Volume 8, No. 3 (1999), pp. 407–425 Allerton Press, Inc.

MATHEMATICAL METHODS OF STATISTICS

CURRENT STATUS REGRESSION

S. A. MURPHY[1], A. W. VAN DER VAART[2], AND J. A. WELLNER[3]

[1]University of Michigan
1440 Mason Hall, Ann Arbor, MI 48109-1027, USA

[2]Faculty of Sciences, Free University Amsterdam
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

[2]University of Washington, Statistics
Box 354322 Seattle, WA 98195-4322, USA

We consider the asymptotic properties of maximum likelihood and penalized maximum likelihood estimators of the slope parameter in the linear regression model under current status type censoring. For the penalized maximum likelihood estimator we present the full result, while for the ordinary maximum likelihood estimator we give a partial treatment only.

Key words: Survival analysis, profile likelihood, semiparametric model, penalized maximum likelihood.

AMS 1991 Subject Classification: 62G15, 62G20, 62F25.

1. Introduction

In the simple linear regression model the dependent variable Y is a linear transformation $Y = \theta Z + \varepsilon$ for an observed independent variable Z and an unobserved error ε. Under the "current status censoring" mechanism, we do not observe Y directly, but only the variable $X = (\mathbf{1}\{Y \leq C\}, Z, C) = (\Delta, Z, C)$ for a censoring variable C. We assume that (Z, C) is independent of ε. Then the distribution of X is determined by the slope parameter θ, the distribution function F of ε, and the distribution of (Z, C). The distribution function F is considered completely unknown (and hence can contain an intercept), except for possibly qualitative smoothness assumptions.

[1]Research partially supported by NSF grant DMS-9307255.

The likelihood for the pair (θ, F) can be defined as the conditional density of X given (Z, C), which is equal to

$$p_{\theta,F}(x) = F(c - \theta z)^{\delta}\big(1 - F(c - \theta z)\big)^{1-\delta}.$$

In this paper we consider estimation of θ based on a random sample $X_1, \ldots, X_n$ from the distribution of X. More precisely, we are interested in likelihood based procedures. The distribution of (Z, C) is assumed not to depend on (θ, F), and hence does not appear in the likelihood. It is considered fixed throughout the paper, but need not be known. We assume that the distribution of (Z, C) is continuous and has compact support.

2. History and Related Work

The above model has a long history in the econometrics literature, where it is known as the binary response model. See, e.g., Manski [20, 21], Cosslett [5, 6], Horowitz [11, 12], and Klein and Spady [19]. Then C is not interpreted as a censoring variable, but as another covariate besides Z. The coefficient of this covariate is set to unity in order to make the parameter identifiable. One can assume that ε and Z are independent, as in the present paper, or that the distribution of (ε, Z) ranges over a bigger class. When (Z, ε) is allowed to have an arbitrary distribution, then the efficient information for θ is zero; see, e.g., Chamberlain [4] or Pollard [23]. This implies that there is no hope of constructing a $n^{1/2}$-consistent estimator in this case.

Manski [20, 21] proposed and studied the "maximum score estimator." Kim and Pollard [18] obtained the limiting distribution of the "maximum score estimator" under some appropriate conditions. Interestingly and somewhat disappointingly, the convergence rate is $n^{1/3}$ instead of the usual $n^{1/2}$ rate. Horowitz [11, 12] has studied "smoothed maximum score estimators" that obtain (optimal) rates of convergence $n^{-k/(2k+1)}$ in larger models in which ε and Z are not assumed independent. Cosslett [6] calculated the efficient information for θ and showed that this is positive in the model where the covariate Z and the random error ϵ are independent, as we assume in this paper. Efficient estimators (with a $\sqrt{n}$-rate) are constructed by Klein and Spady [19], who allow some dependence of ε and Z through a parametric heteroscedasticity factor. A special case of their method is to replace $F(u) = \mathrm{E}(\Delta | C - \theta Z = u)$ in the likelihood by a kernel smoother (using a known value of θ) and maximize the resulting function over θ.

Concerning the maximum likelihood estimator (the estimator $(\tilde{\theta}, \tilde{F})$) nothing more seems to be known than its consistency, which was proved by Cosslett [5]. The purpose of this paper is to strengthen this to a rate result and to study a related, penalized likelihood estimator. The rate $n^{-1/3}$ derived for the maximum likelihood estimator in this paper is the best result obtainable by our methods, but in many other situations these would yield suboptimal results. We do not exclude the possibility that in the present case this rate is actually sharp. More precise results are under study, but appear not easy to obtain.

Han [10], in studying a more general regression model with the binary choice model as a special case, proposed the maximum rank correlation estimator. Sherman [28] proved that under the assumption of Z and ϵ independent, Han's estimator is $n^{1/2}$-consistent and satisfies a central limit theorem, but is not efficient.

Huang [14] studied the maximum score estimator in the context of the current status regression model when Z and ϵ are not independent.

Other regression models with current status or interval-censored data have been studied recently by several authors. Huang [15, 16] studies maximum likelihood estimators for the proportional odds and Cox proportional hazard models with current status data. Rossini and Tsiatis [25] consider efficient estimators based on smoothing methods in the case of the proportional odds regression model. For regression models with ("case 2") interval-censored data, see Huang, Rossini, and Wellner [17], Rabinowitz, Tsiatis, and Aragon [24], and Satten [26].

3. Estimators

The first idea is to estimate (θ, F) by the maximizer $(\widetilde{\theta}, \widetilde{F})$ of the log likelihood

$$L_n(\theta, F) = \frac{1}{n}\sum_{i=1}^{n}\Big(\delta_i \log F(c_i - \theta z_i) + (1-\delta_i)\log\big(1 - F(c_i - \theta z_i)\big)\Big).$$

In Theorem 3.4 we show that, under some regularity conditions, this estimator is consistent (and give a rate). We do not present results on the asymptotic distribution of $\widetilde{\theta}$, which remains an open problem.

It is possible that the maximum likelihood estimator is suboptimal, perhaps even asymptotically, particularly if F is *a priori* thought to be smooth. The roughness of the profile likelihood function

$$\theta \mapsto \sup_F L_n(\theta, F)$$

can be taken as a suggestion that the asymptotic behaviour of the maximum likelihood estimator may be nonstandard. This suggestion is somewhat supported by the fact that the behaviour of the maximum likelihood estimator for known error distributions F is different in the cases that F is smooth (when the model is a standard smooth parametric model) or discrete (when the model is more like sampling from a uniform distribution).

The roughness of the profile likelihood function is caused by the fact that the maximum likelihood estimator $\widetilde{F}$ for F is "not smooth". More precisely, we note the following properties of its support. From the form of the likelihood it is clear that $\widetilde{F}$ is not unique, since only the values $\widetilde{F}(c_i - \widetilde{\theta} z_i)$ matter. It is convenient to take the maximum likelihood estimator discrete with support points in the set of values $c_i - \widetilde{\theta} z_i$, augmented with the value ∞ if necessary. Then the point masses at the points $c_i - \widetilde{\theta} z_i$ are essentially uniquely determined by the likelihood. Alternatively, the point masses can be smoothed out over the intervals between the values $c_i - \widetilde{\theta} z_i$ in many ways, so as to change $\widetilde{F}$ into a continuous distribution, without changing the value of the likelihood. However, even though the maximum likelihood estimator can be taken infinitely often differentiable, the freedom in smoothing $\widetilde{F}$ is severely limited. In line with results by Groeneboom [7, 8] (see Groeneboom and Wellner [9]) for current status data without regression, it may be expected that the point masses of the discrete version of $\widetilde{F}$ are actually zero at most of the points $c_i - \widetilde{\theta} z_i$. Then any smoothed version of $\widetilde{F}$ will necessarily have a density that vanishes on many intervals between the values $c_i - \widetilde{\theta} z_i$ and is high on other intervals.

We can control for the possible roughness of the estimator of F by adding a penalty term to the likelihood. Our main results concern estimators $(\widehat{\theta}, \widehat{F})$ that maximize the *penalized likelihood*

$$L_n(\theta, F) - \widehat{\lambda}_n^2 J^2(F).$$

Here the penalty $J(F)$ is defined as

$$J^2(F) = \int_D F''(u)^2 \, du,$$

where the domain D of the integral is taken to be a finite interval that contains the support of $C - \theta Z$ for every θ. The size of the smoothing parameter $\widehat{\lambda}_n$ determines the importance of the penalty. This parameter may be data-dependent, but should satisfy

$$\widehat{\lambda}_n^2 = o_P\Big(\frac{1}{n^{1/2}}\Big), \qquad \frac{1}{\widehat{\lambda}_n} = O_P(n^{2/5}). \tag{3.1}$$

In view of the preceding discussion, the role of the penalty is not so much to force $\widehat{F}$ to be (twice) differentiable, but rather to upper bound the weights that $\widehat{F}$ can allocate to the intervals between the values $c_i - \widehat{\theta} z_i$. Presently, reversing the discussion in the preceding paragraph, the (at least) twice differentiable penalized likelihood estimator $\widehat{F}$ could be discretized by moving its mass into the points $c_i - \widehat{\theta} z_i$ without changing the value of the likelihood. Since the penalty forces the masses of $\widehat{F}$ to be (more) evenly distributed over the intervals between the points $c_i - \widehat{\theta} z_i$, the resulting point masses are bounded above. Thus, inserting a penalty in the likelihood can also be viewed as a device to allocate the total mass 1 to the points $c_i - \widehat{\theta} z_i$ more evenly than is done by the unpenalized estimator $\widetilde{F}$.

Condition (3.1) leaves some freedom in choosing $\widehat{\lambda}_n$. Any choice satisfying (3.1) will result in an asymptotically efficient estimator $\widehat{\theta}$. The best convergence rate for the estimator $\widehat{F}$ is obtained by choosing $\widehat{\lambda}_n$ exactly of the order $n^{-2/5}$, but this may not be optimal (in terms of higher order properties) for estimating θ. Of course, other penalty terms could be used as well, for instance the L_2-norm of a higher derivative. Using the second derivative appears to yield the minimal smoothness of the estimators $\widehat{F}$ needed to make our arguments go through. (Note that the efficient score function given in (6.3) involves $F' = f$.)

Throughout the paper we assume that the distribution of (C, Z) is smoothly supported on a compact set. More precisely, the following assumptions are made throughout the paper, without further reference.

The support of $C - \theta Z$ is strictly contained in the interval D for every $\theta \in \Theta$. The support of $C - \theta_0 Z$ is the closure of its interior. The support of (C, Z) contains an interior point. The variables $C - \theta Z$ and (C, Z) have densities that are uniformly bounded (also in θ).

Furthermore, we assume throughout that Θ is compact and that θ_0 is an interior point of Θ.

We prove the following theorems. The main result is the asymptotic efficiency of the penalized maximum likelihood estimator $\widehat{\theta}$. Let $P_{\theta,F}$ be the distribution of (Δ, C, Z) under (θ, F), and abbreviate P_{θ_0,F_0} to P_0.

Theorem 3.1. *Both $(\widehat{\theta}, \widehat{F})$ and $(\widetilde{\theta}, \widetilde{F})$ exist (but are not unique).*

Theorem 3.2. *Suppose that the conditions listed previously hold, that the functions $u \mapsto \mathrm{E}(Z \mid C - \theta Z = u)$ can be chosen three times continuously differentiable and form a P_0-Donsker class when θ ranges over Θ, and that F_0 is three times continuously differentiable on the interval D with nonzero derivative. Then $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ is asymptotically normal with mean zero and variance the inverse of the efficient information $\widetilde{I}_{\theta_0,F_0} = P_{\theta_0,F_0}\widetilde{\ell}^2_{\theta_0,F_0}$, for $\widetilde{\ell}_{\theta,F}$ given by* (6.3).

Our third theorem is used in the proof of the preceding theorem, but is also of independent interest. Let $\|\cdot\|_2$ denote the L_2-norm under the product of counting measure on $\{0,1\}$ and $P^{C,Z}$, the natural dominating measure for the densities $p_{\theta,F}$. In particular, with $P_0 f$ denoting the mean of $f = f(X)$ under the law of $X = (\Delta, Z, C)$ under $(\theta_0, \mathcal{F}_0)$,

$$\|p_{\theta,F} - p_{\theta_0,F_0}\|_2^2 = 2\int \left|F(c - \theta z) - F_0(c - \theta_0 z)\right|^2 dP^{C,Z}(c,z) = 2P_0(p_{\theta,F} - p_{\theta_0,F_0})^2.$$

Theorem 3.3. *Under the conditions of the preceding theorem:*

(i) *$J(\widehat{F}) = O_P(1)$ and $\|p_{\widehat{\theta},\widehat{F}} - p_{\theta_0,F_0}\|_2 = O_P(\widehat{\lambda}_n)$;*

(ii) *if the conditional distribution of Z given $C - \theta_0 Z$ is nondegenerate, then this implies that both $|\widehat{\theta} - \theta_0|$ and $\|\widehat{F}(c - \theta_0 z) - F(c - \theta_0 z)\|_2$ are $O_P(\widehat{\lambda}_n)$.*

The last theorem gives an upper bound on the rate of convergence of the (unpenalized) maximum likelihood estimator, also in the L_2-distance.

Theorem 3.4. *Suppose that the conditions listed previously hold and that the map $\theta \mapsto \mathrm{E}(Z \mid C - \theta Z)$ is continuous in square mean at θ_0. Then*

(i) *$\|p_{\widetilde{\theta},\widetilde{F}} - p_{\theta_0,F_0}\|_2 = O_P(n^{-1/3})$;*

(ii) *both $|\widehat{\theta} - \theta_0|$ and $\int_E (\widehat{F} - F)^2(x)\,dx$ are $O_P(n^{-1/3})$ for any interval $E \subset D$ on which $f^{C-\theta Z}$ is bounded away from zero, uniformly for θ in a neighbourhood of θ_0.*

The remainder of the paper consists of proof of the preceding theorems, the proof of each theorem being in a different section, in the order Theorem 3.1, 3.3, 3.2, and 3.4. The last three sections start with an outline of the proof.

Our proofs use entropy methods and maximal inequalities for the empirical process, such as described in Van der Vaart and Wellner [29]. The role of the penalty $J(F)$ is made clear by the following technical lemmas. The first is due to Birman and Solomjak [3] and extends the entropy bound for Hölder classes by Kolmogorov. Given a class $\mathcal{F}$ of functions $f : \mathcal{X} \mapsto \mathbb{R}$ defined on some set $\mathcal{X}$ let $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$ be the minimal number of brackets of size ε relative to the norm needed to cover $\mathcal{F}$. Here given two functions $l, u \colon \mathcal{X} \mapsto \mathbb{R}$ a bracket $[l, u]$ consists of all functions f such that $l \le f \le u$, and its size is the norm $\|u - l\|$.

Lemma 3.5. *Let $\mathcal{F}$ be a class of functions $f: \mathcal{D} \mapsto \mathbb{R}$ on an interval $D \subset \mathbb{R}$ such that $\|f\|_\infty \le M$ and such that the $(k-1)$th derivative is absolutely continuous with $\int f^{(k)}(x)^2\,dx \le M$, for some constant M. Then there exists a constant C such that*

$$\log N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) \le C\Big(\frac{M}{\varepsilon}\Big)^{1/k}, \qquad 0 < \varepsilon \le M.$$

Lemma 3.6. *Let F be a distribution function with $J(F) < \infty$. Then*

(i) $\big|F'(s) - F'(s_0)\big| \le J(F)|s - s_0|^{1/2}$ *for every* $s, s_0 \in D$;

(ii) $\sup_{s \in D}\big|F'(s)\big| \le 1 + J(F)$.

Proof. By the Cauchy–Schwarz inequality $\big|F'(s) - F'(s_0)\big| = \big|\int_{s_0}^{s} F''(s)\,ds\big| \le J(F)|s - s_0|^{1/2}$ for every $s, s_0 \in D$. Integrating this with respect to s we see that $\big|F(s) - F(s_0) - F'(s_0)(s - s_0)\big| \le J(F)|D|^{3/2}$. Since F takes on values in the unit interval only, we conclude that $|F'(s_0)|$ and hence $\|F'\|_\infty$ is bounded by a multiple of $1 + J(F)$. □

We use the following notation. The empirical distribution and empirical process are denoted by $\mathbb{P}_n = \sum_{i=1}^n \delta_{X_i}$ and $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P_0)$, respectively.

The notations $\gtrsim$ and $\lesssim$ mean greater than, or smaller than, up to a constant that may depend on the true parameter of the model, but not on any other parameter values.

4. Proof of Theorem 3.1

We shall prove the existence of $(\widehat{\theta}, \widehat{F})$ only, the existence (and consistency) of the other estimator following from Cosslett [5].

Write $L_n(\theta, F)$ for the log likelihood (unpenalized). For a given θ and a given vector $p \in \mathbb{R}^n$, let $\mathcal{F}_{\theta,p}$ be the set of all functions obtained by first ordering the points $c_i - \theta z_i$, yielding points $t_1 \le t_2 \le \cdots \le t_n$, and next requiring that $J(F) < \infty$, that $F(t_i) = p_i$ for every i, and that F is monotone on each of the intervals $[t_i, t_{i+1}]$. For $0 \le p_1 \le p_2 \le \cdots \le p_n \le 1$ this is the set of distribution functions with $J(F) < \infty$ and $F(t_i) = p_i$ for every i. Then the supremum of the likelihood over all (θ, F) is equal to

$$\sup_\theta \sup_p \sup_{F \in \mathcal{F}_{\theta,p}} \Big(L_n(\theta, F) - \widehat{\lambda}^2 J^2(F)\Big).$$

Here the second supremum is over all vectors p with $0 \le p_1 \le p_2 \le \cdots \le p_n \le 1$.

The inner supremum is taken for some function $F_{\theta,p} \in \mathcal{F}_{\theta,p}$. To see this, note first that $L_n(\theta, F)$ is constant on $\mathcal{F}_{\theta,p}$, and suppose that $F_m \in \mathcal{F}_{\theta,p}$ is a sequence with

$$J^2(F_m) \to G(p, \theta) := \inf_{F \in \mathcal{F}_{\theta,p}} J^2(F).$$

By the parallelogram law applied to the Hilbert space norm $J(F)$,

$$J^2(F_m - F_n) + J^2(F_m + F_n) = 2J^2(F_m) + 2J^2(F_n).$$

Since $\frac{1}{2}(F_m + F_n) \in \mathcal{F}_{\theta,p}$, we have $J^2\big(\frac{1}{2}(F_m + F_n)\big) \ge G(\theta, p)$. Combined with the preceding display, this shows that $J(F_m - F_n) \to 0$. Thus F_m' is a Cauchy sequence

in $L_2(D)$ and hence has a converging subsequence. By Lemma 3.6 (ii) (which uses the fact that F is a distribution function only to infer that F is bounded at two points) and the Ascoli–Arzela theorem, the sequence F_m also has a subsequence that converges uniformly to a function $F_{\theta,p}$. Conclude that $F_{\theta,p} \in \mathcal{F}_{\theta,p}$ and $J^2(F_{\theta,p}) = G(\theta,p)$.

The function $p \mapsto J(F_{\theta,p})$ is convex. Indeed, since $\frac{1}{2}(F_{\theta,p_1} + F_{\theta,p_2}) \in \mathcal{F}_{\theta,(p_1+p_2)/2}$ and the semi-norm $F \mapsto J(F)$ is convex,

$$J\big(F_{\theta,(p_1+p_2)/2}\big) \le J\big(\tfrac{1}{2}(F_{\theta,p_1} + F_{\theta,p_2})\big) \le \tfrac{1}{2}J(F_{\theta,p_1}) + \tfrac{1}{2}J(F_{\theta,p_2}).$$

We conclude that $p \mapsto J(F_{\theta,p})$ is continuous on $\mathbb{R}^n$ and in particular on the compact set $0 \le p_1 \le \cdots \le p_n \le 1$. Therefore, the function $p \mapsto L_n(\theta, F_{p,\theta}) - \widehat{\lambda}^2 J^2(p)$ is continuous as well and hence attains its maximum. It follows that the supremum on the right side of

$$g(\theta) := \sup_F \Big(L_n(\theta,F) - \widehat{\lambda}^2 J^2(F)\Big)$$

is taken for some F_θ. Since $L_n(\theta,F) \le 0$ and $\inf_\theta g(\theta) > -\infty$ there must exist a finite constant M such that

$$g(\theta) = \sup_{F:\ J(F)\le M} \Big(L_n(\theta,F) - \widehat{\lambda}^2 J^2(F)\Big).$$

The functions $\theta \mapsto L_n(\theta,F)$ are equicontinuous when F satisfies $J(F) \le M$, since

$$\big|F(c-\theta_1 z) - F(c-\theta_2 z)\big| \le \|F'\|_\infty |\theta_1-\theta_2||z| \lesssim \big(1 + J(F)\big)|\theta_1-\theta_2|,$$

in view of Lemma 3.6 (ii). It follows that the functions $\theta \mapsto g(\theta)$ are continuous and hence attain their maximum at some point $\widehat{\theta}$. Now $(\widehat{\theta}, F_{\widehat{\theta}})$ maximizes the likelihood.

5. Proof of Theorem 3.3

5.1. Proof of Theorem 3.3 (ii). We shall use the following proposition, due to Murphy and Van der Vaart [22].

The proposition concerns minimum contrast estimators in a general setting. The purpose is to estimate a "parameter" η using a criterion function $\eta \mapsto \mathbb{P}_n m_{\eta,\widehat{\lambda}_n}$, for $\mathbb{P}_n$ the empirical distribution of a random sample of size n in a measurable space $(\mathcal{X},\mathcal{A})$ and $m_{\eta,\lambda} : \mathcal{X} \mapsto \mathbb{R}$ given measurable functions. The random variables λ_n are defined on the same probability space as the random sample and assumed to take their values in a set $\Lambda_n \subset \mathbb{R}$. They play the role of a random "smoothing parameter". For a given n we consider estimators $\widehat{\eta}_n$ that are restricted to a given set H_n and satisfy

$$\mathbb{P}_n m_{\widehat{\eta}_n,\widehat{\lambda}_n} \ge \mathbb{P}_n m_{\eta_0,\widehat{\lambda}_n}.$$

This is valid, for example, for $\widehat{\eta}_n$ equal to the maximizer of the function $\eta \mapsto \mathbb{P}_n m_{\eta,\widehat{\lambda}_n}$ over H_n, if this set contains η_0.

Assume that the following conditions are satisfied for every $\lambda \in \Lambda_n$, every $\eta \in H_n$, and every $\delta > 0$

$$P_0 m_{\eta,\lambda} - P_0 m_{\eta_0,\lambda} \lesssim -d_\lambda^2(\eta,\eta_0) + \lambda^2, \tag{5.1}$$

$$\mathrm{E}^* \sup_{\substack{d_\lambda(\eta,\eta_0)<\delta \\ \lambda\in\Lambda_n,\eta\in H_n}} \big|\mathbb{G}_n(m_{\eta,\lambda} - m_{\eta_0,\lambda})\big| \lesssim \phi_n(\delta). \tag{5.2}$$

Here $d_\lambda^2(\eta, \eta_0)$ may be thought of as the square of a distance, but the following theorem is true for arbitrary functions $\eta \mapsto d_\lambda^2(\eta, \eta_0)$. (Contrary to what the notation suggests, this function may even take on negative values. In the latter case, set $d_\lambda(\eta, \eta_0) = (d_\lambda^2(\eta, \eta_0) \vee 0)^{1/2}$.) The function ϕ_n may be arbitrary, except for the condition in the following theorem.

Proposition 5.1 *Suppose that* (5.1)–(5.2) *are valid for functions* ϕ_n *such that* $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ *is decreasing for some* $\alpha < 2$ *and sets* $\Lambda_n \times H_n$ *such that* $P(\widehat{\lambda} \in \Lambda_n, \widehat{\eta} \in H_n) \to 1$. *Then* $d_{\widehat{\lambda}}(\widehat{\eta}, \eta_0) \le O_P^*(\delta_n + \widehat{\lambda})$ *for any sequence of positive numbers* δ_n *such that* $\phi_n(\delta_n) \le \sqrt{n}\delta_n^2$ *for every* n.

Proof. For each $n \in \mathbb{N}$, $j \in \mathbb{Z}$, and $M > 0$ define the set

$$S_{n,j,M} = \left\{(\lambda, \eta) \in \Lambda_n \times H_n : 2^{j-1}\delta_n < d_\lambda(\eta, \eta_0) \le 2^j \delta_n, \lambda \le 2^{-M} d_\lambda(\eta, \eta_0)\right\}.$$

Then the intersection of the events $\widehat{\lambda} \in \Lambda_n$, $\widehat{\eta}_{\widehat{\lambda}} \in H_n$, and $d_{\widehat{\lambda}}(\widehat{\eta}_{\widehat{\lambda}}, \eta_0) \ge 2^M(\delta_n + \widehat{\lambda})$ is contained in the union of the events $\{(\widehat{\lambda}, \widehat{\eta}_{\widehat{\lambda}}) \in S_{n,j,M}\}$ over $j \ge M$. By the definition of $\widehat{\eta}_{\widehat{\lambda}}$, the variable $\sup_{(\lambda,\eta)\in S_{n,j,M}} \mathbb{P}_n(m_{\eta,\lambda} - m_{\eta_0,\lambda})$ is nonnegative on the event $\{(\widehat{\lambda}, \widehat{\eta}_{\widehat{\lambda}}) \in S_{n,j,M}\}$. Conclude that, for every $\delta > 0$,

$$\begin{aligned} \mathrm{P}^*\left(d_{\widehat{\lambda}}(\widehat{\eta}_{\widehat{\lambda}}, \eta_0) \ge 2^M(\delta_n + \widehat{\lambda}), \widehat{\lambda} \in \Lambda_n, \widehat{\eta}_{\widehat{\lambda}} \in H_n\right) \\ \le \sum_{j \ge M} \mathrm{P}^*\Big(\sup_{(\lambda,\eta)\in S_{j,n,M}} \mathbb{P}_n(m_{\eta,\lambda} - m_{\eta_0,\lambda}) \ge 0\Big). \end{aligned}$$

For every j involved in the sum, we have, for every $(\lambda, \eta) \in S_{j,n,M}$ and every sufficiently large M,

$$\begin{aligned} P_0(m_{\eta,\lambda} - m_{\eta_0,\lambda}) &\lesssim -d_\lambda^2(\eta, \eta_0) + \lambda^2 \\ &\lesssim -(1 - 2^{-2M})\, d_\lambda^2(\eta, \eta_0) \lesssim -2^{2j-2}\delta_n^2. \end{aligned}$$

Thus, using Markov's inequality, we see that the series is bounded by

$$\begin{aligned} &\sum_{j \ge M} \mathrm{P}^*\Big(\sup_{(\lambda,\eta)\in S_{j,n,M}} \left|\mathbb{G}_n(m_{\eta,\lambda} - m_{\eta_0,\lambda})\right| \gtrsim \sqrt{n}2^{2j-2}\delta_n^2\Big) \\ &\lesssim \sum_{j \ge M} \frac{\phi_n(2^{j+1}\delta_n)}{\sqrt{n}\delta_n^2 2^{2j}} \lesssim \sum_{j \ge M} 2^{j\alpha - 2j}, \end{aligned}$$

in view of the definition of δ_n and the fact that $\phi_n(c\delta) \le c^\alpha \phi_n(\delta)$ for every $c > 1$ by the assumption on ϕ_n. The expression on the right converges to zero for every $M = M_n \to \infty$. □

For the proof of Theorem 3.3, we apply this proposition with $\eta = (\theta, F)$ and

$$m_{\eta,\lambda} = \log \frac{p_\eta + p_{\eta_0}}{2p_{\eta_0}} - \frac{1}{2}\lambda^2\big(J^2(F) - J^2(F_0)\big).$$

By the concavity of the logarithm, and the defining property of $\widehat{\eta}$,

$$\begin{aligned} \mathbb{P}_n m_{\widehat{\eta},\widehat{\lambda}} &\ge \frac{1}{2}\mathbb{P}_n \log \frac{p_{\widehat{\eta}}}{p_{\eta_0}} - \frac{1}{2}\widehat{\lambda}^2\big(J^2(\widehat{F}) - J^2(F_0)\big) \\ &\ge \frac{1}{2}\mathbb{P}_n \log \frac{p_{\eta_0}}{p_{\eta_0}} - \frac{1}{2}\widehat{\lambda}^2 0 = 0 = \mathbb{P}_n m_{\eta_0,\widehat{\lambda}}. \end{aligned}$$

Thus, the preceding proposition applies to $\widehat{\eta}$.

By the usual inequalities relating the Kullback–Leibler divergence and the Hellinger distance (see, e.g., Van der Vaart and Wellner [29], Theorem 3.4.4)

$$\begin{aligned} P_0 m_{\eta,\lambda} - P_0 m_{\eta_0,\lambda} &\lesssim -h^2(p_\eta, p_{\eta_0}) - \lambda^2\big(J^2(F) - J^2(F_0)\big) \\ &\lesssim -\|p_\eta - p_{\eta_0}\|_2^2 - \lambda^2 J^2(F) + \lambda^2, \end{aligned}$$

since p_{η_0} is bounded away from zero on D. Thus, (5.1) is satisfied for the choice

$$d_\lambda^2(\eta, \eta_0) = \|p_\eta - p_{\eta_0}\|_2^2 + \lambda^2 J^2(F).$$

To verify (5.2) for a suitable function ϕ_n, we apply a maximal inequality to the empirical process $\mathbb{G}_n(m_{\eta,\lambda} - m_{\eta_0,\lambda}) = \mathbb{G}_n m_{\eta,0}$. The functions $m_{\eta,0}(x)$ are uniformly bounded in x and η. Furthermore, since the derivative of the function $p \mapsto \log(p + p_0)$ is bounded uniformly in p_0 that are bounded away from zero, their L_2-norm satisfies

$$P_0 m_{\eta,0}^2 \lesssim \|p_\eta - p_{\eta_0}\|_2^2.$$

Since $\widehat{\lambda}^{-1} = O_P(n^{2/5})$ by (3.1) it is not a loss of generality to assume that $\widehat{\lambda} \geq \lambda_n$ for λ_n a small multiple of $n^{-2/5}$. In other words, we may restrict λ to the set $\Lambda_n = \{\lambda : \lambda \geq \lambda_n\}$. Then $d_\lambda(\eta, \eta_0) \leq \delta$ and $\lambda \in \Lambda_n$ implies that

$$\|p_\eta - p_{\eta_0}\|_2 \lesssim \delta, \qquad J(F) \lesssim \frac{\delta}{\lambda} \lesssim \frac{\delta}{\lambda_n}.$$

By Lemma 3.4.2 of Van der Vaart and Wellner [29] conclude that (5.2) is satisfied for

$$\phi_n(\delta) = J(\delta)\Big(1 + \frac{J(\delta)}{\delta^2\sqrt{n}}\Big),$$

where $J(\delta)$ is the entropy-with-bracketing integral

$$J(\delta) = \int_0^\delta \sqrt{1 + \log N_{[]}\big(\varepsilon, \{m_{\eta,0} : \theta \in \Theta, J(F) \leq \delta/\lambda_n\}, L_2(P_0)\big)}\, d\varepsilon.$$

This is bounded up to a constant by $\delta^{3/4} + \delta/\lambda_n^{1/4}$ in view of Lemma 5.2 below. We conclude by Proposition 5.1 that $d_{\widehat{\lambda}}(\widehat{\eta}, \eta_0) = O_P\big(n^{-2/5} + \widehat{\lambda}_n\big)$. This implies the assertion (i) of Theorem 3.3, in view of (3.1).

Lemma 5.2. *For every $M \geq 1$,*

$$\log N_{[]}\big(\varepsilon, \{F(c - \theta z) : \theta \in \Theta, J(F) \leq M\}, L_2(P_0)\big) \lesssim \Big(\frac{M}{\varepsilon}\Big)^{1/2}.$$

Proof. Let $F_{|D}$ be the restriction of F to D. By Lemmas 3.5 and 3.6,

$$\log N\big(\varepsilon, \{F_{|D} : J(F) \leq M\}, \|\cdot\|_\infty\big) \lesssim \Big(\frac{M}{\varepsilon}\Big)^{1/2}.$$

Now we may construct a net over the class of functions of interest, by first choosing an ε/M-net $\theta_1, \ldots, \theta_p$ over Θ (for the Euclidean distance), next choosing an ε-net $F_1, \ldots, F_q$ over the functions $F_{|D}$ (for the supremum metric), and finally forming all functions $F_i(c - \theta_j z)$. Then for every (θ, F) there exists (θ_j, F_i) such that

$$\left|F(c-\theta z) - F_i(c-\theta_j z)\right| \le \|F'\|_\infty |\theta - \theta_j| \|z\|_\infty + \|F_{|D} - F_i\|_\infty \lesssim M\frac{\varepsilon}{M} + \varepsilon \lesssim \varepsilon.$$

We need at most $2|\Theta|M/\varepsilon$ points θ_j, and at most $\exp(C(M/\varepsilon)^{1/2})$ points F_i for some C. Thus, the entropy for the uniform norm of the class of functions in the lemma is bounded by a multiple of $(M/\varepsilon)^{1/2} + \log(M/\varepsilon) + 1$. Consequently, the bracketing entropy for the L_2-norm is bounded similarly. □

5.2. Consistency. By Theorem 3.3 (i) the density estimator $p_{\widehat{\theta},\widehat{F}}$ is consistent for p_{θ_0,F_0} for the $\|\cdot\|_2$-norm. Furthermore, $J(\widehat{F}) = O_P(1)$. In this section we prove that these statements carry over into the consistency of $\widehat{\theta}$ and $\widehat{F}$ separately. For $\widehat{\theta}$ we use the Euclidean norm. Since the true likelihood evaluates the functions F only at the points $c - \theta_0 z$, we do not have control over F off the support of the variable $C - \theta_0 Z$. To assert that F is consistent, we may use the norm $\|\cdot\|_{D_0}$, for D_0 the support of $C - \theta_0 Z$, and, for a given set D,

$$\|F\|_D = \sup_{y\in D}\left|F(y)\right| + \sup_{y\in D}\left|F'(y)\right|.$$

For the consistency of the derivative $\widehat{F}'$, we assume that D_0 is the closure of its interior, which is true, for instance, if D_0 is an interval.

Lemma 5.3. *For every fixed M, the set of restrictions $F_{|D}$ of distributions functions F with $J(F) \le M$ is precompact relatively to $\|\cdot\|_D$.*

Proof. By Lemma 3.6 the class of functions F'_D is uniformly Lipschitz of order $\frac{1}{2}$, hence equicontinuous, and $\|F'_D\|$ is uniformly bounded, as soon as $J(F)$ is uniformly bounded. Applying the Ascoli–Arzela theorem, we see that every sequence of distribution functions F_m with $J(F_m) = O(1)$ has a subsequence such that both F_m and F'_m converge uniformly on D to limits. The limit of F'_m must necessarily be the derivative of the limit of F_m. □

Lemma 5.4. *If $\|p_{\theta,F} - p_{\theta_0,F_0}\|_2 = 0$ for a distribution function with $J(F) < \infty$, then $\theta = \theta_0$ and $F = F_0$ on the support of $C - \theta_0 Z$.*

Proof. The condition implies that $F(c - \theta z) = F_0(c - \theta_0 z)$ almost surely under the distribution of (C, Z). By continuity the functions must be equal on the support of (C, Z). Partially differentiating the identity with respect to z and c, we find

$$f(c-\theta z) = f_0(c-\theta_0 z), \qquad -\theta f(c-\theta z) = -\theta_0 f_0(c-\theta_0 z).$$

These identities are valid on the interior of the support of (C, Z). Since f_0 is nonzero, conclude that $\theta = \theta_0$.

Next conclude that $F = F_0$ almost surely under the distribution of $C - \theta_0 Z$ and hence $F = F_0$ on the support of $C - \theta_0 Z$. □

Lemma 5.5. $\widehat{\theta} \xrightarrow{P} \theta_0$ *and* $\|\widehat{F} - F_0\|_{D_0} \xrightarrow{P} 0$.

Proof. Suppose that $p_{\theta_m,F_m} \to p_{\theta_0,F_0}$ in $\|\cdot\|_2$ and $J(F_m) = O(1)$. By the first lemma every subsequence of (θ_m, F_m) has a further subsequence such that $\theta_m \to \theta$ and $\|F_m - F\|_D \to 0$ for some θ and F. Then $\|p_{\theta_m,F_m} - p_{\theta,F}\|_2 \to 0$ by the continuity of the map $(\theta, F) \mapsto p_{\theta,F}$. Thus, $\|p_{\theta,F} - p_{\theta_0,F_0}\|_2 = 0$ and hence $\theta = \theta_0$ and $F = F_0$ on the support of $C - \theta_0 Z$ by the second lemma. Under the assumption that D_0 is the closure of its interior, this implies that F' and F_0' agree on D_0 as well. It follows that $F_m \to F_0$ and $F_m' \to F_0'$ uniformly on D_0.

Combined with the preceding lemmas and Theorem 3.3 (i), this yields the lemma. □

5.3. Proof of Theorem 3.3 (ii). To see that the rate of convergence of $\widehat{F}(c - \widehat{\theta}z)$ in the $\|\cdot\|_2$-norm carries over into a rate for $\widehat{F}$ in the $L_2(P^{C-\theta_0 Z})$-distance, we start with proving the differentiability of $F(c - \theta z)$ in (θ, F).

Lemma 5.6.

$$P_0\Big[F(c-\theta z) - F_0(c-\theta_0 z) - \big(-zF_0'(c-\theta_0 z)(\theta-\theta_0) + (F-F_0)(c-\theta_0 z)\big)\Big]^2 \lesssim |\theta-\theta_0|^{5/2} J(F) + |\theta-\theta_0|^2 P_0(F'-F_0')^2(c-\theta_0 z).$$

Proof. The left side is equal to

$$P_0\Big[F(c-\theta z) - F(c-\theta_0 z) + zF_0'(c-\theta_0 z)(\theta-\theta_0)\Big]^2 \lesssim (\theta-\theta_0)^2 P_0\Big[z\big(F'(c-\xi z) - F'(c-\theta_0 z)\big)\Big]^2 + (\theta-\theta_0)^2 P_0 z^2 (F'-F_0')^2(c-\theta_0 z),$$

for $\xi = \xi(c, z)$ between θ and θ_0. Now $|z|$ is bounded and $\big|F'(c-\xi z) - F'(c-\theta_0 z)\big| \lesssim J(F)|z||\xi - \theta_0|^{1/2}$. The result follows. □

Since we already know that $|\widehat{\theta} - \theta_0| \xrightarrow{P} 0$, that $P_0(\widehat{F}' - F_0')^2(c - \theta_0 z) \xrightarrow{P} 0$, and that $J(\widehat{F})$ is bounded, we see that

$$P_0\Big[\widehat{F}(c-\widehat{\theta}z) - F_0(c-\theta_0 z)\Big]^2 \gtrsim P_0\Big[-zF_0'(c-\theta_0 z)(\widehat{\theta}-\theta_0) + (\widehat{F}-F_0)(c-\theta_0 z)\Big]^2 - o_P(1)|\widehat{\theta}-\theta_0|^2.$$

By the assumptions that the conditional distribution of Z given $C - \theta_0 Z$ is nondegenerate and $F_0' = f_0$ is nonzero, the expectation on the right is bounded (below) by a constant times $|\widehat{\theta} - \theta_0|^2 + P_0(\widehat{F} - F_0)^2(c - \theta_0 z)$, by the lemma below, applied with $g_1 = zF_0'(c - \theta_0 z)(\widehat{\theta} - \theta_0)$ and $g_2 = (\widehat{F} - F_0)(c - \theta_0 z)$. Indeed, by the Cauchy–Schwarz inequality, for any function g,

$$\big(P_0 zF_0'(c-\theta_0 z)g(c-\theta_z)\big)^2 = \big(\mathrm{E}_0\,\mathrm{E}_0(Z|\,C-\theta_0 Z)F_0'(C-\theta_0 Z)g(C-\theta_0 Z)\big)^2 \le \mathrm{E}_0\,\mathrm{E}_0(Z|\,C-\theta_0 Z)^2 (F_0')^2(C-\theta_0 Z)\,\mathrm{E}_0\,g^2(C-\theta_0 Z).$$

The first term on the right is strictly smaller than $\mathrm{E}_0 Z^2 f^2(C-\theta_0 Z)$ unless $Zf(C-\theta_0 Z)$ is a function of $C-\theta_0 Z$, which is excluded by our assumptions. This concludes the proof of Theorem 3.3 (ii). □

Lemma 5.7. *Let g_1 and g_2 be measurable functions such that $(Pg_1g_2)^2 \leq cPg_1^2Pg_2^2$ for a constant $c<1$. Then*

$$P(g_1+g_2)^2 \geq (1-\sqrt{c})(Pg_1^2+Pg_2^2).$$

6. Proof of Theorem 3.2

We shall use the following proposition, which is proved in Van der Vaart [30].

Suppose that the observations are an i.i.d. sample from a density $p_{\theta,\eta}$ indexed by a Euclidean parameter θ and an arbitrary parameter η. For every parameter (θ,η) let $\tilde{\ell}_{\theta,\eta}$ be an arbitrary measurable vector-valued function such that $\tilde{\ell}_{\theta_0,\eta_0}$ is the efficient score function for the parameter θ at (θ_0,η_0). We consider estimators $(\hat{\theta}_n,\hat{\eta}_n)$ such that

$$\frac{1}{n}\sum_{i=1}^{n}\tilde{\ell}_{\hat{\theta}_n,\hat{\eta}_n}(X_i) = o_P\Big(\frac{1}{n^{1/2}}\Big). \tag{6.1}$$

The following proposition yields the asymptotic normality of the sequence $\sqrt{n}(\hat{\theta}_n-\theta_0)$ under regularity conditions and the structural "no bias"-condition

$$P_{\hat{\theta}_n,\eta_0}\tilde{\ell}_{\hat{\theta}_n,\hat{\eta}_n} = o_P\Big(\frac{1}{n^{1/2}}\Big). \tag{6.2}$$

Proposition 6.1. *Suppose that the model $\theta \mapsto p_{\theta,\eta_0}$ is differentiable in quadratic mean at θ_0, that $(P_{\theta_0,\eta_0}+P_{\theta,\eta_0})\|\tilde{\ell}_{\theta,\eta}\|^2 = O(1)$, and that $(\theta,\eta)\mapsto\tilde{\ell}_{\theta,\eta}$ is continuous in P_{θ_0,η_0}-probability at (θ,η_0). Furthermore, suppose that the class of functions $\tilde{\ell}_{\theta,\eta}$ is P_{θ_0,η_0}-Donsker for (θ,η) ranging over a neighbourhood of (θ_0,η_0). If $(\hat{\theta}_n,\hat{\eta}_n)$ is consistent for (θ_0,η_0) and (6.1) and (6.2) are satisfied, then the sequence $\sqrt{n}(\hat{\theta}_n-\theta_0)$ is asymptotically normal with mean zero and covariance the inverse of $\tilde{I}_{\theta_0,\eta_0}$.*

We shall apply this theorem with $(\theta,\eta)=(\theta,F)$ and $\tilde{\ell}_{\theta,\eta}$ the efficient score function for the model, for every (θ,η). We construct suitable one-dimensional submodels in order to show that the *efficient score equation* (6.1) is satisfied. For this choice of functions $\tilde{\ell}_{\theta,F}$ the bias condition (6.2) is satisfied trivially, with the left side vanishing, as will follow from the direct calculation later in this section, but is also explained by the linearity of the model in F.

6.1. EFFICIENT SCORE FUNCTION. The ordinary score function for θ of the model is the function

$$\dot{\ell}_{\theta,F}(x) = -zf(c-\theta z)Q_{\theta,F}(x),$$

for

$$Q_{\theta,F}(x) = \frac{\delta}{F(c-\theta z)} - \frac{1-\delta}{1-F(c-\theta z)}.$$

The score function for the submodel given by $F_t = F + tB$ is equal to

$$A_{\theta,F}B(x) = B(c - \theta z)Q_{\theta,F}(x).$$

Of course, the path $F_t = F + tB$ defines a true submodel only for perturbations B such that F_t is nondecreasing and $B(-\infty) = B(\infty) = 0$. If we restrict the model by requiring that $J(F) < \infty$, as we do for Theorem 3.2, then B should also have $J(B) < \infty$. Comparing the formulas for $\dot{\ell}_{\theta,F}$ and $A_{\theta,F}B$, we see that minimizing $P_{\theta,F}(\dot{\ell}_{\theta,F} - A_{\theta,F}B)^2$ over B is a weighted least squares problem that is solved by

$$B_{\theta,F}(u) = f(u)h_\theta(u),$$

for

$$h_\theta(u) = \frac{\mathrm{E}_{\theta,F}(ZQ^2_{\theta,F}(X)|\, C - \theta Z = u)}{\mathrm{E}_{\theta,F}(Q^2_{\theta,F}(X)|\, C - \theta Z = u)} = \mathrm{E}(Z|\, C - \theta Z = u).$$

(The last equality follows by direct calculation.) Since the support of $C - \theta Z$ is contained in the interval D, the function h_θ can be defined arbitrarily outside D. It follows that this function $B_{\theta,F}$ certainly corresponds to a true submodel F_t provided the functions f and h_θ are sufficiently regular. Then the efficient score function for θ is given by

$$\tilde{\ell}_{\theta,F}(x) = -\big[z - h_\theta(z)\big]f(c - \theta z)Q_{\theta,F}(x). \tag{6.3}$$

Note, however, that the present function $B_{\theta,F}$ satisfies $J(B_{\theta,F}) < \infty$ only if F is three times differentiable, which is more than we initially assume for every F in the model. Therefore, in the following subsection we use a more complicated type of path F_t, which is well-defined as soon as $J(F) < \infty$. From this it is clear that $\tilde{\ell}_{\theta,F}$ is the efficient score function already under the condition that $J(F) < \infty$. The construction of this path is somewhat complicated, but it is necessary, because our proof of asymptotic normality of $\hat{\theta}$ uses a perturbation of $\hat{F}$, for which the finiteness of $J(\hat{F})$ is quaranteed by definition, but possibly not a smooth third derivative.

6.2. LEAST FAVORABLE SUBMODEL. By assumption the support D_θ of the variables $C - \theta Z$ (under P_0) is contained strictly within the interval D, for every θ. Therefore, for every (θ, t) such that $|\theta - t|$ is sufficiently close to zero, there exists a strictly increasing, infinitely often differentiable function $u \mapsto \psi_{\theta,t}(u)$ with

$$\begin{aligned} \psi_{\theta,t}(u) &= u, \qquad u \in D_\theta, \\ \psi_{\theta,t}\big(u + (\theta - t)h_\theta(u)\big) &= u, \qquad u \in \delta D. \end{aligned}$$

Moreover, we can ensure that $(u,t) \mapsto \psi_{t,\theta}(u)$ is infinitely often differentiable at $u \in D$, $t = \theta$ as well. The second identity in the preceding display, which, unfortunately, rules out the identity function, ensures that $\psi_{u,t}(D) = D$ and will be used to control the partial derivative of $J\big(F_t(\theta, F)\big)$ with respect to t in the argument below.

For a given pair (θ, F), we now define a least favorable submodel as

$$F_t(\theta, F)(u) = F \circ \psi_{\theta,t}\big(u + (\theta - t)h_\theta(u)\big).$$

Then $F_\theta(\theta, F)(c - \theta z) = F(c - \theta z)$ for every (c, z) in the support of (C, Z), and, with a dot denoting differentiation with respect to t,

$$\begin{aligned} F_t(\theta, F)'(u) &= F' \circ \psi_{\theta,t}\big(u + (\theta - t)h_\theta(u)\big) \\ &\qquad \times \psi'_{\theta,t}\big(u + (\theta - t)h_\theta(u)\big)\big(1 + (\theta - t)h'_\theta(u)\big), \\ \dot F_t(\theta, F)(u) &= F' \circ \psi_{\theta,t}\big(u + (\theta - t)h_\theta(u)\big) \\ &\qquad \times \Big[\dot\psi_{\theta,t}\big(u + (\theta - t)h_\theta(u)\big) - \psi'_{\theta,t}\big(u + (\theta - t)h_\theta(u)\big)h_\theta(u)\Big], \\ \frac{\partial}{\partial t}\log p_{t,F_t(\theta,F)}(x) &= -\Big[F_t'(\theta, F)(c - tz)z + \dot F_t(\theta, F)(c - tz)\Big]Q_{t,F_t(\theta,F)}(x). \end{aligned}$$

Evaluated at $t = \theta$ this yields the efficient score function $\tilde\ell_{\theta,F}$. Next, with $\phi_{\theta,t}(u) = \psi_{\theta,t}\big(u + (\theta - t)h_\theta(u)\big)$,

$$F_t(\theta, F)''(u) = F'' \circ \phi_{\theta,t}(u)\phi'_{\theta,t}(u)^2 + F' \circ \phi_{\theta,t}(u)\phi''_{\theta,t}(u).$$

For sufficiently small $|\theta - t|$ the map $\phi_{\theta,t}$ is a strictly increasing, three times differentiable bijection on D. Therefore,

$$J^2\big(F_t(\theta, F)\big) = \int_D \Big[F''(v)\big(\phi'_{\theta,t} \circ \phi_{\theta,t}^{-1}(v)\big)^2 + F'(v)\big(\phi''_{\theta,t} \circ \phi_{\theta,t}^{-1}(v)\big)\Big]^2 \frac{dv}{\phi'_{\theta,t} \circ \phi_{\theta,t}^{-1}(v)}.$$

It follows that $J\big(F_t(\theta, F)\big) < \infty$ whenever $J(F) < \infty$ and $|\theta - t|$ is sufficiently close to zero. Furthermore, some tedious calculus shows that this quantity is partially differentiable with respect to t in a neighbourhood of θ, with derivative at $t = \theta$ bounded in absolute value by a multiple of $\int_D \big(F''(v)^2 + F'(v)^2\big)\,dv \lesssim J^2(F)$.

Since $(\hat\theta, \hat F)$ maximizes the likelihood and $F_{\hat\theta}(\hat\theta, \hat F) = \hat F$, the value $\hat\theta$ maximizes the function $t \mapsto \log \prod p_{t,F_t(\hat\theta,\hat F)}(x_i) - \hat\lambda^2 J^2(F_t(\hat\theta, \hat F))$. It follows that

$$\frac{1}{n}\sum_{i=1}^n \tilde\ell_{\hat\theta,\hat F}(X_i) - \hat\lambda^2 \frac{\partial}{\partial t}_{|t=\hat\theta} J^2\big(F_t(\hat\theta, \hat F)\big) = 0.$$

In view of (3.1) and the fact that $J(\hat F) = O_P(1)$,

$$\frac{1}{n}\sum_{i=1}^n \tilde\ell_{\hat\theta,\hat F}(X_i) = o_P(n^{-1/2}).$$

By the linearity of the model in F, or by direct calculation, it follows that

$$P_{\theta,\eta_0}\tilde\ell_{\theta,\eta}(X) = 0, \qquad \text{every} \quad \theta, \eta, \eta_0.$$

Thus conditions (6.1) and (6.2) have been verified.

6.3. Regularity conditions of Proposition 6.1. In order to verify the "regularity conditions" of Proposition 6.1, note first that

$$\begin{aligned} \big|\hat F'(c - \hat\theta z) - \hat F'(c - \theta_0 z)\big| &\lesssim J(\hat F)|\hat\theta - \theta_0|^{1/2}, \\ \big|\hat F(c - \hat\theta z) - \hat F(c - \theta_0 z)\big| &\lesssim \big(J(\hat F) + 1\big)|\hat\theta - \theta_0|. \end{aligned}$$

Since $\widehat{\theta}_n \xrightarrow{P} \theta_0$, the right sides converge to zero in probability. Combined with the convergence $\widehat{F}_n \xrightarrow{P} F_0$ with respect to the uniform norm on the closure of the support of $C - \theta_0 Z$, and the assumption that F_0 is bounded away from zero and one on D, the functions $Q_{\widehat{\theta},\widehat{F}}(x)$ are seen to be bounded with probability tending to one. Furthermore, the functions $\widehat{F}'(c - \widehat{\theta}z)$ are uniformly bounded. Since also the functions $h_{\widehat{\theta}}$ are uniformly bounded, it follows that the functions $\widetilde{\ell}_{\widehat{\theta},\widehat{F}}(x)$ are uniformly bounded with probability tending to one. Thus $(P_{\theta_0,F_0} + P_{\theta,F_0})\|\widetilde{\ell}_{\theta,F}\|^2 = O(1)$ is bounded trivially. Furthermore, $\widetilde{\ell}_{\widehat{\theta},\widehat{F}}(x) \to \widetilde{\ell}_{\theta_0,F_0}(x)$ for P_{θ_0,F_0}-almost every x.

It is straightforward to check that the model $\theta \mapsto p_{\theta,F_0}$ is differentiable in quadratic mean at θ_0 with score function $\dot{\ell}_{\theta_0,F_0}$ as given previously.

By Lemma 5.2 and the bracketing-central-limit-theorem of Ossiander (Cf. Theorem 2.5.6 of Van der Vaart and Wellner [29]), the class of functions $F(c-\theta z)$, with F ranging over the distribution functions with $J(F) \le M$, and $\theta \in \Theta$, is P_0-Donsker. For the functions $F(c - \theta z)$ restricted to be bounded away from zero and one, the functions $Q_{\theta,F}(x)$ are Lipschitz transformations of the functions $\big(F(c - \theta z), \delta\big)$. Thus, under this restriction, this class is P_0-Donsker by Theorem 2.10.6 of Van der Vaart and Wellner [29]. It is also uniformly bounded.

By Lemma 6.2 (below) and the bracketing-central-limit theorem, the class of all functions $F'(c - \theta z)$ with $J(F) \le M$ is P_0-Donsker. It is also uniformly bounded.

The class of functions $z - h_\theta(z)$ is P_0-Donsker by assumption.

Combining these results, we conclude by Theorem 2.10.6 of Van der Vaart and Wellner [29] that the class of functions $\widetilde{\ell}_{\theta,F}$ with θ ranging over Θ and F over the distribution functions such that $J(F) \le M$ and such that $\|F - F_0\|_D$ is sufficiently small, is P_0-Donsker.

Lemma 6.2. *For every $M \ge 1$,*

$$\log N_{[]}\big(\varepsilon, \{F'(c - \theta z) \colon \theta \in \Theta, J(F) \le M\}, L_2(P_0)\big) \lesssim \Big(\frac{M}{\varepsilon}\Big).$$

Proof. By Lemma 3.6 the class of derivatives F' of functions F with $J(F) \le M$ is uniformly bounded by a multiple of $J(F)+1 \lesssim M$ on D. Clearly, $\int_D (F')'^2(u)\, du = J^2(F) \le M^2$. Therefore, by Lemma 3.5,

$$\log N\big(\varepsilon, \{F'_{|D} \colon J(F) \le M\}, \|\cdot\|_\infty\big) \lesssim \Big(\frac{M}{\varepsilon}\Big).$$

Furthermore, by Lemma 3.6 $\big|F'(s) - F'(s_0)\big| \le |s - s_0|^{1/2} J(F)$ for every $s, s_0 \in D$.

Now we can construct a net over the class of functions of interest, by first choosing an $(\varepsilon/M)^2$-net $\theta_1, \ldots, \theta_p$ over Θ (for the Euclidean distance), next choosing an ε-net $G_1, \ldots, G_q$ over the functions $F'_{|D}$ (for the supremum metric), and finally forming all functions $G_i(c - \theta_j z)$. Then for every (θ, F) there exists (θ_j, G_i) such that

$$\big|F'(c - \theta z) - G_i(c - \theta_j z)\big| \lesssim M|\theta - \theta_j|^{1/2} + \|F'_{|D} - G_i\|_\infty \lesssim M\frac{\varepsilon}{M} + \varepsilon \lesssim \varepsilon.$$

We need at most $2|\Theta|M^2/\varepsilon^2$ points θ_j, and at most a power of (M/ε) points G_i. Thus, the entropy for the uniform norm of the class of functions in the lemma is bounded by a multiple of $(M/\varepsilon)+\log(M/\varepsilon)$. Consequently, the bracketing entropy for the L_2-norm is bounded similarly. □

7. Proof of Theorem 3.4

7.1. Proof of Theorem 3.4 (i). We use the following proposition to obtain first a rate of convergence for $p_{\tilde{\theta},\tilde{F}}$ in the Hellinger distance h. Since

$$\|p_{\theta,F}-p_{\theta_0,F_0}\|_2^2 \le 4\int (p_{\theta,F}^{1/2}-p_{\theta_0,F_0}^{1/2})^2,$$

this rate translates immediately into a rate for the L_2-norm. The proposition is due to Birgé and Massart [2] and Shen and Wong [27], and is also a consequence of combining Theorems 3.4.1 and 3.4.4 of Van der Vaart and Wellner [29] (see bottom page 328).

Proposition 7.1. *Let $\mathcal{P}$ be a set of probability densities on a fixed measurable space such that for functions ϕ_n such that $\delta \mapsto \phi_n(\delta)/\delta^\alpha$ is nondecreasing for some $\alpha < 2$, and every $\delta > 0$,*

$$\int_0^\delta \sqrt{\log N_{[]}(\varepsilon,\mathcal{P},h)}\,d\varepsilon \le \phi_n(\delta).$$

Suppose that δ_n are positive numbers with $\phi_n(\delta_n) \le \sqrt{n}\delta_n^2$ for every n. Then any estimator $\widehat{p} \in \mathcal{P}$ such that $\mathbb{P}_n \log(\widehat{p}/p_0) \ge 0$ satisfies $h(\widehat{p},p_0) = O_P(\delta_n)$.

Thus, it suffices to compute the entropy-with-bracketing of the class of densities $p_{\theta,F}$ for the Hellinger distance. For simplicity assume that $Z \ge 0$ almost surely. (The general case can be treated by considering brackets for the values $z \ge 0$ and $z < 0$ separately.) Then $F_1 \le F \le F_2$ and $\theta_2 \le \theta \le \theta_1$ imply

$$F_1(c-\theta_1 z) \le F(c-\theta z) \le F_2(c-\theta_2 z).$$

Thus brackets $[F_1,F_2]$ and $[\theta_2,\theta_1]$ for F and θ, respectively, yield brackets

$$\Big[F_1(c-\theta_1 z)^\delta\big(1-F_2(c-\theta_2 z)\big)^{1-\delta}, F_2(c-\theta_2 z)^\delta\big(1-F_1(c-\theta_1 z)\big)^{1-\delta}\Big]$$

for the densities $p_{\theta,F}$. The sizes of these brackets in the squared Hellinger distance are equal to the sums of two terms, corresponding to the integrals over $\delta = 1$ and $\delta = 0$, respectively. The first of these two terms is

$$\begin{aligned}
&P^{C,Z}\big(F_2^{1/2}(c-\theta_2 z)-F_1^{1/2}(c-\theta_1 z)\big)^2\\
&\quad\lesssim P^{C,Z}\big(F_2^{1/2}(c-\theta_2 z)-F_1^{1/2}(c-\theta_2 z)\big)^2 + P^{C,Z}\big|F_1(c-\theta_2 z)-F_1(c-\theta_1 z)\big|\\
&\quad\lesssim \int (F_2^{1/2}-F_1^{1/2})^2\,dP^{C-\theta_2 Z} + \int \mathrm{P}\big(C-\theta_1 Z < s \le C-\theta_2 Z\big)\,dF_1(s)\\
&\quad\lesssim \int_D (F_2^{1/2}-F_1^{1/2})^2\,dy + |\theta_1-\theta_2|,
\end{aligned}$$

by the assumptions on the distribution of (C, Z). The second of the two terms can be bounded similarly, and we obtain that the squared Hellinger distance of the brackets is bounded above by a multiple of

$$\int (F_2^{1/2} - F_1^{1/2})^2 + \left((1-F_2)^{1/2} - (1-F_1)^{1/2}\right)^2 d\lambda + |\theta_1 - \theta_2|$$

with λ the Lebesgue measure on D. The bracket is of size proportional to ε if the L_2-distances between $F_1^{1/2}$ and $F_2^{1/2}$ and between $(1-F_1)^{1/2}$ and $(1-F_2)^{1/2}$ are smaller than ε, and the Euclidean distance $|\theta_1 - \theta_2|$ is smaller than ε^2. We conclude that, with h the Hellinger distance, d the distance with square

$$d^2(f,g) = \int \left(\sqrt{f} - \sqrt{g}\right)^2 + \left(\sqrt{1-f} - \sqrt{1-g}\right)^2 d\lambda,$$

$\mathcal{F}$ the set of all distribution functions, and $\mathcal{F}_D$ their restrictions to D,

$$\log N_{[]}\left(\varepsilon, \{p_{\theta,F} : \theta \in \Theta, F \in \mathcal{F}\}, h\right) \lesssim \log N_{[]}\left(\varepsilon, \mathcal{F}_D, d\right) + \log\left(\frac{1}{\varepsilon}\right).$$

The first term on the right is bounded by a constant times $1/\varepsilon$. This follows from the fact that the bracketing entropy of the set of uniformly bounded, monotone functions is bounded by a constant times $1/\varepsilon$ (cf. Theorem 2.7.5 of Van der Vaart and Wellner [29]). Indeed, since the square roots $F^{1/2}$ of the functions $F \in \mathcal{F}_D$ are monotone and take their values in $[0,1]$, they can be covered by $\exp(C/\varepsilon)$ brackets of size ε in $L_2(\lambda)$. Similarly, the functions $(1-F)^{1/2}$ can be covered (independently) by $\exp(C/\varepsilon)$ brackets. Then construct brackets of the form $[a^2 \vee (1-d^2), b^2 \wedge (1-c^2)]$ for the functions F, if $[a,b]$ and $[c,d]$ are the brackets containing $F^{1/2}$ and $(1-F)^{1/2}$, respectively. As F ranges over $\mathcal{F}_D$, this gives at most $\exp(2C/\varepsilon)$ brackets, that cover $\mathcal{F}$ and have d-size less than 2ε.

Theorem 3.4 now follows from Proposition 7.1, applied with

$$\phi_n(\delta) = \int_0^\delta \sqrt{\frac{1}{\varepsilon} + \log\frac{1}{\varepsilon}}\, d\varepsilon.$$

7.2. CONSISTENCY. The consistency of $(\widetilde{\theta}, \widetilde{F})$ is proved in Cosslett [5] through application of Wald's theorem, as modified by Kiefer and Wolfowitz. Since we need the consistency for the proof of Theorem 3.4 (ii), we rederive this result for completeness, as a consequence of Theorem 3.4 (i).

The set of all subdistribution functions is compact for the vague topology, and by assumption $\widetilde{\theta}$ takes its values in a compact. Hence every subsequence of $(\widetilde{F}, \widetilde{\theta})$ has a (vaguely) converging subsequence. It suffices to show that (θ_0, F_0) is the only possible limit point. We already know that $\widetilde{F}(c - \widetilde{\theta}z) \to F_0(c - \theta_0 z)$ for almost every (c,z), at least along subsequences. By the following lemma we conclude that $(\widetilde{F}, \widetilde{\theta}) \to (\theta_1, F_1)$ implies that $\widetilde{F}(c - \widetilde{\theta}z) \to F_1(c - \theta_1 z)$ for every $c - \theta_1 z$ where F_1 is continuous. We conclude that $F_0(c - \theta_0 z) = F_1(c - \theta_1 z)$ for almost every (c,z). By right continuity we can extend this to every (c,z) in the interior of the support of (C,Z). This implies $(\theta_1, F_1) = (\theta_0, F_0)$.

Lemma 7.2. *If $F_m \rightsquigarrow F$ for the vague topology, $x_m \to x$ and F is continuous at x, then $F_m(x_m) \to F(x)$.*

PROOF OF THEOREM 3.4 (II). By almost the same argument as in Lemma 5.6 (but note that we evaluate the derivative at θ rather than at θ_0),

$$P_0\Big[F(c-\theta z) - F_0(c-\theta_0 z) - \big(-zF_0'(c-\theta z)(\theta-\theta_0) + (F-F_0)(c-\theta z)\big)\Big]^2 = o\big(|\theta-\theta_0|^2\big).$$

By assumption, as $\theta \to \theta_0$,

$$\frac{\mathrm{E}_0\big(\mathrm{E}_0(Z|\,C-\theta Z)F_0'(C-\theta Z)\big)^2}{\mathrm{E}_0\big(ZF_0'(C-\theta Z)\big)^2} \to \frac{\mathrm{E}_0\big(\mathrm{E}_0(Z|\,C-\theta_0 Z)F_0'(C-\theta_0 Z)\big)^2}{\mathrm{E}_0\big(ZF_0'(C-\theta_0 Z)\big)^2} := c < 1.$$

Therefore, by Lemma 5.7, applied with $g_1 = -zF_0'(c-\theta z)(\theta-\theta_0)$ and $g_2 = (F-F_0)(c-\theta z)$,

$$\begin{aligned} P_0\Big[-zF_0'(c-\theta z)(\theta-\theta_0) + (F-F_0)(c-\theta z)\Big]^2 &\gtrsim P_0\big(zF_0'(c-\theta z)(\theta-\theta_0)\big)^2 + P_0(F-F_0)(c-\theta z)^2 \\ &\gtrsim |\theta-\theta_0|^2 + \int (F-F_0)^2\, dP^{C-\theta Z}. \end{aligned}$$

Theorem 3.4 (ii) follows.

References

[1] P. Bickel, C. Klaassen, Y. Ritov, and J. Wellner, *Efficient and Adaptive Estimation for Semiparametric Models*, Johns Hopkins University Press, Baltimore, 1993.

[2] L. Birgé and P. Massart, *Rates of convergence for minimum contrast estimators*, Probab. Theory Rel. Fields, 97 (1993), 113–150.

[3] M. S. Birman and M. Z. Solomjak, *Piecewise-polynomial approximation of functions of the classes W_p*, Mathematics of the USSR Sbornik, 73 (1967), 295–317.

[4] G. Chamberlain, *Asymptotic efficiency in semi-parametric models with censoring*, J. Econometrics, 32 (1986), 189–218.

[5] S. R. Cosslett, *Distribution-free maximum likelihood estimator of the binary choice model*, Econometrica, 51 (1983), 765–782.

[6] S. R. Cosslett, *Efficiency bounds for distribution-free estimators of the binary choice and censored regression modelsa*, Econometrica, 55 (1987), 559–585.

[7] P. Groeneboom, *Asymptotics for interval censored observations*, Report 87-18, Dept. of Math., Univ. of Amsterdam, 1987.

[8] P. Groeneboom, *Nonparametric maximum likelihood estimators for interval censoring and deconvolution*, Report 378, Dept. of Statist., Stanford University, 1991.

[9] P. Groeneboom and J. A. Wellner, *Information Bounds and Nonparametric Maximum Likelihood Estimation*, Birkhäuser, Basel, 1992.

[10] A. K. Han, *Non-parametric analysis of a generalized regression model*, J. Econometrics, 35 (1987), 303–316.

[11] J. L. Horowitz, *A smoothed maximum score estimator for the binary response model*, Econometrica, 60 (1992), 505–531.

[12] J. L. Horowitz, *Optimal rates of convergence of parameter estimates in the binary response model with weak distributional assumptions*, Econometric Theory, 9 (1993), 1–18.

[13] J. L. Horowitz, *Semiparametric estimation of a regression model with an unknown transformation of the dependent variable*, Econometrica, 64 (1996), 103–137.

[14] J. Huang, *Maximum scored likelihood estimation of a linear regression model with interval censored data*, Report 356, Dept. of Statist., Univ. of Washington, 1993.

[15] J. Huang, *Maximum likelihood estimation for proportional odds regression model with current status data*, in: Proc. Internat. Workshop on the Analysis of Censored Data, IMS Lecture Notes—Monograph Series, Vol. 27, pp. 129–145, Inst. Math. Statist., Hayward, CA, 1995.

[16] J. Huang, *Efficient estimation for the Cox model with interval censoring*, Ann. Statist., 24 (199), 540–568.

[17] J. Huang, A. J. Rossini, and J. A. Wellner, *Efficient estimation for the proportional hazards model with Case 2 interval censoring*, Report 251, Univ. of Iowa, Dept. of Statist., 1996.

[18] J. Kim and D. Pollard, *Cube root asymptotics*, Ann. Statist., 18 (1990) 191–219.

[19] R. W. Klein and R. H. Spady, *An efficient semiparametric estimator for binary response models*, Econometrica, 61 (1993), 387–421.

[20] C. F. Manski, *Maximum score estimation of the stochastic utility model of choice*, J. Econometrics, 3 (1975), 205–228.

[21] C. F. Manski, 1985 *Semiparametric analysis of discrete response: asymptotic properties of the maximum score estimator*, J. Econometrics, 27 (1985), 313–333.

[22] S. A. Murphy and A. W. Van der Vaart, *Semiparametric likelihood ratio tests*, Ann. Statist., 25 (1997), 1471–1509.

[23] D. Pollard, *Asymptotics of a binary choice model*, Lecture Notes, Dept. Statist., Yale University, 1994, http://www.stat.yale.edu/pollard/dbp.html.

[24] D. Rabinowitz, A. Tsiatis, and J. Aragon, *Regression with interval-censored data*, Biometrika, 82 (1995), 501–513.

[25] A. J. Rossini and A. A. Tsiatis, *A semiparametric proportional odds regression model for the analysis of current status data*, J. Amer. Statist. Assoc., 91 (1996), 713–721.

[26] G. A. Satten, *Rank-based inference in the proportional hazard model for interval censored data*, Biometrika, 83 (1996), 355–370.

[27] X. Shen and W. H. Wong, *Convergence rate of sieve estimates*, Ann. Statist., 22 (1994), 580–615.

[28] R. P. Sherman, *The limiting distribution of the maximum rank correlation estimator*, Econometrica, 61 (1993), 123–137.

[29] A. W. Van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*, Springer, New York, 1996.

[30] A. W. Van der Vaart, *Efficient estimation in semiparametric models*, Ann. Statist., 24 (1996), 862–878.

[*Received March* 1999]