ADVANCES IN ECONOMETRICS VOLUME 27A

MISSING DATA METHODS: CROSS-SECTIONAL METHODS AND APPLICATIONS

EDITED BY

DAVID M. DRUKKER
Stata, College Station, Texas

United Kingdom – North America – Japan
India – Malaysia – China

CONSISTENT ESTIMATION AND ORTHOGONALITY

Tiemen Woutersen

ABSTRACT

Observations in a dataset are rarely missing at random. One can control for this non-random selection of the data by introducing fixed effects or other nuisance parameters. This chapter deals with consistent estimation the presence of many nuisance parameters. It derives a new orthogonality concept that gives sufficient conditions for consistent estimation of the parameters of interest. It also shows how this orthogonality concept can be used to derive and compare estimators. The chapter then shows how to use the orthogonality concept to derive estimators for unbalanced panels and incomplete data sets (missing data).

Keywords: Missing data; panel data; causal inference; information orthogonality

JEL classifications: C30; C33; C35

INTRODUCTION

When one has a dataset where some data are missing, then it is rarely the case that one can argue that the data are missing at random. In particular, the

Missing Data Methods: Cross-sectional Methods and Applications
Advances in Econometrics, Volume 27A, 155–178
Copyright © 2011 by Emerald Group Publishing Limited
All rights of reproduction in any form reserved
ISSN: 0731-9053/doi:10.1108/S0731-9053(2011)000027A009

individuals or firms in the population of interest may differ in unobserved ways and this unobserved heterogeneity is likely to determine which individuals or firms we observe (see Moffitt & Ridder, 2007). Therefore, controlling for unobserved heterogeneity is a powerful tool to deal with missing data. Unfortunately, controlling for unobserved heterogeneity introduces a large number of nuisance parameters in the model. Suppose one uses a likelihood model. There are several ways to eliminate nuisance parameters from the likelihood. One approach is elimination of the nuisance parameters through integration. Some recent articles, Berger, Liseo, and Wolpert (1999) and Lancaster (2000), have focused on favorable properties of this approach. Another way to eliminate the nuisance parameters is the conditional profile likelihood approach, as discussed in Cox and Reid (1987, 1993). Cox and Reid show that the estimates of the parameters of interest can depend on the parameterization of the likelihood. They argue that the best results are obtained if one chooses a parametrization that is orthogonal in the information matrix sense.

Lancaster (2002) shows that – even with such a parametrization – the conditional profile likelihood may not give consistent estimates of the parameters of interest. Lancaster (2002) gives examples of cases where the profile likelihood does not yield consistent estimates, but the mode of the integrated likelihood does. This seems to suggest that, in terms of consistency, the integrating out method does very well compared to other methods.

In this chapter, we show that orthogonality in the information matrix sense is not enough to ensure consistency of the mode of the integrated likelihood. We give an example of a likelihood that is orthogonal in the information matrix sense but in which the mode of the integrated likelihood is an inconsistent estimator for the parameters of interest. In this example, a score-based, method-of-moments estimator is consistent.

These examples and counterexamples lead naturally to the question: Which properties must the likelihood possess to make consistent estimation possible for the parameters of interest? An established result is that consistent inference is possible when we can choose a parametrization that is exactly orthogonal. It is not always possible to choose such a parametrization. Tibshirani and Wasserman (1994) give an overview of orthogonality concepts. We introduce a new orthogonality concept whose conditions are weaker than exact orthogonality. We call it *exact orthogonality in expectation* (EOE). A parameterization is exactly orthogonal in expectation if the log likelihood has the following property: The cross-derivatives of parameters of interest and the nuisance parameters are zero in expectation for all values of the parameter space. If the parametrization of the likelihood is exactly orthogonal in expectation and if a mild regularity condition is satisfied, then consistent inference is possible. The regularity condition is that the likelihood function is specified in such a way that consistent inference is possible if we would know the true values of the nuisance parameters. So the new orthogonality concept ensures the existence of a consistent estimator. In particular, it formalizes the notion that consistent estimation of the parameter of interest requires some separation between parameters of interest and the nuisance parameters.

The new orthogonality concept can be used to derive new estimators. For example, some of the estimators for duration models with fixed effects in Woutersen (2000a, 2000b) were found with the help of this orthogonality concept. Moreover, we use the new orthogonality concept to explain the consistency problem of the integrating out approach at a more general level than that of example and counterexample.

This chapter is organized as follows. The second section gives terminology. The third section gives the new orthogonality concept and relates it to two existing ones. The fourth section discusses properties of the likelihood that ensure the existence of a consistent estimator for the parameters of interest. The fifth section gives an example where the mode of the integrated likelihood is an inconsistent estimate – although the properties of the likelihood are such that a consistent estimate exists. The consistent method-of-moments estimator is provided as well. The sixth section concludes.

PRELIMINARIES AND NOTATION

Let $L(\theta)$ denote the log likelihood, which is a function of θ for given data x. Assume that only a part of the entire parameter vector θ is of interest to us and that we therefore choose a parameterization $\theta=(\beta, \lambda)$, where β is the parameter of interest, and λ is the nuisance parameter. The nuisance parameter can play a role in determining which observations we can observe. For example, it could be that observations for which $\lambda<0$ are missing. The parameter λ could also determine which observations are censored, for example, that observations for which $0<\lambda<1$ are censored with probability 1/2. This chapter shows when we can recover β despite these missing data problems that are covered by λ. In order to be completely flexible in this respect, we assume that the dimension of λ is large. In particular, suppose that the number of nuisance parameters is proportional to the number of observations, N. Neyman and Scott (1948) describe that in such cases the incidental parameter problem arises: Given that the number of nuisance parameters is proportional to N, it is impossible to estimate all nuisance parameters consistently. Therefore, the maximum likelihood estimator for β is inconsistent. The nuisance parameters cause the problems of the maximum likelihood method. Several authors have proposed to remove the nuisance parameters from the likelihood before maximizing with respect to the parameters of interest. We discuss three ways to eliminate nuisance parameters.

One approach is to eliminate the nuisance parameters by integration; this gives the *integrated likelihood*:

$$\mathscr{L}^{I}(\beta) = \int \exp\{L(\beta, \lambda)\} d\lambda$$

where $\exp\{L(\beta, \lambda)\}$ denotes the likelihood. The mode of the integrated likelihood can be used as an estimator for the parameters of interest. We refer to this estimator as the *integrated likelihood estimator*. A good reference is Berger et al. (1999).

The *conditional profile likelihood* approach eliminates the nuisance parameters by replacing them with some value $\hat{\lambda}_{\beta}$, which is sometimes referred to as the conditional MLE value of λ:

$$\mathscr{L}^{P}(\beta) = \sup_{\lambda} L(\beta, \lambda)$$

Cox and Reid (1987) develop this approach and prefer to apply it to parametrizations that are orthogonal in the information matrix sense.

A third way to eliminate the nuisance parameters is by differentiation with respect to the parameter of interest, β. This approach requires that the parametrization of the log likelihood, $L(\beta, \lambda)$, is such that the expectation of L_{β} does not depend on the nuisance parameters. In that case the "score" L_{β} can be used as a moment function. We call the resulting estimator the *orthogonal score estimator*. We discuss orthogonality concepts in the third section and use these concepts to explain this approach in the fourth section;

THE LIKELIHOOD FUNCTION: THREE ORTHOGONALITY CONCEPTS

The idea behind orthogonality concepts is to "separate" the parameters of interest from the nuisance parameters. In this chapter, we discuss three degrees of "separation." The strongest form of separation is obtained when it is possible to write the log likelihood function as the sum of two functions:

the first function depends on the parameters of interest and the second on the nuisance parameters:

$$L(\beta,\lambda) = G(\beta) + H(\lambda)$$

If it is possible to write the log likelihood in such a way, then the parameterization is called *exactly orthogonal*. The term likelihood factorization is used as well (see, Tibshirani & Wasserman, 1994). We use the following notation for the derivatives and cross-derivatives. Assume that the parameter of interest, β, is a vector with K elements and that the nuisance parameter, λ, is a vector with M elements. Then the derivative of the likelihood function, $L(\beta,\lambda)$, with respect to the parameter of interest, is a vector with K elements. With an abuse of notation we write it as $L_{\beta}(\beta, \lambda)$. The cross-derivatives can be written as the $K \times M$ matrix, $L_{\beta\lambda}$. A parameterization of the likelihood is exactly orthogonal if this matrix consists only of zeros:

$$L_{\beta\lambda}(\beta,\lambda) = 0 \quad \text{for all } \beta, \lambda \tag{1}$$

An established result is that consistent inference is possible when we choose a parametrization that is exactly orthogonal (see, e.g., Anscombe, 1964). It is not always possible to choose this parametrization. Therefore, we discuss two orthogonality concepts with weaker conditions.

In a weaker concept of orthogonality, the latter condition holds in expectation,

$$EL_{\beta\lambda}(\beta,\lambda) = 0 \quad \text{for all } \beta, \lambda \tag{2}$$

i.e.

$$\int_{t_{\min}}^{t_{\max}} L_{\beta\lambda}(\beta,\lambda)e^{L(\beta_0,\lambda_0)}dt = 0 \quad \text{for all } \beta, \lambda$$

where t denotes the dependent variable, and $t \in [t_{\min}, t_{\max}]$. We call this *exactly orthogonal in expectation* and denote it by EOE. We use this new orthogonality concept extensively in the next section. A third concept is orthogonality in the information matrix sense, also called *information orthogonality*, which requires that the information matrix be block diagonal. That is, the cross-derivatives of the parameters of interest and the nuisance parameters are zero in expectation if we evaluate these cross-derivatives at the true values of the parameter.

$$EL_{\beta\lambda}(\beta,\lambda) = 0 \text{ at } \beta_0, \lambda_0 \tag{3}$$

i.e.

$$\int_{t_{\min}}^{t_{\max}} L_{\beta\lambda}(\beta_0, \lambda_0) e^{L(\beta_0, \lambda_0)} dt = 0$$

where t denotes the dependent variable, and $t \in [t_{\min}, t_{\max}]$ and β_0, λ_0 denote the true value of the parameters. Cox and Reid (1987) use this concept and refer to it as "orthogonality." We prefer the term information orthogonality to distinguish it from the other two orthogonality concepts and to stress that it is defined in terms of the properties of the information matrix. Jeffreys (1961) provides an extensive discussion of this orthogonality concept.

The relationship between the orthogonality concepts of this section is as follows: Exact orthogonality implies EOE, and EOE implies information orthogonality. If a parametrization is not orthogonal, one can try a reparametrization. Lancaster (2000) shows how a fixed effect model of Poisson count data can be reparameterized such that it is exactly orthogonal. Jeffreys (1961) and Lancaster (1999, 2000, 2002) give examples of models where it is possible to find a parameterization that is information orthogonal. It is not always possible to find a parameterization of the likelihood that is orthogonal in the sense of any of the above orthogonality concepts. We show in Appendix A how a parameterization is found that is information orthogonal for the single index model with fixed effects. Two key assumptions for this result are that the model have a single index form and that unobserved heterogeneity be a nonstochastic function of the parameters of interest, the nuisance parameters, and the strictly exogenous covariates. Tibshirani and Wasserman (1994) give an overview of the literature that is related to orthogonality and reparametrizations.

INFERENCE BASED ON THE SCORE

In this section, we show that the existence of an EOE parametrization of the likelihood ensures that a score-based estimator is consistent. We compare this estimator with the integrated likelihood estimator and show how the EOE concept can be used to derive a new estimator for the exponential hazard model.

An established result is that the expectation of the score equals zero, that is,

$$EL_\theta(\theta_0) = EL_\theta(\beta_0, \lambda_0) = 0$$

where L_θ is a vector with $K + M$ elements. As noted before, β_0 and λ_0 have lengths K and M, respectively. In the case that the nuisance parameter λ_0

denotes a fixed effect, the dimension of λ_0 equals N. Let the first K elements of L_θ be denoted by L_β. With a slight abuse of notation, we refer to L_β as the "score." Obviously, for these first K elements, the zero expectation property holds as well:

$$EL_\beta(\beta_0, \lambda_0) = 0$$

An important property of EOE parametrizations is that the expectation of L_β does not vary with λ.

$$\frac{\partial EL_\beta(\beta, \lambda)}{\partial \lambda} = E\frac{\partial L_\beta(\beta, \lambda)}{\partial \lambda} = EL_{\beta\lambda}(\beta, \lambda) = 0 \qquad (4)$$

This implies that the "zero score" property also holds at values other than the true λ. That is

$$EL_\beta(\beta_0, \lambda) = 0 \quad \forall \lambda$$

That the expectation of $L_\beta(\beta_0, \lambda)$ does not depend on the value of λ suggests that we could plug any value for λ into the moment function and then equate the moment function to zero. For example, elements of can be replaced with guesses of the value of λ. That these guesses are not consistent estimates is not a problem because the expected value of $L_\beta(\beta_0, \lambda)$ is zero for values of λ. Intuitively, better guesses for should yield a more efficient estimator for β. But lacking good information on λ, we could give equal weight to all possible values of λ, giving rise to the weighted score procedure discussed in subsection "Weighted Score." A method that avoids integration is discussed in subsection "Consistency with Exact Orthogonality in Expectation."

Weighted Score

The *weighted score* is the derivative of the likelihood with respect to the parameter of interest and gives equal weight to all possible values of the unknown, but bounded, λ. We denote the weighted score by $S^W(\beta)$ and calculate it by an M-dimensional integration:

$$S^W(\beta) = \int L_\beta(\beta, \lambda)\omega d\lambda \qquad (5)$$

In this subsection, we assume that either the bounds for λ are finite and known or that the integral in 5 has an analytical form.[1] The value of the constant ω is such that $\int \omega d\lambda = 1$. The expectation of the weighted score, $ES^W(\beta)$, is zero at the true parameter value β_0.

$$\begin{aligned}ES^W(\beta_0) &= E\int L_\beta(\beta_0,\lambda)\omega d\lambda \\ &= \int\left\{\int L_\beta(\beta_0,\lambda)\omega d\lambda\right\}e^{L(\beta_0,\lambda_0)}dt\end{aligned}$$

Note that $e^{L(\beta_0,\lambda_0)}$ does not depend on λ. Therefore, we can change the order of integration:

$$\begin{aligned}ES^W(\beta_0) &= \int\left\{\int L_\beta(\beta_0,\lambda)e^{L(\beta_0,\lambda_0)}dt\right\}\omega d\lambda \\ &= \int\{EL_\beta(\beta_0,\lambda)\}\omega d\lambda \\ &= 0 \text{ since } EL_\beta(\beta_0,\lambda) = 0\ \forall\lambda \qquad (6)\end{aligned}$$

Using the weighted score as a moment function gives the following estimating equation:

$$\int L_\beta(\beta,\lambda)\omega d\lambda = 0 \qquad (7)$$

In the next subsection, we show that the weighted score gives consistent estimate for β. Since the weighted score resembles the integrated likelihood, some remarks about their differences seems justified. The integrated likelihood uses the mode of the integrated likelihood as an estimator:

$$\tilde{\beta}^I = \arg\max_\beta L^I(\beta) = \arg\max_\beta\left\{\ln\int e^{L(\beta,\lambda)}d\lambda\right\}$$

This gives the following first-order condition (FOC):

$$\frac{\partial L^I(\beta)}{\partial\beta} = 0$$

That is

$$\frac{\partial\ln\int e^{L(\beta,\lambda)}d\lambda}{\partial\beta} = \int\frac{L_\beta(\beta,\lambda)e^{L(\beta,\lambda)}}{\int e^{L(\beta,\lambda)}d\lambda}d\lambda = 0$$

or equivalently

$$\int L_\beta(\beta,\lambda)\frac{e^{L(\beta,\lambda)}}{\int e^{L(\beta,\lambda)}d\lambda}d\lambda = 0 \qquad (8)$$

Comparing Eqs. (7) and (8) gives some insight into the difference between the weighted score and the integrated likelihood approach. The estimating equation of the weighted score, Eq. (7), gives equal weight to all values of λ; that is, the "weighting function" ω is flat and the weighted score is obtained by uniform integration over λ. The estimating function of the integrated likelihood, however, does use a nontrivial "weighting function" that varies with λ. From Eq. (8), it follows that the weighting function equals $(e^{l(\beta,\lambda)})/(\int e^{l(\beta,\lambda)}d\lambda)$. This weighting function is always positive and integrates to one. It can be interpreted as the posterior for λ for a given value of β (using flat priors for all elements of λ and β). In this interpretation, we condition on β without knowing its true value. In other words, we use the same data to derive both the score of beta, $L_\beta(\beta, \lambda)$, and the weighting function for the unknown λ. In the next section, we show algebraically that the interaction between the score of beta and the weighting function causes the inconsistency of the integrated likelihood estimator. There is no such interaction for the weighted score and the next subsection shows that consistent inference is a general property of that method.

Consistency with Exact Orthogonality in Expectation

As noted earlier, it is not always possible to find a parametrization that is EOE. However, if we have such an EOE parametrization, we only need a mild regularity condition to make consistent estimation possible. The regularity condition is that the likelihood function is specified in such a way that consistent inference is possible if we would know the true values of the nuisance parameters. The only function of this condition is to preclude some of the likelihood functions that are formulated in terms of unknown expectation of random variables. The regularity condition ensures that inference based on $L_\beta(\beta, \lambda_0)$ gives a consistent estimate for β. We do not know λ_0, but fortunately there are other functions that have the same expectation as $L_\beta(\beta, \lambda_0)$. The parametrization is EOE. Therefore,

$$EL_\beta(\beta, \lambda) = EL_\beta(\beta, \lambda_0) \quad \forall\lambda \tag{9}$$

and for bounded λ,

$$ES^W(\beta) = EL_\beta(\beta, \lambda_0) \tag{10}$$

See Appendix B for details. So the expectations of $L_\beta(\beta, \lambda)$ and $S^W(\beta)$ do not depend on knowledge of λ_0. The incidental parameter problem inhibits consistent estimation of λ_0. Therefore, it is attractive to base inference on a moment function whose expectation does not depend on λ_0. To understand these moment functions, an analogue with the theory of adaptive estimation can be helpful. Adaptive estimation is concerned with the efficient estimation of the parameter of interest, β, in the presence of the nuisance parameter, λ. The dimension of λ is either fixed or increases at a slower rate than the number of observations so that β and λ can be consistently estimated. Adaptive estimation is possible if the Cramer–Rao bound for β does not change if the true values of the nuisance parameters, λ_0, are replaced by consistent estimates, $\hat{\lambda}$. The condition for the Cramer–Rao bound for parameters of interest to be the same when nuisance parameters are known as when they are estimated is information orthogonality. Indeed, information orthogonality is a necessary condition for adaptive estimation (see, e.g., Newey, 1990, Theorem 3.3 and Stein, 1956). This chapter does not want to restrict itself to the cases, where λ can be consistently estimated. Therefore, we need a stronger orthogonality concept than the one that is used for adaptive estimation. Indeed, Eqs. (9) and (10) do not depend on estimates of the nuisance parameters being close to the truth, and therefore a stronger orthogonality concept is needed. For this stronger orthogonality concept, EOE, we can derive the following theorem. Theorem 1 proves that inference based on the sample analogues of Eqs. (9) and (10) gives consistent estimates for the parameters of interest. Let $L_\beta(\beta,\lambda)$ be a score of a likelihood that is EOE. Suppose that we could evaluate this score at the true value of the nuisance parameter λ_0 and could use $L_\beta(\beta,\lambda_0)$ as a moment function. In particular, suppose $\hat{\beta} = \arg\max_\beta L_\beta(\beta,\lambda_0)' L_\beta(\beta,\lambda_0)$ is a consistent estimator for β_0. In that case, we could use the fact that $L_\beta(\beta,\lambda)$ and $S^W(\beta)$ have the same expectation as $L_\beta(\beta,\lambda_0)$ and use them as moment functions. Theorem 1 proves that the resulting moment estimator is consistent. It thereby proves that EOE is a sufficient condition for the existence of a consistent estimator. We assume the following.

Assumption 1. Let (i) $w_i (i = 1, 2, \ldots)$ be i.i.d. (ii) $\beta_0 \in \Theta$, which is compact and, for all i, $\lambda_{0,i} \in \Theta_\lambda$, which is compact; (iii) $E\{L_\beta(\beta,\lambda_{0,i},w_i))\} = 0$ only if $\beta = \beta_0$; (iv) $L_\beta(\beta,\lambda_i,w_i)$ be continuous with probability one for all $\beta \in \Theta$ and $\lambda_i \in \Theta_\lambda$; (iv) $\mathrm{E}[\sup_{\beta\in\Theta,\lambda_i\in\Theta_\lambda} |L_\beta(\beta,\lambda_i,w_i)| < \infty]$ for all i.

Note that Assumption 1 allows for fixed or random effects and allows that the data on an individual is missing as a function of the random or fixed effect λ_i. Note that the identification condition, $E\{L_\beta(\beta,\lambda_{0,i},w_i))\} = 0$ only if $\beta = \beta_0$, only needs to hold for a known realization of the fixed or random effect λ_i. In particular, it is instructive to compare this condition to the condition for identification of a model with no heterogeneity, $\lambda_1 = \lambda_2 = \ldots = \bar{\lambda}$: In that case, the condition for identification is $E\{L_\beta(\beta, \bar{\lambda}, w_i)\} = 0$ only if $\{\beta = \beta_0$ and $\bar{\lambda} = \bar{\lambda}_0\}$ while the other conditions are the same. Thus, Assumption 1 is strictly weaker than identifying $\{\beta, \bar{\lambda}\}$ through their score equations. This is possible through the parameter separation of EOE. Moreover, the estimator estimates minimizes over β, γ, while the conditions imply that this yields a unique β_0 but not necessarily a unique γ_0. In other words, β_0 is point identified while γ_0 is not required to be point identified. The score function $L_\beta(\beta, \gamma)$ uses $\lambda_1 = \lambda_2 = \ldots = \gamma$ and $L_\gamma(\beta, \gamma) = \sum_i L_{\lambda_i}(\beta, \lambda_i)|_{\lambda_i = \gamma}$. This gives,

Theorem 1. Let Assumption 1 hold and let $L(\beta, \lambda)$ be a log likelihood whose parametrization is EOE. Let

$$\{\hat{\beta}_{\mathrm{EOE}}, \hat{\gamma}_{\mathrm{EOE}}\} = \underset{\beta\in\Theta, \gamma\in\Theta_\gamma}{\arg\min}\{L_\beta(\beta, \gamma)' L_\beta(\beta, \gamma) + L_\gamma(\beta, \gamma)' L_\gamma(\beta, \gamma) \quad \text{and}$$

$$\hat{\beta}_{\mathrm{IntScore}} = \underset{\beta\in\Theta}{\arg\min}\, S^W(\beta)' S^W(\beta)$$

Then

$$\hat{\beta}_{\mathrm{EOE}} = \beta_0 + o_p(1)$$

Moreover, if $\sup_{\beta\in\Theta}|S^W(\beta)| < \infty$, then

$$\hat{\beta}_{\mathrm{IntScore}} = \beta_0 + o_p(1)$$

Proof: See Appendix C

Let $L_\beta(\beta, \lambda_0)$ be a moment function where the true value of the nuisance parameter, λ_0, is known. If inference based on the moment function $L_\beta(\beta, \lambda_0)$ gives a consistent estimate for β, then inference based on $L_\beta(\beta, \lambda)$ gives a consistent estimate for β under regularity conditions. Moreover, under the additional condition that the integral of the score is well behaved, the moment estimator based on $S^W(\beta)$ is also consistent.

We illustrate the weighted score method with an example.

Example

Consider the exponential hazard model with fixed effects and exogenous regressors. Suppose we observe T, possibly censored, spells for N individuals and that the hazard has the following form:

$$\theta_{is}(t) = e^{x_{is}\beta + v_i}, \quad i = 1, \ldots, N, \quad s = 1, \ldots, T$$

where x_{is} denotes the vector of regressors and v_i the fixed effect. We suppress the subscript i of the fixed effect. Suppose the observed duration y_{is} is the minimum of the duration t_{is} and the censoring time c_{is}. That is,

$y_{is} = \min(t_{is}, c_{is})$. Let the indicator d_{is} be zero if censoring took place and otherwise be equal to 1. The log likelihood of this model is

$$L(\beta, \nu) = \frac{1}{N} \sum_i L_i(\beta, \nu)$$

where

$$L_i(\beta, \nu) = \sum_s d_{is}(\nu + x_{is}\beta) - \sum_s e^{\nu + x_{is}\beta} y_{is}$$

This parametrization of the likelihood does not satisfy any of the orthogonality conditions of the third section. However, we can reparameterize the likelihood. In particular, we need a parametrization $\{\beta, \lambda\}$ for which the condition of EOE holds:

$$E\frac{\partial^2 L}{\partial \lambda_i \partial \beta} = 0 \quad \text{for all values of } \beta \text{ and } \lambda$$

The easiest way to solve this equation is to write the old fixed effect as a function of the new fixed effect, beta and the regressors; that is, $\nu = \nu(\beta, \lambda, x_i)$. So $L(\beta, \nu)$ is a function of ν, β, and the omitted x_i and ν is a function of β, λ, and x_i, that is, $L(\beta, \nu(\beta, \lambda, x_i))$. Similarly, we can think of λ as a function of β, ν, x_i, that is, $\lambda(\beta, \nu, x_i)$. For the general linear model, we can solve the last differential equation. In Appendix A we show that an implicit solution for $\lambda(\beta, \nu, x_i)$ is:

$$\lambda = \sum_s \int_{-\infty}^{\nu + x_{is}\beta} EL_{\mu\mu} d\mu$$

where $\mu = \nu + x\beta$, and $L_{\mu\mu} = \partial^2 L/(\partial\mu)^2$. Differentiating gives the following two expressions:

$$\frac{\partial \lambda}{\partial \nu} = \sum_s EL_{\mu\mu}$$

and

$$\frac{\partial \nu}{\partial \beta} = -\frac{\sum_s x_{is} EL_{\mu\mu}}{\sum_s EL_{\mu\mu}}$$

In our model, $EL_{\mu\mu} = e^{\nu + x_{is}} Ey_{is}$. After the differential equation is solved, we can base our inference on either the score, L_β, whose expected value does not vary with λ, or on the weighted score, $S_\beta = \sum_i \int L_\beta d\lambda / N$.

Differentiating $L_i(\beta, v(\beta, \lambda, x_i))$ with respect to β yields the following:

$$\begin{aligned} L_{i\beta} &= \sum_s d_{is}\left(\frac{\partial v}{\partial \beta} + x_{is}\right) - \sum_s \left(\frac{\partial v}{\partial \beta} + x_{is}\right) e^{v + x_{is}\beta} y_{is} \\ &= \frac{\partial v}{\partial \beta}\left\{\sum_s d_{is} - \sum_s e^{v + x_{it}\beta} y_{is}\right\} + \sum_s d_{is} x_{is} - \sum_s x_{is} e^{v + x_{is}\beta} y_{is} \end{aligned} \quad (11)$$

Appendix D shows that, for the exponential hazard model,

$$S^W = \frac{1}{N}\sum_i S_i^W$$

where

$$S_i^W = \int L_{i\beta} d\lambda$$

Using Eq. (11) gives the following expression for S_i^W:

$$\begin{aligned} S_i^W &= \int \left[\left\{\sum_s d_{is}\right\} \sum_s x_{is} e^{v + x_{it}\beta} E y_{is} + \left\{\sum_s d_{is} x_{is}\right\} \sum_s e^{v + x_{it}\beta} E y_{is}\right] \frac{d\lambda}{dv} dv \\ &= \left\{\sum_s d_{is}\right\} \sum_s x_{is} e^{v + x_{it}\beta} E y_{is} + \left\{\sum_s d_{is} x_{is}\right\} \sum_s e^{v + x_{it}\beta} E y_{is} \end{aligned} \quad (12)$$

The last expression involves Ey_{is}. Since we usually do not know Ey_{is} we replace them with unbiased estimators, for example, the realization of y_{is}. Chamberlain (1985) argues that the relevant limiting distribution has the number of individuals increasing but not the time dimension. The resulting estimating function gives a consistent estimate for β.

INCONSISTENCY OF THE INTEGRATED LIKELIHOOD ESTIMATOR

Berger et al. (1999) review the profile likelihood methods and conclude that the "use of these methods tends to be restricted to rather special frameworks." Like Lancaster (1999 and 2000), Berger et al. argue in favor of the integrated likelihood approach. However, consistency is *not* a general property of this method. Consider the same problem as in the last subsection, estimating the fixed effect exponential hazard model with censored observations. The log likelihood of this model is

$$L(\beta, \nu) = \frac{1}{N}\sum_i L_i(\beta, \nu) \tag{13}$$

where

$$L_i(\beta, \nu) = \sum_s d_{is}(\nu + x_{is}\beta) - \sum_s e^{\nu + x_{is}\beta} y_{is}$$

We can integrate the likelihood with respect to the EOE fixed effect, λ_i. Appendix E shows that the integrated likelihood has the following form:

$$L^I = \frac{1}{N}\sum_i \left\{ \sum_s (x_{is}\beta) d_{is} + \ln\left(\sum_s e^{x_{is}\beta} E y_{is}\right) - \left(\sum_s d_{is} + 1\right) \ln\left(\sum_s e^{x_{is}\beta} y_{is}\right)\right\} \tag{14}$$

Differentiating the integrated likelihood with respect to β gives the following FOC:

$$\frac{\partial L^I(\beta)}{\partial \beta} = \frac{1}{N}\sum_i \left\{ \sum_s x_{is} d_{is} + \frac{\sum_s x_{is} e^{x_{is}\beta} E y_{is}}{\sum_s e^{x_{is}\beta} E y_{is}} - \left(\sum_s d_{is} + 1\right) \frac{\sum_s x_{is} e^{x_{is}\beta} y_{is}}{\sum_s e^{x_{is}\beta} y_{is}} \right\} \tag{15}$$

This FOC resembles to some extent the FOC of the weighted score. However, in this case, Eq. (15), the stochastic y_{is}, $s = 1, \ldots, T$ appears in the denominator. This causes the expectation of $(\partial L^I(\beta))/(\partial \beta)$ to be nonzero at the truth (see Appendix F for details). This nonzero expectation at the truth causes the inconsistency of the integrated likelihood. We summarize these findings in a theorem.

Theorem 2. EOE is *not* a sufficient condition for consistency of the integrated likelihood estimator.

Corollary. Information orthogonality is *not* a sufficient condition for consistency of the integrated likelihood estimator.

Proof Corollary: EOE implies information orthogonality.

The example of the exponential hazard model is analytically tractable. However, simulations with the panel probit model show that the inconsistency problem is not confined to the exponential hazard model.[3] Note that, by Theorem 1, EOE ensures the existence of a consistent estimator.

Indeed, the orthogonal score estimator of the last section is consistent for the exponential hazard model of Eq. (13). So consistency is not a general property of the integrated likelihood, but EOE ensures that the score estimator is consistent. Since the expectation of the orthogonal score estimator, EL_{β}, is not a function of the nuisance parameter, we can say, intuitively, that the orthogonal score estimator "ignores" the nuisance parameters. However, "ignoring" the nuisance parameters is only possible if we can find a parametrization that is EOE. Suppose, however, that we can only find an information orthogonal parametrization.[3] In that case "ignoring" the nuisance parameters is not possible and we have to rely on either the integrating out method or the profile likelihood approach. The consistency problems of the profile likelihood approach are well documented, and this chapter shows that consistency is not a general property of the integrating out method either. Characterizing the inconsistencies of the latter while using an information orthogonal parametrization seems an interesting research area.

CONCLUSION

The integrated likelihood technique is an elegant tool to eliminate nuisance parameters. This chapter, however, shows that the mode of the integrated likelihood need not be a consistent estimator for the parameter of interest. We introduced a new orthogonality concept that ensures consistency and also introduce the concept of the "weighted score." In our view, both are useful tools to think about consistency and the integrating out approach. Moreover, the new orthogonality concept can suggest consistent score estimators and formalizes the notion that we need some "degree of separation" to ensure consistent estimation of the parameters of interest in the presence of nuisance parameters.

NOTES

[1] These assumptions are dropped in the next subsection.

[2] An information orthogonal parametrization was used; depending on the values of the fixed effects and the regressors, the ratio $\hat{\beta}/\beta_0$ varied between 0 and 3. The magnitude of the inconsistency of this and other models will be explored in a separate paper.

[3] For scalar dependent variables with fixed effects, it is always possible to find an information orthogonal parametrization, see Jeffreys (1961).

ACKNOWLEDGMENT

I would like to thank Moshe Buchinsky, Guido Imbens, and an anonymous referee for helpful comments and discussions. All errors are mine. This chapter is based on the second chapter of my Brown University dissertation.

REFERENCES

Anscombe, F. J. (1964). Normal likelihood functions. *Annals of the Institute of Statistical Mathematics*, *16*(1), 1–19.

Berger, J. O., Liseo, B., & Wolpert, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science*, *14*, 1–28.

Chamberlain, G. (1985). Heterogeneity, omitted variable bias, and duration dependence. In: J. J. Heckman & B. Singer (Eds.), *Longitudinal analysis of labor market data*. Cambridge: Cambridge University Press.

Cox, D. R., & Reid, N. (1987). Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society, Series B*, *49*, 1–39.

Jeffreys, H. (1961). *Theory of probability* (3rd ed). Oxford: Clarendon Press.

Lancaster, T. (1999). *Some econometrics of scarring*. Brown manuscript. Brown University, Providence, RI, USA.

Lancaster, T. (2000). The incidental parameters since 1948. *Journal of Econometrics*, *95*, 391–413.

Lancaster, T. (2002). Orthogonal parameters and panel data. *Review of Economic Studies*, *69*(3), 647–666.

Moffitt, R. A., & Ridder, G. (2007). Econometric methods for data combination. In: J. Heckman & E. Leamer (Eds.), *Handbook of econometrics* (Vol. 6B). North-Holland, Amsterdam, The Netherlands: Elsevier

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, *5*(2), 99–135.

Newey, W. K., & McFadden, D. (1994). Large sample estimation and hypothesis testing. In: R. F. Engle & D. MacFadden (Eds.), *Handbook of econometrics* (Vol. 4). Amsterdam: North-Holland.

Neyman, J., & Scott, E. L. (1948). Consistent estimation from partially consistent observations. *Econometrica*, *16*, 1–32.

Stein, C. (1956). Efficient nonparametric testing and estimation. In: *Proceedings of the third Berkeley Symposium on mathematical statistics and probability* (Vol. 1). Berkeley, CA: University of California Press.

Tibshirani, R., & Wasserman, L. (1994). Some aspects of reparametrization of statistical models. *The Canadian Journal of Statistics*, *22*, 163–173.

Woutersen, T. (2000a). *Estimating the hand of the past: New estimators for duration models with endogenous regressors and endogenous censoring*. UWO working paper. Brown University, Providence, RI, USA.

Woutersen, T. (2000b). *Essays on the integrated hazard and orthogonality concepts*. Unpublished dissertation, University of Western Ontario, London, ON, Canada.

APPENDIX A: ORTHOGONALITY IN THE SINGLE INDEX MODEL

Consider the following classes of panel data models:

1. Let the observations for a single agent be stochastically independent and depend on unknowns only through $\mu_{it} = v_i + x_{it}\beta$, where v_i denotes an individual specific fixed effect and x_{it} a vector of strictly exogenous regressors. The panel Poisson, logit and probit models are of this single index form.
2. Let $y_{it} = G(x_{it}\beta + v_i) + \varepsilon_{it}$, where $\varepsilon_{it} \sim N(0, \sigma^2)$; $G(\ )$ is twice continuously differentiable and v_i and x_{it} are defined as above.

We show how information orthogonality can be obtained for these classes of panel data models; for some models, the information orthogonal parametrization suggests (or is) an EOE parametrization. The parameters β and λ_i are information orthogonal if the following condition is satisfied:

$$E\frac{\partial^2 L}{\partial\lambda_i\partial\beta}|_{\beta=\beta_0,\lambda=\lambda_0} = 0$$

where L denotes the conditional log likelihood function (conditional on x) and λ_i the individual parameter in information orthogonal reparametrization. The information matrix is evaluated at the true value; therefore, $E(\partial^2 L)/(\partial\lambda_i\partial\beta)$ has to hold at $\{\beta_0, \lambda_0\}$. We can rewrite $L(\beta, v_i)$ as $L(\mu_{i1}, \ldots, \mu_{iT})$ where $\mu_{it} = x_{it}\beta + v_i$; then (we omit the subscript i):

$$\frac{\partial L}{\partial\lambda} = \frac{\partial v}{\partial\lambda}\frac{\partial L}{\partial v} = \frac{\partial v}{\partial\lambda}\sum_s \frac{\partial L}{\partial\mu_s}\frac{\partial\mu_s}{\partial v} = \frac{\partial v}{\partial\lambda}\sum_s \frac{\partial L}{\partial\mu_s}$$

We can rephrase the information orthogonality condition as:

$$\begin{aligned} E\frac{\partial^2 L}{\partial\beta\partial\lambda} &= \frac{\partial v}{\partial\lambda}E\frac{\partial^2 L}{\partial\beta\partial v} = \frac{\partial v}{\partial\lambda}E\frac{\partial\left\{\sum_s \frac{\partial L}{\partial\mu_s}\right\}}{\partial\beta} \\ &= \frac{\partial v}{\partial\lambda}E\sum_s \frac{\partial\{\partial L/\partial\mu_s\}}{\partial\mu_s}\frac{\partial\mu_s}{\partial\beta} = \frac{\partial v}{\partial\lambda}E\left\{\sum_s L_{\mu_s\mu_s}\frac{\partial\mu_s}{\partial\beta}\right\} \\ &= 0 \end{aligned}$$

This gives

$$E\left\{\sum_s L_{\mu_s\mu_s}\frac{\partial\mu_s}{\partial\beta}\right\} = E\left\{\sum_s L_{\mu_s\mu_s}\left(x_s + \frac{\partial v}{\partial\beta}\right)\right\} = 0$$

which gives, omitting subscripts for μ:

$$\frac{\partial v}{\partial\beta} = -\frac{E\sum_s L_{\mu\mu}x_{is}}{E\sum_s L_{\mu\mu}} = -\frac{\sum_s x_{is}EL_{\mu\mu}}{\sum_s EL_{\mu\mu}}$$

The solution for this differential equation is the following:

$$\lambda = \sum_s \int_{-\infty}^{v+x\beta} EL_{\mu\mu}d\mu$$

This solution can be easily checked by total differentiation. We calculate (numerically) the integrated log likelihood:

$$L^I = \sum_i L_i^I = \sum_i \ln\int_0^{\bar{\lambda}} e^L d\lambda = \sum_i \ln\int_{-\infty}^{\infty} e^L \frac{\partial\lambda}{\partial v}dv$$

where $\bar{\lambda} = \sum_s \int_{-\infty}^{\infty} EL_{\mu\mu}d\mu$.

If the regressor x_{is} is a $(K \times 1)$ vector, then we want to (information) orthogonalize the fixed effects (incidental parameters) to all common parameters. This gives us the following differential equations:

$$E\frac{\partial^2 L}{\partial\lambda_i\partial\beta_j} = 0 \quad \text{for } j = 1, \ldots K$$

And the solution is similar to above (but now x_{is} is a vector):

$$\lambda = \sum_s \int_{-\infty}^{v_i + x_{is}\beta} EL_{\mu\mu}d\mu$$

and

$$\frac{\partial\lambda}{\partial v_i} = \sum_s EL_{\mu\mu}(v_i + x_{is}\beta)$$

For some models, this procedure yields a parametrization that is EOE. We illustrate this with the example in the text, the exponential hazard model with fixed effects (as discussed in section "Consistency with Exact Orthogonality in Expectation"). Using $\lambda = \sum_s \int_{-\infty}^{v_i + x_{is}\beta} EL_{\mu\mu}d\mu$ gives a parametrization that is EOE; in particular,

$$\lambda = \sum_s \int_{-\infty}^{v+x_{is}\beta} EL_{\mu\mu} d\mu = \sum_s \int_{-\infty}^{v+x_{is}\beta} e^{v+x_{is}\beta} Ey_{is} d\mu$$

where we omitted the subscripts for v_i. Note that

$$\frac{\partial v}{\partial \beta} = -\frac{E\sum_s L_{\mu\mu} x_{is}}{E\sum_s L_{\mu\mu}} = -\frac{\sum_s x_{is} e^{x_{is}\beta} Ey_{is}}{\sum_s e^{x_{is}\beta} Ey_{is}}$$

Therefore, $\partial^2 v/(\partial\beta\partial\lambda) = 0$.

We now check whether the parametrization is indeed EOE. As was shown in the text, the log likelihood contribution of individual i has the following expression:

$$L_i(\beta, v) = \sum_s d_{is}(v + x_{is}\beta) - \sum_s e^{x_{is}\beta + v} y_{is}$$

Differentiating with respect to β gives

$$\begin{aligned} L_{i\beta} &= \sum_s d_{is}\left(\frac{\partial v}{\partial \beta} + x_{is}\right) - \sum_s \left(\frac{\partial v}{\partial \beta} + x_{is}\right) e^{v+x_{is}\beta} y_{is} \\ &= \frac{\partial v}{\partial \beta}\sum_s d_{is} + \sum_s d_{is}x_{is} - \frac{\partial v}{\partial \beta}\sum_s e^{v+x_{is}\beta} y_{is} - \sum_s x_{is} e^{v+x_{is}\beta} y_{is} \end{aligned}$$

Therefore,

$$\begin{aligned} L_{i\beta} = &-\frac{\sum_s x_{is} e^{x_{is}\beta} Ey_{is}}{\sum_s e^{x_{is}\beta} Ey_{is}} \sum_s d_{is} + \sum_s d_{is} x_{is} \\ &+ \frac{\sum_s x_{is} e^{x_{is}\beta} Ey_{is}}{\sum_s e^{x_{is}\beta} Ey_{is}} \sum_s e^{v+x_{is}\beta} y_{is} - \sum_s x_{is} e^{v+x_{is}\beta} y_{is} \end{aligned}$$

$$L_{i\beta\lambda} = \frac{\partial v}{\partial \lambda}\left\{\frac{\sum_s x_{is} e^{x_{is}\beta} Ey_{is}}{\sum_s e^{x_{is}\beta} Ey_{is}} \sum_s e^{v+x_{is}\beta} y_{is} - \sum_s x_{is} e^{v+x_{is}\beta} y_{is}\right\}$$

where $\partial v/\partial\lambda = \sum_s \int_{-\infty}^{v+x_{is}\beta} EL_{\mu\mu} d\mu = \sum_s e^{v+x_{is}\beta} Ey_{is}$, which is not stochastic. Therefore,

$$EL_{i\beta\lambda} = \frac{\partial v}{\partial \lambda}\left\{\frac{\sum_s x_{is}e^{x_{is}\beta}Ey_{is}}{\sum_s e^{x_{is}\beta}Ey_{is}}\sum_s e^{v+x_{is}\beta}Ey_{is} - \sum_s x_{is}e^{v+x_{is}\beta}Ey_{is}\right\}$$

$$= \frac{\partial v}{\partial \lambda}\left\{\sum_s x_{is}e^{v+x_{is}\beta}Ey_{is} - \sum_s x_{is}e^{v+x_{is}\beta}Ey_{is}\right\} = 0$$

This yields

$$EL_{\beta\lambda} = \frac{1}{N}\sum_i EL_{i\beta\lambda} = 0 \quad \text{for all } \beta, \lambda$$

Therefore, the parametrization is EOE.

APPENDIX B

First, note that it follows from EOE that

$$EL_\beta(\beta, \lambda) = EL_\beta(\beta, \lambda_0) \quad \forall\lambda$$

Next, also note that

$$ES^W(\beta) = E\int L_\beta(\beta, \lambda)\omega d\lambda = EL_\beta(\beta, \lambda_0) \quad \forall\lambda$$

$$ES_i^W(\beta) = E\int_{\lambda\_\min}^{\lambda \max} L_\beta(\beta, \lambda)\omega d\lambda$$

$$= \int\left\{\int_{\lambda\_\min}^{\lambda \max} L_\beta(\beta, \lambda)\omega d\lambda\right\}e^{L(\beta_0,\lambda_0)}dt$$

$$= \int_{\lambda\_\min}^{\lambda \max} \{L_\beta(\beta, \lambda)e^{L(\beta_0,\lambda_0)}dt\}\omega d\lambda$$

$$= \int_{\lambda\_\min}^{\lambda \max} \{EL_\beta(\beta, \lambda)\}\omega d\lambda$$

$$= \int_{\lambda\_\min}^{\lambda \max} \{EL_\beta(\beta, \lambda_0)\}\omega d\lambda$$

$$= EL_\beta(\beta, \lambda_0) \text{ since } \int_{\lambda\_\min}^{\lambda \max} \omega d\lambda = 1$$

APPENDIX C (PROOF OF THEOREM 1)

$$\{\hat{\beta}_{\mathrm{EOE}}, \hat{\gamma}_{\mathrm{EOE}}\} = \underset{\beta\in\Theta, \gamma\in\Theta,}{\arg\min}\{L_\beta(\beta,\gamma)'L_\beta(\beta,\gamma) + L_\gamma(\beta,\gamma)'L_\gamma(\beta,\gamma)\} \text{ and}$$

$$\hat{\beta}_{\mathrm{IntScore}} = \underset{\beta\in\Theta}{\arg\min}\, S^{II}(\beta)'S^{II}(\beta)$$

Note that by Newey and McFadden (1994, Lemma 2.4), we have that $L_\beta(\beta,\gamma)$, $L_\gamma(\beta,\gamma)$, and $S^{II}(\beta)$ converge uniformly to their expectations. Also note that Assumption 1 part (iii) and EOE imply that $E\{L_\beta(\beta,\gamma)\} = 0$ only if $\beta = \beta_0$. Also note that the expectation of $L_\gamma(\beta,\gamma)$ does not depend on β so that the limit of the objective function $\{L_\beta(\beta,\gamma)'L_\beta(\beta,\gamma) + L_\gamma(\beta,\gamma)'L_\gamma(\beta,\gamma)\}$ is minimized at $\beta = \beta_0$. The remainder of the proof follows Newey and McFadden (1994, Theorem 2.6).

APPENDIX D

$$S_i^{II} = \int L_{i\beta}\,d\lambda = \int L_\beta \frac{d\lambda}{dv}dv$$
$$= \int \left[\frac{\partial v}{\partial \beta}\left\{\sum_s d_{is} - \sum_s e^{v+x_{is}\beta}y_{is}\right\} + \left\{\sum_s d_{is}x_{is} - \sum_s x_{is}e^{v+x_{is}\beta}y_{is}\right\}\right]\frac{d\lambda}{dv}dv$$

Note that

$$\frac{\partial v}{\partial \beta} = -\frac{\sum_s x_{is}EL_{\mu\mu}}{\sum_s EL_{\mu\mu}} = -\frac{\sum_s x_{is}e^{v+x_{is}\beta}Ey_{is}}{\sum_s e^{v+x_{is}\beta}Ey_{is}}$$

and

$$\frac{\partial \lambda}{\partial v_i} = \sum_s EL_{\mu\mu} = \sum_s e^{v+x_{is}\beta}Ey_{is}$$

Therefore,

$$S_i^{II} = \int \left[\left\{\sum_s d_{is}\right\}\sum_s x_{is}e^{v+x_{is}\beta}Ey_{is} + \left\{\sum_s d_{is}x_{is}\right\}\sum_s e^{v+x_{is}\beta}Ey_{is}\right]dv$$
$$= \left\{\sum_s d_{is}\right\}\sum_s x_{is}e^{v+x_{is}\beta}Ey_{is} + \left\{\sum_s d_{is}x_{is}\right\}\sum_s e^{v+x_{is}\beta}Ey_{is}$$

Note that $\sum_s e^{v+x_{is}\beta} y_{is} \sum_s x_{is} e^{v+x_{is}\beta} Ey_{is} = \sum_s x_{is} e^{v+x_{is}\beta} y_{is} \sum_s e^{v+x_{is}\beta} Ey_{is}$. Therefore,

$$\begin{aligned} S_i^W &= \int \left[\left\{ \sum_s d_{is} \right\} \sum_s x_{is} e^{v+x_{is}\beta} Ey_{is} + \left\{ \sum_s d_{is} x_{is} \right\} \sum_s e^{v+x_{is}\beta} Ey_{is} \right] dv \\ &= \left\{ \sum_s d_{is} \right\} \sum_s x_{is} e^{v+x_{is}\beta} Ey_{is} + \left\{ \sum_s d_{is} x_{is} \right\} \sum_s e^{v+x_{is}\beta} Ey_{is} \end{aligned}$$

One can derive the same moment function by using the score $L_{i\beta}$:

$$L_{i\beta} = -\frac{\sum_s x_{is} e^{v+x_{is}} Ey_{is}}{\sum_s e^{v+x_{is}} Ey_{is}} \left\{ \sum_s d_{is} - \sum_s e^{v+x_{is}\beta} y_{is} \right\} + \sum_s d_{is} x_{is} - \sum_s x_{is} e^{v+x_{is}\beta} y_{is}$$

Note that $EL_{i\beta}(\beta_0) = 0$. The zero expectation property is maintained if one multiplies $L_{i\beta}$ by $\sum_s e^{v+x_{is}} Ey_{is}$.

APPENDIX E

$$L^I = \frac{1}{N} \sum_i L_i(\beta, \lambda)$$

$$\begin{aligned} L_i^I &= \ln \int e^{L_i(\beta,\lambda)} d\lambda \\ &= \ln \int e^{L_i(\beta,\lambda)} \frac{d\lambda}{dv} dv \end{aligned}$$

where $L_i(\beta, \lambda) = \sum_s (v + x_{is}\beta) d_{is} - \sum_s e^{v+x_{is}\beta} y_{is}$

Use $d\lambda/dv = \sum_s e^{x_{is}\beta+v} Ey_{is}$

$$\begin{aligned} L_i^I &= \ln \int e^{Li(\beta,\lambda)} \left(\sum_s e^{x_{is}\beta+v} Ey_{is} \right) dv \\ &= \ln \int e^{Li(\beta,\lambda)} \left(\sum_s e^{x_{is}\beta+v} Ey_{is} \right) \frac{dv}{df} df \end{aligned}$$

where $f = e^v$ and $dv/df = 1/f$; use $L_i(\beta, \lambda) = \sum_s (\ln f + x_{is}\beta) d_{is} - f \sum_s e^{x_{is}\beta} y_{is}$

$$L_i^l = \ln \int \frac{1}{f} e^{L(\beta,\lambda)}(f \sum_s e^{x_{is}\beta} Ey_{is}) df$$

$$= \ln \int f^{\sum_s d_{is}} e^{\sum_s (x_{is}\beta) d_{is}} e^{f \sum_s e^{x_{is}\beta} y_{is}} \left(\sum_s e^{x_{is}\beta} Ey_{is} \right) df$$

$$e^{L_i^l} = e^{\sum_s (x_{is}\beta) d_{is}} \int f^{\sum_s d_{is}} e^{\sum_s (x_{is}\beta) d_{is}} e^{f \sum_s e^{x_{is}\beta} y_{is}} \left(\sum_s e^{x_{is}\beta} Ey_{is} \right) df$$

$$= e^{\sum_s (x_{is}\beta) d_{is}} \left(\sum_s e^{x_{is}\beta} Ey_{is} \right) \int f^{\sum_s d_{is}} e^{f \sum_s e^{x_{is}\beta} y_{is}} df$$

Note that the expression under the integral sign implies that f has a gamma distribution.

Therefore,

$$e^{L_i^l} = e^{\sum_s (x_{is}\beta) d_{is}} \left(\sum_s e^{x_{is}\beta} Ey_{is} \right) \left(\sum_s e^{x_{is}\beta} y_{is} \right)^{-(\sum_s d_{is}+1)}$$

This gives

$$L_i^l = \sum_s (x_{is}\beta) d_{is} + \ln \left(\sum_s e^{x_{is}\beta} Ey_{is} \right) - \left(\sum_s d_{is} + 1 \right) \ln \left(\sum_s e^{x_{is}\beta} y_{is} \right)$$

$$\frac{\partial L_i^l(\beta)}{\partial \beta} = \sum_s x_{is} d_{is} + \frac{\sum_s x_{is} e^{x_{is}\beta} Ey_{is}}{\sum_s e^{x_{is}\beta} Ey_{is}} - \left(\sum_s d_{is} + 1 \right) \frac{\sum_s x_{is} e^{x_{is}\beta} y_{is}}{\sum_s e^{x_{is}\beta} y_{is}}$$

APPENDIX F

$$\frac{\partial L_i^l(\beta)}{\partial \beta} = \sum_s x_{is} d_{is} + \frac{\sum_s x_{is} e^{x_{is}\beta} Ey_{is}}{\sum_s e^{x_{is}\beta} Ey_{is}} - \left(\sum_s d_{is} + 1 \right) \frac{\sum_s x_{is} e^{x_{is}\beta} y_{is}}{\sum_s e^{x_{is}\beta} y_{is}}$$

This FOC resembles the FOC of the weighted score. However, in this case we have stochasts, y_{is}, $s = 1, \ldots, T$ in the denominator. This causes $(\partial L^l(\beta))/(\partial \beta)$ to be nonzero at the truth. We define

$$h(\beta) = \sum_s e^{x_{is}\beta} y_{is} \sum_s x_{is} d_{is} + \sum_s e^{x_{is}\beta} y_{is} \frac{\sum_s x_{is} e^{x_{is}\beta} E y_{is}}{\sum_s e^{x_{is}\beta} E y_{is}}$$
$$- \left(\sum_s d_{is} + 1 \right) \sum_s x_{is} e^{x_{is}\beta} y_{is}$$

The expectation of $h(\beta)$ equals the expectation of the weighted score: $Eh(\beta) = Eg(\beta)$. This equality also holds at the true value: $Eh(\beta_0) = Eg(\beta_0) = 0$ However, $h(\beta)$ is correlated with $1/(\sum_s e^{x_{is}\beta} E y_{is})$. Therefore, $E\{g(\beta_0) 1/(\sum_s e^{x_{is}\beta_0} y_{is})\} \neq 0$. Thus, in general, $E\{(\partial L^I(\beta_0))/(\partial\beta)\} \neq 0$.