Nonparametric Imputation of Missing Values for Estimating Equation Based Inference[1]

DONG WANG and SONG XI CHEN

Dep rt ent of St tistics, o St te ni ersity, A es, A

d ng i st te ed **and** *songchen i st te ed*

MMA Y

We propose a nonparametric imputation procedure for data with missing values and establish an empirical likelihood inference for parameters de ned by general estimating equations. The imputation is carried out multiple times via a nonparametric estimator of the conditional distribution of the missing variable given the always observable variable. The empirical likelihood is used to construct a pro le likelihood for the parameter of interest. We demonstrate that the propo the8d05ametric aluesthe7eterthat thealuesthat the likelihoforableapdc

Examples include the e ort by $genenet_{,}or \quad org$ **and other data depositories to combine genetics,**

advantage of being more robust against model misspeci cation, although a correct model for the conditional distribution of the missing variable given the observed variable is often needed to attain the semiparametric e ciency bound.

Nonparametric methods have also been proposed for missing data. Titterington (1977) and Titterington and Mill (1983) considered kernel estimation of a multivariate density for data with incomplete observations. When the parameter of interest is the mean of a response variable which is subject to missingness, Cheng (1994) proposed using the kernel conditional mean estimator to impute the missing values. Hahn (1998) and Hirano, Imbens and Ridder (2003) studied the estimation of average treatment e ects using nonparametrically estimated propensity scores. Since in treatment e ect problems, the response of any unit can only be observed under one treatment, it is seen as a missing data problem. In survey statistics, Kim and Fuller (2004) proposed the fractional hot deck imputation method, in which multiple values are drawn from the same imputation cell as the missing observation, and a weight is assigned to each imputed value.

In this paper we consider estimation of parameters de ned by a set of estimating equations in the presence of missing values. Estimating equation (Godambe, 1991; Boos, 1992) is a very general framework for statistical inference. When the parameter of interest is not directly related to the mean, the commonly used conditional mean based imputation via either a parametric (Yates, 1933) or nonparametric (Cheng, 1994) regression estimator may results in either biased estimation or reduced e ciency. This is especially the case for data with missing covariates.

We propose in this paper a nonparametric imputation procedure that generates multiple copies of missing values from a kernel estimator of the conditional distribution of the missing variables given the fully observable variables. This model-free (nonparametric) imputation is particularly suited for the wide range of parameters de ned by estimating

equations. We then employ Owen (1988, 1990)'s empirical likelihood to formulate a non-parametric pro le likelihood based on an extended sample which consists of the original data and the nonparametrically imputed values. Our use of empirical likelihood is largely encouraged by its attractive inferential features for estimating equations when there is no missing values by Qin and Lawless (1994), as well as its inference for a

Let $\delta_i = 1$ if Y_i is observed and $\delta_i = 0$ if Y_i is missing. Like Cheng (1994), Wang and Rao (2002) and others, we assume that δ and Y are conditionally independent given X, namely the strongly ignorable missing at random proposed by Rosenbaum and Rubin (1983). As a result,

$$P(\delta = 1 \mid Y; X) = P(\delta = 1 \mid X) =: p(X)$$

where p(x) is called the propensity score and prescribes selection bias in the missingness if p(x) is not a constant function.

Let $F(y|X_i)$ be the conditional distribution of Y given $X = X_i$. Let

$$\hat{F}(y|X_i) = \sum_l \frac{\delta_l W(\frac{X_l - X_i}{h}) I(Y_i \le y)}{\sum_j^n \delta_j W(\frac{X_j - X_i}{h})} \tag{1}$$

be a kernel estimator of $F(y|X_i)$ based on the completely observed portion (no missing values) of the sample. Here $W(\cdot)$ is a d_x-dimensional kernel function; h is a smoothing bandwidth satisfying $\sqrt{n}h \to 0$ and $nh^{d_x} \to \infty$ as $n \to \infty$; and $I(\cdot)$ is the d_y-dimensional indicator function. Here we concentrate on the case that both X and Y are continuous random variables. Extension to discrete random variables can be readily made; see Section 5 for a case of binary random variable.

We propose to impute a missing Y_i with a $\tilde{Y}_i$ randomly generated from the estimated conditional distribution $\hat{F}(y|X_i)$. E ectively $\tilde{Y}_i$ has a discrete distribution where the probability of selecting a Y_l with $\delta_l = 1$ is

$$\frac{W\{(X_l - X_i)/h\}}{\sum_j^n \delta_j W\{(X_j - X_i)/h\}}. \tag{2}$$

To control the variability of the estimating functions with imputed values, we make independent draws $\{\tilde{Y}_i\}$ from $\hat{F}(y|X_i)$ and use

$$\tilde{g}(Z_i; \theta) = \delta_i g(Z_i; \theta) + (1 - \delta_i) \bar{} \sum g(X_i; \tilde{Y}_i; \theta) \tag{3}$$

as the estimation function for the i-th observation.

A popular method of imputation is to impute a missing Y_i by the conditional mean of Y given $X = X_i$ as proposed in Yates (1933) under a parametric regression model and in Cheng (1994) and Wang and Rao (2002) via the Nadaraya-Watson kernel estimator for the conditional mean. However, it may not work for other parameters, for instance, quantiles or correlation coe cients. Nor is it generally applicable to the case of missing covariates in a regression context. The proposed nonparametric imputation is generally applicable for any parameter de ned by estimating equations. We are to show that when is a mean related parameter, the proposed imputation method leads to a parameter estimator that has the same e ciency as that obtained by conditional mean imputation.

\Curse of dimensiona" is an issue with kernel estimators. Indeed, the estimation accuracy of $\hat{F}(y|X_i)$ deteriorates as d_x increases. However, as demonstrated in Section 4, the curse of dimension does not pose any leading order e ect on the estimation of as long as the bias of the kernel estimator is controlled by letting $\sqrt{n}h \to 0$. There is one e ect of the dimension though: when $d_x \geq 4$, controlling the bias requires a higher order kernel. Using a higher order kernel causes $\hat{F}(y|X_i)$ not being a bona de conditional distribution and can cause a minor problem for the imputation. Wang and Chen (2006) propose a re nement of the imputation procedure to allow high order kernels and hence $d_x \geq 4$. To simplify our exposition, we con ne ourself in this paper to $d_x \leq 3$.

EMP AL L EL OOD

The nonparametric imputation produces an extended sample $\{Z_i^*\}_1^n$ where

$$Z_i^* = \begin{cases} Z_i; & \text{if } \ _i = 1; \\ (X_i; \{Y_i^* \} \quad); & \text{if } \ _i = 0: \end{cases} \tag{4}$$

Let p_i represents the probability weight allocated to Z_i^*. The empirical likeliho

is

$$L(\ ) = \sup \left\{ \prod_i^n p_i \ p_i \geq 0; \sum_i^n p_i = 1; \sum_i^n g(Z_i; \ ) = 0 \right\};$$

where g is the adjustment to the original estimating function given in (3). This is the formulation of Qin and Lawless (1994) on adjusted estimating functions. By following the standard derivation of empirical likelihood (Qin and Lawless, 1994), the optimal p_i is

$$p_i = \frac{1}{n} \frac{1}{1 + t(\ )g(Z_i; \ )};$$

where $t(\ )$ is the Lagrange multiplier that satisfies

$$\frac{1}{n} \sum_i \frac{g(Z_i; \ )}{1 + t(\ )g(Z_i; \ )} = 0: \tag{5}$$

Let $`(\ ) = -\log\{L(\ )=n^{-n}\}$ be the log empirical likelihood ratio and $\hat{}$ be the maximum empirical likelihood estimator that maximizes $L(\ )$.

The efficiency of $\hat{}$ is studied in the next section which also includes a proposal for constructing confidence regions for based on the empirical likelihood ratio. The latter is largely motivated by the attractive features (natural shape and orientation as well as range respecting) of empirical likelihood confidence regions in the absence of missing values; see Hall and La Scala (1990) and Chen and Cui (2006).

MAIN RESULTS

Let denote the true parameter value. Write $g(Z) =: g(Z; \ )$. We define

$$\begin{aligned} \tilde{} &= E[p(X)Cov\{g(Z)|X\} + E\{g(Z)|X\}E\{g(Z)|X\}]; \\ &= E\left[p^{-}(X)Cov\{g(Z)|X\} + E\{g(Z)|X\}E\{g(Z)|X\}\right] \end{aligned}$$

and $V = \{E\left(\frac{@g}{@}\right) \tilde{}^{-} E\left(\frac{@g}{@}\right)\}^{-}$ at = .

eore *nder the conditions gi en in the Appendi* **s n** $\to \infty$ *nd* $\to \infty$

$$\sqrt{n}(\hat{} - \quad) \to N(0; \quad);$$

in distri tion, with $= V E(\frac{@g}{@}) \tilde{}^{-} \quad \tilde{}^{-} E(\frac{@g}{@})V$

The estimator ^ is consistent for and the potential selection bias in the missingness as measured by the propensity score p(x) has been ltered out. If there is no missing values, ~ = = E(gg), which means that

$$= \left\{ E(\frac{@g}{@}) \ (Egg \ )^{-} E(\frac{@g}{@}) \right\}^{-} :$$

This is the asymptotic variance of the maximum empirical likelihood estimator based on full observations given in Qin and Lawless (1994). Comparing the forms of with and without missing values shows that the e ciency of the maximum empirical likelihood estimator based on the proposed imputation will be close to that based on full observations if either the proportion of missing data is low, that is when p(X) is close to 1, or if E{p^{-}

To appreciate the proposal of letting the number of imputation $\to \infty$, we note that when is xed, the and ~ matrices used to de ne have forms:

$$= E\left\{p^{-}(X) + {}^{-}(1-p(X))\right\}Cov(g|X) + E(g|X)E(g|X)_{\lrcorner} \quad \text{and}$$

$$\sim = E\left\{p(X) + {}^{-}(1-p(X))\right\}Cov(g|X) + E(g|X)E(g|X)_{\lrcorner}:$$

Hence, a larger will reduce the terms in and ~ which are due to a single nonparametric imputation. Our numerical experience suggests that = 20 is su cient for most situations.

Let us now turn our attention to the log empirical likelihood ratio

$$\mathcal{R}( ) = 2`( ) - 2`(\hat{}):$$

Let I_r be the r-dimensional identity matrix. The next theorem shows that the log empirical likelihood ratio converges to a linear combination of independent

parameter is replaced by a plugged-in estimator as revealed by Hjort, McKeague and Van Keilegom (2004).

When $= EY$, $\mathcal{R}(\ ) \to \{V(\ )/V(\ )\}$ in distribution where

$$V(\ ) = E\{\ (X)/p(X)\} + Var\{m(X)\}$$

and $V(\ ) = E\{\ (X)p(X)\} + Var\{m(X)\}$. This is the limiting distribution given in Wang and Rao (2002).

As con dence regions can be readily transformed to test statistics for testing a hypothesis regarding , we shall focus on con dence regions. There are potentially several methods for the construction of a con dence region for . One is based on an estimation of the covariance matrix and the asymptotic normality

method where the estimation of the conditional distribution is based on χ_{nc}^{*}.

3. Let $`^{*}(\hat{})$ be the empirical likelihood ratio based on the re-imputed data set χ_{n}^{*}, $\hat{}^{*}$ be the corresponding maximum empirical likelihood estimator, and $\mathcal{R}^{*}(\hat{}) = 2$

it is based on the fact that

$$E\left[g(Z_i; \ ) \frac{P(\ _i = 1)}{p(X_i)} \ _i = 1 \right\} = 0:$$

Based on the usual formulation of the generalized method of moments (GMM, Hansen, 1982), the weighted-GMM estimator for   considered in our simulation is

$$\tilde{} = \arg\min \left\{ \frac{1}{n_c} \sum_i^n \ _i g(Z_i; \ ) \frac{1}{\hat{p}(X_i)} \right\} A_T \left\{ \frac{1}{n_c} \sum_i^n \ _i g(Z_i; \ ) \frac{1}{\hat{p}(X_i)} \right\};$$

where n_c is the number of complete observations, A_T is a nonnegative de nite weighting matrix, and $\hat{p}(X_i)$ is a consistent estimator for $p(X_i)$. and The 9R013328 GMM(estimator) Tj 49 0 Td (eigene

- $p(x) \equiv 0.65$ for all x;
- $p(x) = 0.5 I(x > 0) + I(x \leq 0)$.

The missing mechanism (b) is missing completely at random; whereas the other two are missing at random and prescribe selection bias in the missingness.

Let $_x$ and $_y$ be the means, and $_x$ and $_y$ be the variances of X and Y, respectively. In the construction of the empirical likelihood for (Owen, 1990), ($_x$; $_y$; $_x$; $_y$) are treated as nuisance parameters.

Table 1 contains the bias and standard error of the four estimators considered based on 1000 simulations with the sample size $n = 100$ and 200 respectively. It also contains the empirical likelihood con dence intervals using the full observations, complete observations only, and the proposed nonparametric imputation method at a nominal level of 95% . They are all based on the proposed bootstrap calibration method with $B = 2000$. When using the nonparametric imputation method, = 20 imputations were made for each missing Y_i. The con dence intervals based on the weighted-GMM are calibrated using the asymptotic normal approximation with the covariance matrix estimated by the kernel method.

The results in Table 1 can be summarized as follows. The nonparametric imputation method signi cantly reduces the bias compared to inference based only on complete observations when the data are missing at random but not missing completely at random. The estimator based on the completely observed data su ers quite severe bias under missing mechanisms (a) and (c). The proposed imputation reduces the variance relative to the estimator matrix 5019.7776 0 Tdre)0).

0.imputatc19.

former is due to the selection bias and the latter is due to the normal approximation. The proposed con dence intervals have satisfactory coverages which are quite close to the nominal level 0:95.

5.2 *Generalized linear models with missing covariates*

In the second simulation study we consider missing covariates in a generalized linear model (GLM). We also take the opportunity to discuss an extension of the proposed imputation procedure to binary random variables. Commonly used methods in dealing with missing data

smaller than that based on the complete observations only; the proposed method also leads to a reduction in the mean square error by as much as 20% relative to the weighted-GMM. All three methods give similar mean squared errors for the parameter . The con dence intervals based on only complete observations or the weighted-GMM tend to show notable undercoverage, while the proposed con dence intervals have satisfactory coverage levels for all parameters.

EMP AL DY

Microarray technology provides an powerful tool in molecular biology by measuring the expression level of thousands of genes simultaneously. One problem of interest is to test whether the expression level of genes is related to a traditional trait like body weight, food consumption, or bone density. This is usually the rst step in uncovering roles that a gene plays in important pathways. The BXD recombinant inbred strains of mouse were derived from crosses between C57BL/6J (B6 or B) and DBA/2J (D2 or D) strains (Williams, Gu, Qi and Lu, 2001). Around one hundred BXD strains have been established by researchers at University of Tennessee and the Jackson Laboratory. A variety of phenotype data are accumulated for BXD mouse over the years (Pierce et al., 2004).

The trait that we consider is the fresh eye weight measured on 83 BXD strains by Zhai, Lu, and Williams (ID 10799, BXD phenotype data base). The Hamilton Eye Institute Mouse Eye M430v2 RMA Data Set contains measures of expression in the eye on 39,000 transcripts. It is of interest to test whether the fresh eye weight is related to the expression level of certain genes. However, the microarray data are only available for 45 out of the 83 BXD mouse strains for which fresh eye weights are all available. The most common practice is to use only complete observations and ignore missing values in the statistical inference. As demonstrated in our simulation, this approach can lead to inconsistent parameter estimators if there is a selection bias in the missingness. Even in the absence of

selection bias, the estimators are not e cient as only those complete observations are used.

We conduct four separate simple linear regression analysis of the eye weight on the expression level of four genes respectively. The genes are $H3071E5$, $Slc26a8$, $Tex9$, and $Rps16$, which are identi ed by the corresponding probe names in the microarray dataset. Here we have missing covariates in our analysis. The missing gene expression levels are imputed from a kernel estimator of the conditional distribution of the gene expression level given the fresh eye weight. The smoothing bandwidths were selected based on the cross-validation method, which is 1.5 for the rst three genes in Table 3 and 1.8 for gene $Rps16$.

Table 3 reports empirical likelihood estimates of the intercept and slope parameters and their 95% con dence intervals based on the proposed nonparametric imputation and empirical likelihood. It also contains results fromresults

The following conditions are needed in the proofs of the theorems.

C1: The functions $p(x)$, $f(x)$ and $m_g(x)$ all have bounded second partial derivatives, and $\inf_x p(x) \geq c$ for some $c > 0$.

C2: The estimating function $g(x; y;)$ has bounded second partial derivative with regard to x, and $E\ g(Z;) < \infty$. In addition, $@\ g(z;)=@\ @$ is continuous in in a neighborhood of the true value ; $@g(z;)=@$, $g(z;)$, and $@\ g(z;)=@\ @$ are all bounded by some integrable functions in the neighborhood.

C3: The matrices and ~ are, respectively, positive de nite with the smallest eigenvalue bounded away from zero, and $E[@g(z;)=@]$ has full column rank p.

C4: The kernel function W is a non-negative, symmetric and bounded probability density function with nite second moments.

C5: The smoothing bandwidth h satis es $nh^{d_x} \to \infty$, $\sqrt{n}h \to 0$ as $n \to \infty$, and $d_x \leq 3$.

Assuming p(x) bounded away from zero in C1 implies that data cannot be missing with probability 1 anywhere in the domain of the X variable. Conditions C2 and C3 are required for empirical likelihood based inference with estimating equations. Conditions C4 and C5 are standard in kernel estimation, and that $\sqrt{n}h \to 0$ is to

distribution, and then use the Cramer-Wold device to prove Lemma 1. De ne

$$m_{g_u}(x) = E(g_u(X;Y;\ )|X = x) \text{ and } \hat{m}_{g_u}(x) = \frac{\sum_i^n\ _i W(\frac{x-X_i}{h})g_u(x;Y_i;\ )}{\sum_i^n\ _i W(\frac{x-X_i}{h})}:$$

Now we have

$$\begin{aligned}
&\frac{1}{n}\sum_i^n \left\{ {}_i g_u(X_i;Y_i;\ ) + (1-\ _i)\ ^{-}\sum g_u(X_i;\tilde{Y}_i;\ )\right\} \\
=\ &\frac{1}{n}\sum_i^n\ _i\{g_u(X_i;Y_i;\ ) - m_{g_u}(X_i)\} \\
&+\frac{1}{n}\sum_i^n(1-\ _i)\left\{\ ^{-}\sum g_u(X_i;\tilde{Y}_i;\ ) - \hat{m}_{g_u}(X_i)\right\} \\
&+\frac{1}{n}\sum_i^n(1-\ _i)\{\hat{m}_{g_u}(X_i) - m_{g_u}(X_i)\} + \frac{1}{n}\sum_i^n m_{g_u}(X_i) \\
:=\ &S_n + A_n + T_n + R_n:
\end{aligned}$$

Note that S_n and R_n are sums of independent and identically distributed random variables. De ne $(x) = p(x)f(x)$ and $\hat{}(x) = \frac{}{n}\sum_j^n\ _j W_h(X_j - x)$ as its kernel estimator, where $W_h(u) = h^{-d_x}W(u=h)$. Then,

$$\begin{aligned}
T_n\ =\ &\frac{1}{n}\sum_i^n(1-\ _i)\frac{\frac{}{n}\sum_j^n\ _j W_h(X_j - X_i)\{g_u(X_i;Y_j;\ ) - m_{g_u}(X_j)\}}{(X_i)} \\
&+\frac{1}{n}\sum_i^n(1-\ _i)\{\hat{m}_{g_u}(X_i) - m_{g_u}(X_i)\}\frac{(X_i) - \hat{}(X_i)}{(X_i)} \\
&+\frac{1}{n}\sum_i^n(1-\ _i)\left\{\frac{\frac{}{n}\sum_j^n\ _j W_h(X_j - X_i)(m_{g_u}(X_j) - m_{g_u}(X_i))}{(X_i)}\right\} \\
:=\ &T_n\ + T_n\ + T_n\ :
\end{aligned}$$

De ne

$$T_n\ = \sum_j^n E\{T_n\ |(X_j;Y_j;\ _j)\} = \sum_j^n\ _j E\{T_n\ |(X_j;Y_j;\ _j = 1)\} \qquad \text{(A2)}$$

to be a projection of T_n . Then write $T_n\ = T_n\ + (T_n\ - T_n\ )$: As

$$T_n\ =\ \frac{1}{n}\sum_i^n(1-\ _i)\frac{\frac{}{n}\sum_j^n\ _j W_h(X_j - X_i)\{g_u(X_i;Y_j;\ ) - m_{g_u}(X_j)\}}{(X_i)}$$

$$= \frac{1}{n}\sum_{j}^{n} {}_{j}\{g_u(X_i;Y_j;\ ) - m_{g_u}(X_j)\}\left\{\frac{1}{n}\sum_{i}^{n}(1 - {}_{i})\frac{W_h(X_i - X_j)}{(X_i)}\right\};$$

$$T_n = \frac{1}{n}\sum_{j}^{n} {}_{j}E\ \{g_u(X_i;Y_j;\ ) \mid X_j\}$$

suggests that $T_n = T_n + o_p(n^{-=})$. By standard argument, we can show that $T_n = o_p(n^{-\frac{1}{2}})$. Derivations similar to those for T_n can be used to establish $T_n = o_p(n^{-=})$. Thus, we have

$$\sqrt{n}T_n \to N\left[0; E\{(1-p(X)) \quad _{g_u}(X)=p(X)\}\right]; \tag{A4}$$

in distribution, where $_{g_u}(X) = Var\{g_u(X;Y; \ ) \mid X\}$.

Also note $\sqrt{n}S_n \to N[0; E\{p(X) \ _{g_u}(X)\}]$ and $\sqrt{n}R_n \to N[0; Var\{m_{g_u}(X)\}]$ both in distribution. Further, it is straight forward to show that $nCov(S_n;T_n) = E\{(1-p(X)) \ _{g_u}(X)\} + o(1)$, $nCov(R_n;S_n) = 0$ and $nCov(R_n;T_n) = o(1)$. It readily follows that

$$\sqrt{n}(S_n + T_n + R_n) \to N\left[0; E\{ \ _{g_u}(X)=p(X)\} + Var\{m_{g_u}(X)\}\right]; \tag{A5}$$

in distribution.

Now we consider the asymptotic distribution of

$$A_n = \frac{1}{n}\sum_i^n (1- \ _i) \quad ^- \sum g_u(X_i;\hat{Y}_i \ ; \ ) - \hat{m}_{g_u}(X_i)\Big\}:$$

Given all the original observations, $n^{-=}(1- \ _i)\{ \ ^- \sum \ g_u(X_i;\hat{Y}_i \ ; \ ) - \hat{m}(X_i)\}$, $i = 1;2;::::;n$; are independent with conditional mean zero and conditional variance $(n \ )^-(1- \ _i)\{\hat{}_{g_u}(X_i) - \hat{m}_{g_u}(X_i)\}$. Here $\hat{}_{g_u}(x) = \sum_j^n \ _j W_h(x-X_j)g_u(x;Y_j; \ )=\hat{}(x)$ is a kernel estimator of $_{g_u}(x) = E\{g_u(X;Y; \ )|X = x\}$. By verifying Lyapounov's condition, we can show that conditioning on the original observations,

$$\sqrt{n}A_n \to N\ 0;(n \ )^- \sum_i^n (1- \ _i)\{\hat{}_{g_u}(X_i) - \hat{m}_{g_u}(X_i)\}_{\lrcorner}; \tag{A6}$$

in distribution. The conditional variance =

By Lemma 1 of Schenker and Welsh (1988), as n $\to \infty$ and $\to \infty$, $\sqrt{}$

$$:= \quad T_{n\,a} + T_{n\,b} + T_{n\,c} + T_{n\,d}:$$

$\blacksquare$

Proof of Theorem 1: Using argument similar to that of Qin and Lawless (1994), it can be shown that as $n \to \infty$ and $\to \infty$, with probability tending to 1, $L(\ )$ attains its maximum value at some point $\hat{}$ within the open ball $- < n^{-}$, and $\hat{}$ and $\hat{t} = t(\hat{})$ satisfy

$$Q_n(\hat{}; \hat{t}) = 0; \quad Q_n(\hat{}; \hat{t}) = 0:$$

Taking the derivatives with regard to and t ,

$$\frac{@Q_n(\ ;0)}{@} = \frac{1}{n}\sum_i \frac{@g(Z_i;\ )}{@}; \quad \frac{@Q_n(\ ;0)}{@t} = -\frac{1}{n}\sum_i g(Z_i;\ )g\ (Z_i;\ );$$

$$\frac{@Q_n(\ ;0)}{@} = 0; \qquad \frac{@Q_n(\ ;0)}{@t} = \frac{1}{n}\sum_i \left\{ \frac{@g(Z_i;\ )}{@} \right\} :$$

Expanding $Q_n(\hat{}; \hat{t})$, $Q_n(\hat{}; \hat{t})$ at $(\ ;0)$, we have

$$\begin{aligned} 0 &= Q_n(\hat{}; \hat{t}) \\ &= Q_n(\ ;0) + \frac{@Q_n(\ ;0)}{@}(\hat{} - \ ) + \frac{@Q_n(\ ;0)}{@t}(\hat{t} - 0) + o_p(\ _n); \\ 0 &= Q_n(\hat{}; \hat{t}) \\ &= Q_n(\ ;0) + \frac{@Q_n(\ ;0)}{@}(\hat{} - \ ) + \frac{@Q_n(\ ;0)}{@t}(\hat{t} - 0) + o_p(\ _n); \end{aligned}$$

where $_n = \hat{} - \ + \hat{t}$: Then we can write

$$\begin{pmatrix} \hat{t} \\ \hat{} - \end{pmatrix} = S_n^{-} \begin{pmatrix} -Q_n(\ ;0) + o_p(\ _n) \\ o_p(\ _n) \end{pmatrix};$$

where

$$S_n = \begin{pmatrix} @0 \\ \end{pmatrix}$$

Proof of Theorem 2: **Notice that**

$$\mathcal{R}(\quad) = 2 \sum_{i}$$

$$\rightarrow \quad N\left[0; E_*\{ \ _{g_u}(X;\hat{\ })/p(X)\} + Var_*\{m_{g_u}(X;\hat{\ })\}\right];$$

in distribution, where $E_*(\cdot)$ and $Var_*(\cdot)$ represent the conditional expectation and variance given the original data respectively. De ne

$$\hat{m}_{gu}(x;\hat{\ }) = \frac{\sum_i^n \ _i W(\frac{x-X_i}{h}) g_u(x; Y_i;\hat{\ })}{\sum_i^n \ _i W(\frac{x-X_i}{h})} \text{ and } \hat{m}^*_{g_u}(x;\hat{\ }) = \frac{\sum_i^n \ _i^* W(\frac{x-X_i^*}{h}) g_u(x; Y_i^*;\hat{\ })}{\sum_i^n \ _i^* W(\frac{x-X_i^*}{h})}:$$

Then

$$\begin{aligned}
& S_n^*(\hat{\ }) + T_n^*(\hat{\ }) + R_n^*(\hat{\ }) - S_n(\hat{\ }) - T_n(\hat{\ }) - R_n(\hat{\ }) \\
= \ & \frac{1}{n}\sum_i^n \ _i^*\{g_u(Z_i^*;\hat{\ }) - m_{gu}(X_i^*;\hat{\ })\} - \frac{1}{n}\sum_j^n \ _j\{g_u(Z_j;\hat{\ }) - m_{g_u}(X_j;\hat{\ })\} \\
+ \ & \frac{1}{n}\sum_i^n [(1 - \ _i^*)\{\hat{m}^*_{g_u}(X_i^*) - \hat{m}_{g_u}(X_i^*)\}] \\
+ \ & \frac{1}{n}\sum_i^n \ (1 - \ _i^*)\{\hat{m}_{g_u}(X_i^*;\hat{\ }) - m_{g_u}(X_i^*;\hat{\ })\} - \frac{1}{n}\sum_j^n (1 - \ _j)\{\hat{m}_{g_u}(X_j;\hat{\ }) - m_{g_u}(X_j;\hat{\ })\} \\
+ \ & \frac{1}{n}\sum_i^n \ m_{g_u}(X_i^*;\hat{\ }) - \frac{1}{n}\sum_j^n m_{g_u}(X_j;\hat{\ })\Big\} \\
:= \ & B \ + B \ + B \ + B :
\end{aligned}$$

For both B and B , we can apply the central limit theorem for bootstrap samples (e.g. Shao and Tu, 1985) to derive

$$\sqrt{n}B \ \rightarrow N\left[0; E_*\{p(X) \ _{g_u}(X;\hat{\ })\}\right] \text{ and } \sqrt{n}B \ \rightarrow N\left[0; Var_*\{m_{g_u}(X;\hat{\ })\}\right]; \qquad \text{(A9)}$$

in distribution. Also it can be shown B = $o_p(n^{-\ =})$. Use similar argument to (A3) to show

$$\begin{aligned}
B \quad = \quad & \frac{1}{n}\sum_i^n \ _i^*\{g_u(Z_i^*;\hat{\ }) - m_{g_u}(X_i^*;\hat{\ })\}\frac{1 - p(X_i^*)}{p(X_i^*)} \\
& -\frac{1}{n}\sum_j^n \ _j\{g_u(Z_j;\hat{\ }) - m_{g_u}(X_j;\hat{\ })\}\frac{1 - p(X_j)}{p(X_j)} \ + o_p(n^{-\ =}):
\end{aligned}$$

Then follow the proof for Lemma 1 and apply the bootstrap central limit theorem to conclude (A8).

For $A_n^*(\hat{\ })$, given the observations in the bootstrap sample that are not imputed, we have

$$\sqrt{}$$

Cheng, P. (1994). Nonparametric estimation of mean functionals with data missing at random. *o rn of the A eric n St tistic Associ tion* **, 81{87.**

Dempster, A. P., A. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *o rn of the Roy St tistic Society er B* **, 1{38.**

Devroye, L. P. and T. J. Wagner (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *the Ann s of St tistics* **, 231{239.**

Godambe, V. P. (1991). *Esti ting F nctions***. Oxford, U.K.: Oxford University Press.**

Hahn, J. (1998). On the role of the propensity score in e cient semiparametric estimation of average treatment e ects. *Econo etric* **, 315{331.**

Hall, P. and B. La Scala (1990). Methodology and algorithms of empirical likelihood. *ntern tion St tistic Re ie* **, 109{127.**

Hansen, L. (1982). Large sample properties of generalized method of moment estimators. *Econo etric* **, 1029{1084.**

Hirano, K., G. W. Imbens, and G. Ridder (2003). E cient estimation of average treatment e ects using the estimated propensity score. *Econo etric* **, 1161{1189.**

Hjort, N., I. McKeague, and I. Van Keilegom (2004). Extending the scope of empirical likelihood. *he Ann s of St tistics***. Under revision.**

Ibrahim, J. G., M. Chen, S. R. Lipsitz, and A. H. Herring (2005). Missing-data methods for generalized linear models: a comparative review. *o rn of the A eric n St tistic Associ tion* **, 332{346.**

Kim, J. K. and W. Fuller (2004). Fractional hot deck imputation. *Bio etri* , 559{578.

Kolaczyk, E. D. (1994). Empirical likelihood for generalized linear models. *St tistic Sinic* , 199{218.

Little, R. J. A. and D. B. Rubin (2002). *St tistic An ysis ith Missing D t nd edition*. Hoboken, NJ, USA: Wiley.

McCullagh, P. and J. A. Nelder (1983). *Gener ized Line r Mode s*. New York: Chapman and Hall.

Owen, A. (1988). Empirical likelihood ratio con dence intervals for a single functional. *Bio etri* , 237{249.

Owen, A. (19[illegible]

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *o rn of the A eric n t tistic Associ tion* **, 106{121.**

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal e ects. *Bio etri* **, 41{55.**

Rubin, D. B. (1987). *M tip e p t tion for Nonresponse in r eys***. New York: Wiley.**

Schenker, N. and A. H. Welsh (1988). Asymptotic results for multiple imputation. *he Ann s of t tistics* **, 1550{1566.**

Shao, J. and R. R. Sitter (1996). Bootstrap for imputed survey data. *o rn of the A eric n t tistic Associ tion* **, 1278{1288.**

Shao, J. and D. Tu (1985). *he c nife nd Bootstr p***. New York: Springer Verlag.**

Stone, C. J. (1977). Consistent nonparametric regression. *he Ann s of t tistics* **, 595{620.**

Titterington, D. M. (1977). Analysis of incomplete multivariate binary data by kernel method. *Bio etri* **, 455{460.**

Titterington, D. M. and G. M. Mill (1983). Kernel-based density estimates from incomplete data. *o rn of the Roy t tistic ociety erie B* **, 258{266.**

Wang, D. and S. X. Chen (2006). Nonparametric imputation of missing values for estimating equation based inference. Technical report, Iowa State University.

Wang, J., R. W. Williams, and K. F. Manly (2003). WebQTL: Web-based complex trait analysis. *Ne roinfor tics* **, 299{308.**

Wang, Q. and J. N. K. Rao (2002). Empirical likelihood-based inference under imputation for missing response data. *he Ann s of St tistics* **, 896{924.**

Williams, R. W., J. Gu, S. Qi, and L. Lu (2001). The genetic structure of recombinant inbred mice: high-resolution consensus maps for complex trait analysis. *Geno e Bio ogy* **, research0046.1{0046.18.**

Yates, F. (1933). The analysis of replicated experiments when the eld results are incomplete. *E pori o rn of Experi ent Agric t re* **, 129{142.**

n = 100					
Methods	Bias	Std. Err.	MSE	Coverage	Length of CI
Full Observations	0.0018	0.0899	0.0081	0.937	0.3543
Missing Mechanism					
Complete Obs.	0.0597	0.1262	0.0195	0.863	0.4896
Weighted-GMM	0.0057	0.1105	0.0123	0.761	0.3568
N. Imputation	-0.0050	0.1013	0.0103	0.948	0.4888
Missing Mechanism δ					
Complete Obs.	-0.0046	0.1160	0.0160	0.921	0.4562
Weighted-GMM	-0.0097	0.1068	0.0115	0.740	0.3057
N. Imputation	-0.0077	0.0991	0.0099	0.936	0.4239
Missing Mechanism c					
Complete Obs.	-0.1123	0.1480	0.0345	0.831	0.3594
Weighted-GMM	-0.0226	0.1141	0.0135	0.780	0.3569
N. Imputation	-0.0217	0.1050	0.0115	0.944	0.4264
n = 200					
Methods	Bias	Std. Err.	MSE	Coverage	Length of CI
Full Observations	0.0047	0.0605	0.0037	0.949	0.2514
Missing Mechanism					
Complete Obs.	0.0727	0.0776	0.0113	0.849	0.3269
Weighted-GMM	0.0072	0.0755	0.0058	0.800	0.2479
N. Imputation	0.0076	0.0695	0.0049	0.953	0.3239
Missing Mechanism δ					
Complete Obs.	-0.0007	0.0753	0.0057	0.957	0.3146
Weighted-GMM	-0.0067	0.0688	0.0048	0.793	0.2338
N. Imputation	-0.0004	0.0648	0.0041	0.957	0.2841
Missing Mechanism c					
Complete Obs.	-0.0905	0.1000	0.0182	0.782	0.3955
Weighted-GMM	-0.0035	0.0747	0.0060	0.773	0.2751
N. Imputation	-0.0055	0.0677	0.0046	0.946	0.2862

Table 1: Inference for the correlation coe cient with missing values. The four methods considered are empirical likelihood using full observations, empirical likelihood using

Methods	Bias	Std. Err.	MSE	Coverage	Length of CI
		n = 150			
		= −1			
Full Observations	-0.0035	1.244	1.549	0.967	5.380
Complete Obs.	-1.622	1.489	4.847	0.901	6.429
Weighted-GMM	-0.3113	1.402	2.061	0.932	5.107
N. Imputation	0.0645	1.279	1.640	0.953	5.368
		= 1			
Full Observations	0.0270	0.4270	0.1831	0.965	1.835
Complete Obs.	0.4070	0.4995	0.4152	0.908	2.308
Weighted-GMM	0.1200	0.4795	0.2443	0.935	1.722
N. Imputation	-0.0030	0.4346	0.1889	0.951	1.828
		= −1:5			
Full Observations	-0.0766	0.5009	0.2568	0.976	2.172
Complete Obs.	-0.0664	0.5539	0.3112	0.975	2.506
Weighted-GMM	-0.0663	0.5589	0.3168	0.837	

Gene	Probe	Complete Observations Only (parametric)		Nonparametric Imputation (with empirical likelihood)	
		Intercept			
H3071E5	1444597_at	-21.99	(-40.97, -2.998)	-15.69	(-37.02, 5.209)
Slc26a8	1441747_at	73.59	(49.45, 97.73)	67.28	(38.34, 95.87)
Tex9	1453360_a_at	-23.81	(-46.12, -1.507)	-14.66	(-38.57, 8.776)
Rps16	1455835_at	-13.52	(-31.08, 4.041)	-8.090	(-26.76, 10.18)
		Slope			
H3071E5	1444597_at	10.16	(5.720, 14.59)	8.736	(2.688, 14.21)
Slc26a8	1441747_at	-6.352	(-9.294, -3.411)	-5.561	(-9.431, -1.471)
Tex9	1453360_a_at	5.101	(2.588, 7.613)	4.094	(0.8753, 6.979)
Rps16	1455835_at	6.766	(3.371, 10.16)	5.754	(1.948, 9.236)
		Correlation Coefficient			
H3071E5	1444597_at	0.5757	(0.3395, 0.7436)	0.4426	(0.1321, 0.6977)
Slc26a8	1441747_at	-0.5533	(-0.7285, -0.3102)	-0.4319	(-0.6809, -0.0761)
Tex9	1453360_a_at	0.5296	(0.2996, 0.7124)	0.4024	(0.1013, 0.6846)
Rps16	1455835_at	0.5256	(0.2744, 0.7097)	0.4151	(0.0755, 0.6613)

Table 3: Parameter estimates and confidence intervals (shown in parentheses) based on a simple linear regression model using the parametric method with complete observations only and the empirical likelihood method using the proposed nonparametric imputation. For the parametric inference, the confidence intervals for the intercept and slope are obtained using quantiles of the t-distribution, and the confidence intervals for the correlation coefficient are obtained by Fisher's z transformation. The four different genes are identified by the probe names.