Jerry, Bronwyn H. Hall, and Zvi Griliches, "Econometric Models for Count Data with an Application to the Patents–R&D Relationship," *Econometrica* 52 (1984), 909–938.

Jones, R. M., "Absenteeism," Manpower Papers 4. (London: Department of Employment, 1971).

Kahn, Shulamit, and Kevin Lang, "Efficient Estimation of Structural Hedonic Systems," *International Economic Review* 29 (1988), 157–166.

Kniesner, Thomas J., and John D. Leeth, *Simulating Workplace Safety Policy* (Boston and Dordrecht: Kluwer Academic Publishers, 1995).

Lawless, J. F., "Negative Binomial and Mixed Poisson Regression," *Canadian Journal of Statistics* 15 (1987), 209–225.

McCullagh, P., and J. A. Nelder, *Generalized Linear Models,* 2nd ed. (London: Chapman and Hall, 1989).

Norman, L., and F. M. Spratling, "Health in Industry: A Contribution to the Study of Sickness Absence," in *Experience in London Transport* (London: Butterworth, 1956).

Patil, G. P., *Random Counts in Scientific Work,* vol. 1 (University Park, PA, and London: Pennsylvania State University Press, 1970).

Robinson, Peter M., "Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form," *Econometrica* 55 (1987a), 875–891.

——— "Adaptive Estimation of Heteroskedastic Regression Models," *Revista de Econometrica* 7 (1987b), 5–28.

Rosen, Sherwin, "The Theory of Equalizing Differences," in Orley Ashenfelter and Richard Layard (eds.), *The Handbook of Labor Economics* (Amsterdam: North-Holland, 1986), 641–692.

Stone, C. J. (with discussion), "Consistent Nonparametric Regression," *Annals of Statistics* 5 (1977), 595–645.

White, Halbert, "A Heteroskedasticity-Consistent Covariance Matrix Estimator, and a Direct Test for Heteroskedasticity," *Econometrica* 48 (1980), 817–838.

——— "Maximum Likelihood Estimation of Misspecified Models," *Econometrica* 50 (1982), 1–25.

APPENDIX A

k Nearest Neighbor (nn) Weights

Let X_{ir} be the rth element of X_i and define

$$s_r^2 = (n-1)^{-1} \sum_i (X_{ri} - \overline{X}_r)^2 \tag{A.1}$$

$$\overline{X}_r = n^{-1} \sum_i X_{ri}, \qquad 1 \le r \le q \tag{A.2}$$

$$\rho_{ij} = \left[\sum_r (X_{ri} - X_{rj})^2 / s_r \right]^{1/2}, \qquad i, j = 1, \ldots, n; i \ne j \tag{A.3}$$

for a sequence $k_n = k$ such that $k < n$ and $1/k + k/n \to 0$ as $n \to \infty$.

In the absence of ties among the X_i's the $k$nn weights are defined as

$$w_{ij} = \begin{cases} 1/k; & \rho_{ij} \le \rho_{ik},\ i \ne j \\ 0; & \rho_{ij} > \rho_{ik}. \end{cases} \tag{A.4}$$

The uniform $k$nn are intuitively appealing because all the nonparametric estimates can be viewed as local averages around the point at which one evaluates the regression. With the $k$nn estimates one decides how many points to use in the local averages. Note that the own observation does not enter in the local average, so the estimate is known as a "leave one out."

Our data set does not have ties because age has been scaled in days to avoid the ties problem. When ties are present in the data, a tie-breaking rule must be introduced, as suggested by Stone (1977) and amplified by Robinson (1987a).

APPENDIX B

Variable Dictionary

Absences	Number of absence spells of seven days or less in 1985
*Abs*83	Number of absence spells in 1983
*Abs*84	Number of absence spells in 1984
Age	Years of age
Doctor	Proportion of times absent during 1981–1984 that worker showed a doctor's certificate; equals zero if worker has not been absent during the four years
Driver	Dummy variable; worker is driver
Employees	Number of people working in garage
Family	Dummy variable; worker has a spouse or dependent
Garage	Distance from garage to center of London (index)
Home	Dummy variable; distance from garage to home is in 99th percentile
LongAbs	Number of absence spells greater than seven days in 1985
*Lost*83	Days absent in 1983
*Lost*84	Days absent in 1984
Male	Dummy variable; worker is a man
Service	Years of service with firm
ShortAbs	Proportion of absences that were one-day duration during 1983–1984; equals zero if worker had no absences