Sri Lankan Journal of Applied Statistics (Special Issue)
Modern Statistical Methodologies in the Cutting Edge of Science

Maximum Empirical Likelihood Estimation In A Heteroscedastic Linear Regression Model With Possibly Missing Responses

Anton Schick* **and Yilin Zhu**

Department of Mathematical Sciences, Binghamton University,
Binghamton, New York, USA
*Corresponding Author: anton@math.binghamton.edu

Received: 4, September 2013 / Revised: 3, February 2014 / Accepted: 11, March 2014

ABSTRACT

A heteroscedastic linear regression model is considered where responses are allowed to be missing at random and with the conditional variance modeled as a function of the mean response. Maximum empirical likelihood estimation is studied for an empirical likelihood with an increasing number of estimated constraints. The resulting estimator is shown to be asymptotically normal and can outperform the ordinary least squares estimator.

Keywords: Weighted least squares estimator, Missing at random, Estimated constraints, Increasing number of constraints.

1. Introduction

Consider the heteroscedastic linear regression model in which the response variable Y is linked to the q-dimensional covariate vector X by the formula

$$Y = \theta^T Z + \varepsilon,$$

where Z is $m(X)$ for a known measurable function m from $\mathbb{R}^q$ into $\mathbb{R}^p$, θ is an unknown vector in $\mathbb{R}^p$, the error variable ε is conditionally centered, i.e., $E[\epsilon|X] = 0$, and its conditional variance

$$\sigma^2(X) = E[\epsilon^2|X]$$

is bounded and bounded away from zero. In order to identify the regression parameter, it is assumed that the matrix

$$M = E[ZZ^\top]$$

is well defined and positive definite. For $q = 1$, a possible choice of m is

$$m(x) = (1, x, x^2, \dots, x^{p-1})^\top, \quad x \in \mathbb{R},$$

which corresponds to polynomial regression. The case $p = 2$ yields the classical heteroscedastic simple linear regression model. For $q = 2$, a possible choice of m is given by

$$m(x_1, x_2) = (1, x_1, x_2, x_1^2, x_1 x_2, x_2^2)^\top, \quad x_1, x_2 \in \mathbb{R}.$$

In the ideal situation one observes the pair (X, Y). In real life data sets, however, one frequently encounters missing values. Here we allow the response Y to be missing. Then one observes $(\delta, X, \delta Y)$ with δ an indicator random variable. The interpretation is that for $\delta = 1$ one observes the full pair (X, Y), while for $\delta = 0$ one observes only the covariate X. We make the common assumption that the response is *missing at random*. This means that the conditional probability of $\delta = 1$ given (X, Y) depends on X alone,

$$P(\delta = 1 | X, Y) = P(\delta = 1 | X).$$

Monographs on missing data are Little and Rubin (2002) and Tsiatis (2006). We assume throughout that the conditional probability

$$\pi(X) = P(\delta = 1 | X)$$

is bounded away from zero. This implies that $E[\delta]$ is positive.

The data in our model are $(\delta_1, X_1, \delta_1 Y_1), \dots (\delta_n, X_n, \delta_n Y_n)$ which are independent copies of the triple $(\delta, X, \delta Y)$. We set

$$Z_j = m(X_j) \quad \text{and} \quad \varepsilon_j = Y_j - \theta^\top Z_j, \quad j = 1, \dots, n,$$

and let $N = \delta_1 + \cdots + \delta_n$ denote the number of complete observations. A possible estimator of θ is the least squares estimator $\hat{\theta}_L$ based on the complete observations,

$$\hat{\theta}_L = \operatorname*{arg\,min}_{\vartheta \in \mathbb{R}^p} \sum_{j=1}^{n} \delta_j (Y_j - \vartheta^\top Z_j)^2.$$

If σ^2 were known, we could use the weighted least squares estimator $\hat{\theta}_W$ to estimate the regression parameter θ. This estimator minimizes the weighted sum of squares

$$Q(\vartheta) = \sum_{j=1}^{n} \delta_j \frac{(Y_j - \vartheta^\top Z_j)^2}{\sigma^2(X_j)}, \quad \vartheta \in \mathbb{R}^p,$$

and satisfies the stochastic expansion

$$\hat{\theta}_W = \theta + \frac{1}{n}\sum_{j=1}^{n} H^{-1}\frac{\delta_j \varepsilon_j}{\sigma^2(X_j)} Z_j + o_P(n^{-1/2}) \tag{1.1}$$

with

$$H = E\Big[\frac{\delta\varepsilon^2}{\sigma^4(X)} ZZ^\top\Big] = E\Big[\frac{\pi(X)}{\sigma^2(X)} ZZ^\top\Big].$$

Note that H is invertible because M is invertible and π/σ^2 is bounded and bounded away from zero. Thus $n^{1/2}(\hat{\theta}_W - \theta)$ is asymptotically normal with mean vector 0 and dispersion matrix H^{-1}. Since we treat σ^2 as unknown, the weighted least squares estimator is no longer available. For this reason we call $\hat{\theta}_W$ the oracle weighted least squares estimator. A natural approach is to minimize instead of $Q(\vartheta)$ the weighted sum of squares

$$\hat{Q}(\vartheta) = \sum_{j=1}^{n} \frac{\delta_j (Y_j - \vartheta^\top Z_j)^2}{\hat{\sigma}^2(X_j)}, \quad \vartheta \in \mathbb{R}^d,$$

in which an estimator $\hat{\sigma}^2$ replaces the unknown σ^2. Working with a kernel estimator of $\hat{\sigma}^2$ based on least squares residuals, Müller and Van Keilegom (2012) show that under additional assumptions the resulting estimator is asymptotically equivalent to the oracle weighted least squares estimator. This generalizes the work of Carroll (1982) who was the first to obtain such a result without missing responses, i.e., in the case when δ is identically 1. Similar results without missing responses were obtained by Müller and Stadtmüller (1987), Robinson (1987) and Schick (1987) for various nonparametric estimators of σ^2 under differing conditions.

Schick (2013) has shown that one can construct an estimator that is asymptotically equivalent to the oracle weighted least squares estimator without constructing an estimator of the variance function σ^2. He treated the case $q = 1$ with missing responses. He avoided estimation of the variance function σ^2 by working with a *guided* maximum empirical likelihood estimator associated with an empirical likelihood with an increasing number of estimated constraints,

$$\mathscr{S}_n(\vartheta) = \sup\Big\{ \prod_{j=1}^{n} n\pi_j : \pi_1 \geq 0, \ldots, \pi_n \geq 0, \sum_{j=1}^{n} \pi_j = 1, \\ \sum_{j=1}^{n} \pi_j \delta_j (Y_j - \vartheta^\top Z_j) v_{r_n}(G_j)) = 0\Big\}$$

with

$$G_j = \frac{1}{N}\sum_{i=1}^{n} \delta_i \mathbf{1}[X_i \le X_j], \quad j = 1, \dots, n,$$

r_n a positive integers tending to infinity with n, and v_r a function from $[0,1]$ into $\mathbb{R}^{r+1}$ for each positive integer r. His estimator maximizes the restriction of $\mathscr{S}_n$ to the random ball centered at the least squares estimator of radius $C(\log n/n)^{1/2}$ for some constant C,

$$\hat\theta_S = \operatorname*{arg\,max}_{n^{1/2}\|\vartheta - \hat\theta_L\| \le C \log^{1/2} n} \mathscr{S}_n(\vartheta).$$

He obtained the asymptotic equivalence of this estimator and the oracle weighted least squares estimator (by establishing the expansion (1.1) with $\hat\theta_S$ in place of $\hat\theta_W$) under a growth condition on r_n and mild assumptions on the functions v_r. With $\mathscr{U}$ denoting the uniform distribution on [0,1], he required these functions to satisfy the following conditions.

(C1) There are positive constants c_0, c_1, c_2, c_3 such that the inequalities

$$\begin{gathered} \|v_r(x)\|^2 \le c_0 r, \\ \|v_r(x) - v_r(y)\|^2 \le c_1 r^3 |y - x|^2, \\ c_2 \le \int (u^\top v_r)^2 \, d\mathscr{U} \le c_3, \end{gathered}$$

hold for all x and y in $[0,1]$ and all unit vectors u in $\mathbb{R}^{r+1}$.

(C2) For every g in $L_2(\mathscr{U})$,

$$\inf_{b \in \mathbb{R}^{r+1}} \int (b^\top v_r - g)^2 \, d\mathscr{U} \to 0 \quad \text{as } r \to \infty.$$

Guided maximum likelihood estimation was introduced and studied by Peng and Schick (2012) for a fixed number of constraints. As shown in Schick (2013) possible choices for the function v_r are

(a) $v_r = (1, \varphi_1, \dots, \varphi_r)^\top$ with

$$\varphi_k(x) = \sqrt{2}\cos(k\pi x), \quad 0 \le x \le 1, k = 1, 2, \dots,$$

(b) $v_r = (v_{r,0}, \dots, v_{r,r})^\top$ with

$$v_{r,i}(x) = r^{1/2}\max(0, 1 - |rx - i|), \quad i = 0, \dots, r, \ 0 \le x \le 1.$$

The first choice consists of the first $1+r$ elements of the trigonometric basis of $L_2(\mathscr{U})$. The components of the second choice form a basis of the linear splines with knots at the points $0/r, \dots, r/r$.

In many situations the conditional variance function can be modeled as a function of the mean response as expressed in the following assumption.

(A0) There is a measurable function τ from $\mathbb{R}$ into some compact subset of $(0,\infty)$ such that
$$\sigma^2(X) = \tau^2(\theta^\top Z).$$

We now make this assumption and study the guided maximum likelihood estimator
$$\hat{\theta} = \underset{n^{1/2}\|\vartheta-\hat{\theta}_*\| \le C\log^{1/2} n}{\arg\max} \mathscr{R}_n(\vartheta) \tag{1.2}$$

associated with the modified empirical likelihood

$$\mathscr{R}_n(\vartheta) = \sup\Big\{ \prod_{j=1}^n n\pi_j : \ \pi_1 \ge 0, \dots, \pi_n \ge 0, \sum_{j=1}^n \pi_j = 1, \\ \sum_{j=1}^n \pi_j \delta_j (Y_j - \vartheta^\top Z_j) v_{r_n}(\bar{R}_j(\hat{\theta}_*)) = 0 \Big\}, \quad \vartheta \in \mathbb{R}^p,$$

which is obtained from $\mathscr{S}_n(\vartheta)$ by replacing G_j by $\bar{R}_j(\hat{\theta}_*)$. Here

$$\bar{R}_j(\vartheta) = \frac{1}{N}\sum_{i=1}^n \delta_i \mathbf{1}[\vartheta^\top Z_i \le \vartheta^\top Z_j], \quad \vartheta \in \mathbb{R}^p, \ j = 1, \dots, n,$$

and $\hat{\theta}_*$ is a discretized version of $\hat{\theta}_L$ which is obtained by matching $\hat{\theta}_L$ with the closest point in the grid

$$\{c(i_1, \dots, i_p)^\top/\sqrt{n} : i_1, \dots, i_p = \dots, -2, -1, 0, 1, 2, \dots\}$$

where c is a positive constant. We work with a discretized version to simplify our proofs. Indeed, discretized estimators can be treated as non-stochastic sequences in the proofs. Discretization was introduced by Le Cam (1960) for precisely this reason and has become a popular tool in the construction of efficient estimators in semiparametric models, see Bickel (1982) and Schick (1986, 1987). We need the following assumptions.

(A1) There is a neighborhood $\mathcal{N}$ of θ, a finite constant K and a positive α, $\alpha \le 1$, such that

$$\Big|E[\delta\mathbf{1}[\vartheta^\top Z \le y]] - E[\delta\mathbf{1}[\theta^\top Z \le z]]\Big| \le K\Big(\|\vartheta - \theta\|^\alpha + |y-z|^\alpha\Big)$$

holds for all $\vartheta \in \mathcal{N}$ and all $y, z \in \mathbb{R}$.

(A2) The matrix

$$H_* = E\Big[\frac{\delta}{\tau^2(\theta^\top Z)} W W^\top\Big]$$

is positive definite with

$$W = \nu(\theta^\top Z) = E(Z|\theta^\top Z, \delta = 1).$$

We are ready to state our main result.

Theorem 1. *Suppose (A0), (A1), (A2), (C1) and (C2) hold. Assume also that the error variable ε has a finite fourth moment and r_n satisfies $r_n^5 \log n = o(n)$ and $r_n^5 = o(n^\alpha)$. Then the guided maximum empirical likelihood estimator $\hat\theta$ defined in (1.2) satisfies the expansion*

$$\hat\theta = \theta + \frac{1}{n}\sum_{j=1}^n H_*^{-1} \frac{\delta_j \varepsilon_j}{\tau^2(\theta^\top Z_j)} \nu(\theta^\top Z_j) + o_P(n^{-1/2}). \tag{1.3}$$

Therefore $n^{1/2}(\hat\theta - \theta)$ is asymptotically a centered multivariate normal random vector with dispersion matrix H_*^{-1}.

The theorem shows that the new estimator is no longer equivalent to the oracle weighted least squares estimator. Simulations in Section 2 show that the new estimator can significantly outperform the complete case least squares estimator $\hat\theta_L$ in certain situations, but may be worse in others.

A proof of the theorem is given in Section 4. A discussion of our assumption and some preparatory results are in Section 3. The results of a simulation study are in Section 2.

2. Simulations

In order to assess our method, we performed a small simulation study using R. We compared several versions of our guided maximum likelihood estimator with the least squares estimator (OLSE) and the oracle weighted least squares estimator (WLSE). We took $p = q = 2$, $\theta = (1,1)^\top$, $\delta = 1$, $Z = m(X) = X$,

X a bivariate normal random vector with mean vector $(1, -2)^\top$ and diagonal dispersion matrix with diagonal entries 1 and 4, and $\varepsilon = \tau(\theta^\top X)\zeta$, with ζ standard normal and independent of X. For this choice, assumption (A1) holds with $\alpha = 1$; see Remark 3 below.

We looked at four cases for τ, namely $\tau = \tau_A$, $\tau = \tau_B$, $\tau = \tau_C$ and $\tau = \tau_D$, where

$$\tau_A^2(x) = 3 * \mathbf{1}[x \leq -1] + .2 * \mathbf{1}[x > -1].$$

$$\tau_B^2(x) = 0.1 + 2 * \exp(-\pi(x+1)^2/2).$$

$$\tau_C^2(x) = \min\{0.3 * |x+1| + 0.1, 10\}.$$

and

$$\tau_D^2(x) = \sin(x) + 1.1.$$

We ran simulations for these choices of τ with sample sizes $n = 100$ and $n = 200$, and 2000 repetitions. In the tables we report 100 times the simulated mean square error for each estimator. The simulated mean square error for an estimator $\tilde{\theta}$ with 2000 repetitions is defined by

$$\frac{1}{2000} \sum_{i=1}^{2000} \|\tilde{\theta}_i - \theta\|^2$$

with $\tilde{\theta}_i$ the result of the i-th repetition. It estimates $E[\|\tilde{\theta} - \theta\|^2]$.

We write GT(r) for our estimator when using the trigonometric basis and $r_n = r$, and GS(r) for our estimator when using the spline basis and $r_n = r$. Table 1 reports the results for OLSE, WLSE and GT(r) with $r = 1, \ldots, 5$, while Table 2 reports the results for OLSE, WLSE and GS(r) with $r = 1, \ldots, 5$. We used the same sample to construct each of the twelve estimators. Thus the columns OLSE and WLSE are the same for both tables.

We can see our proposed estimator performs better than the OLSE for the choices τ_A and τ_B in all cases when r is at least 3, while for the choices τ_C and τ_D is does perform worth. Also, the performance of our estimator is influenced by the choice of r although the choice $r = 4$ performs quite will in all cases. There seems little difference between the choice of bases.

Table 2.1: Simulated Mean Square Errors with Trigonometric Basis

τ	n	OLSE	WLSE	GT(1)	GT(2)	GT(3)	GT(4)	GT(5)
τ_A	100	1.067	0.369	0.892	0.764	0.619	0.581	0.580
τ_A	200	0.528	0.177	0.454	0.394	0.321	0.300	0.298
τ_B	100	0.281	0.118	0.351	0.222	0.179	0.179	0.179
τ_B	200	0.137	0.058	0.161	0.100	0.086	0.086	0.086
τ_C	100	0.803	0.603	0.933	0.875	0.878	0.886	0.902
τ_C	200	0.393	0.299	0.436	0.409	0.402	0.403	0.409
τ_D	100	1.118	0.579	1.567	1.506	1.478	1.399	1.381
τ_D	200	0.543	0.269	0.712	0.679	0.664	0.621	0.614

Each entry is 100 times the simulated mean square error of the corresponding estimator, for two sample sizes and two choices of τ.

Table 2.2: Simulated Mean Square Errors with Spline Basis

τ	n	OLSE	WLSE	GS(1)	GS(2)	GS(3)	GS(4)	GS(5)
τ_A	100	1.067	0.369	0.887	0.717	0.593	0.579	0.582
τ_A	200	0.528	0.177	0.451	0.370	0.305	0.299	0.298
τ_B	100	0.281	0.118	0.333	0.212	0.180	0.181	0.182
τ_B	200	0.137	0.058	0.154	0.098	0.086	0.087	0.087
τ_C	100	0.803	0.603	0.964	0.905	0.896	0.895	0.917
τ_C	200	0.393	0.299	0.450	0.422	0.407	0.406	0.412
τ_D	100	1.118	0.579	1.544	1.493	1.495	1.393	1.378
τ_D	200	0.543	0.269	0.706	0.678	0.671	0.620	0.605

Each entry is 100 times the simulated mean square error of the corresponding estimator, for two sample sizes and two choices of τ.

3. Comments and Remarks

We begin with some comments on our assumptions, and then discuss implications of these assumptions.

Remark 1. For $\alpha = 1$, the requirements on r_n are equivalent to $r_n^5 \log n = o(n)$, while for $\alpha < 1$, they are equivalent to $r_n^5 = o(n^\alpha)$. For example, if α equals $1/2$, then we need $r_n^{10} = o(n)$.

The next two remarks address assumption (A1).

Remark 2. Let F denote the distribution of $\theta^\top Z$. We shall see that (A1) holds if F is Hölder. For real numbers y and z and a vector ϑ in $\mathbb{R}^p$, we set

$$J(y,z,\vartheta) = |E[\delta \mathbf{1}[\vartheta^\top Z \le y]] - E[\delta \mathbf{1}[\theta^\top Z \le z]]|.$$

We derive the bound

$$J(y,z,\vartheta) \le |F(z) - F(y)| + F(y+B) - F(y-B) + P(\|\vartheta - \theta\|\|Z\| > B)$$

valid for positive B. Indeed, for random variables S and T we have

$$\mathbf{1}[S+T \le y] - \mathbf{1}[S \le z] = \mathbf{1}[S+T \le y] - \mathbf{1}[S \le y] + \mathbf{1}[S \le y] - \mathbf{1}[S \le z]$$

and

$$\begin{aligned} |\mathbf{1}[S+T \le y] - \mathbf{1}[S \le y]| &= \mathbf{1}[S+T \le y < S] + \mathbf{1}[S \le y < S+T] \\ &\le \mathbf{1}[y - |T| < S \le y + |T|] \\ &\le \mathbf{1}[y - B < S \le y + B] + \mathbf{1}[|T| > B]. \end{aligned}$$

Applying this with $S = \theta^\top Z$ and $T = (\vartheta - \theta)^\top Z$, we obtain the inequality. Now assume that F is Hölder with exponent κ, $0 < \kappa \le 1$. Since F is bounded, it is also Hölder for any exponent in the interval $(0, \kappa)$. Using the Hölder property of F, we derive the following results using the above inequality.

(i) If $\|Z\|$ is bounded by say C_*, we take $B = \|\vartheta - \theta\| C_*$ and obtain (A1) with $\alpha = \kappa$.

(ii) If $\|Z\|$ has a finite moment of order $\nu \ge 2$, then we have

$$P(\|\vartheta - \theta\|\|Z\| > B) \le \|\vartheta - \theta\|^\nu E[\|Z\|^\nu]/B^\nu$$

and with $B = \|\vartheta - \theta\|^\beta$ and $\beta = \nu/(\kappa + \nu)$

$$J(y,z,\vartheta) \le |F(y) - F(z)| + 2L\|\vartheta - \theta\|^{\kappa\beta} + E[\|Z\|^\nu]\|\vartheta - \theta\|^{\kappa\beta},$$

where L is the Hölder constant. Since F is also Hölder with exponent $\kappa\beta$, we obtain (A1) with $\alpha = \kappa\beta = \kappa\nu/(\kappa + \nu)$. Note that α increases with ν and tends to κ as ν increases to infinity. For $\kappa = 1$ and $\nu = 2$, we have $\alpha = 2/3$.

Remark 3. Assume that Z has an elliptically contoured distribution conditionally given $\delta = 1$. This means that given $\delta = 1$, the random vector Z has the same distribution as $\mu + \Sigma\eta$ where μ is a vector in $\mathbb{R}^p$, Σ is a positive definite $p \times p$ matrix, and η is a spherically symmetric random vector, i.e., $Q\eta$ has the same distribution as η for each orthogonal $p \times p$ matrix Q. Then the random variables $\{u^\top \eta : \|u\| = 1\}$ have the same distribution function F. Thus we derive

$$\begin{aligned} E[\delta \mathbf{1}[\vartheta^\top Z \le y]] &= E[\delta P(\vartheta^\top Z \le y | \delta = 1)] \\ &= E[\delta] P(\vartheta^\top \mu + \vartheta^\top \Sigma\eta \le y) \\ &= E[\delta] F((y - \vartheta^\top \mu)/\|\Sigma\vartheta\|) \end{aligned}$$

for every non-zero vector ϑ in $\mathbb{R}^p$. It is now easy to see that (A1) holds with $\alpha = 1$ if F has a density f satisfying

$$\sup_{y \in \mathbb{R}} (1 + |y|) f(y) < \infty$$

and if θ is not the zero vector. This shows that in the setting of our simulations (A1) holds with $\alpha = 1$.

Remark 4. Let us briefly mention an example when (A2) fails to hold. Take δ identical to 1 and $Z = X = (X_1, X_2)^\top$, where X is a bivariate standard normal random vector. Then, for $\theta = (1,1)^\top$, the conditional distribution of X_1 given $\theta^\top X = X_1 + X_2$ is the same as that of X_2 given $X_1 + X_2$, and $E[X_1 | X_1 + X_2] = E[X_2 | X_1 + X_2] = (X_1 + X_2)/2$. Consequently (A2) is not met in this case.

Let us now discuss implications of our assumptions and in the process derive important quantities and results needed in the proof of the theorem. For $\vartheta \in \mathbb{R}^p$, we let G_ϑ denote the conditional distribution function of $\vartheta^\top Z$ given $\delta = 1$, i.e.,

$$G_\vartheta(s) = P(\vartheta^\top Z \le s | \delta = 1) = \frac{E[\delta \mathbf{1}[\vartheta^\top Z \le s]}{E[\delta]}, \quad s \in \mathbb{R},$$

and set

$$R_j(\vartheta) = G_\vartheta(\vartheta^\top Z_j), \quad j = 1, \dots, n.$$

Then (A1) implies

$$|R_j(\vartheta) - R_j(\theta)| \le \frac{K\|\vartheta - \theta\|^\alpha}{E[\delta]} (1 + \|Z_j\|^\alpha), \quad \vartheta \in \mathcal{N},$$

and we have

$$E[\delta_j(\bar{R}_j(\vartheta) - R_j(\vartheta))^2|X_j, \delta_1, \ldots, \delta_n] \le \frac{\delta_j}{N}$$

as the left hand side equals

$$\frac{\delta_j}{N^2}\Big(E[(1 - R_j(\vartheta))^2|X_j, \delta_j = 1] + \sum_{i\ne j}^{n} \delta_i E[R_j(\vartheta)(1 - R_j(\vartheta))|X_j, \delta_j = 1]\Big).$$

Thus, for every sequence θ_n such that $n^{1/2}(\theta_n - \theta)$ is bounded, we have

$$\frac{1}{n}\sum_{j=1}^{n} E[\delta_j(\bar{R}_j(\theta_n) - R_j(\theta))^2] = O(n^{-\alpha}). \tag{3.1}$$

Remark 5. It follows from (A1) that G_θ is continuous. Thus, given $\delta = 1$, the random variable $R = G_\theta(\theta^\top Z)$ is uniformly distributed on $[0, 1]$. This will be used in the next remark.

Remark 6. Let us now look at the dispersion matrix V_n of $\zeta_n = \delta\varepsilon v_{r_n}(R)$ and the covariance matrix A_n of ζ_n and $\xi = \delta\varepsilon W/\sigma^2(X)$. We have

$$V_n = E[\zeta_n\zeta_n^\top] = E[\delta\varepsilon^2 v_{r_n}(R)v_{r_n}^\top(R)] = E[\delta\sigma^2(X)v_{r_n}(R)v_{r_n}^\top(R)]$$

and

$$A_n = E[\zeta_n\xi^\top] = E[\frac{\delta\varepsilon^2}{\sigma^2(X)}v_{r_n}(R)W^\top] = E[\delta v_{r_n}(R)W^\top] = E[\delta v_{r_n}(R)Z^\top].$$

Since σ^2 takes values in a closed subinterval $[a, b]$ of $(0, \infty)$, we have

$$aE[\delta(u^\top v_{r_n}(R))^2] \le u^\top V_n u \le bE[\delta(u^\top v_{r_n}(R))^2]$$

and obtain from the third part of (C1) the inequality

$$ac_2E[\delta] \le \inf_{\|u\|=1} u^\top V_n u \le \sup_{\|u\|=1} u^\top V_n u \le bc_3E[\delta]. \tag{3.2}$$

In view of the identity $A_n = E[\delta v_{r_n}(R)W^\top]$, it follows from (A2) and the properties of V_n that A_n is eventually of full rank p. Finally, using condition (C2), one verifies as in Schick (2013) that

$$E[\|A_n^\top V_n^{-1}\zeta_n - \xi\|^2 \to 0$$

and obtains

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n} A_n^\top V_n^{-1}\delta_j\varepsilon_j v_{r_n}(R_j) = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{\delta_j\varepsilon_j}{\sigma^2(X_j)}\nu(\theta^\top Z_j) + o_P(1)$$

and the convergence

$$A_n^\top V_n^{-1} A_n \to H_*.$$

4. Proof of Theorem 1

Note that the random vector

$$\Gamma_n = \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{\delta_j\varepsilon_j}{\sigma^2(X_j)}Z_j$$

is asymptotically normal with mean vector zero and dispersion matrix H. As demonstrated in Peng and Schick (2012), the desired (1.1) now follows if one shows that the local log-empirical likelihood ratio

$$\mathscr{L}_n(t) = \log\frac{\mathscr{R}_n(\theta + n^{-1/2}t)}{\mathscr{R}_n(\theta)}, \quad t\in\mathbb{R}^p,$$

satisfies the expansion

$$\sup_{\|t\|\le 2C\log^{1/2} n}\frac{|\mathscr{L}_n(t) - t^\top\Gamma_n + (1/2)t^\top Ht|}{(1+|t|)^2} = o_P(1). \tag{4.1}$$

Thus we just need to prove (4.1). It suffices to prove the desired expansion with $\hat\theta_*$ replaced by a sequence θ_n which satisfies $n^{1/2}(\theta_n-\theta)$ is bounded. Indeed, if the result holds for each such sequence, it also holds for any discretized root-n consistent estimator of θ such as $\hat\theta_*$.

The empirical likelihood that takes the above into consideration is

$$\mathscr{R}_n(\vartheta) = \sup\Big\{\prod_{j=1}^{n} n\pi_j : \pi_1\ge 0,\dots,\pi_n\ge 0, \sum_{j=1}^{n}\pi_j = 1,$$
$$\sum_{j=1}^{n}\pi_j\delta_j(Y_j - \vartheta^\top Z_j)v_{r_n}(\bar R_{nj}) = 0\Big\}$$

with $\bar R_{nj} = \bar R_j(\theta_n)$. Abbreviate $R_j(\theta)$ by R_j. From (3.1) and the second inequality in (C1) we derive

$$\frac{1}{n}\sum_{j=1}^{n}E[\delta_j\|v_{r_n}(\bar R_{nj}) - v_{r_n}(R_j)\|^2] = O(r_n^3 n^{-\alpha}). \tag{4.2}$$

For $t \in \mathbb{R}^p$, we set

$$\mathbb{U}_{n,t} = \frac{1}{n}\sum_{j=1}^{n} \delta_j \Big(\varepsilon_j - \frac{t^\top Z_j}{\sqrt{n}}\Big) v_{r_n}(\bar{R}_{nj})$$

and

$$\mathbb{V}_{n,t} = \frac{1}{n}\sum_{j=1}^{n} \delta_j \Big(\varepsilon_j - \frac{t^\top Z_j}{\sqrt{n}}\Big)^2 v_{r_n}(\bar{R}_{nj}) v_{r_n}^\top(\bar{R}_{nj}).$$

In addition, we introduce

$$U_n = \frac{1}{\sqrt{n}}\sum_{j=1}^{n} \varepsilon_j v_{r_n}(R_j)$$

Let $C_n = 2C\log^{1/2} n$. As in the proof of Theorem 2 in Schick (2013), we derive that (4.1) is implied by the following three statements.

$$\sup_{\|t\|\le C_n} \frac{\|\mathbb{U}_{n,t} - U_n + A_n t\|}{1+\|t\|} = o_P(r_n^{-1/2}), \tag{4.3}$$

$$\sup_{\|t\|\le C_n} \sup_{\|u\|=1} |u^\top \mathbb{V}_{n,t} u - u^\top V_n u| = o_P(1/r_n), \tag{4.4}$$

$$\sup_{\|t\|\le C_n} \frac{|-2\log \mathscr{R}_n(\theta + n^{-1/2}t) - \mathbb{U}_{n,t}^\top \mathbb{V}_{n,t}^{-1} \mathbb{U}_{n,t}|}{(1+\|t\|)^2} = o_P(1). \tag{4.5}$$

Here A_n and V_n are as in Remark 6. These statements will be proved next.

Proof of (4.3). Let us set

$$\tilde{\mathbb{A}}_n = \frac{1}{n}\sum_{j=1}^{n} \delta_j v_{r_n}(\bar{R}_{nj}) Z_j^\top \quad \text{and} \quad \mathbb{A}_n = \frac{1}{n}\sum_{j=1}^{n} \delta_j v_{r_n}(R_j) Z_j^\top.$$

One verifies

$$\mathbb{U}_{n,t} - U_n + A_n t = \mathbb{U}_{n,0} - U_n - (\tilde{\mathbb{A}}_n - A_n)t$$

and finds, by conditioning on $\delta_1, X_1, \dots, \delta_n, X_n$,

$$\begin{aligned} E[\|\mathbb{U}_{n,0} - U_n\|^2] &= E\Big[\Big\|\frac{1}{\sqrt{n}}\sum_{j=1}^{n} \delta_j \varepsilon_j (v_{r_n}(\bar{R}_{nj}) - v_{r_n}(R_j))\Big\|^2\Big] \\ &= \frac{1}{n}\sum_{j=1}^{n} E[\delta_j \varepsilon_j^2 \|v_{r_n}(\bar{R}_{nj}) - v_{r_n}(R_j)\|^2] \\ &= \frac{1}{n}\sum_{j=1}^{n} E[\delta_j \sigma^2(X_j) \|v_{r_n}(\bar{R}_{nj}) - v_{r_n}(R_j)\|^2]. \end{aligned}$$

Since σ^2 is bounded, (4.2) yields

$$E[\|\mathbb{U}_{n,0} - U_n\|^2] = O(r_n^3 n^{-\alpha}) = o(1/r_n^2).$$

Thus, (4.3) follows if we verify $\|\tilde{\mathbb{A}}_n - A_n\| = o_P(r_n^{-1/2})$. We have

$$\begin{aligned}\|\tilde{\mathbb{A}}_n - \mathbb{A}_n\| &\le \frac{1}{n}\sum_{j=1}^{n} \delta_j \|(v_{r_n}(\bar{R}_{nj}) - v_{r_n}(R_j))Z_j^\top\| \\ &= \frac{1}{n}\sum_{j=1}^{n} \delta_j \|v_{r_n}(\bar{R}_{nj}) - v_{r_n}(R_j)\|\|Z_j\|.\end{aligned}$$

Thus an application of the Cauchy–Schwarz inequality and (4.2) yield

$$\|\tilde{\mathbb{A}}_n - \mathbb{A}_n\| = O_P((r_n^3 n^{-\alpha})^{1/2}) = o_P(1/r_n).$$

Finally, we have

$$\begin{aligned}E[\|\mathbb{A}_n - A_n\|^2] &\le E\Big[\Big\|\frac{1}{n}\sum_{j=1}^{n}(\delta_j v_{r_n}(R_j)Z_j^\top - E[\delta_j v_{r_n}(R_j)Z_j^\top\Big\|^2\Big] \\ &\le \frac{1}{n}E[\delta\|v_{r_n}(R)\|^2\|Z\|^2] \le \frac{c_0 r_n}{n}E[\|Z\|^2].\end{aligned}$$

Combining the above yields $\|\tilde{\mathbb{A}}_n - A_n\| = o_P(r_n^{-1})$.

Proof of (4.4). Note that V_n is the expected value of the random matrix

$$\mathbb{V}_n = \frac{1}{n}\sum_{j=1}^{n} \delta_j \varepsilon_j^2 v_{r_n}(R_j) v_{r_n}^\top(R_j).$$

Thus we derive

$$E[\|\mathbb{V}_n - V_n\|^2] \le \frac{1}{n}E[\delta\varepsilon^4\|v_{r_n}(R)\|^4] \le \frac{c_0^2 r_n^2}{n}E[\delta\varepsilon^4]$$

using the first inequality in (C1).

Let u be a unit vector in $\mathbb{R}^{r_n}$ and t be a vector in $\mathbb{R}^p$ with $\|t\| \le C_n$. Then we have

$$|u^\top \mathbb{V}_n u - u^\top V_n u| = |u^\top(\mathbb{V}_n - V_n)u| \le \|\mathbb{V}_n - V\| = O_p(r_n n^{-1/2})$$

and

$$\begin{aligned}|u^\top \mathbb{V}_{nt} u - u^\top \mathbb{V}_n u| &\le \frac{1}{n}\sum_{j=1}^{n}\delta_j\Big|\Big(\varepsilon_j - \frac{t^\top Z_j}{\sqrt{n}}\Big)^2 w_n^2(\bar{R}_{nj}) - \varepsilon_j^2 w_n^2(R_j)\Big| \\ &\le W_n + \frac{1}{n}\sum_{j=1}^{n}\delta_j\Big(\frac{2C_n}{\sqrt{n}}|\varepsilon_j|\|Z_j\| + \frac{C_n^2}{n}\|Z_j\|^2\Big)c_0 r_n\end{aligned}$$

with $w_n = u^\top v_{r_n}$ and

$$W_n = \frac{1}{n}\sum_{j=1}^n \delta_j \varepsilon_j^2 \left| w_n^2(\bar{R}_{nj}) - w_n^2(R_j) \right|.$$

Using the identity $a^2 - b^2 = (a-b)^2 + 2b(a-b)$ and then the Cauchy–Schwarz inequality we derive the bound

$$W_n \le D_n + (u^\top \mathbb{V}_n u D_n)^{1/2} \le D_n + (bc_3 + \|\mathbb{V}_n - V_n\|)^{1/2} D_n^{1/2}$$

with

$$D_n = \frac{1}{n}\sum_{j=1}^n \delta_j \varepsilon_j^2 (w_n(\bar{R}_{nj}) - w_n(R_j))^2 \le \frac{1}{n}\sum_{j=1}^n \delta_j \varepsilon_j^2 \|v_{r_n}(\bar{R}_{nj}) - v_{r_n}(R_j)\|^2.$$

With the aid of (4.2) we derive $E[D_n] = O_p(r_n^3 n^{-\alpha})$. The above show

$$\sup_{\|t\| \le C_n} \sup_{\|u\|=1} |u^\top \mathbb{V}_{n,t} u - u^\top V_n u| = O_p(C_n r_n n^{-1/2}) + O_p(r_n^{3/2} n^{-\alpha/2})$$

which is the desired result as $r_n^5 = o(n^\alpha)$.

Proof of (4.5). From (3.2) and (4.4) we derive that

$$\lambda_n = \inf_{\|t\| \le C_n} \inf_{\|u\|=1} u^\top \mathbb{V}_{nt} u \quad \text{and} \quad \Lambda_n = \sup_{\|t\| \le C_n} \sup_{\|u\|=1} u^\top \mathbb{V}_{nt} u$$

satisfy

$$P(\lambda_n > ac_2 E[\delta]/2) \to 1 \quad \text{and} \quad P(\Lambda_n < 2bc_3 E[\delta]) \to 1.$$

Since ε has a finite fourth moment and $\|Z\|$ has a finite second moment, we obtain the rates

$$M_{n1} = \max_{1 \le j \le n} |\varepsilon_j| = o_P(n^{1/4}) \quad \text{and} \quad M_{n2} = \max_{1 \le j \le n} |Z_j| = o_P(n^{1/2}).$$

It is easy to verify the rate

$$\sup_{\|t\| \le C_n} \frac{\|U_n - A_n t\|}{1 + \|t\|} \le \|U_n\| + \|A_n\| = O_p(r_n^{1/2}).$$

From this and (4.3) we then obtain

$$\mathbb{U}_{n,*} = \sup_{\|t\| \le C_n} \frac{\|\mathbb{U}_{n,t}\|}{1 + \|t\|} = O_p(r_n^{1/2}).$$

With

$$S_{jt} = \Big(\varepsilon_j - \frac{t^\top Z_j}{\sqrt{n}}\Big) v_{r_n}(\bar{R}_{nj}),$$

we derive

$$M_n = \sup_{\|t\| \le C_n} \max_{1 \le j \le n} \|S_{jt}\| \le (c_0 r_n)^{1/2}(M_{n1} + C_n n^{-1/2} M_{n2}) = o_P(r_n^{1/2} n^{1/4}),$$

$$S_n = \sup_{\|t\| \le C_n} \Big\| \frac{1}{n} \sum_{j=1}^n S_{jt} \Big\| \le \frac{1 + C_n}{\sqrt{n}} \mathbb{U}_{n,*} = O_p(C_n r_n^{1/2} n^{-1/2}),$$

$$T_n = \sup_{\|t\| \le C_n} \frac{1}{n} \sum_{j=1}^n \|S_{jt}\|^4 \le \frac{1}{n} \sum_{j=1}^n 8 c_0^2 r_n^2 \Big(|\varepsilon_j|^4 + \frac{C_n^4 M_{n2}^2}{n^2} \|Z_j\|^2 \Big) = O_p(r_n^2).$$

The above yield

$$P(\lambda_n - 5 M_n S_n > a c_2 E[\delta]/4) \to 1.$$

Thus the event $\{\lambda_n > 5 M_n S_n\}$ has probability tending to 1. On this event, we obtain as in Schick (2013) that the left-hand side of (4.5) is bounded by

$$\mathbb{U}_{n,*}^2 \left[\frac{S_n (\Lambda_n T_n)^{1/2}}{(\lambda_n - M_n S_n)^3} + \frac{4 \Lambda_n^2 S_n^2 T_n}{\lambda_n^2 (\lambda_n - M_n S_n)^4} \right],$$

which is of order $O_P(C_n r_n^{5/2} n^{-1/2} + C_n^2 r_n^4 / n) = o_P(1)$. This gives the desired result (4.5).

References

1. Bickel, P. J. (1982) On adaptive estimation. *Annals of Statistics* 10: 647-671. DOI: 10.1214/aos/1176345863

2. Carroll, R. J. (1982) Adapting for heteroscedasticity in linear models. *Annals of Statistics* 10: 1224-1233. DOI: 10.1214/aos/1176345987

3. LeCam, L. (1960) Locally asymptotically normal families of distributions. *University of California Publications in Statistics* 3: 37-98.

4. Little, R.J.A and Rubin, D.B. (2002) Statistical Analysis with Missing Data. Second edition. Wiley Series in Probability and Statistics. Hoboken: Wiley. DOI: 10.1002/0471704091.scard

5. Müller, H. G. and Stadtmüller, U. (1987) Estimation of heteroscedasticity in regression analysis. *Annals of Statistics* 15: 610-625.
DOI: 10.1214/aos/1176350364

6. Müller, U.U. and Van Keilegom, I. (2012) Efficient parameter estimation in regression with missing responses. *Electronic Journal of Statistics* 6: 1200-1219. DOI: 10.1214/12-EJS708

7. Peng, H. and Schick, A. (2012) Maximum empirical likelihood estimation an related topics. Preprint available at
http://www.math.binghamton.edu/anton/preprint.html.

8. R Core Team (2012) *R: A Language and environment for statistical computing*. R Fundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

9. Robinson, P. M. (1987) Asymptotically efficient estimation in the presence of heteroscedasticity of unknown form. *Econometrica* 55: 875-891.
DOI:10.2307/1911033

10. Schick, A. (1986) On asymptotically efficient estimation in semiparametric models. *Annals of Statistics* 14: 1139-1151. DOI: 10.1214/aos/1176350055

11. Schick, A. (1987) A note on the construction of asymptotically linear estimators. *Journal of Statistical Planning and Inference* 16: 89-105. Correction (1989) 22: 269-270. DOI: 10.1016/0378-3758(87)90059-0 and DOI: 10.1016/0378-3758(89)90118-3

12. Schick, A. (2013) Weighted least squares estimation with missing responses: An empirical likelihood approach. *Electronic Journal of Statistics* 7: 932-945. DOI: 10.1214/13-EJS793

13. Tsiatis, A.A. (2006) Semiparametric Theory and Missing Data. Springer Series in Statistics. New York: Springer. DOI: 10.1007/0-387-37345-4