Working Paper 94-49
Statistics and Econometrics Series 17
December 1994

Departamento de Estadística y Econometría
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

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 examined an econometric model of counts of worker absences due to illness. The underlying theoretical model is of a sluggishly adjusting hedonic labor market. We compared results from three parametric estimators, nonlinear least squares plus Poisson and negative binomial pseudo maximum likelihood, to generalized least squares using nonparametric estimates of the conditional variance. Our data support the hedonic model of worker absenteeism. Semiparametric generalized least squares coefficients are similar in sign, magnitude, and statistical significance to their econometric analogs where the mean and variance of the errors were specified ex ante. Overdispersion test reject the Poisson specification. Robustness checks confirm that in our data parameter estimates are sensitive to regressor list but are not sensitive to econometric technique, including how we corrected for possible heteroskedasticity of unknown form.

Key Words

Count Data Models; Heteroskedasticity of Unknown Form; Semiparametric Estimation; Efficiency; Sickness Absence; Hedonic Model.

*Delgado, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid; Kneisner, Department of Economics, Indiana University. Research supported by the Spanish Dirección General de Investigación Científica y Técnica (DGICYT), reference number: PB92-0247. We thank London Regional Transport for providing the data and 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.

ABSTRACT

We examined an econometric model of counts of worker absences due to illness. The underlying theoretical model is of a sluggishly adjusting hedonic labor market. We compared results from three parametric estimators, nonlinear least squares plus Poisson and negative binomial pseudo maximum likelihood, to generalized least squares using nonparametric estimates of the conditional variance. Our data support the hedonic model of worker absenteeism. Semiparametric generalized least squares coefficients are similar in sign, magnitude, and statistical significance to their econometric analogs where the mean and variance of the errors were specified ex ante. Overdispersion tests reject the Poisson specification. Robustness checks confirm that in our data parameter estimates are sensitive to regressor list but are not sensitive to econometric technique, including how we corrected for possible heteroskedasticity of unknown form.

JEL Classification Code: C14, C25, I1, J2.

1. 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 investigated 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 (Hausman, Hall, and Griliches 1984; Cameron and Trivedi 1986; 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 estimators with the empirical performance of popular parametric estimators of count data models.

The theoretical model underlying our absence count regressions is sluggish adjustment to hedonic labor market equilibrium. We examined four models: nonlinear least squares, Poisson and negative binomial pseudo maximum likelihood, and generalized least squares with heteroskedasticity of unknown form. Regression coefficients and standard errors are generally similar across the four 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 the specification correcting for subtle heteroskedasticity.

2. A Microeconometric Model of Worker Absenteeism

The ideal microeconometric model of worker absenteeism has several distinguishing properties (Avery and Hotz 1984; Barmby, Orme, and Treble 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). Ordinary 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 parameterize 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 the desirability to allow 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, Kahn and Lang 1988). Alternatively, a researcher can numerically simulate hedonic equilibrium over a set of cost and utility function parameters (Kniesner and Leeth 1988). 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

$$W_i = f(a_i, C_i; S_i, D_i, E_i, \varepsilon_i) \qquad (1)$$

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 is positive or negative 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 equation (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

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

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

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

where $j=1,\ldots,T$ indexes 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).

3. Count Data Models — Econometric Background

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

[1]We acknowledge that lagged absences a_{-ji} may not be independent of 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.

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

where X_i' is the vector of independent 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 (Gourieroux, Montfort, and Trognon 1984a,b; Hausman, Hall, and Griliches 1984; Cameron and Trivedi 1986; 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 -- each estimated coefficient is the proportionate change in absenteeism associated with a unit change in an independent variable.

3.1 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 $E(a_i|X_i) = \exp(X_i'\beta_0) = Var(a_i|X_i)$. The choice of $g(X_i) = X_i$ in (5) produces the Poisson pseudo maximum likelihood estimator of β_0, which solves

$$\Sigma_i(a_i - \exp(X_i'\beta))X_i = 0, \quad (6)$$

the sample analog of (5).[2]

The asymptotic variance of the Poisson pseudo maximum likelihood estimator in (5) and (6) is

[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 applies. 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 pseudo maximum likelihood estimator is consistent if the mean is correctly specified (Gourieroux, Montfort, and Trognon 1984a; Cameron and Trivedi 1986; and McCullagh and Nelder 1989).

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

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

$$\Sigma_i\ (a_i - \exp(X_i'\beta))X_i\ \exp(X_i'\beta)\ W(X_i) = 0, \quad (8)$$

which has asymptotic variance

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

Among the class of GLS estimators the most efficient is that using $W(X_i) = \sigma_i^{-2}$, which is termed infeasible GLS. Notice that among the infeasible GLS the Poisson pseudo maximum likelihood estimators are equally efficient ((7) and (9) are identical) when the underlying model is Poisson.

The equality between the mean and variance under the Poisson assumption is restrictive in economic applications and is why researchers have proposed more general count data models (Cameron and Trivedi 1986, Lawless 1987, Gurmu and Trivedi 1993). A popular generalization involves assuming that absenteeism follows a compound Poisson

$$Pr(a_i=\delta_i\,|\,X_i)=\int\exp[-\exp(X_i'\beta_0+\varepsilon_i)](\exp(X_i'\beta_0+\varepsilon_i)^{\delta}/\delta_i)h(\varepsilon_i)d\varepsilon_i,\ \delta_i=0,1,2..,\ (10)$$

where $h(\varepsilon_i)$, the marginal density of the error term ε_i, is assumed to be Gamma so that although the conditional mean of absences remains $E(a_i\,|\,X_i) = \exp(X_i'\beta_0)$ the conditional variance of absences is $Var(a_i\,|\,X_i) = (\exp(X_i'\beta_0)(1+(1/t_i)\exp(X_i'\beta_0))$. One can study a range of econometric models by allowing $t_i = \exp(X_i'\beta_0)^k/\alpha$ where $\alpha>0$ is a parameter to be estimated, and k is an arbitrary constant (Cameron and Trivedi 1986). The negative binomial model with a quadratic variance function, common in empirical research, sets $k \equiv 0$ so that $Var(a_i\,|\,X_i) = \exp(X_i'\beta_0)(1 + \alpha\exp(X_i'\beta_0))$ (Hausman, Hall,

and Griliches 1984).[3] To cover a range of specifications representative of the count data literature we studied linear exponential absenteeism parameter estimates from nonlinear least squares, Poisson pseudo maximum likelihood, negative binomial pseudo maximum likelihood, and an optimal feasible GLS that we now explain.

3.2 Variance of Unknown Form

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

$$\sigma_i^2 = \mathrm{Var}(a_i | X_i), \tag{11}$$

which permits heteroskedasticity of unknown form. The conditional mean from (4) and the conditional variance in (11) form the semiparametric count data model we applied to absenteeism data for London bus drivers. We also estimated the hedonic absenteeism locus with semiparametric nonlinear generalized least squares using as estimated conditional variances

$$\hat{\sigma}_i^2 = \Sigma_j (a_j - \exp(X_j'\tilde{\beta}))^2 w_{ij}, \tag{12}$$

where $\tilde{\beta}$ is a preliminary root-n consistent parameter estimate, and w_{ij} are nonparametric k nearest neighbors (k-nn) probabilistic weights (see Appendix A). Specifically, we used so-called uniform k-nn weights to estimate the variances in (12) (Robinson 1987a).[4]

The semiparametric weighted least squares estimator we applied estimates β_0 via the solution to

---

[3]Another popular model, which we did not estimate, specifies $\sigma_i^2 = \alpha\exp(X_i'\beta)$.

[4]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_js, 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_js such that the corresponding X_j is one of the k X_js closest to x, according to the scaled Euclidean distance. As the number of observations increases the number of X_js close to x also increase, which intuitively explains why the number of terms in the weighted average (the number of nearest neighbors) must increase with the sample size. So, weights in the estimated conditional variances in (12) 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).

$$\Sigma_i \exp(X_i'\beta)X_i(a_i - \exp(X_i'\beta))\hat{\sigma}_i^{-2} = 0. \qquad (13)$$

Under regularity conditions the vector $\hat{\beta}$ that solves the first-order condition (13) has asymptotic variance

$$\text{Asy Var } \{n^{1/2}(\hat{\beta}-\beta_0)\} \equiv I_0^{-1} = E(X_iX_i'\exp(2X_i'\beta)\sigma_i^{-2})^{-1}. \qquad (14)$$

The semiparametric efficiency bound in (14) 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 generalized least squares are similar to the moment conditions needed for asymptotic normality of generalized least squares. Our nearest neighbor weights require that the smoothing parameter, k ≡ number of nearest neighbors, increases with the sample size but at a slower rate (Robinson 1987a, Delgado 1992). We examined two different nearest neighbors 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 estimated the covariance matrix implied by (14) 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

$$\tilde{I}_0^{-1} = [n^{-1}\Sigma_iX_iX_i'\hat{\sigma}_i^{-2}]^{-1}[n^{-1}\Sigma_iX_iX_i'\tilde{u}_i^2\hat{\sigma}_i^{-2}][n^{-1}\Sigma_iX_iX_i'\hat{\sigma}_i^{-2}]^{-1} \qquad (15)$$

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

Recapitulation. In examining count models of worker absences we first estimated nonlinear least squares then Poisson and negative binomial pseudo maximum likelihood. For maximum generality we then provide semiparametric GLS estimates using the initial $\hat{\beta}$s from nonlinear least squares.[5]

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

4. 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.[6] Information on absences includes the starting and returning dates, reason, and justification. Also in our data are 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.[7]

4.1 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 regressions is the number of absence spells of one week or less.[8]

We estimated 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 sluggish adjustment hedonic labor market outcomes described theoretically in equations (3) and (4). Given the regressor vector containing lagged

---

[6]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 1953-57.

[7]Regression variables are defined in Appendix B.

[8]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, Orme, and Treble (1991).

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 worker's personal and workplace characteristics regressors include age, sex, marital status, health, length of service with the bus company, length of journey to work, and plant size as metered by number of people working in the bus garage (Jones 1971).

We present results from four estimators: nonlinear least squares plus Poisson and negative binomial pseudo maximum likelihood, and a semiparametric generalized least squares estimator that was iterated until convergence from the nonlinear least squares coefficients 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 nonrobust standard errors (Eicker 1963; White 1980a, 1982).

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 examined 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 traffic accidents while commuting to and from work. Our emphasis is on whether two core results from the microeconometric literature on absenteeism appeared in our count data regressions. Specifically, did we find a substantial negative impact of age on absences coupled with statistically insignificant effects of other demographic characteristics (Allen 1981a,b)?

4.2 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.[9] The coefficients of the two lagged dependent variables, Abs84 and Abs83, are significantly positive across models, and their sum is in the range 0.13 to 0.20 so that the estimated long-run effects of regressors on absenteeism are about 15 to 25 percent 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. **[Insert Table 1 here.]**

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 percent older than average will have five to nine percent fewer short-term absence spells.[10] 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 7–10 percent 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

[9]We note that the dummy variable for 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.

[10]To facilitate estimation age, number of employees, and years of service are in logarithms so that all variables are scaled similarly. Thus, the coefficient of age is an estimated elasticity.

research, and the coefficient estimates conform to the pattern appearing in previous microeconomic research on worker absenteeism.

4.3 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 pseudo maximum likelihood model in the first column of Table 1. To elaborate, the robust standard errors of the Poisson pseudo maximum likelihood estimator in column 1 tend to be about 10 percent smaller than the robust standard errors of the coefficients of the regression models in Table 1 that permit ex ante unspecified heteroskedasticity. However, the relatively small difference between the robust and unrobust standard errors in Table 1 indicates that the information matrix equivalence holds approximately.[11] Our results are in contrast to Cameron and Trivedi (1986) who found that specialized count data models such as Poisson or negative binomial pseudo maximum likelihood dominated nonlinear least squares judged by the economic significance of the estimated parameters.[12] Where coefficients' signs, magnitudes, and statistical significance are concerned it makes little difference in our data whether we used nonlinear least squares, either Poisson or negative binomial pseudo maximum likelihood, or semiparametric generalized least squares.

A convenient 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

---

[11]As the sample size increases robust and nonrobust standard errors converge whenever k, the number of nearest neighbors, increases at the appropriate rate. In finite samples robust and nonrobust standard errors will not be the same. It also happens in the parametric case; robust and nonrobust standard errors of feasible generalized least squares coefficients are never identical in finite samples.

[12]Cameron and Trivedi (1986) generally reported nonrobust standard errors, which with overdispersion leads to underestimated standard errors for the Poisson model, suggesting why they found in favor of specialized count data models such as the Poisson.

the first column of Table 1 by testing the null hypothesis that $\hat{\alpha} = 0$ in the artificial regression with error term υ_i

$$[(a_i - \exp(X_i'\hat{\beta}))^2/\exp(X_i'\hat{\beta}) - 1] = \alpha\exp(X_i'\hat{\beta}) + \upsilon_i. \qquad (16)$$

Rejecting the null hypothesis that $\hat{\alpha} = 0$ 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 (16) $\hat{\alpha} = 0.19$ with $|t| = 15.4$ so that our data reject the Poisson specification against the more general (negative binomial) alternative where $\sigma_i^2 = \exp(X_i'\beta)[1 + \alpha\exp(X_i'\beta)]$.[13]

4.4 Robustness Checks

We also examined the robustness of our results to an increase in the number of nearest neighbors and to two 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 nonlinear least squares coefficient estimates, generally declining in absolute value.[14] The coefficient of (Male·Family) is 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, Abs83 and Abs84, the partial effects of the other regressors on absences,

---

[13]To elaborate, we rejected the null hypothesis of equidispersion by rejecting the null hypothesis $\alpha = 0$ in the ancillary regression (16). The 95 percent 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.

[14]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.

particularly sex and age, are magnified as expected. The conclusion to be drawn from our robustness checks is that in our data the choice of 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 pseudo maximum likelihood, or whether to permit a general form of heteroskedasticity. **[Insert Table 2 here.]**

4.5 Goodness of Fit

Poisson is nested inside the negative binomial specification (when $\alpha = 0$ in the ancillary regression (16) both models are the same). A likelihood ratio test rejects the null hypothesis that $\hat{\alpha} = 0$ (equidispersion) in both Tables 1 and 2. However, we have noted that $|\hat{\alpha}|$ is small, so the distinction between Poisson and negative binomial pseudo maximum likelihood should not be overemphasized.

We report the sum of squared residuals and R^2s for all weighted least squares procedures, NLLS and S(NLLS), in Tables 1 and 2. Nonlinear least squares based results can be interpreted as iterated GLS (see equation (6)). 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 0.30 and in Table 2, where the lagged dependent variables are deleted, R^2 is about 0.14 -- which are commonly appearing values in cross-section regression contexts.

5. 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, binomial pseudo maximum likelihood

and nonlinear 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 generalized least squares 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 parameterize 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 econometric 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 generalized least squares coefficients are similar in sign, magnitude, and significance to parallel regression coefficients estimated with ex ante variance specifications.

Appendix A

K Nearest Neighbor Weights

Let X_{ir} be the r^{th} element of X_i, and define

$$s_r = (n-1)^{-1} \Sigma_i (X_{ri} - \bar{X}_r)^2, \quad \textbf{(A1)}$$

$$\bar{X}_r = n^{-1} \Sigma_i X_{ri},\ 1 \le r \le q, \text{ and} \quad \textbf{(A2)}$$

$$\rho_{ij} = [\Sigma_r (X_{ri} - X_{rj})^2 / s_r]^{1/2},\ i,j = 1,\ldots,n;\ i \ne j. \quad \textbf{(A3)}$$

For a sequence $k_n = k$ such that $k < n$ and $1/k + k/n \to 0$ as $n \to \infty$, let c_{in}, $1 \le i \le n$, be constants satisfying $c_{in} \ge \ldots. \ge c_{kn} > 0$, $c_{in} = 0$, $k < i \le n$, $\Sigma_i c_{in} = 1$, and $\lim_{n\to\infty} \max_{1\le i\le k} kc_{in} < \infty$.

The k-nn weights are defined as

$$w_{jn}(X_i) = (c_{\upsilon n} + \ldots + c_{\upsilon + \lambda - 1, n})/\lambda \quad \textbf{(A4)}$$

where l is the indicator function of an event with

$$\upsilon = 1 + \#\{l\colon 1 \le l \le n,\ l \ne j \text{ and } \rho_{il} < \rho_{ij}\} \text{ and} \quad \textbf{(A5)}$$

$$\lambda = 1 + \#\{l\colon 1 \le l \le n,\ l \ne j, \text{ and } \rho_{il} = \rho_{ij}\}. \quad \textbf{(A6)}$$

The uniform weights use $c_{in} = k^{-1}$ for $1 \le i \le k$ and $c_{in} = 0$ for $i > k$. See also Stone (1977) and Delgado and Robinson (1992) for other weight functions.

Appendix B

Variable Dictionary

Variable	Description
Absences	Number of absence spells of seven days or less in 1985.
Abs83	Number of absence spells in 1983.
Abs84	Number of absence spells in 1984.
Age	Years of age.
Doctor	Proportion of times absent during 1981-84 that the worker showed a doctor's certificate; equals zero if the worker has not been absent during the four years.
Driver	Dummy variable; worker is driver.
Employees	Number of people working in the garage.
Family	Dummy variable; worker has a spouse or dependent.
Garage	Distance from the garage to the center of London (index).
Home	Dummy variable; distance from garage to home is in the 99th percentile.
LongAbs	Number of absence spells greater than seven days in 1985.
Lost83	Days absent in 1983.
Lost84	Days absent in 1984.
Male	Dummy variable; worker is a man.
Service	Years of service with the firm.
ShortAbs	Proportion of absences that were one-day duration during 1983-84; equals zero if the worker had no absences.

References

Allen, S.G. (1981a), "An empirical model for work attendance," *Review of Economics and Statistics* 63(1), 77-87.

Allen, S.G. (1981b), "Compensation, safety, and absenteeism: evidence from the paper industry," *Industrial and Labor Relations Review* 34(2), 207-18.

Avery, R.B. and V.J. Hotz (1984), "Statistical models for analyzing absentee behavior," in Paul S. Goodman, Robert S. Atkin, and Associates, *Absenteeism*. San Francisco: Jossey-Bass Inc., Publishers, 158-93.

Barmby, T.A., C.D. Orme, and J.G. Treble (1991), "Worker absenteeism: an analysis using micro data," *The Economic Journal* 101(405), 214-29.

Brown, J.N. and H.S. Rosen (1982), "On the estimation of structural hedonic price models," *Econometrica* 50(3), 765-68.

Cameron, A.C. and P.K. Trivedi (1986), "Econometric models based on count data: comparisons and applications of some estimators and tests," *Journal of Applied Econometrics* 1(1), 29-53.

Cameron, A.C. and P.K. Trivedi (1990), "Regression-based tests for overdispersion in the Poisson model," *Journal of Econometrics* 46(3), 347-64.

Cameron, A.C., P.K. Trivedi, F. Milne, and J. Piggott (1988), "A microeconomic model of the demand for health care and health insurance in Australia," *Review of Economic Studies* 55(1), 85-106.

Carroll, R.J. (1982), "Adapting for heteroskedasticity in linear models," *Annals of Statistics* 10(4), 1224-33.

Chamberlain, G. (1987), "Asymptotic efficiency in estimation with conditional moment restrictions," *Journal of Economics* 34(3), 305-34.

Cornwall, C.J. and P.A. Raffle (1961), "Sickness absence of women bus conductors in London transport 1953-57," *British Journal of Industrial Medicine* 18, 197-212.

Delgado, M.A. (1992), "Semiparametric generalized least squares in the multivariate nonlinear regression model," *Econometric Theory* 8(2), 203-22.

Delgado, M.A. (1993), "Computing nonparametric functional estimates in semiparametric problems," *Econometric Reviews* 12(1), 25-128.

Delgado, M.A. and P.A. Robinson (1992), "Nonparametric and semiparametric methods for econometric research," *Journal of Economic Surveys* 6(1), 1-50.

Eicker, F. (1963), "Asymptotic normality and consistency of the least squares estimators for families of linear regressions," *The Annals of Mathematical Statistics* 34, 447-56.

Epple, D. (1987), "Hedonic prices and implicit markets: estimating demand and supply functions for differentiated products," *Journal of Political Economy* 95(1), 59-80.

Gourieroux, C., A. Montfort, and A. Trognon (1984a), "Pseudo maximum likelihood methods: theory," *Econometrica* 52(3), 681-700.

Gourieroux, C., A. Montfort, and A. Trognon (1984b), "Pseudo maximum likelihood methods: applications to Poisson models," *Econometrica* 52(3), 701-20.

Gurmu, S. and P. K. Trivedi (1993), "Recent developments in models of event counts: a survey," Working Paper, Department of Economics, Indiana University, Bloomington, May.

Härdle, W. (1990), *Applied nonparametric regression*. Cambridge: Cambridge University Press.

Hausman, J., B.H. Hall, and Z. Griliches (1984), "Econometric models for count data with an application to the patents-R & D relationship," *Econometrica* 52(4), 909-38.

Jones, R.M. (1971), "Absenteeism," Manpower papers #4. Department of Employment, London, U.K.

Kahn, S. and K. Lang (1988), "Efficient estimation of structural hedonic systems," *International Economic Review* 29(1), 157-66.

Kniesner, T.J. and J.D. Leeth (1988), "Simulating hedonic labor market models," *International Economic Review* 29(4), 755-89.

Lawless, J.F. (1987), "Negative binomial and mixed Poisson regression," *Canadian Journal of Statistics* 15(3), 209-25.

McCullagh, P. and J. A. Nelder, (1989), *Generalized Linear Models*, Second Edition. London: Chapman and Hall.

Norman, L. and F.M. Spratling (1956), "Health in industry: a contribution to the study of sickness absence," in *Experience in London Transport*, London: Butterworth and Company.

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

Robinson, P.M. (1987a), "Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form," *Econometrica* 55(4), 875-91.

Robinson, P.M. (1987b), "Adaptive estimation of heteroskedastic regression models," *Revista de Econometria* 7, 5-28.

Rosen, S. (1986), "The theory of equalizing differences." In Orley Ashenfelter and Richard Layard, Eds., *The Handbook of Labor Economics*, North-Holland Publishing Co., Amsterdam, 641-92.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

Stone, C.J. (1977) (with discussion), "Consistent nonparametric regression," *Annals of Statistics* 5(4), 595-645.

White, H. (1980a), "A heteroskedasticity-consistent covariance matrix estimator, and a direct test for heteroskedasticity," *Econometrica* 48(4), 817-38.

White, H. (1980b), "Using least squares to approximate unknown regression functions," *International Economic Review* 21(2), 149-70.

White, H. (1982), "Maximum likelihood estimation of misspecified models," *Econometrica* 50(1), 1-25.

Table 1

Estimates of Sluggish Adjustment Models of Worker Absenteeism[a]
(n = 5501)

Regressor {Mean/SD}	PMLE (Poisson)[b]	PMLE (NegBin)[c]	NLLS[d]	SEMIPAR[e] ($k=n^{1/2}$)	SEMIPAR ($k=n^{3/5}$)
Intercept	4.1779 (.6202)** [.7939]**	4.1374 (.7545)** [1.2818]**	4.1781 (.6579)** [.9449]**	3.8483 (.7334)** [.8548]**	4.0758 (.7231)** [.8534]**
Abs84 {2.866/2.710}	.1318 (.0040)** [.0061]**	.1492 (.0056)** [.0203]**	.1116 (.0036)** [.0068]**	.1308 (.0045)** [.0066]**	.1249 (.0042)** [.0067]**
Abs83 {2.862/2.944}	.0373 (.0036)** [.0074]**	.0502 (.0051)** [.0291]	.0206 (.0032)** [.0095]*	.0265 (.0041)** [.0106]*	.0271 (.0040)** [.0107]*
log(Age) {3.749/0.234}	-.7445 (.0503)** [.0658]**	-.8087 (.0606)** [.1214]**	-.5211 (.0530)** [.0789]**	-.8701 (.0596)** [.0734]**	-.8096 (0.588)** [.0733]**
Doctor {0.551/0.305}	-.3517 (.0474)** [.0589]**	-.3444 (.0593)** [.1034]**	-.3239 (.0504)** [.0754]**	-.3776 (.0553)** [.0651]**	-.3687 (.0548)** [.0655]**
Driver {0.645/0.475}	-.0669 (.0236)** [.0262]*	-.0653 (.0255)** [.0484]	-.0737 (.0209)** [.0322]*	-.0939 (.0234)* [.0288]**	-.0889 (.0231)** [.0289]**
log(Employees) {5.531/0.363}	.1094 (.0253)** [.0367]**	.0794 (.0323)* [.0999]	.1920 (.0264)** [.0551]**	.1621 (.0290)** [.0455]**	.1572 (.0288)** [.0475]**
Family {0.774/0.418}	.0749 (.0247)** [.0312]*	.0972 (.0308)** [.0556]	.0156 (.0254) [.0371]	.1033 (.0295)** [.0340]**	.0736 (.0289)* [.0327]*
Garage {4.474/0.074}	-.4887 (.1294)** [.1638]**	-.4308 (.1595)** [.2995]	-.6631 (.1399)** [.2074]**	-.4440 (.1518)** [.1777]*	-.4873 (.1505)** [.1771]**
Home {0.099/0.299}	.0689 (.0292)* [.0463]	.0882 (.0373)* [.1331]	.0454 (.0286) [.0675]	.0772 (.0333)* [.0592]	.0713 (.0330)* [.0619]
LongAbs {0.616/0.914}	-.0228 (.0099)* [.0148]	-.0265 (.0126)* [.0389]	-.0306 (.0099)** [.0173]	-.0243 (.0117)* [.0176]	-.0255 (.0117)* [.0179]
Lost84 {23.212/40.015}	-.0009 (.0003)** [.0003]**	-.0012 (.0003)** [.0009]	-.0007 (.0003)* [.0006]	-.0009 (.0003)** [.0006]	-.0007 (.0003)* [.0006]

Table 1 (continued)

Regressor {Mean/SD}	PMLE (Poisson)[b]	PMLE (NegBin)[c]	NLLS[d]	SEMIPAR[e] ($k=n^{1/2}$)	SEMIPAR ($k=n^{3/5}$)
Lost83	.0015	.0016	.0010	.0014	.0012
{21.222/40.784}	(.0002)**	(.0002)**	(.0002)**	(.0003)**	(.0003)**
	[.0003]**	[.0004]**	[.0005]*	[.0005]**	[.0005]*
Male	-.0485	-.0644	-.0229	-.0497	-.0612
{0.930/0.255}	(.0369)	(.0479)	(.0363)	(.0409)	(.0407)
	[.0482]	[.0944]	[.0609]	[.0546]	[.0515]
log(Service)	.1523	.1783	.0553	.2819	.2189
{2.215/1.226}	(.0161)**	(.0235)**	(.0188)**	(.0175)**	(.0178)**
	[.0167]**	[.0411]**	[.0207]**	[.0167]**	[.0164]**
ShortAbs	.4021	.4176	.3127	.3937	.3758
{0.412/0.285}	(.0524)**	(.0659)**	(.0583)**	(.0612)**	(.0609)**
	[.0629]**	[.1156]**	[.0790]**	[.0704]**	[.0709]**
ESS =	--	--	16978.8	18393.9	17696.0
LL =	-9634.7	-9386.9	--	--	--
R^2 =	--	--	0.39	0.33	0.35

[a]Absences = $\exp(X'\beta + \varepsilon_i)$. 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 ( ), and robust standard errors are in square brackets [ ]. ** Indicates significance at the 0.01 level, and * indicates significance at the 0.05 level

[b]Poisson pseudo maximum likelihood

[c]Negative binomial pseudo maximum likelihood

[d]Nonlinear least squares

[e]Semiparametric generalized least squares using the nonlinear least squares residuals

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 (.6106)** [.9011]**	7.0901 (.8944)** [1.9711]**	6.776 (.8186)** [.9599]**	6.3134 (.9205)** [1.0077]**
log(Age)	-1.2638 (.0496)** [.0704]**	-1.2943 (.0948)** [.1699]**	-1.1637 (.0686)** [.0762]**	-1.4997 (.0745)** [.0769]**
Doctor	-.4314 (.0467)** [.0654]**	-.4982 (.0686)** [.1452]**	-.3299 (.0648)** [.0697]**	-.4196 (.0705)** [.0723]**
Driver	-.0689 (.0204)** [.0301]*	-.0805 (.0298) [.0678]	-.0506 (.0272) [.0328]	-.0801 (.0301)** [.0329]*
log(Employees)	.1803 (.0258)** [.0388]**	.1548 (.0372)** [.0808]	.2159 (.0352)** [.0414]**	.1833 (.0379)** [.0427]**
Family	-.0336 (.0684) [.1026]	-.0587 (.1136) [.3174]	.0199 (.0827) [.1115]	.0363 (.0973) [.1268]
Garage	-.7362 (.1270)** [.1775]**	-.7325 (.1920)** [.4146]	-.7359 (.1720)** [.1879]**	-.4722 (.1868)* [.1962]*
Home	.1525 (.0291)** [.0432]**	.1477 (.0470)** [.1203]	.1649 (.0355)** [.0475]**	.2114 (.0411)** [.0487]**
LongAbs	.1161 (.0094)** [.0142]**	.1176 (.0159)** [.0476]**	.1032 (.0114)** [.0162]**	.1281 (.0129)** [.0161]**
Lost84	.0021 (.0002)** [.0003]**	.0033 (.0003)** [.0014]*	.0013 (.0002)** [.0004]**	.0021 (.0003)** [.0004]**
Lost83	.0025 (.0002)** [.0003]**	.0036 (.0003)** [.0018]**	.0018 (.0002)** [.0004]**	.0022 (.0002)** [.0004]**
Male	-.2251 (.0488)** [.0662]**	-.2186 (.0714)** [.1741]	-.2230 (.0575)** [.0697]**	-.2693 (.0677)** [.0735]**

Table 2 (continued)

Regressor	PMLE (Poisson)[b]	PMLE (NegBin)[c]	NLLS[d]	SEMIPAR[e] ($k=n^{1/2}$)
(Male * Family)	.1555 (.0734) [.1110]	.1926 (.1190) [.3272]	.0883 (.0905) [.1200]	.1404 (.1072) [.1352]
log(Service)	.2724 (.0164)** [.0198]**	.3234 (.0282)** [.0573]**	.1856 (.0252)** [.0250]**	.4101 (.0228)** [.0149]**
ShortAbs	.5504 (.0499)** [.0680]**	.6749 (.0767)** [.1640]**	.4173 (.0694)** [.0716]**	.5268 (.0749)** [.0750]**
ESS =	--	--	22910.8	23891.3
LL =	-10,947	-10,140	--	--
R^2 =	--	--	0.16	0.13

[a]Absences = $\exp(X'\beta + \varepsilon_i)$. 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 ( ), and robust standard errors are in square brackets [ ]. ** Indicates significance at the 0.01 level, and * indicates significance at the 0.05 level

[b]Poisson pseudo maximum likelihood

[c]Negative binomial pseudo maximum likelihood

[d]Nonlinear least squares

[e]Semiparametric generalized least squares using the nonlinear least squares residuals