Cox-McFadden Semiparametric Estimation for a Class of Clustered Proportional Hazards

Jan Ondrich
Center for Policy Research
426 Eggers Hall
Syracuse University
Syracuse, New York 13244

February 8, 2010

Abstract

Cox's marginal likelihood approach to estimating proportional hazards for independent durations is semiparametric in that it eliminates the duration baseline. Successive marginal likelihood contributions are conditional logit probabilities that the next rank-ordered failure is the next observed rank-ordered failure. This paper examines the problem of constructing a valid baseline-free marginal likelihood for the semiparametric estimation of proportional hazards when observations are clustered. It turns out that the problem is isomorphic to the problem of relaxing the assumption of independent errors in an extreme-value stochastic utility model. McFadden characterizes all discrete choice probability models with univariate extreme-value disturbances that are consistent with stochastic utility maximization. He presents sufficient conditions for the joint distribution to be consistent with stochastic utility maximization. The sufficient conditions describe the set of GEV models. This paper characterizes all marginal likelihoods for clustered proportional hazards in which the duration baseline is eliminated from the estimation. It shows further that a sufficient condition for the elimination of the duration baseline is that the probability that the first rank-ordered failure is the first observed rank-ordered failure can be modeled as a GEV probability.

* Professor of Economics, Syracuse University, and Senior Research Associate, Center for Policy Research, Syracuse University. I would like to

thank Gary Chamberlain for teaching me survival analysis, James Follain for teaching me logistic choice models, and Michael Rothschild for introducing me to the writings of Feller. None of these individuals are involved with this study, so all errors are my own. Finally, I would like to thank Gert G. Wagner, Director of the German Socio-Economic Panel (GSOEP) for providing resources and time. The theoretical work for this paper was completed while the author was Visiting Scholar at DIW/GSOEP. This paper was presented at a GSOEP seminar and at the 2006 New York Camp Econometrics.

1 Introduction

Cox (1972, 1975) develops the proportional hazard model of durations and suggests semiparametric estimation using a partial likelihood approach. Durations are rank-ordered and contributions to the partial likelihood are provided at each failure time by the subset of the sample at risk immediately before the failure time. All partial likelihood contributions are conditional logit probabilities. Because the duration baseline is canceled out of the partial likelihood, the approach has the advantage of being baseline-free: duration-dependence parameters, frequently viewed as nuisance parameters, do not have to be estimated. Researchers interested in duration dependence can recover the duration baseline in a second step. The case for partial likelihood was strengthened with the later finding by Tsiatis (1981) that partial likelihood estimation is equivalent to marginal likelihood estimation.

The introduction of stratified partial likelihood estimation (see Chamberlain 1985; Gross and Huber 1987; Andersen, Borgan, Gill, and Keiding 1993; and Ridder and Tunali 1999) allows for models with group-specific duration baselines. The group-specific duration baselines can be recovered in a second stage, but the coefficients of covariates invariant within groups cannot be recovered. Stratified partial likelihood estimation, therefore, does not allow hazard prediction.

This paper investigates a class of models for baseline-free marginal likelihood for clustered proportional hazards that allows hazard prediction and the estimation of coefficients of covariates invariant within groups. The starting point is the work of Hougaard (1986a, 1986b) and the analysis of McFadden (1978) generalizing the conditional logit model. The class of models represent an alternative to the class of Generalized Accelerated Failure Time models, on which interesting new analysis has recently been completed by Khan and Tamer (2007,2010) and Khan, Shin, and Tamer (2010).

When the durations are independent, the mathematical form of the partial (marginal) likelihood contributions is identical to that of the log-likelihood contributions for the logit model, proposed by Luce (1959) to estimate the probability that an item is selected from a choice set of alternatives. McFadden (1974) presents a formal econometric analysis of the conditional logit model. The model assumes that the stochastic utility of each choice is the sum of a deterministic component and an extreme-value error term. The model has the property that the log-odds of any two choices are independent of the availability or attributes of other alternatives. While the independence of irrelevant alternatives (IIA) property simplifies the econometric estimation, it is an undesirable feature in choice settings in which alternatives have close substitutes.

McFadden (1978) introduces a class of multivariate extreme-value distributions (called generalized extreme-value or simply GEV) that allows the IIA property to be relaxed. His GEV discrete choice models are consistent with stochastic utility maximization in the sense that choice probabilities are unchanged when all utilities in the choice set are increased by the same amount.

It turns out that the problem of constructing marginal likelihood approaches to estimating proportional hazards for clustered observations is isomorphic to the problem of relaxing the IIA property in an extreme-value stochastic utility model. Specifically, a joint distribution of extreme-value utility shocks is consistent with stochastic utility maximization in a discrete choice model if and only the joint distribution is a multivariate survivor function for a sample of clustered proportional hazard durations allowing cancellation of the duration baseline in the partial (marginal) likelihood. I first characterize all partial likelihoods for clustered proportional hazards in which the the duration baseline is eliminated from the estimation. I show next that a sufficient condition for the elimination of the duration baseline is that the probability that the first rank-ordered failure is the first observed rank-ordered failure can be modeled as a GEV probability.

Section 2 describes the conditional logit model, the IIA property, and the GEV class of models developed by McFadden. Section 3 presents Cox's proportional hazard model and the main propositions of this study. Examples of Cox-McFadden proprtional hazard models are presented in section 4. The main propositions are proved for the marginal likelihood case in section 5. In section 6 the case of tied data is discussed. The recovery of the baseline hazard is described in section 7. Conclusions are given in section 9. An accompanying paper (Ondrich 2010) discusses asymptotic inference.

2 IIA and the GEV Model

The discrete choice model specification that is used most often in applied econometric applications is the conditional logit model. The conditional logit model provides a simple closed form for the choice probabilities; in contrast, the calculation of the choice probabilities in the multinomial probit model requires multivariate integration that can only be accomplished through numerical approximation. The likelihood function for the conditional logit specification is globally concave, which eases the computational burden of obtaining maximum likelihood estimates.

In the multinomial logit model, the probability that an individual chooses choice i from a choice set C consisting of N choices is given by

$$P(i|C, \mathbf{Z}, \boldsymbol{\beta}) = \theta_i / \sum_{j \in C} \theta_j \quad ,$$

where $\mathbf{Z} = (\mathbf{Z}_1, ..., \mathbf{Z}_J)$ gives the attributes of C, $\mathbf{Z}_j$ is a K-vector of explanatory variables describing the attributes of alternative j (perhaps interacted or moderated by the characteristics of the decision-maker), $\boldsymbol{\beta}$ is a K-vector of taste parameters, and θ_i stands for $e^{\mathbf{Z}_i \boldsymbol{\beta}}$.

The conditional logit model is characterized by the independence of irrelevant alternatives (IIA) property, namely, the ratio of probabilities (relative odds) of choosing any two alternatives is independent of the availability of a third alternative:

$$P(i|C, \mathbf{Z}, \boldsymbol{\beta}) = P(i|C_0, \mathbf{Z}, \boldsymbol{\beta}) P(C_0|C, \mathbf{Z}, \boldsymbol{\beta}), \tag{2.1}$$

where $i \in C_0 \subseteq C$ and

$$P(C_0|C, \mathbf{Z}, \boldsymbol{\beta}) = \sum_{j \in C_0} P(j|C, \mathbf{Z}, \boldsymbol{\beta}). \tag{2.2}$$

A famous example has a commuter choosing between a car and a bus for a commute. When he is late for work, which happens randomly 1/3 of the time, he drives (choice A); otherwise he chooses a bus. There are two bus companies, a red bus company and a blue bus company, indistinguishable but for color. When he is not late and is waiting for a bus, the first bus to arrive is equally likely to be blue (choice BB) or red (choice RB). From this information it is clear that with choice set $C = \{A, RB, BB\}$,

$$P(A) = P(RB) = P(BB) = \frac{1}{3} \quad . \tag{2.3}$$

Now suppose that the blue bus company suspends operations. The choice set becomes $C_0 = \{A, RB\}$, which has probability 2/3 by equations (2.2) and (2.3). With choice set $C_0 = \{A, RB\}$, the multinomial logit model predicts that

$$P(A) = P(RB) = \frac{1}{2} \quad ,$$

using equations (2.1) and (2.3). But this prediction is not likely to be validated. The commuter will continue to choose the car whenever he is late, 1/3 of the time, and will choose the red bus 2/3 of the time, whenever he is not.

It is clear from this example that models with the IIA property are inadequate in describing choice from a set of alternatives with different degrees of substitutability or complementarity. The red bus and blue bus are perfect substitutes, whereas the car and the red bus (or the car and the blue bus) are not. Several studies (McFadden, Train, and Tye 1977, Hausman and McFadden 1984, Small and Hsiao 1985, and McFadden 1987) discuss methods of testing whether IIA is violated in a given econometric application. The problem is to construct an alternative model, preferably one with closed forms for the choice probabilities.

It was solved by McFadden (1978) making using of results derived by Williams (1977) and Daly and Zachary (1978) on the compatibility of a given probabilistic choice model with utility maximization (see Daly and Zachary 1978, or Börsch-Supan 1987).

Theorem 1 (McFadden): *Suppose $M(\theta_1, ..., \theta_N)$ is a non-negative function defined on the non-negative real numbers with the following three properties:*
1) alternating distinct partials, i.e., for any distinct $\{j_1, ..., j_Q\}$ from the choice set $\{1, ..., N\}$, the Qth partial $\partial^Q M / \partial\theta_{j_1} ... \partial\theta_{j_Q}$ is non-negative if Q is odd and non-positive if Q is even;
2) infinite limits, i.e., $\lim_{\theta_i \to \infty} M(\theta_1, ..., \theta_N) = \infty$, $i = 1, ..., N$; and
3) homogeneity of degree $\mu \geq 0$.
Then, the probabilities
$P(i|C, Z, \beta) = \theta_i \dfrac{\partial M(\theta_1, ..., \theta_N)}{\partial\theta_i} / M(\theta_1, ..., \theta_N), \;\; i = 1, ..., N$
define a probabilistic choice model on the choice set $\{1, ..., N\}$ that is consistent

with utility maximization.

The function M is McFadden's negative log copula. A copula is a function that assigns the value of the joint distribution function to each n-tuple of values of the marginal distributions. (Andersen 2004 uses copulas to construct a two-stage semi-parametric estimator for multivariate failure-time data.) I define a negative log copula to be a function that assigns the value of the negative log of the joint distribution function to each n-tuple of values of the negative log of the univariate marginal distributions. McFadden's negative log copula will be shown to play a crucial role in specifying the baseline-free partial likelihood and marginal likelihood for clustered proportional hazards.

3 Proportional Hazards and Two Propositions

The duration or failure time T of a stochastic process is its random age at termination or failure. The assumption in this study is that durations are continuous random variables: they possess an absolutely continuous distribution function $F(t)$. The distribution function is non-defective, i.e., $F(\infty) = 1$, and has density $f(t)$. The unitary complement of the distribution function of a continuous duration is its survivor function

$$S(t) \equiv P(T \geq t) = 1 - F(t).$$

The survivor function represents the probability that the process survives up to age t and only fails at time t or later.

One of the fundamental concepts in the analysis of continuous durations is the hazard rate, denoted by h and defined by

$$h(t) \equiv f(t)/(1 - F(t)).$$

The quantity $h(t)dt$ represents the probability that the process fails in the interval $[t, t + dt)$ conditional on survival to age t. It is well known that for a specific $h(t)$, the survivor function and density are given by

$$S(t) = \exp(-\int_0^t h(u)du) \equiv \exp(-H(t))$$

and

$$f(t) = h(t)\exp(-\int_0^t h(u)du).$$

For a sample of N spells, Cox's proportional hazard specification assigns to spell i a hazard rate of the form

$$h_i(t|\mathbf{Z},\boldsymbol{\beta}) = \exp(\mathbf{Z}_i\boldsymbol{\beta})h_0(t) \equiv \theta_i h_0(t),$$

where $\mathbf{Z}_i$ is the covariate vector for spell i, $\mathbf{Z} = (\mathbf{Z}_1, ..., \mathbf{Z}_N)$, $\boldsymbol{\beta}$ is the coefficient vector, and h_0 is the (unspecified) baseline hazard rate. (I will assume that the covariate vector is time-invariant, i.e., it does not change with process age. The principal results of this study allow time-varying covariates, as discussed in Ondrich 2006.) The integrated baseline hazard rate is defined by

$$H_0(t) \equiv \int_0^t h_0(u)du,$$

so that the survivor function for spell i can be written simply as

$$S_i(t) = \exp(-\theta_i H_0(t)).$$

In a proportional hazard model the hazard elasticity with respect to any continuous positive covariate depends only on the value of the covariate and its coefficient, and does not require additional knowledge of the process age t.

The sample survivor function for the sample of N spells is defined as

$$S(u_1, ..., u_N|\mathbf{Z},\boldsymbol{\beta}) \equiv P(T_1 \geq u_1, ..., T_N \geq u_N).$$

It will also be necessary to define marginal survivor functions. The marginal survivor function of a subset of the N sample spells is derived from the sample survivor function by setting $u_i = 0$ for all i not in the subset. Alternatively, denote the subset by A and for each i define Y_i^A to be the indicator equal to one if i is an element of A. Then, letting $\mathbf{u}$ be the vector $(u_1, ..., u_N)$, the marginal survivor function is given by

$$S_A(\mathbf{u}|\mathbf{Z},\boldsymbol{\beta}) = S(Y_1^A u_1, ..., Y_N^A u_N|\mathbf{Z},\boldsymbol{\beta}).$$

Of particular interest will be the marginal survivor function $S_{R(t)}(t\iota|\mathbf{Z},\boldsymbol{\beta})$, for which $\mathbf{u}$ is the constant vector $t\boldsymbol{\iota}$, where $\boldsymbol{\iota}$ is the N-dimensional unitary vector, and the subset of interest is the risk set at time t, denoted $R(t)$, the subset of sample spells that survive to age t. Note that $R(0)$ is the complete sample of durations.

If the N sample durations are statistically independent, the sample survivor function is

$$S(\mathbf{u}|\mathbf{Z},\boldsymbol{\beta}) = \exp(-\sum_{i=1}^{N} \theta_i H_0(u_i)),$$

defined on non-negative real N-tuples $\mathbf{u}$. The main propositions in this study involve samples of clustered durations for which the survivor function takes the form

$$S(\mathbf{u}|\mathbf{Z},\boldsymbol{\beta}) = \exp(-M(\theta_1 H_0(u_1), ..., \theta_N H_0(u_N))),$$

again defined on the non-negative real N-tuples $\mathbf{u}$. For all subsets A of $\{1, ..., N\}$, the marginal survivor function is

$$S_A(\mathbf{u}|\mathbf{Z},\boldsymbol{\beta}) = \exp(-M(Y_1^A \theta_1 H_0(u_1), ..., Y_N^A \theta_N H_0(u_N))).$$

If $\mathbf{u}$ is a constant vector and M is homogeneous of degree one, then

$$M(Y_1^A \theta_1 H_0(t), ..., Y_N^A \theta_N H_0(t)), A) = H_0(t) M(Y_1^A \theta_1, ..., Y_N^A \theta_N),$$

so that for all constant vectors $\mathbf{u}$ and sets A,

$$S_A(t\boldsymbol{\iota}|\mathbf{Z},\boldsymbol{\beta}) = \exp(-H_0(t) M(Y_1^A \theta_1, ..., Y_N^A \theta_N))).$$

The final definition that is required for the propositions is that of the hazard function, which is distinct from a hazard rate. The (multivariate) hazard function is simply the negative log of the survivor function, while a marginal hazard function is the negative log of the marginal survivor function.

Proposition 1 *If the hazard function M of a sample of N clustered proportional hazards is differentiable and homogeneous of positive degree in its univariate integrated hazard rates $\theta_i H_0(t)$, $i = 1, ..., N$, then the marginal likelihood is baseline-free and the probability that duration i is the failure at time t from the set $R(t)$ of durations at risk is given by*

$$P(i|R(t),\mathbf{Z},\boldsymbol{\beta}) = \theta_i \frac{\partial M(Y_1^{R(t)} \theta_1, ..., Y_N^{R(t)} \theta_N)}{\partial \theta_i} / M(Y_1^{R(t)} \theta_1, ..., Y_N^{R(t)} \theta_N).$$

Proposition 2 *The non-negative function $M(\theta_1, ..., \theta_N)$ is the hazard function of a sample of N clustered proportional hazards with a baseline-free marginal likelihood if it satisfies the following three conditions:*
1) alternating distinct partials, i.e., for any distinct $\{j_1, ..., j_Q\}$ from the set

$\{1, ..., N\}$, the Qth partial $\partial^Q M/\partial\theta_{j_1}...\partial\theta_{j_Q}$ is non-negative if Q is odd and non-positive if Q is even;
2) infinite limits, i.e., $\lim_{\theta_i \to \infty} M(\theta_1, ..., \theta_N) = \infty$, $i = 1, ..., N$; and
3) homogeneity of degree $\mu \geq 0$.

Before proving Propositions 1 and 2, I will present examples of Cox-McFadden proportional hazard models in the next section.

4 Cox-McFadden Models

This section presents examples of Cox-McFadden models. I assume that the sample of N individuals can be divided into G independent groups or clusters. Group g is composed of N_g individuals and is associated with its own negative log survivor function M_g. Each M_g satisfies the conditions of Propositions 1 and 2, and therefore $\sum_{g=1}^{G} M_g$ also satisfies these conditions.

The first example comes from Hougaard (1986a, 1986b) and results from positive stable mixing. (See Cardell 1997 for a discussion of positive stable and related mixing distributions.) Suppose $X_1, X_2, ..., X_n, ...$ are independent and identically distributed. Their common distribution is stable if, for each n, there exists a constant c_n such that $c_n X_1$ and $\sum_{i=1}^{n} X_i$ follow the same distribution. Any stable distribution has constants c_n of the form $n^{1/\alpha}$, where the characteristic exponent $\alpha \in (0, 2]$. Normal distributions have $\alpha = 2$ and are the only stable distributions with finite variance. The positive stable distributions, *i.e.*, those which have support on the positive real numbers, all have $\alpha \in (0, 1)$ with Laplace transforms, apart from scaling factors, of the form $\omega(\lambda) = \exp(-\lambda^\alpha)$, for $\lambda \geq 0$.

If group g shares a common positive stable random effect with characteristic exponent α, then

$$M_g(\theta_1, ..., \theta_{N_g}) = \left(\sum_{i=1}^{N_g} \theta_i^{1/\alpha} \right)^\alpha,$$

which satisfies the conditions of Propositions 1 and 2. The function M_g has the form of the negative log copula for the utility shocks from a single nest in a nested logit choice model (see McFadden 1978).

Feller (1971) shows that if X_1 and X_2 are independent stable distributions with characteristic exponents α_1 and α_2 ($\alpha_2 < 1$), then $X_1 X_2^{1/\alpha_1}$ is stable with characteristic exponent $\alpha_1\alpha_2$. Therefore, if X_1 and X_2 are both positive stable, $X_1 X_2^{1/\alpha_1}$ is positive stable as well. Hougaard (1986b) uses this to construct a nested frailty model in which three siblings share a family effect and the twins share a "twin" effect.

Sastry (1997) analyzes a nested frailty (using gamma distributions) for child survival in Brazil , where the data are clustered at both the family and community levels. Following Hougaard and using positive stable distributions to construct the nests, the negative log survivor function for community g composed of individuals j, each a member of a family i, is given by

$$M_g(\theta_1, ..., \theta_{N_g}) = \left(\sum_{i \in g} \left(\sum_{j \in i} \theta_{ij}^{1/\alpha_2} \right)^{\alpha_2/\alpha_1} \right)^{\alpha_1}$$

where $\alpha_1 \geq \alpha_2$. McFadden (1978) presents the negative log copula for the utility shocks from a doubly nested set of choices in a nested logit choice model that has an identical form. Hierarchies with more than two levels of nesting can be easily constructed, and non-nested models are also possible.

5 Marginal Likelihood

The discovery that maximization of the marginal likelihood yields the partial likelihood estimator when durations are independent is important because the marginal likelihood function is a proper likelihood function to which the usual asymptotic theory of maximum likelihood directly applies.

Initially, it is assumed that the sample spells are uncensored. Let $T_i, i = 1, ..., N$, represent the failure times of the N sample spells. Further, let $T_0^o < T_1^o < ... < T_N^o$ be the ordered failure times and let (i) denote the anti-rank, namely the label of the spell failing at T_i^o. Construct two vectors, $\mathbf{O} = (T_1^o, ..., T_N^o)$, the vector of order statistics, and $\mathbf{r} = ((1), ..., (N))$, the vector of rank statistics. Note that the vector of sample failure times, $\mathbf{T} = (T_1, ..., T_N)$ can be reconstructed from knowledge of $\mathbf{O}$ and $\mathbf{r}$.

Kalbfleisch and Prentice (1980) present an example in which $N = 4$ and $\mathbf{T} = (5, 17, 12, 15)$. The vector of order statistics for this data is $\mathbf{O} = (5, 12, 15, 17)$ and the vector of rank statistics is $\mathbf{r} = (1, 3, 4, 2)$. If the value of the jth component of $\mathbf{r}$ equals i, then T_i is the jth smallest sample failure time, with value given by the jth component of $\mathbf{O}$.

The fact that the vector of rank statistics carries the sample information about $\boldsymbol{\beta}$ when the baseline hazard rate h_0 is completely unspecified can be demonstrated by a simple argument. The hazard rate for duration i, T_i, in the proportional hazard model is given by $\theta_i h_0(t)$.For all i, define $V_i = g^{-1}(T_i)$, where g is an arbitrary element of G, the group of differentiable and strictly increasing transformations of $(0, \infty)$ into $(0, \infty)$. Then, given $\mathbf{Z}$ and $\boldsymbol{\beta}$, the hazard rate for V_i is given by $\theta_i h_0^*(v)$,where $h_0^*(v) = h_0(g(v))g'(v)$.This shows that when the baseline hazard is unspecified, the vector of order statistics can be modified arbitrarily as long as the vector of rank statistics is unchanged, and the problem of inference about $\boldsymbol{\beta}$ has not changed. The estimation problem for $\boldsymbol{\beta}$, given an unspecified baseline, is invariant to (continuous) monotonic transformations of duration.

The estimation of $\boldsymbol{\beta}$ can therefore be based on the marginal likelihood of $\mathbf{r}$. Sample values of the random ordered failure times are $(T_1^o, ..., T_N^o) = (t_1, ..., t_N)$. When sample durations are independent, the marginal likelihood of $\mathbf{r}$ is given by

$$\begin{aligned} P \quad (\mathbf{r} = \quad & ((1), ..., (N))|\mathbf{Z}, \boldsymbol{\beta}) \\ = \quad & P(T_{(1)} < ... < T_{(N)}|\mathbf{Z}, \boldsymbol{\beta}) \\ = \quad & \int_0^\infty \int_{t_1}^\infty \cdots \int_{t_{N-1}}^\infty \prod_{i=1}^N f(t_i|\mathbf{Z}_{(i)}, \boldsymbol{\beta}) dt_N \cdots \ dt_1. \end{aligned}$$

When durations are clustered, the density for T_i must also be conditioned on A_i, the event $\{T_{(j)} > t_i|\, j = i+1, ..., N\}$,where A_N is the null event. Therefore, in the case of clustered durations,

$$\begin{aligned} P(\mathbf{r} \quad = \quad & ((1), ..., (N))|\mathbf{Z}, \boldsymbol{\beta}) = P(T_{(1)} < ... < T_{(N)}|\mathbf{Z}, \boldsymbol{\beta}) \\ = \quad & \int_0^\infty \int_{t_1}^\infty \cdots \int_{t_{N-1}}^\infty \prod_{i=1}^N f(t_i|A_i, \mathbf{Z}_{(i)}, \boldsymbol{\beta}) dt_N \cdots \ dt_1. \end{aligned} \tag{5.1}$$

The multiple integral in equation (5.1) is evaluated recursively, as given by

$$\begin{aligned} & \int_0^\infty f(t_1|A_1, \mathbf{Z}_{(1)}, \boldsymbol{\beta}) \ \cdots \int_{t_{N-2}}^\infty \ f(t_{N-1}|A_{N-1}, \mathbf{Z}_{(N-1)}, \boldsymbol{\beta}) \\ & \quad \int_{t_{N-1}}^\infty f(t_N|A_N, \mathbf{Z}_{(N)}, \boldsymbol{\beta}) dt_N \, dt_{N-1} \cdots \ dt_1. \end{aligned}$$

It is required to prove that this integral equals

$$\prod_{i=1}^{N} \frac{\theta_{(i)} M^{[(i)]}(Y_1^{R(t_i)}\theta_1, ..., Y_N^{R(t_i)}\theta_N)}{M(Y_1^{R(t_i)}\theta_1, ..., Y_N^{R(t_i)}\theta_N)},$$

where the superscript $[(i)]$ denotes the partial derivative with respect to the argument given by the anti-rank (i). I will prove that the marginal likelihood equals

$$\left[\prod_{i=1}^{N} \frac{\theta_{(i)} M^{[(i)]}(Y_1^{R(t_i)}\theta_1, ..., Y_N^{R(t_i)}\theta_N)}{M(Y_1^{R(t_i)}\theta_1, ..., Y_N^{R(t_i)}\theta_N)}\right] (S_0(0))^{M(Y_1^{R(0)}\theta_1, ..., Y_N^{R(0)}\theta_N)},$$

where $S_0(t)$ is the baseline survivor function $\exp(-H_0(t))$. The desired result then follows from the fact that $S_0(0) = 1$.

To simplify the notation, I will write $\boldsymbol{\theta}$ for the vector $(\theta_1, ..., \theta_N)$, and, for all subsets A of $\{1, ..., N\}$, define

$$M(\boldsymbol{\theta}, A) = M(Y_1^A\theta_1, ..., Y_N^A\theta_N),$$

and, for all $i = 1, ..., N$,

$$M^{[(i)]}(\boldsymbol{\theta}, A) = M^{[(i)]}(Y_1^A\theta_1, ..., Y_N^A\theta_N),$$

The proof is by induction on the number of integrations performed. Because A_N is the null event, the first integration is simply the probability that the (N)th duration survives to t_{N-1}:

$$\left[\frac{\theta_{(N)} M^{[(N)]}(\boldsymbol{\theta}, R(t_N))}{M(\boldsymbol{\theta}, R(t_N))}\right] (S_0(t_{N-1}))^{M(\boldsymbol{\theta}, R(t_N))}. \tag{5.2}$$

Euler's Theorem states that if $M(\theta_1, ..., \theta_N)$ is homogeneous of degree k, then $kM(\theta_1, ..., \theta_N) = \sum_{i=1}^{N} \theta_i \frac{\partial M}{\partial \theta_i}(\theta_1, ..., \theta_N)$ (see Friedman 1971). Therefore, because only one of the N arguments of $M(\theta_1, ..., \theta_N)$ is nonzero when the risk set is $R(t_N)$, the expression in brackets in equation (5.2) equals one.

The induction hypothesis is that the result for the first j integrations is

$$\left[\prod_{i=N-j+1}^{N} \frac{\theta_{(i)} M^{[(i)]}(\boldsymbol{\theta}, R(t_i))}{M(\boldsymbol{\theta}, R(t_i))}\right] (S_0(t_{N-j}))^{M(\boldsymbol{\theta}, R(t_{N-j+1}))}.$$

The proof is complete if I show that the result after $j+1$ integrations is

$$\left[\prod_{i=N-j}^{N} \frac{\theta_{(i)} M^{[(i)]}(\boldsymbol{\theta},R(t_i))}{M(\boldsymbol{\theta},R(t_i))}\right] (S_0(t_{N-j-1}))^{M(\boldsymbol{\theta},R(t_{N-j}))} .$$

Therefore, it must be shown that

$$\int_{t_{N-j-1}}^{\infty} f(t_{N-j}|A_{N-j}, \mathbf{Z}_{(N-j)}, \boldsymbol{\beta})\, (S_0(t_{N-j}))^{M(\boldsymbol{\theta},R(t_{N-j+1}))}\, dt_{N-j}$$

$$= \frac{\theta_{(N-j)} M^{[(N-j)]}(\boldsymbol{\theta},R(t_{N-j}))}{M(\boldsymbol{\theta},R(t_{N-j}))} (S_0(t_{N-j-1}))^{M(\boldsymbol{\theta},R(t_{N-j}))} . \tag{5.3}$$

The first task is to evaluate $f(t_{N-j}|A_{N-j}, \mathbf{Z}_{(N-j)}, \boldsymbol{\beta})$.Note that the probability that spell i survives to u_i given that spell j survives to u_j for all $j \neq i$ is given by

$$\frac{S(u_1, ..., u_{i-1}, u_i, u_{i+1}, ..., u_N|\mathbf{Z}, \boldsymbol{\beta})}{S(u_1, ..., u_{i-1}, 0, u_{i+1}, ..., u_N|\mathbf{Z}, \boldsymbol{\beta})}.$$

Hence, the probability that spell $(N-j)$ survives to t_{N-j} given that the remaining spells in risk set $R(t_{N-j})$ exceed t_{N-j} is given by

$$S_{R(t_{N-j})}(t_{N-j}\boldsymbol{\iota}|\mathbf{Z}, \boldsymbol{\beta})/S_{R(t_{N-j+1})}(t_{N-j}\boldsymbol{\iota}|\mathbf{Z}, \boldsymbol{\beta}). \tag{5.4}$$

The density $f(t_{N-j}|A_{N-j}, \mathbf{Z}_{(N-j)}, \boldsymbol{\beta})$ is obtained by deriving the numerator in (5.4) with respect to argument $(N-j)$ and changing the sign:

$$f(t_{N-j}|A_{N-j}, \mathbf{Z}_{(N-j)}, \boldsymbol{\beta}) = -\frac{S^{[(N-j)]}_{R(t_{N-j})}(t_{N-j}\boldsymbol{\iota}|\mathbf{Z}, \boldsymbol{\beta})}{S_{R(t_{N-j+1})}(t_{N-j}\boldsymbol{\iota}|\mathbf{Z}, \boldsymbol{\beta})}. \tag{5.5}$$

Since the denominator on the right-hand side of equation (5.5) equals $(S_0(t_{N-j}))^{M(\boldsymbol{\theta},R(t_{N-j+1}))}$,the integral in equation (5.3) equals

$$-\int_{t_{N-j-1}}^{\infty} S^{[(N-j)]}_{R(t_{N-j})}(t_{N-j}\boldsymbol{\iota}|\mathbf{Z}, \boldsymbol{\beta}) dt_{N-j}. \tag{5.6}$$

The partial derivative inside the integral of (5.6) equals

$$\theta_{(N-j)} M^{[(N-j)]}(\boldsymbol{\theta},R(t_{N-j})) h_0(t_{N-j}) \exp(-H_0(t_{N-j}) M(\boldsymbol{\theta},R(t_{N-j})). \tag{5.7}$$

Substituting (5.7) into (5.6), multiplying inside the integral by $M(\boldsymbol{\theta},R(t_{N-j}))$

and outside the integral by its reciprocal, yields

$$-\frac{\theta_{(N-j)}M^{[(N-j)]}(\boldsymbol{\theta},R(t_{N-j}))}{M(\boldsymbol{\theta},R(t_{N-j}))}\cdot$$

$$\int_{t_{N-j-1}}^{\infty} h_0(t_{N-j})M(\boldsymbol{\theta},R(t_{N-j}))\exp(-H_0(t_{N-j})M(\boldsymbol{\theta},R(t_{N-j}))dt_{N-j}. \quad (5.8)$$

The integrand in (5.8) equals

$$\frac{d\left(S_0(t_{N-j})\right)^{M(\boldsymbol{\theta},R(t_{N-j}))}}{dt_{N-j}}, \quad (5.9)$$

and therefore the integral equals $-\left(S_0(t_{N-j-1})\right)^{M(\boldsymbol{\theta},R(t_{N-j}))}$.Substituting the evaluated integral into (5.8) yields the right-hand side of equation (5.3). The proof is complete for the case of no censoring.

When sample spells can be censored, the data vector for the $i$ith duration is $(X_i, \delta_i, \mathbf{Z}_i)$, where again $X_i = \min(T_i, U_i)$ for uninformative censoring time U_i independent of T_i, and δ_i is the censoring indicator equal to one when $X_i = U_i < T_i$ and zero otherwise. Let $X_1^o < ... < X_N^o$ represent the ordered observation times, and define $\mathbf{O}^* = (X_1^o, ..., X_N^o)$. Let $\mathbf{r}^* = ((1)^*, ..., (N)^*)$ denote the vector of corresponding anti-ranks, and let $\boldsymbol{\delta}^* = (\delta_{(1)^*}, ..., \delta_{(N)^*})$ denote the vector of ordered censoring indicators. Just as in the uncensored case where $\mathbf{T} = (T_1, ..., T_N)$ can be reconstructed from knowledge of $(\mathbf{O}, \mathbf{r})$, here in the case where spells may be censored, $(\mathbf{X}, \boldsymbol{\delta})$, where $\mathbf{X} = (X_1, ..., X_N)$ and $\boldsymbol{\delta} = (\delta_1, ..., \delta_N)$, can be reconstructed from knowledge of $(\mathbf{O}^*, \mathbf{r}^*, \boldsymbol{\delta}^*)$.

As an example, suppose $\mathbf{X} = (5, 17, 12, 15)$ and $\boldsymbol{\delta} = (0, 1, 1, 0)$. Then, $\mathbf{O}^* = (5, 12, 15, 17)$, $\mathbf{r}^* = (1, 3, 4, 2)$, and $\boldsymbol{\delta}^* = (0, 1, 0, 1)$. If the value of the jth component of $\mathbf{r}^*$ equals i, then X_i is the jth smallest sample failure time, with value given by the jth component of $\mathbf{O}^*$. Similarly, the ith component of $\boldsymbol{\delta}$ equals one if and only if the value of the jth component of $\mathbf{r}^*$ equals i and the value of the jth component of $\boldsymbol{\delta}^*$ equals one. The value of the ith component of $\boldsymbol{\delta}$ equals zero otherwise.

Kalbfleisch and Prentice (1980) explain that some modification to the marginal likelihood is necessary in the presence of general uninformative independent censoring. The censored model will not in general possess the group invariance properties. When censoring occurs in the sample, the rank information is incomplete. In the example above, the vector of rank statistics, $\mathbf{r}$, is known to be either $(1, 3, 4, 2)$, $(1, 4, 3, 2)$, or $(1, 4, 2, 3)$; it seems reasonable to estimate $\boldsymbol{\beta}$

using the marginal likelihood that the vector of rank statistics is one of those observationally possible. Doing so ignores the exact time of censoring, but the invariance property of the uncensored model demonstrates that the time between failures is irrelevant. Therefore, in a model with L failures, the marginal likelihood is adjusted as follows:

$$\begin{aligned} P(\mathbf{r} &= ((1),...,(L))|\mathbf{Z},\boldsymbol{\beta}) \\ &= P(T_{(1)} < ... < T_{(L)}|\mathbf{Z},\boldsymbol{\beta}) \\ &= \int_0^\infty \int_{t_1}^\infty \cdots \int_{t_{L-1}}^\infty \prod_{i=1}^L f(t_i|A_i,\mathbf{Z}_{(i)},\boldsymbol{\beta})dt_L \cdots\ dt_1. \end{aligned}$$

It is clear from the demonstration in the uncensored case that the marginal likelihood equals

$$\prod_{i=1}^L \frac{\theta_{(i)} M^{[(i)]}(\boldsymbol{\theta},R(t_i))}{M(\boldsymbol{\theta},R(t_i))}\quad .$$

The proof of Proposition 1 for the marginal likelihood in the presence of censoring is complete.

To prove Proposition 2 for the marginal likelihood, it suffices to show that

$$S(u_1,...,u_N|\mathbf{Z},\boldsymbol{\beta}) = \exp(-M(\theta_1 H_0(u_1),...,\theta_N H_0(u_N)))$$

is a valid multivariate survivor function. There are three conditions that must be satisfied. First, $S(0,...,0|\mathbf{Z},\boldsymbol{\beta}) = 0$. Second, for any $i = 1,...,N$, as u_i limits to infinity, $S(u_1,...,u_N|\mathbf{Z},\boldsymbol{\beta})$ limits to zero.

The third condition is more technical. Define $X_i = -T_i$ and $x_i = -u_i$, for $i = 1,...,N$. Then

$$\begin{aligned} S(u_1,...,u_N|\mathbf{Z},\boldsymbol{\beta}) &= P(T_1 \geq u_1,...,T_N \geq u_N) \\ &= P(X_1 \leq -u_1,...,X_N \leq -u_N) \\ &= P(X_1 \leq x_1,...,X_N \leq x_N). \end{aligned}$$

The last line is a cumulative distribution function,which is known to have non-negative distinct partials (see, for example, McFadden 1978). Distinct partials of the cumulative distribution function with respect to $-u_i$ are also non-negative. Therefore, the third condition is that distinct partials of the multivariate survivor function with respect to u_i alternate in sign, with first partials being non-positive.

To prove the first condition for a multivariate survivor function, note that as $(u_1, ..., u_N)$ limits to the zero vector, so does $(H_0(u_1), ..., H_0(u_N))$. Therefore, by the homogeneity of degree $\mu \geq 0$ of M, it follows that $S(0, ..., 0|\mathbf{Z}, \boldsymbol{\beta}) = 0$.

To prove the second condition for a multivariate survivor function, note that as u_i limits to infinity, the univariate marginal survivor function $\exp(-\theta_i H_0(u_i))$ must limit to zero. Then the second condition of Proposition 2 guarantees that $S(u_1, ..., u_N|\mathbf{Z}, \boldsymbol{\beta})$ limits to zero as well.

Because $dH_0(u_i)/du_i$ is non-negative, to prove the third condition for a multivariate survivor function, it is sufficient to demonstrate that the distinct partials of $\exp(-M(\theta_1, ..., \theta_N))$ with respect to the θ_i alternate in sign, with first partials being non-positive. The proof is by induction on Q. It is obviously true for $Q = 1$. Any Qth partial derivative is a sum of terms for which the ith term is of the form $\chi_i e^{-M}$, where χ_i is some product of distinct partial derivatives of M. Thus, any $(Q + 1)$th partial derivative is the sum of terms of the form $\chi_i D_j e^{-M} + e^{-M} D_j \chi_i$, where D_j is the operator for the partial derivative with respect to the jth argument. Clearly, $\chi_i D_j e^{-M}$ is of opposite sign to $\chi_i e^{-M}$ because the first partials of e^{-M} are nonpositive. On the other hand, $D_j \chi_i$ is evaluated by the chain rule, and is of opposite sign to χ_i, because taking the partial derivative of each partial of M involves a sign change by the first condition of Proposition 2. Therefore, the third condition for a multivariate survivor function is satisfied, and Proposition 2 is proved.

6 Ties in the Data

Although durations are continuous, the recording of durations will always involve some measurement error, and ties may result. This is problematic because both the partial likelihood and marginal likelihoods require the data to be completely rank-ordered. To incorporate tied data into the analysis, the same approach can be used as in the case of censoring.

Suppose that there are m_i spells ($m_i \geq 1$) at each of the L ordered observed failure times, t_i, where $\sum_{i=1}^{L} m_i = N$. Assuming the ties to result from the grouping of durations in the continuous model, the information available on the rank vector is incomplete. While it is known that the ranks of durations failing at t_i are lower than those failing at t_j whenever $i < j$, the ranks of the m_i durations failing at t_i cannot be known. The marginal likelihood in this case should specify the likelihood that the rank vector is one of those possible.

In their discussion of the case of independent durations, for which $M(\theta_1, ..., \theta_N) = \sum_{i=1}^{N} \theta_i$, Kalbfleisch and Prentice (1980) point out that the calculation can be simplified somewhat by recognizing that the ranks assigned to the m_i durations failing at t_i do not depend on the ranks assigned to the m_j durations failing at t_j. The sum then becomes the product of L weighted sums. Let Ξ_i be the set of permutations of the labels of the m_i durations failing at t_i and let $\boldsymbol{\xi} = (\xi_1, ..., \xi_{m_i})$ be an element of Ξ_i. As before, $R(t_i)$ is the risk set at time t_i. Define $R(t_i, \xi^r)$ to be the set difference $R(t_i) - \{\xi_1, ..., \xi_{r-1}\}$ and $D(t_i) = R(t_i) - R(t_i+)$ to be the set of durations failing at t_i.

Then, the marginal likelihood for $\boldsymbol{\beta}$ can be expressed as

$$\prod_{i=1}^{L} \left(\prod_{j \in D(t_i)} \theta_j \sum_{\boldsymbol{\xi} \in \Xi_i} \left(\prod_{r=1}^{m_i} \left[\sum_{l \in R(t_i, \xi^r)} \theta_l \right]^{-1} \right) \right) . \tag{6.1}$$

Because the summation in the marginal likelihood is over all permutations of labels of the tied durations, its computation may be burdensome if there are a large number of ties at any failure time. When the number of durations failing at each t_i is small relative to the number in the corresponding risk set $R(t_i)$, Peto (1972) and Breslow (1974) claim that (6.1) can be approximated using

$$\prod_{i=1}^{L} \left(\frac{\prod_{j \in D(t_i)} \theta_j}{\left(\sum_{l \in R(t_i)} \theta_l \right)^{m_i}} \right) .$$

Efron (1977) suggests an alternative approximation that takes into account that distinct summations $\sum_{l \in R(t_i, \xi^r)} \theta_l$ will have greater multiplicity the lower is the value of r:

$$\prod_{i=1}^{L} \left(\frac{\prod_{j \in D(t_i)} \theta_j}{\prod_{r=1}^{m_i} \left(\left(\sum_{l \in R(t_i)} \theta_l \right) - \frac{(r-1)}{m_i} \left(\sum_{l \in D(t_i)} \theta_l \right) \right)} \right) .$$

Kalbfleisch and Prentice (1980) suggest using a semi-parametric model formed by grouping failure times whenever the ratio of m_i to the size of the risk set

$R(t_i)$ is high for any failure time (see Prentice and Gloeckler 1978, and Meyer 1990).

The case in which durations are clustered is more complicated. The marginal likelihood for $\boldsymbol{\beta}$ becomes

$$\prod_{i=1}^{L} \sum_{\boldsymbol{\xi} \in \Xi_i} \left(\prod_{r=1}^{m_i} \theta_{\xi_r} M^{[\xi_r]}(\boldsymbol{\theta}, R(t_i, \xi^r)) \left[M(\boldsymbol{\theta}, R(t_i, \xi^r)) \right]^{-1} \right). \tag{6.2}$$

When the number of durations failing at each t_i is small relative to the number of spells in the corresponding risk set $R(t_i)$, (6.2) can be approximated using

$$\prod_{i=1}^{L} \left(\frac{\prod_{j \in D(t_i)} \theta_j M^{[j]}(\boldsymbol{\theta}, R(t_i))}{(M(\boldsymbol{\theta}, R(t_i)))^{m_i}} \right).$$

Finally, the following alternative approximation takes into account that distinct $M(\boldsymbol{\theta}, R(t_i, \xi^r))$ in (6.2) will have greater multiplicity the lower is the value of r:

$$\prod_{i=1}^{L} \left(\frac{\prod_{j \in D(t_i)} \theta_j M^{[j]}(\boldsymbol{\theta}, R(t_i))}{\prod_{r=1}^{m_i} \left(M(\boldsymbol{\theta}, R(t_i)) - \frac{(r-1)}{m_i} \left(\sum_{l \in D(t_i)} \theta_l M^{[l]}(\boldsymbol{\theta}, R(t_i)) \right) \right)} \right).$$

7 Recovering the Baseline

Breslow (1972) develops a methodology for recovering the duration baseline from the marginal likelihood estimates for a sample of independent durations. Breslow explains that the Kaplan-Meier estimate can be derived in a maximum likelihood framework by assuming that the hazard is constant between successive observed failure times:

$$h_0(t) = \rho_i, \quad t_{i-1} < t \leq t_i, \quad i = 1, ..., L. \tag{7.1}$$

He notes that this approach is used by Grenander (1956) to derive maximum likelihood estimates for the monotone hazard class. Breslow next adopts the convention of considering all censored durations as censored at the previous uncensored failure time. Breslow's estimator for ρ_i is the maximum likelihood estimator for the resulting likelihood (see Kalbfleisch and Prentice 1980):

$$\prod_{i=1}^{L} h_0(t_i)^{m_i} \left(\prod_{j \in D(t_i)} \theta_j \right) \exp(-\int_0^t h_0(u)du \sum_{j \in \Omega(t_i)} \theta_j) \quad ,$$

where $\Omega(t_i)$ is the set of durations either failing or censored at t_i. Substituting in from equation (7.1) and rearranging terms gives

$$\prod_{i=1}^{L} \rho_i^{m_i} \left(\prod_{j \in D(t_i)} \theta_j \right) \exp(-\rho_i(t_i - t_{i-1}) \sum_{j \in R(t_i)} \theta_j) \quad .$$

Since $\theta_j = \exp(\mathbf{Z}_j \boldsymbol{\beta})$, the maximum likelihood estimator of ρ_i for any value of $\boldsymbol{\beta}$ is therefore

$$\widehat{\rho}_i = \frac{m_i}{(t_i - t_{i-1}) \sum_{j \in R(t_i)} \exp(\mathbf{Z}_j \boldsymbol{\beta})} \quad ,$$

and the estimate of the cumulative baseline hazard $H_0(t) = \int_0^t h_0(u)du$, evaluated at t_i is

$$\widehat{H}_0(t_i) = \sum_{l=1}^{i} \frac{m_l}{\sum_{j \in R(t_l)} \exp(\mathbf{Z}_j \boldsymbol{\beta})} \quad .$$

The estimators of ρ_i and $H_0(t)$ can both be evaluated at the value of $\boldsymbol{\beta}$ that maximizes the marginal likelihood.

When durations are clustered, the likelihood becomes

$$\prod_{i=1}^{L} \rho_i^{m_i} \left(\prod_{j \in D(t_i)} \theta_j M^{[j]}(\boldsymbol{\theta}, R(t_i)) \right) \exp(-\rho_i(t_i - t_{i-1}) M(\boldsymbol{\theta}, R(t_i))) \quad .$$

The maximum likelihood estimator of ρ_i for any value of $\boldsymbol{\beta}$ is

$$\widehat{\rho}_i = \frac{m_i}{(t_i - t_{i-1}) M(\boldsymbol{\theta}, R(t_i))} \quad ,$$

and the estimate of the cumulative baseline hazard evaluated at t_i is

$$\widehat{H}_0(t_i) = \sum_{j=1}^{i} \frac{m_j}{M(\boldsymbol{\theta}, R(t_j))} \quad .$$

8 Conclusions

Cox (1972, 1975) develops the proportional hazard model of durations and suggests semiparametric estimation that does not specify a duration baseline using a partial likelihood approach. Contributions to the partial likelihood are provided at each failure time by the subset of the sample at risk immediately before the failure time. For researchers interested in duration dependence, the duration baseline can be recovered in a second step. The case for partial likelihood was strengthened with the later finding by Tsiatis (1981) that partial likelihood estimation is equivalent to marginal likelihood estimation.

This paper examines the problem of estimating model parameters in a clustered proportional hazard model, leaving the baseline hazard unspecified. It turns out that the problem is isomorphic to the problem of relaxing the assumption of independent errors in an extreme-value stochastic utility model. McFadden characterizes all discrete choice probability models with univariate extreme-value disturbances that are consistent with stochastic utility maximization. He presents sufficient conditions for the joint distribution to be consistent with stochastic utility maximization. The sufficient conditions describe the set of GEV models. This paper characterizes all marginal likelihoods for clustered proportional hazards in which the the duration baseline is eliminated from the estimation. It shows further that a sufficient condition for the elimination of the duration baseline is that the probability that the first rank-ordered failure is the first observed rank-ordered failure can be modeled as a GEV probability.

The marginal likelihoods allow independent censoring and I discuss approximations to the marginal likelihoods in the presence of ties.

An appendix on asymptotic inference (Ondrich 2010) can be found at http://faculty.maxwell.syr.edu/jondric This appendix makes three contributions. First, the theory of multiplicative intensity models supports the incorporation of time-varying covariates. Second, consistency and asymptotic normality of the model parameters follow directly from the previous work of Andersen and Gill (1982) for the partial likelihood with independent observations. With independent observations the marginal likelihood is globally concave, which is not the case with clustered observations. However, the results carry over to the case of clustered observations if one considers an open ball containing the true value of the relevant parameter vector, over which ball the marginal likelihood is strictly concave. Third, an asymptotically correct variance matrix for the marginal likelihood estimator of the vector $\boldsymbol{\gamma}$ of model parameters (excluding duration-baseline parameters) is

$\mathbf{I}^{-1}(\widehat{\gamma})\mathbf{O}(\widehat{\gamma})\mathbf{I}^{-1}(\widehat{\gamma})$, where $\mathbf{O}(\widehat{\gamma})$ is a weighted outer product of scores and $\mathbf{I}(\widehat{\gamma})$ is the empirical information matrix.

References

[1] Andersen, Elisabeth Wreford. 2004. "Composite Likelihood and Two-Stage Estimation in Family Studies." Biostatistics 5 (1): 15-30.

[2] Andersen, Per Kragh, Ørnulf Borgan, Richard D. Gill, and Niels Keiding. 1993. Statistical Models Based on Counting Processes. Springer Series in Statistics New York: Springer-Verlag.

[3] Andersen, Per Kragh, and Richard D. Gill. 1982. "Cox's Regression Model for Counting Processes: A Large Sample Study." Annals of Statistics 10 (4) (December): 1100-1120.

[4] Arjas, Elja, and Pentti Haara. 1984. "A Marked Point Process Approach to Censored Failure Data with Complicated Covariates." Scandinavian Journal of Statistics 11 (4): 193-209.

[5] Bernstein, Sergei. 1928. "Sur les fonctions absolument monotones." Acta Mathematica 51: 1-66.

[6] Börsch-Supan, Axel. 1987. Econometric Analysis of Discrete Choice: With Applications on the Demand for Housing in the U.S. and West Germany. Lecture Notes in Economics and Mathematical Systems No. 296. Berlin: Springer-Verlag.

[7] Breslow, N. 1974. "Covariance Analysis of Censored Survival Data." Biometrics 30 (1) (March): 89-99.

[8] Cardell, N. Scott. 1997. "Variance Components Structures for the Extreme-Value and Logistic Distributions with Application to Models of Heterogeneity." Econometric Theory 13 (2): 185-213.

[9] Chamberlain, Gary. 1985. "Heterogeneity, Omitted Variable Bias, and Duration Dependence." Chapter 1 In Longitudinal Analysis of Labor Market Data, edited by James J. Heckman and Burton Singer. Cambridge: Cambridge University Press, 3-38.

[10] Cox, D. R. 1972. "Regression Models and Life-Tables." Journal of the Royal Statistical Society, Series B 34 (2): 187-220.

[11] Cox, D. R. 1975. "Partial Likelihood." Biometrika 62 (2) (August): 269-276.

[12] Daly, Andrew J., and Stanley Zachary. 1978. "Improved Multiple Choice Models." Chapter 10 In Determinants of Travel Choice, edited by David A. Hensher and Quasim Dalvi. Farnborough, England: Saxon House, Teakfield Limited, 335-357.

[13] Efron, Bradley. 1977. "The Efficiency of Cox's Likelihood Function for Censored Data." Journal of the American Statistical Association 72 (359) (September): 557-565.

[14] Feller, William. 1971. An Introduction to Probability Theory and Its Applications. Second Edition. New York: Wiley.

[15] Fleming, Thomas R., and David P. Harrington. 1991. Counting Processes and Survival Analysis. New York: John Wiley & Sons.

[16] Friedman, Avner. 1971. Advanced Calculus. New York: Holt, Rinehart and Winston.

[17] Grenander, Ulf. 1956. "On the Theory of Mortality Measurement, I and II." Skandinavisk Aktuarietidskrift 39: 70-96; 125-153.

[18] Gross, Shulamith T., and Catherine Huber. 1987. "Matched Pair Experiments: Cox and Maximum Likelihood Estimation." Scandinavian Journal of Statistics 14: 27-41.

[19] Hausman, Jerry, and Daniel. McFadden. 1984. "Specification Tests for the Multinomial Logit Model." Econometrica 52 (5) (September): 1219-1240.

[20] Hougaard, Philip. 1986a. "Survival Models for Heterogeneous Populations Derived from Stable Distributions." Biometrika 73 (2): 387-396.

[21] Hougaard, Philip. 1986b. "A Class of Multivariate Failure Time Distributions." Biometrika 73 (3): 671-678.

[22] Kalbfleisch, J. D., and Ross L. Prentice. 1980. The Statistical Analysis of Failure Time Data.Wiley Series in Probability and Mathematical Statistics New York: Wiley.

[23] Khan, Shakeeb, Youngki Shin, and Elie Tamer. 2010. "Heteroskedastic Transformation Models with Covariate Dependent Censoring." Journal of Business Economics and Statistics, forthcoming.

[24] Khan, Shakeeb, and Elie Tamer. 2007. "Partial Rank Estimation of Duration Models with General Forms of Censoring." Journal of Econometrics 25: 251-280.

[25] Khan, Shakeeb, and Elie Tamer. 2010. "Inference on Randomly Censored Regression Models Using Conditional Moment Inequalities." Journal of Econometrics, forthcoming.

[26] McFadden, Daniel. 1974. "Conditional Logit Analysis of Qualitative Choice Behavior." Chapter 4 In Frontiers in Econometrics, edited by Paul Zarembka. New York: Academic Press.

[27] McFadden, Daniel. 1978. "Modelling the Choice of Residential Location." Paper No. 25 In Spatial Interaction Theory and Planning Models, edited by Anders Karlqvist, Lars Lundqvist, Folke Snickars, and Jörgen W. Weibull. Amsterdam: North Holland, 75-96.

[28] McFadden, Daniel. 1987. "Regression-Based Specification Tests for the Multinomial Logit Model." Journal of Econometrics 34 (1/2) (January-February): 63-82.

[29] McFadden, Daniel, Kenneth Train, and William B. Tye. 1977. "An Application of Diagnostic Tests for the Independence from Irrelevant Alternatives Property of the Multinomial Logit Model." In Transportation Research Record No. 637: Forecasting Passenger and Freight Travel Washington, DC: Transportation Research Board, National Academies of Science, 39-46.

[30] Meyer, Bruce D. 1990. "Unemployment Insurance and Unemployment Spells." Econometrica 58 (4) (July): 757-782.

[31] Ondrich, Jan. 2010. "Asymptotic Inference for Cox-McFadden Partial and Marginal Likelihoods." Working Paper, Center for Policy Research, Syracuse University.

[32] Peto, Richard. 1972. "Contribution to the discussion of 'Regression Models and Life-Tables" by D. R. Cox." Journal of the Royal Statistical Society, Series B 34 (2): 205-207.

[33] Prentice, Ross L., and Lynn A. Gloeckler. 1978. "Regression Analysis of Grouped Survival Data with Application to Breast Cancer Data." Biometrics 34 (1) (March): 57-67.

[34] Ridder, Geert, and Insan Tunali. 1999. "Stratified Partial Likelihood Estimation." Journal of Econometrics 92: 193-232.

[35] Sastry, Narayan. 1997. "A Nested Frailty Model for Survival Data, With an Application to the Study of Child Survival in Northeast Brazil." Journal of the American Statistical Association 92 (438) (June): 426-435.

[36] Small, Kenneth A., and Cheng Hsiao. 1985. "Multinomial Logit Specification Tests." International Economic Review 26 (3) (October): 619-627.

[37] Tsiatis, A. A. 1981. "A Large Sample Study of Cox's Regression Model." Annals of Statistics 9: 93-108.

[38] Williams, Huw. 1977. "On the Formation of Travel Demand Models and Economic Evaluation Measures of User Benefit." Environment and Planning 9A (3): 285-344.

[39] Wong, Wing Hung. 1986. "Theory of Partial Likelihood." Annals of Statistics 14 (1) (March): 88-123.