Electronic Journal of Statistics
Vol. 7 (2013) 932–945
ISSN: 1935-7524
DOI: 10.1214/13-EJS793

# Weighted least squares estimation with missing responses: An empirical likelihood approach

**Anton Schick**[*]

*Binghamton University*
*Department of Mathematical Sciences*
*Binghamton, NY 13902-6000, USA*
*e-mail:* anton@math.binghamton.edu
*url:* www.math.binghamton.edu/anton

**Abstract:** A heteroscedastic linear regression model is considered where responses are allowed to be missing at random. An estimator is constructed that matches the performance of the weighted least squares estimator without the knowledge of the conditional variance function. This is usually done by constructing an estimator of the variance function. Our estimator is a maximum empirical likelihood estimator based on an increasing number of estimated constraints and avoids estimating the variance function.

**AMS 2000 subject classifications:** Primary 62F12, 62G05; secondary 62J05.
**Keywords and phrases:** Heteroscedastic linear regression, missing at random, maximum empirical likelihood estimation, estimated constraints, increasing number of constraints, efficiency.

Received November 2012.

## 1. Introduction

We consider a heteroscedastic linear regression model in which the response variable $Y$ is linked to a (one-dimensional) covariate $X$ by the formula

$$Y = \theta^\top m(X) + \varepsilon,$$

where $\theta$ is an unknown vector in $\mathbb{R}^d$, $m$ is a known measurable function from $\mathbb{R}$ to $\mathbb{R}^d$, the error variable $\varepsilon$ is conditionally centered, i.e., $E[\varepsilon|X] = 0$, and its conditional variance $\sigma^2(X) = E[\varepsilon^2|X]$ is bounded and bounded away from zero. We assume that the matrix

$$M = E[m(X)m^\top(X)]$$

is well defined and positive definite. This identifies $\theta$ as $M^{-1}E[m(X)Y]$ and implies that $E[\|m(X)\|^2]$ is finite. The model contains as special cases,

1. regression through the origin: $m(X) = X$;
2. simple linear regression: $m(X) = (1, X)^\top$;

[*] Supported by NSF Grant DMS 0906551.

3. polynomial regression: $m(X) = (1, X, \dots, X^{d-1})^\top$;
4. linear regression with a change in the slope at a known point $a$: $m(X) = (1, X, \max(0, X - a))^\top$; and
5. linear regression with a change in intercept and slope at a known point $a$: $m(X) = (\mathbf{1}[X \le a], X\mathbf{1}[X \le a], \mathbf{1}[X > a], X\mathbf{1}[X > a])^\top$.

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 to be missing. Then one observes $(\delta, X, \delta Y)$ with $\delta$ 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 [6]) and Tsiatis (2006 [17]). 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 denote the unobserved errors by

$$\varepsilon_j = Y_j - \theta^\top m(X_j), \quad j = 1, \dots, n.$$

A possible estimator of the regression parameter $\theta$ is the weighted least squares estimator $\hat\theta_w$ which minimizes the weighted sum of squares

$$Q_w(\vartheta) = \sum_{j=1}^n \delta_j w(X_j)(Y_j - \vartheta^\top m(X_j))^2, \quad \vartheta \in \mathbb{R}^d,$$

for some nonnegative measurable weight function $w$ which we require to be bounded and bounded away from zero. The ordinary least squares estimator corresponds to the choice $w = 1$. It is easy to see that the estimator $\hat\theta_w$ satisfies the stochastic expansion

$$\hat\theta_w = \theta + \frac{1}{n} \sum_{j=1}^n M_w^{-1} \delta_j w(X_j) m(X_j) \varepsilon_j + o_P(n^{-1/2})$$

with

$$M_w = E[\delta w(X) m(X) m^\top(X)] = E[\pi(X) w(X) m(X) m^\top(X)].$$

Note that this matrix is well defined and positive definite by the properties of $\pi$, $w$ and $M$. This implies that $n^{1/2}(\hat\theta_w - \theta)$ is asymptotically normal with mean vector 0 and dispersion matrix $D_w = M_w^{-1} S_w M_w^{-1}$ with

$$S_w = E[\delta w^2(X) \varepsilon^2 m(X) m^\top(X)] = E[\pi(X) w^2(X) \sigma^2(X) m(X) m^\top(X)].$$

With $\zeta = \delta m(X)\varepsilon/\sigma^2(X)$ and $\xi = \delta w(x) m(X)\varepsilon$, we can write $M_w = E[\zeta\xi^\top]$ and $S_w = E[\xi\xi^\top]$. It follows from the Cauchy-Schwarz inequality that the difference

$E[\zeta\zeta^\top] - D_w^{-1}$ is nonnegative definite and equals the zero matrix for $\xi = \zeta$. This shows that the asymptotic dispersion matrix $D_w$ is minimal for the choice $w = 1/\sigma^2$. Thus the (asymptotically) best estimator $\hat{\theta}_*$ in the class of weighted least squares estimators minimizes the weighted sum of squares

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

and satisfies the stochastic expansion

$$\hat{\theta}_* = \theta + \frac{1}{n} \sum_{j=1}^{n} H^{-1} \delta_j h(X_j) \varepsilon_j + o_P(n^{-1/2})$$

with

$$h(X) = \frac{1}{\sigma^2(x)} m(X) \quad \text{and} \quad H = E[\delta \varepsilon^2 h(X) h^\top(X)].$$

This implies that $n^{1/2}(\hat{\theta}_* - \theta)$ is asymptotically normal with mean vector 0 and dispersion matrix $H^{-1}$. Since $\sigma^2$ is unknown, the best weighted least squares estimator $\hat{\theta}_*$ is not available. For this reason we call $\hat{\theta}_*$ the oracle weighted least squares estimator.

Since $\sigma^2$ is unknown, a natural approach is to minimize instead of $Q(\vartheta)$ the weighted sum of squares

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

in which an estimator $\hat{\sigma}^2$ replaces the unknown $\sigma^2$. What is the behavior of such an estimator?

Carroll (1982 [1]) was the first to consider this problem. He treated the case of simple linear regression and with the responses fully observed, i.e, the case with $d = 2$, $m(X) = (1, X)^\top$, and $\delta = 1$. He used a regression kernel estimator based on the squared residuals from a least squares fit and showed that the resulting estimator $\hat{\theta}$ also satisfies the asymptotic expansion

$$\hat{\theta} = \theta + \frac{1}{n} \sum_{j=1}^{n} (E[\varepsilon^2 h(X) h^\top(X)])^{-1} h(X_j) \varepsilon_j + o_P(n^{-1/2}),$$

and hence is asymptotically equivalent to the oracle weighted least squares estimator $\hat{\theta}_*$ in the case $\delta = 1$. He proved this result under the assumptions that the covariate $X$ has a density which has compact support and is positive and twice continuously differentiable on its support, that $\sigma^2$ is continuously differentiable on this support, and that $\varepsilon$ has a finite sixth moment. Similar results were obtained by Müller and Stadtmüller (1987 [7]), Robinson (1987 [15]) and Schick (1987 [16]) for different estimators of $\sigma^2$ and under weaker conditions, all for the case $\delta = 1$. By the transfer principle for asymptotically linear estimators of Koul, Müller and Schick (2012 [5]) these results immediately carry to the present case with responses missing at random and yield that a minimizer $\hat{\theta}$ of $\vartheta \mapsto \hat{Q}(\vartheta)$ satisfies the expansion

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

Müller and Van Keilegom (2012 [8]) have demonstrated this expansion in the present setting by a direct argument using a kernel estimator of $\hat{\sigma}^2$. Their results show that an estimator $\hat{\theta}$ that satisfies (1.1) is even semiparametrically efficient in the sense of being a least dispersed regular estimator. This was already known in the case without missing responses, see Chamberlain (1987 [2]).

The goal of this paper is to show 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 $\sigma^2$. Indeed such an estimator can be constructed as a maximum empirical likelihood estimator of the empirical likelihood introduced by Peng and Schick (2012a [12]) modified to allow for missing responses. Let $\varphi_0, \varphi_1, \dots$ denote the trigonometric basis of $L_2(\mathscr{U})$ with $\mathscr{U}$ the uniform distribution on $[0,1]$ defined by $\varphi_0(x) = 1$ and

$$\varphi_k(x) = \sqrt{2}\cos(k\pi x), \quad x \in \mathbb{R}, k = 1, 2, \dots.$$

The modified version of the empirical likelihood of Peng and Schick (2012a [12]) is

$$\begin{aligned}\mathscr{R}_n(\vartheta) = \sup\Big\{ &\prod_{j=1}^{n} n\pi_j : \pi_1 \geq 0, \dots, \pi_n \geq 0, \sum_{j=1}^{n}\pi_j = 1,\\ &\sum_{j=1}^{n}\pi_j\delta_j(Y_j - \vartheta^\top m(X_j))v_n(\mathbb{G}(X_j)) = 0\Big\}, \quad \vartheta \in \mathbb{R}^d,\end{aligned}$$

where $v_n = (\varphi_0, \varphi_1, \dots, \varphi_{r_n})^\top$ is the vector consisting of the first $1 + r_n$ basis functions, $r_n \geq d$ is an integer tending to infinity slowly with the sample size $n$, and $\mathbb{G}$ is the empirical distribution function based on the covariates with observed responses only, i.e.,

$$\mathbb{G}(x) = \frac{\sum_{j=1}^{n}\delta_j \mathbf{1}[X_j \leq x]}{\sum_{j=1}^{n}\delta_j}, \quad x \in \mathbb{R}.$$

Note that $\mathbb{G}$ is an estimator of the conditional distribution function $G_1$ of $X$ given $\delta = 1$. The constraints in the above empirical likelihood try to capture the model assumption $E[\varepsilon|X] = 0$ which is equivalent to the linear constraints $E[\delta\varepsilon a(X)] = 0$, $a \in L_2(G_1)$. The latter is equivalent to the countably many constraints $E[\delta\varepsilon a_i(X)] = 0$, $i = 0, 1, 2, \dots$, with $a_0, a_1, \dots$ an orthonormal basis of $L_2(G_1)$. If $G_1$ is continuous, such a basis is given by $\{a_i = \varphi_i \circ G_1,\ i = 0, 1, 2, \dots\}$. The above empirical likelihood uses the empirical versions of the first $1+r_n$ linear constraints, with the unknown $G_1$ replaced by the estimator $\mathbb{G}$. By our assumption on $\pi$, the continuity of $G_1$ is equivalent to that of the distribution function $G$ of $X$.

To be precise, our estimator $\hat\theta$ is a *guided* maximum empirical likelihood estimator defined as a maximizer of the restriction of $\mathscr{R}_n$ to the random ball centered at the least squares estimator $\hat\theta_{\mathrm{L}}$ of radius $C(\log n/n)^{1/2}$ for some constant $C$,

$$\hat\theta = \mathop{\arg\max}_{n^{1/2}\|\vartheta-\hat\theta_{\mathrm{L}}\|\le C\log^{1/2}n} \mathscr{R}_n(\vartheta).$$

Guided maximum empirical likelihood estimation was introduced and studied by Peng and Schick (2012b [13]). We shall prove the following result.

**Theorem 1.** *Suppose the distribution function $G$ of $X$ is continuous, the error variable $\varepsilon$ has a finite fourth moment, and $r_n$ satisfies $r_n^5\log n = o(n)$. Then the guided maximum empirical likelihood estimator $\hat\theta$ satisfies the expansion (1.1) and hence is asymptotically efficient.*

As demonstrated in Peng and Schick (2012b [13]), this result follows if one shows that the local log-empirical likelihood ratio

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

satisfies the expansion

$$\sup_{\|t\|\le 2C\log^{1/2}n} \frac{|\mathscr{L}_n(t) - t^\top\Gamma_n + (1/2)t^\top Ht|}{(1+\|t\|)^2} = o_P(1), \tag{1.2}$$

with

$$\Gamma_n = \frac{1}{\sqrt{n}}\sum_{j=1}^n \delta_j h(X_j)\varepsilon_j.$$

Indeed, as $\Gamma_n$ is asymptotically normal with mean vector zero and dispersion matrix $H$, we obtain the desired result (1.1) as in Peng and Schick (2012b [13]). Thus Theorem 1 is a consequence of the following result.

**Theorem 2.** *Under the assumptions of Theorem 1, the uniform expansion (1.2) holds for every $C$.*

The empirical likelihood was introduced by Owen (1988 [9], 1990 [10]) to construct nonparametric confidence regions and tests, see also Owen (2001 [11]). Its scope has recently been extended. Hjort, McKeague and Van Keilegom (2009 [4]) treat constraints with nuisance parameters and also investigate the case when the number of constraints goes to infinity, but without nuisance parameters. The latter case was also considered by Chen, Peng and Qin (2009 [3]). Peng and Schick (2012a [12]) allow for nuisance parameters and for the number of constraints to go to infinity. Maximum empirical likelihood estimation was studied by Qin and Lawless (1994 [14]) and extended by Peng and Schick (2012b [13]) to allow for irregular constraints and nuisance parameters.

**Remark 1.** Fix $\vartheta$ in $\mathbb{R}^d$. Set

$$W_j(\vartheta) = (Y_j - \vartheta^\top m(X_j)) v_n(\mathbb{G}(X_j)), \quad j = 1, \dots, n.$$

It follows from Owen's work that $\mathscr{R}_n(\vartheta)$ equals

$$\prod_{j=1}^n \frac{1}{1 + \zeta(\vartheta)^\top \delta_j W_j(\vartheta)}$$

if there is a random vector $\zeta(\vartheta)$ for which $1+\zeta(\vartheta)^\top \delta_j W_j(\vartheta)$ is positive for every $j$ and the identity

$$\frac{1}{n}\sum_{j=1}^n \frac{\delta_j W_j(\vartheta)}{1 + \zeta(\vartheta)^\top \delta_j W_j(\vartheta)} = 0$$

holds, and equals zero otherwise. Moreover, such a random vector $\zeta(\vartheta)$ exists precisely on the event where the convex hull of the random vectors $\delta_1 W_1(\vartheta), \dots, \delta_n W_n(\vartheta)$ contains the origin as an interior point. For a subset $S$ of $\{1, \dots, n\}$, let $S_n$ denote the event that $\delta_j$ equals 1 for $j$ in $S$ and 0 for $j$ not in $S$. On the event $S_n$, $\mathscr{R}_n(\vartheta)$ equals

$$\prod_{j \in S} \frac{1}{1 + \zeta(\vartheta)^\top W_j(\vartheta)}$$

if the convex hull of $W_j(\vartheta)$, $j \in S$, contains the origin as interior point and equals zero otherwise. Thus, on $S_n$, the empirical likelihood $\mathscr{R}_n(\vartheta)$ equals

$$\mathscr{R}_S(\vartheta) = \sup\{\prod_{j \in S} n\pi_j : \pi_j \geq 0, \sum_{j \in S} \pi_j = 1, \sum_{j \in S} \pi_j W_j(\vartheta)\}.$$

On the event $S_n$, we also have $\mathbb{G}(x) = \sum_{j \in S} \mathbf{1}[X_j \leq x]/\mathrm{card}(S)$. This shows that the present empirical likelihood is the complete case version (as defined in Koul, Müller and Schick (2012 [5])) of the empirical likelihood in Peng and Schick (2012a [12]). Thus it can be calculated using the same program as in the case $\delta = 1$, but using only the fully observed data, i.e., the pairs $(X_j, Y_j)$ with $\delta_j = 1$.

**Remark 2.** One can show that the conclusions of our theorems remain true if we replace $\mathbb{G}$ by the empirical distribution function based on all the covariates. However, simulations not reported here show that our estimator behaves better in moderate sample sizes with the present choice $\mathbb{G}$.

**Remark 3.** The conclusions of our theorems remain true for choices of $v_n$ other than $v_n = (\phi_0, \dots, \phi_{r_n})^\top$. Indeed, in the proofs we rely only on the following properties of $v_n$.

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

$$\|v_n(x)\|^2 \leq c_0 r_n,$$
$$\|v_n(x) - v_n(y)\|^2 \leq c_1 r_n^3 |y - x|^2,$$
$$c_2 \leq \int (u^\top v_n)^2 \, d\mathscr{U} \leq c_3,$$

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

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

$$\inf_{b\in\mathbb{R}^{r_n+1}} \int (b^\top v_n - g)^2 \, d\mathscr{U} \to 0.$$

Thus any other choice of $v_n$ with these properties can be used. One possible choice is $v_n = (\psi_{r_n,0}, \dots, \psi_{r_n,r_n})^\top$, where

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

For this choice, (C1) holds with $c_0 = 1$, $c_1 = 2$, $c_2 = 1/6$ and $c_3 = 1$. For example, to obtain the last inequality in (C1), we observe that $\int \psi_{r,i}^2 \, d\mathscr{U}$ equals $1/3$ for $i = 0, r$ and $2/3$ for $i = 1, \dots, r-1$, and $\int \psi_{r,i}\psi_{r,j} d\mathscr{U}$ equals $1/6$ for $|i-j| = 1$ and $0$ for $|i-j| > 1$. From this we conclude the identity

$$6 \int (u^\top v_n)^2 \, d\mathscr{U} = u_0^2 + u_{r_n}^2 + \sum_{i=1}^{r_n-1} 2u_i^2 + \sum_{i=1}^{r_n} (u_{i-1} + u_i)^2,$$

for any vector $u = (u_0, \dots, u_{r_n})^\top$ in $\mathbb{R}^{r_n+1}$, and see that $\int (u^\top v_n)^2 \, d\mathscr{U}$ is bounded from above by $|u|^2$ and from below by $|u|^2/6$. Note that the functions $\psi_{r,0}, \dots, \psi_{r,r}$ form a basis for the set of linear splines with knots at the equidistant points $0/r, 1/r, \dots, 1$. We obtain (C2) because the continuous functions are dense in $L_2(\mathscr{U})$ and because for every continuous function $g$ on $[0,1]$ the inequality

$$\inf_{b\in\mathbb{R}^{r+1}} \Big| g(x) - \sum_{i=0}^{r} b_i \psi_{r,i}(x) \Big| \le \sup\{|g(y) - g(x)| : |y - x| \le 1/r\}$$

holds for all $x$ in $[0,1]$. To see the latter chose $b_i$ such that $\sum_{i=0}^{r} b_i \psi_{r,i}(x) = g(x)$ holds for $x = 0, 1/r, \dots, 1$.

As an illustration of our method we performed a small simulation study using R and compared several estimators, the OLSE, the oracle WLSE, and the GMELE for the choices $r_n = 2, 3, 4$, which we denote by GS(2), GS(3) and GS(3). We report the results for $v_n = (\psi_{r_n,0}, \dots, \psi_{r_n,r_n})^\top$ as they are slightly better than those for the choice $v_n = (\varphi_0, \dots, \varphi_{r_n})^\top$. Following Müller and Van Keilegom (2012 [8]), we took $X$ uniformly distributed on $(-1,1)$, $m(X) = X$, $\theta = 2$, $\pi(X) = 1/(1 + \exp(-X))$, and $\varepsilon = \sigma(X)Z$, with $Z$ standard normal and independent of $X$. In addition to the choices (a) $\sigma^2(X) = .6 - .5X$ and (b) $\sigma^2(X) = (X - .4)^2 + .1$ used in Müller and van Keilegom (2012 [8]), we also looked at (c) $\sigma^2(X) = \exp(-2X)$. We took the constant $C$ in the definition of the GMELE to be the product of $(n/\log n)^{1/2}/(N/\log N)^{1/2}$ with $N = \sum_{j=1}^n \delta_j$ and an estimator of the asymptotic standard deviation of the OLSE, namely

$$\frac{\left(\frac{1}{N}\sum_{j=1}^n \delta_j (Y_j - \hat\theta_{\mathrm{L}} X_j)^2 X_j^2\right)^{1/2}}{\frac{1}{N}\sum_{j=1}^n \delta_j X_j^2}.$$

Table 1
*Simulated Mean Square Errors*

| | $n$ | OLSE | WLSE | GS(2) | GS(3) | GS(4) |
|---|---|---|---|---|---|---|
| (a) | 50 | 0.0597 | 0.0342 | 0.0488 | 0.0520 | 0.0588 |
| | 100 | 0.0284 | 0.0158 | 0.0186 | 0.0189 | 0.0201 |
| | 200 | 0.0138 | 0.0075 | 0.0083 | 0.0082 | 0.0082 |
| (b) | 50 | 0.0796 | 0.0376 | 0.0523 | 0.0543 | 0.0638 |
| | 100 | 0.0387 | 0.0183 | 0.0247 | 0.0223 | 0.0238 |
| | 200 | 0.0187 | 0.0088 | 0.0114 | 0.0097 | 0.0100 |
| (c) | 50 | 0.2079 | 0.0396 | 0.0588 | 0.0655 | 0.0833 |
| | 100 | 0.0980 | 0.0185 | 0.0208 | 0.0212 | 0.0224 |
| | 200 | 0.0491 | 0.0093 | 0.0098 | 0.0098 | 0.0100 |

Each entry is the simulated mean square error of the corresponding estimator, for three sample sizes and three choices of $\sigma^2$: (a) $\sigma^2(X) = .6 - .5X$; (b) $\sigma^2(X) = (X - .4)^2 + .1$; (c) $\sigma^2(X) = \exp(-2X)$.

Table 1 reports the simulated mean square errors for the sample sizes $n = 50$, $n = 100$ and $n = 200$ based on 10000 simulations. We observe that our proposed estimator performs better than the OLSE in all cases considered. This is also true for the estimator with $v_n = (\varphi_0, \dots, \varphi_{r_n})$. The performance of our estimator comes closer to that of the oracle WLSE as the sample sizes increases. As expected, the performance is better for smaller $r_n$ if the sample size is small. Our results for the first two cases are similar to the ones reported by Müller and van Keilegom (2012 [8]). Case (c) shows that the improvements over the OLSE provided by the oracle WLSE estimator and our estimator can be quite dramatic.

## 2. Proof of Theorem 2

Let $G_1$ denote the conditional distribution function of $X$ given $\delta = 1$. It has density $\pi/E[\delta]$ with respect to $G$ and is hence continuous. Since $Y$ is missing at random, the conditional distribution of $Y$ given $X$ and $\delta = 1$ equals the conditional distribution of $Y$ given $X$. Thus

$$E[\varepsilon|X, \delta = 1] = E[\varepsilon|X] = 0 \quad \text{and} \quad E[\varepsilon^2|X, \delta = 1] = E[\varepsilon^2|X] = \sigma^2(X).$$

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

$$\begin{aligned}\hat{Z}_j(t) &= \delta_j(Y_j - (\theta + n^{-1/2}t)^\top m(X_j))v_n(\mathbb{G}(X_j)) \\ &= \delta_j\varepsilon_j v_n(\mathbb{G}(X_j)) - n^{-1/2}\delta_j v_n(\mathbb{G}(X_j))m^\top(X_j)t, \quad j = 1, \dots, n,\end{aligned}$$

$$\hat{U}_{n,t} = \frac{1}{\sqrt{n}}\sum_{j=1}^n \hat{Z}_j(t) \quad \text{and} \quad \hat{V}_{n,t} = \frac{1}{n}\sum_{j=1}^n \hat{Z}_j(t)\hat{Z}_j^\top(t).$$

We also set

$$U_n = \frac{1}{\sqrt{n}}\sum_{j=1}^n \delta_j\varepsilon_j v_n(G_1(X_j)),$$

$$V_n = E[\delta\varepsilon^2 v_n(G_1(X))v_n^\top(G_1(X))] = E[\delta\sigma^2(X)v_n(G_1(X))v_n^\top(G_1(X))]$$

and

$$A_n = E[\delta v_n(G_1(X))m^\top(X)] = E[\delta\varepsilon^2 v_n(G_1(X))h^\top(X)].$$

The functions $\varphi_0, \varphi_1, \dots$ form an orthonormal basis of $L_2(\mathscr{U})$. By the continuity of $G_1$, the random variable $G_1(X)$ has conditional distribution $\mathscr{U}$ given $\delta = 1$. This shows that the matrix $E[\delta v_n(G_1(X))v_n^\top(G_1(X))]$ equals $E[\delta]I_{1+r_n}$, where $I_m$ denotes the $m \times m$ identity matrix. Since $\sigma^2(X)$ is bounded and bounded away from zero, there are constants $0 < k < K < \infty$ such that $ku^\top I_{1+r_n}u < u^\top V_n u < Ku^\top I_{1+r_n}u$ for $u \in \mathbb{R}^{1+r_n}$. Thus we have

$$k \le \inf_{\|u\|=1} u^\top V_n u \le \sup_{\|u\|=1} u^\top V_n u \le K, \tag{2.1}$$

i.e., the eigenvalues of $V_n$ belong to $[k, K]$, and $V_n$ is invertible.

Let $C_n = 2C\log^{1/2} n$. We begin by proving that the desired result follows from the following three statements.

$$\sup_{\|t\|\le C_n} \frac{\|\hat{U}_{n,t} - U_n + A_n t\|}{1 + \|t\|} = O_p(r_n^{3/2}n^{-1/2}), \tag{2.2}$$

$$\sup_{\|t\|\le C_n} \sup_{\|u\|=1} |u^\top \hat{V}_{n,t}u - u^\top V_n u| = o_p(1/r_n), \tag{2.3}$$

$$\sup_{\|t\|\le C_n} \frac{|-2\log\mathscr{R}_n(\theta + n^{-1/2}t) - \hat{U}_{n,t}^\top \hat{V}_{n,t}^{-1}\hat{U}_{n,t}|}{(1+\|t\|)^2} = o_P(1). \tag{2.4}$$

In view of the inequalities $E[\|U_n\|^2] = \operatorname{trace}(V_n) \le K(1 + r_n)$ and

$$|u^\top A_n t|^2 \le E[\|\delta u^\top v_n(G_1(X))\|^2]E[\|m(X)\|^2]\|t\|^2 \le K\|u\|^2 E[\|m(X)\|^2]\|t\|^2,$$

we have the rate

$$\sup_{\|t\|\le C_n} \frac{\|U_n - A_n t\|^2}{(1+\|t\|)^2} = O_P(r_n). \tag{2.5}$$

This and (2.2) give the rate

$$\sup_{\|t\|\le C_n} \frac{\|\hat{U}_{n,t}\|^2}{(1+\|t\|)^2} = O_P(r_n). \tag{2.6}$$

From (2.1) and (2.3) we derive

$$\sup_{\|t\|\le C_n} \sup_{\|u\|=1} |u^\top \hat{V}_{n,t}^{-1}u - u^\top V_n^{-1}u| = o_p(1/r_n). \tag{2.7}$$

The statements (2.6) and (2.7) imply

$$\sup_{\|t\|\le C_n} \frac{|\hat{U}_{n,t}^\top \hat{V}_{n,t}^{-1}\hat{U}_{n,t} - \hat{U}_{n,t}^\top V_n^{-1}\hat{U}_{n,t}|}{(1+\|t\|)^2} = o_P(1). \tag{2.8}$$

From (2.1), (2.2), (2.5) and (2.6) we derive

$$\sup_{\|t\|\le C_n} \frac{|\hat{U}_{n,t}^\top V_n^{-1}\hat{U}_{n,t} - (U_n - A_n t)^\top V_n^{-1}(U_n - A_n t)|}{(1+\|t\|)^2} = o_P(1). \tag{2.9}$$

The statements (2.4), (2.8) and (2.9) yield

$$\sup_{\|t\|\le C_n} \frac{|\mathscr{L}_n(t) - t^\top A_n^\top V_n^{-1}U_n + (1/2)t^\top A_n^\top V_n^{-1}A_n t|}{(1+\|t\|)^2} = o_P(1). \tag{2.10}$$

Let $g$ be in $L_2(G_1)$. Then $g\circ G_1^{-1}$ belongs to $L_2(\mathscr{U})$ where $G_1^{-1}$ is the left-inverse of $G_1$. Conditioning first on $X$ and $\delta$ and then on $\delta$ we find

$$E[(\delta\varepsilon b^\top v_n(G_1(X)) - \delta\varepsilon g(X)))^2] \le BE[\delta]\int (b^\top v_n\circ G_1 - g)^2\, dG_1.$$

Here $B$ is a bound on $\sigma^2$. Since the functions $\varphi_0, \varphi_1, \ldots$ form an orthonormal basis of $L_2(\mathscr{U})$, we have

$$\inf_{b\in\mathbb{R}^{r_n+1}} \int (b^\top v_n\circ G_1 - g)^2\, dG_1 = \inf_{b\in\mathbb{R}^{r_n+1}} \int (b^\top v_n - g\circ G_1^{-1})^2\, d\mathscr{U} \to 0.$$

This implies

$$\inf_{b\in\mathbb{R}^{r_n+1}} E[(\delta\varepsilon b^\top v_n(G_1(X)) - \delta\varepsilon g(X)))^2] \to 0.$$

The left hand side is minimized by $b = V_n^{-1}E[\delta\varepsilon^2 g(X)v_n(G_1(X))]$. Thus we obtain

$$E[\|\delta\varepsilon A_n^\top V_n^{-1}v_n(G_1(X)) - \delta\varepsilon h(X)\|^2] \to 0.$$

From this we conclude $A_n^\top V_n^{-1}U_n = \Gamma_n + o_p(1)$ and $A_n^\top V_n^{-1}A_n \to H$ and obtain the desired result (1.2) in view of (2.10). This completes the proof that the statements (2.2)–(2.4) imply the desired result.

**Proof of (2.2).** It is easy to check that $\|v_n\|^2 \le 1 + 2r_n$ and $\|v_n'\|^2 \le 2\pi^2 r_n^3$. Let $N = \sum_{i=1}^n \delta_i$. We have $E[\delta_j/N] \le 1/n$ for $j = 1, \ldots, n$. Using this we obtain the inequalities

$$E[\delta_j a^2(X_j)(\mathbb{G}(X_j) - G_1(X_j))^2] \le \int a^2\, dG_1/n, \quad j = 1, \ldots, n,$$

valid for $a \in L_2(G_1)$. Indeed, conditioning on $\delta_1, \ldots, \delta_n$ and $X_j$ we derive that the left-hand side equals

$$E[\delta_j a^2(X_j)[(1 - G_1(X_j))^2 + (N - \delta_j)G_1(X_j)(1 - G_1(X_j))]/N^2]$$

and is therefore bounded by $E[\delta_j a^2(X_j)/N] \le \int a^2\, dG_1/n$.

Using these inequalities we verify the statements

$$T_{n1} = \frac{1}{n}\sum_{j=1}^n \delta_j\varepsilon_j^2\|v_n(\mathbb{G}(X_j)) - v_n(G_1(X_j))\|^2 = O_p(r_n^3/n),$$

$$T_{n2} = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \delta_j \varepsilon_j \Big[ v_n(\mathbb{G}(X_j)) - v_n(G_1(X_j)) \Big] = O_p(r_n^{3/2} n^{-1/2}),$$

$$T_{n3} = \frac{1}{n} \sum_{j=1}^{n} \delta_j v_n(\mathbb{G}(X_j)) m^\top(X_j) - A_n = O_p(r_n^{3/2} n^{-1/2}).$$

Indeed, we find $E[\|T_{n2}\|^2] = E[T_{n1}] \le 2\pi^2 r_n^3 \int \sigma^2 \, dG_1/n$ and

$$\begin{aligned} E(\|T_{n3}\|^2) \le 2E\Big[ \frac{1}{n} \sum_{j=1}^{n} \delta_j \|m(X_j)\|^2 \|v_n(\mathbb{G}(X_j)) - v_n(G_1(X_j))\|^2 \Big] \\ + 2E\Big[ \Big\| \frac{1}{n} \sum_{j=1}^{n} \delta_j v_n(G_1(X_j)) m^\top(X_j) - E[\delta v_n(G_1(X) m^\top(X)]) \Big\|^2 \Big] \\ \le 4\pi^2 r_n^3 \frac{1}{n} \sum_{j=1}^{n} E[\delta_j \|m(X_j)\|^2 (\mathbb{G}(X_j)) - G_1(X_j))^2] \\ + \frac{2}{n} E[\delta \|m(X)\|^2 \|v_n(G_1(X))\|^2] \\ \le (4\pi^2 r_n^3 + 2 + 4r_n) \int \|m\|^2 \, dG_1/n. \end{aligned}$$

Verify that $\hat{U}_{n,t} - U_n + A_n t$ equals $T_{n2} + T_{n3} t$. Thus (2.2) follows from the bound

$$\sup_{\|t\| \le C_n} \frac{\|\hat{U}_{n,t} - U_n + A_n t\|^2}{(1 + \|t\|)^2} \le 2\|T_{n2}\|^2 + 2\|T_{n3}\|^2 = O_p(r_n^3/n).$$

**Proof of (2.3).** We verify (2.3) by establishing

$$\sup_{\|u\|=1} |u^\top \bar{V}_n u - u^\top V_n u| = O_p(r_n n^{-1/2}) \tag{2.11}$$

and

$$\sup_{\|t\| \le C_n} \sup_{\|u\|=1} |u^\top \hat{V}_{n,t} u - u^\top \bar{V}_n u| = O_p(r_n^{3/2} n^{-1/2} + C_n r_n^{1/2} n^{-1/2}) \tag{2.12}$$

with

$$\bar{V}_n = \frac{1}{n} \sum_{j=1}^{n} Z_j Z_j^\top \quad \text{and} \quad Z_j = \delta_j \varepsilon_j v_n(G_1(X_j)).$$

We obtain (2.11) since its left-hand side is bounded by the euclidean norm $\|\bar{V}_n - V_n\|$ of $\bar{V}_n - V_n$ and since we have the bound

$$\begin{aligned} nE[\|\bar{V}_n - V_n\|^2] \le \sum_{k=0}^{r_n} \sum_{l=0}^{r_n} E[\delta \varepsilon^4 \varphi_k^2(G_1(X)) \varphi_l^2(G_1(X))] \\ = E[\delta \varepsilon^4 \|v_n(G_1(X))\|^4] \le (1 + 2r_n)^2 E[\delta \varepsilon^4] = O(r_n^2). \end{aligned}$$

In view of the identity $u^\top \hat{V}_{n,t} u - u^\top \bar{V}_n u = \frac{1}{n}\sum_{j=1}^n [(u^\top \hat{Z}_j(t))^2 - (u^\top Z_j)^2]$, we see that the left-hand side of (2.12) is bounded by $2(D_n L_n)^{1/2} + D_n$ with

$$L_n = \sup_{\|u\|=1} u^\top \bar{V}_n u \quad \text{and} \quad D_n = \sup_{\|t\|\le C_n} \frac{1}{n}\sum_{j=1}^n \|\hat{Z}_j(t) - Z_j\|^2.$$

We have $L_n = O_p(1)$ by (2.1) and (2.11), and

$$D_n \le 2T_{n1} + \frac{2C_n^2(1+2r_n)}{n^2}\sum_{j=1}^n \delta_j \|m(X_j)\|^2 = O_p(r_n^3/n) + O_p(C_n^2 r_n/n).$$

Thus (2.12) holds.

**Proof of (2.4).** To verify (2.4) we shall use the following result which is a special case of Lemma 5.2 of Peng and Schick (2012a [12]).

**Lemma 1.** *Let $x_1, \dots, x_n$ be $m$-dimensional vectors. Set*

$$\bar{x} = \frac{1}{n}\sum_{j=1}^n x_j, \quad x^* = \max_{1\le j\le n} \|x_j\|, \quad \nu_4 = \frac{1}{n}\sum_{j=1}^n \|x_j\|^4, \quad S = \frac{1}{n}\sum_{j=1}^n x_j x_j^\top,$$

*and let $\lambda$ denote the smallest and $\Lambda$ the largest eigenvalue of the matrix $S$. Then the inequality $\lambda > 5|\bar{x}|x^*$ implies*

$$\Big| -2\log(\mathscr{R}) - n\bar{x}^\top S^{-1}\bar{x}\Big| \le \frac{n\|\bar{x}\|^3(\Lambda\nu_4)^{1/2}}{(\lambda - \|\bar{x}\|x^*)^3} + \frac{4n\Lambda^2\|\bar{x}\|^4\nu_4}{\lambda^2(\lambda - \|\bar{x}\|x^*)^4} \tag{2.13}$$

*where*

$$\mathscr{R} = \sup\Big\{\prod_{j=1}^n n\pi_j : \ \pi_1 \ge 0, \dots, \pi_n \ge 0, \sum_{j=1}^n \pi = 1, \sum_{i=1}^n \pi_i x_i = 0\Big\}.$$

The bound (2.13) is derived from (5.7) in Peng and Schick (2012a [12]) and the inequalities $(x^{(3)})^2 \le \Lambda x^{(4)}$ and $nx^{(4)} \le \sum_{j=1}^n \|x_j\|^4$.

Let $\hat{\lambda}_{n,t}$ denote the smallest, and $\hat{\Lambda}_{n,t}$ the largest eigenvalue of $\hat{V}_{n,t}$. Let us also set

$$\hat{\lambda}_n = \inf_{\|t\|\le C_n} \hat{\lambda}_{n,t}, \qquad \hat{\Lambda}_n = \sup_{\|t\|\le C_n} \hat{\Lambda}_{n,t}, \quad \text{and} \quad \hat{Z}_n^* = \sup_{\|t\|\le C_n} \max_{1\le j\le n} \|\hat{Z}_j(t)\|.$$

It follows from (2.1) and (2.3) that

$$P(\hat{\lambda}_n > k/2) \to 1 \quad \text{and} \quad P(\hat{\Lambda}_n < 2K) \to 1.$$

Since $\varepsilon$ has a finite fourth moment and $\|m(X)\|$ has a finite second moment, we have

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

and find

$$\hat{Z}_n^* \le (1+2r_n)^{1/2}(M_{n1} + C_n n^{-1/2} M_{n2}) = o_p(r_n^{1/2} n^{1/4}).$$

From (2.6) we obtain

$$\xi_n = \sup_{\|t\|\le C_n} \Big\| \frac{1}{n} \sum_{j=1}^n \hat{Z}_j(t) \Big\| = n^{-1/2} \sup_{\|t\|\le C_n} \|\hat{U}_{n,t}\| = O_p(C_n r_n^{1/2} n^{-1/2}).$$

The bound

$$T_n = \sup_{\|t\|\le C_n} \frac{1}{n} \sum_{j=1}^n \|\hat{Z}_j(t)\|^4 \le (1+2r_n)^2 \frac{1}{n} \sum_{j=1}^n \Big( |\varepsilon_j| + C_n n^{-1/2} \|m(X_j)\| \Big)^4$$

yields $T_n = O_p(r_n^2)$. The above show that

$$P(\hat{\lambda}_n - 5\hat{Z}_n^* \xi_n > k/4) \to 1.$$

Thus the event $\{\hat{\lambda}_n > 5\hat{Z}_n^*\xi_n\}$ has probability tending to 1. On this event the left-hand side of (2.4) is bounded by

$$\sup_{\|t\|\le C_n} \frac{\|\hat{U}_{n,t}\|^2}{(1+\|t\|)^2} \left[ \frac{\xi_n (\hat{\Lambda}_n T_n)^{1/2}}{(\hat{\lambda}_n - \xi_n \hat{Z}_n^*)^3} + \frac{4\hat{\Lambda}_n^2 \xi_n^2 T_n}{\hat{\lambda}_n^2 (\hat{\lambda}_n - \xi_n \bar{Z}_n^*)^4} \right]$$

which is of order $O_P(C_n r_n^{5/2} n^{-1/2} + C_n^2 r^4/n) = o_P(1)$. This gives the desired result (2.4).

## References

[1] Carroll, R. J. (1982). Adapting for heteroscedasticity in linear models. *Ann. Statist.* **10** 1224–1233. MR0673657

[2] Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econometrics* **34** 305–334. MR0888070

[3] Chen, S.X., Peng, L. and Qin, Y.-L. (2009). Effects of data dimension on empirical likelihood. *Biometrika* **96** 711–722. MR2538767

[4] Hjort, N.L., McKeague, I.W. and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *Ann. Statist.* **37** 1079–1111. MR2509068

[5] Koul, H., Müller, U.U. and Schick, A. (2012). The transfer principle: a tool for complete case analysis. *Ann. Statist.* **60** 3031–3047.

[6] Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data.* Second edition. Wiley Series in Probability and Statistics, Wiley, Hoboken. MR1925014

[7] Müller, H. G. and Stadtmüller, U. (1987). Estimation of heteroscedasticity in regression analysis. *Ann. Statist.* **15** 610–625. MR0888429

[8] Müller, U.U. and Van Keilegom, I. (2012). Efficient parameter estimation in regression with missing responses. *Electron. J. Statist.* **6** 1200–1219. MR2988444

[9] Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249. MR0946049

[10] Owen, A. B. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.* **18** 90–120. MR1041387

[11] Owen, A.B. (2001). *Empirical Likelihood*. Chapman & Hall/CRC.

[12] Peng, H. and Schick, A. (2012a). Empirical likelihood approach to goodness of fit testing. To appear in *Bernoulli*.

[13] Peng, H. and Schick, A. (2012b). Maximum empirical likelihood estimation and related topics. Preprint available at `http://math.binghamton.edu/anton/preprint.html`.

[14] Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.* **22** 300–325. MR1272085

[15] Robinson, P. M. (1987). Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica* **55** 875–891. MR0906567

[16] Schick, A. (1987). A note on the construction of asymptotically linear estimators. *J. Statist. Plann. Inference* **16**, 89–105. MR0887419 Correction (1989) **22** 269–270. MR1004355

[17] Tsiatis, A.A. (2006). *Semiparametric Theory and Missing Data.* Springer Series in Statistics. Springer, New York. MR2233926