ELSEVIER

Journal of Statistical Planning and
Inference 79 (1999) 247–258

journal of
statistical planning
and inference

www.elsevier.com/locate/jspi

Incomplete covariates data in generalized linear models

Yi-Hau Chen *, Hung Chen

Graduate Institute of Epidemiology and Department of Mathematics, National Taiwan University,
Taipei 10018, Taiwan, ROC

Received 15 December 1997; accepted 9 November 1998

Abstract

We consider regression analysis when part of covariates are incomplete in generalized linear models. The incomplete covariates could be due to measurement error or missing for some study subjects. We assume there exists a validation sample in which the data is complete and is a simple random subsample from the whole sample. Based on the idea of projection-solution method in Heyde (1997, Quasi-Likelihood and its Applications: A General Approach to Optimal Parameter Estimation. Springer, New York), a class of estimating functions is proposed to estimate the regression coefficients through the whole data. This method does not need to specify a correct parametric model for the incomplete covariates to yield a consistent estimate, and avoids the 'curse of dimensionality' encountered in the existing semiparametric method. Simulation results shows that the finite sample performance and efficiency property of the proposed estimates are satisfactory. Also this approach is computationally convenient hence can be applied to daily data analysis. © 1999 Elsevier Science B.V. All rights reserved.

AMS classification: 62P10; 62P20

Keywords: Estimating function; Generalized linear models; Incomplete data; Measurement error; Missing covariates; Validation sample

1. Introduction

Consider the estimation problem in the context of the generalized linear models (GLMs) with 'incomplete' covariates data. Here the 'incomplete' covariates could be due to measurement error, or missing for some study subjects either by design or by happenstance. To facilitate discussions, we first describe the problem settings and introduce the notations. According to Nelder and Wedderburn (1972), the log likelihood of the response y_i ($i = 1,\ldots,n$), conditional on the q-dimensional covariate vector $\boldsymbol{w}_i$,

* Corresponding author.

PII: S0378-3758(98)00255-9

[illegible]

[illegible]

2. The proposed method

Without loss of generality, let $V=\{1,\dots,m\}$ and $\mathrm{NV}=\{m+1,\dots,n\}$ be the index sets for the validation and nonvalidation samples, respectively. First, consider the following example motivating the proposed method.

Example (*Linear model*). Consider the linear model $y=\beta x+\varepsilon$, with $\varepsilon\sim\mathrm{N}(0,\sigma^2)$. The covariate x is observed only in the validation sample. Instead, the surrogate z of x is always observed. Suppose that $x=\gamma z+\xi$ where $\xi\sim\mathrm{N}(0,\lambda^2)$ and is independent of ε. Recall that the complete-data score function for β is $\sum x_i e_i(\beta)$ with $e_i(\beta)\equiv(y_i-\beta x_i)$, and the maximum likelihood estimate for β can be obtained by using the EM algorithm (Dempster et al., 1977), which provides solution to the estimating equation

$$\sum_{i\in V} x_i(y_i-\beta x_i)+\sum_{i\in\mathrm{NV}} E_{\beta,\theta}\{x_i e_i(\beta)|y_i,z_i\}=0,$$

where $\theta=\{\gamma,\sigma^2,\lambda^2\}$. Another two candidates of the unbiased estimating functions for β are

$$\begin{aligned}&\sum_{i\in V} x_i(y_i-\beta x_i)+\sum_{i\in\mathrm{NV}} E_\gamma(x_i|z_i)\{y_i-E_\gamma(\beta x_i|z_i)\}\\&\quad=\sum_{i\in V} x_i(y_i-\beta x_i)+\sum_{i\in\mathrm{NV}}\gamma z_i(y_i-\beta\gamma z_i)\end{aligned}\tag{1}$$

and

$$\sum_{i\in V} x_i(y_i-\beta x_i)+\sum_{i\in\mathrm{NV}} E_\gamma(x_i|z_i)\{y_i-E_{\beta,\theta}(\beta x_i|y_i,z_i)\}.\tag{2}$$

By the result that

$$E_{\beta,\theta}(x|y,z)=\frac{\beta\lambda^2}{\sigma^2+\beta^2\lambda^2}y+\frac{\sigma^2\gamma}{\sigma^2+\beta^2\lambda^2}z,$$

Eq. (2) can be simplified to

$$\sum_{i\in V} x_i(y_i-\beta x_i)+\frac{\sigma^2}{\sigma^2+\beta^2\lambda^2}\sum_{i\in\mathrm{NV}}\gamma z_i(y_i-\beta\gamma z_i).$$

Recall that $\mathrm{var}_{\beta,\theta}(y|z)=\sigma^2+\beta^2\lambda^2$. It is then expected that Eq. (2) is preferred to Eq. (1) since Eq. (2) takes into account the variance structure among validation and nonvalidation samples. We will give a formal justification for this issue in the next section.

Our proposal is motivated from Eq. (2). Further, following the idea of the projection-solution method (Heyde, 1997), we can replace the conditional expectations in Eq. (2) by least-squares predictors to relax the assumptions about the model relating z to x.

For general GLMs with canonical link, let

$$S_V(\beta)=\sum_{i\in V} w_i(y_i-\mu_i),\tag{3}$$

which is the score function based on the validation data. It would serve as an unbiased estimating function if the validation sample is representative.

In the nonvalidation sample, for each β, we replace the unobserved quantities $\boldsymbol{w}$ and μ in Eq. (3) by their least-squares predictors $\hat{\boldsymbol{w}} \equiv \hat{\Gamma}_{\beta}^{\mathrm{T}}\boldsymbol{h}$ and $\hat{\mu} \equiv \boldsymbol{g}^{\mathrm{T}}\hat{\alpha}_{\beta}$, respectively, where $\boldsymbol{h}=\boldsymbol{h}(\boldsymbol{z})$ is a chosen set of basis functions of $\boldsymbol{z}, \boldsymbol{g}=\boldsymbol{g}(y,\boldsymbol{z}) \equiv (y,\boldsymbol{h}^{\mathrm{T}})^{\mathrm{T}}$, and $\hat{\Gamma}_{\beta}$ and $\hat{\alpha}_{\beta}$ are the corresponding coefficients, which are estimated from the validation sample. Then, define the estimating function based on the nonvalidation data to be

$$S_{\mathrm{NV}}(\beta) = \sum_{i \in \mathrm{NV}} \hat{\boldsymbol{w}}_i(y_i - \hat{\mu}_i) = \sum_{i \in \mathrm{NV}} \hat{\Gamma}_{\beta}^{\mathrm{T}}\boldsymbol{h}_i(y_i - \boldsymbol{g}_i^{\mathrm{T}}\hat{\alpha}_{\beta}).$$

Observed that under some regularity conditions $S_{\mathrm{NV}}(\beta)$ is asymptotically unbiased, i.e., $n^{-1}S_{\mathrm{NV}}(\beta^*) = \mathrm{o_p}(1)$ as $n \to \infty$, where β^* is the true parameter. This follows from the law of large numbers and the identities

$$E\{\boldsymbol{h}(y-\mu^*)\} = 0 \quad \text{and} \quad E\{\boldsymbol{h}(\mu^* - \boldsymbol{g}^{\mathrm{T}}\alpha^*)\} = 0,$$

where μ^* is μ evaluated at β^*, and $\alpha^* \equiv \{E(\boldsymbol{g}\boldsymbol{g}^{\mathrm{T}})\}^{-1}E(\boldsymbol{g}\mu^*)$, the limiting value of $\hat{\alpha}_{\beta}$ evaluated at β^*.

Combining $S_V(\beta)$ and $S_{\mathrm{NV}}(\beta)$, define the estimating function

$$S(\beta) \equiv S_V(\beta) + S_{\mathrm{NV}}(\beta),$$

and the proposed estimator $\hat{\beta}$ is defined to be the solution to $0 = S(\beta)$.

To ensure the existence of a consistent solution to $0 = S(\beta)$, it is required that the derivative of $n^{-1}S(\beta)$ evaluated at the true parameter to be negative-definite asymptotically. This can be achieved if the coefficient $\hat{\Gamma}_{\beta}$ is obtained by using a weighted least-squares method, with $\{d_i, i \in V\}$ as the weights. Assume that $\rho \equiv \lim_n m/n \in (0,1]$, and the usual regularity conditions hold, then for $\beta \in B$, a neighborhood of $\beta^*, n^{-1}(\partial/\partial\beta)S(\beta) \to -\boldsymbol{F}(\beta)$ in probability as $n \to \infty$, where

$$\boldsymbol{F}(\beta) \equiv \rho E(d\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}) + (1-\rho)E(d\boldsymbol{w}\boldsymbol{h}^{\mathrm{T}})\{E(d\boldsymbol{h}\boldsymbol{h}^{\mathrm{T}})\}^{-1}E(d\boldsymbol{h}\boldsymbol{w}^{\mathrm{T}}).$$

The derivation is relegated to Appendix A.

Remark 1. Note that the unbiasedness of $S(\beta)$ does not rely on the choice of the basis $\boldsymbol{h}$, hence the proposed estimating procedure is robust against the choice of $\boldsymbol{h}$ which releases the burden of practitioners to specify a correct parametric model relating $\boldsymbol{z}$ to $\boldsymbol{x}$. Furthermore, the basis $\boldsymbol{h}$ is of finite dimension hence avoids the curse of dimensionality. However, a good choice of the basis $\boldsymbol{h}$ will give better approximation for $\boldsymbol{w}$ and μ, therefore may improve the efficiency.

Remark 2. The estimating function $S(\beta)$ is, however, not the projection of a quasi-score estimating function. Consequently, as commented in Heyde (1997), the optimality properties associated with the quasi-score estimating function will not preserved for $S(\beta)$.

3. Asymptotic theory

We will derive the asymptotic theory for $\hat{\beta}$ under the regularity conditions usually assumed for the maximum likelihood estimation in GLMs, see for example, Fahrmeir and Kaufmann (1985), and additional assumptions ensure the $n^{1/2}$ consistency of the least-squares estimates $\hat{\alpha}_\beta$ and $\hat{\Gamma}_\beta$ at β^*. Detailed assumptions and a sketch of the proof are given in Appendix B.

Theorem. *Under regularity conditions specified in Appendix B and suppose that*

(i) *the validation sample is a simple random subsample from the total sample and the validation fraction* $m/n \to \rho \in (0,1]$ *as* $n \to \infty$;

(ii) $E(\boldsymbol{g}\boldsymbol{g}^{\mathrm{T}})$ *and* $E(d^*\boldsymbol{h}\boldsymbol{h}^{\mathrm{T}})$ *exist and are positive definite, where* d^* *is* d *evaluated at* β^*.

Let α^* *and* Γ^* *be, respectively, the limit of* $\hat{\alpha}_\beta$ *and* $\hat{\Gamma}_\beta$ *at* β^* *as* $n \to \infty$. *Then* $\hat{\beta}$ *is consistent and* $n^{1/2}(\hat{\beta} - \beta^*)$ *is asymptotically normal with mean zero and covariance matrix* Ω *given by*

$$\Omega = \boldsymbol{F}^{*-1}\left\{\rho\Sigma_V + (1-\rho)\Sigma_{\mathrm{NV}} + \frac{(1-\rho)^2}{\rho}\Sigma_\alpha + (1-\rho)(\Sigma_C + \Sigma_C^{\mathrm{T}})\right\}\boldsymbol{F}^{*-1}, \tag{4}$$

where

$$\boldsymbol{F}^* = \rho E(d^*\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}) + (1-\rho)E(d^*\boldsymbol{w}\boldsymbol{h}^{\mathrm{T}})\{E(d^*\boldsymbol{h}\boldsymbol{h}^{\mathrm{T}})\}^{-1}E(d^*\boldsymbol{h}\boldsymbol{w}^{\mathrm{T}}),$$

$$\Sigma_V = \phi E(d^*\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}), \Sigma_{\mathrm{NV}} = \Gamma^{*\mathrm{T}}E\{\boldsymbol{h}(y - \boldsymbol{g}^{\mathrm{T}}\alpha^*)^2\boldsymbol{h}^{\mathrm{T}}\}\Gamma^*,$$

$$\Sigma_\alpha = \Gamma^{*\mathrm{T}}E\{\boldsymbol{h}(\mu^* - \boldsymbol{g}^{\mathrm{T}}\alpha^*)^2\boldsymbol{h}^{\mathrm{T}}\}\Gamma^*,$$

$$\Sigma_C = \alpha_y^*\phi E(d^*\boldsymbol{w}\boldsymbol{h}^{\mathrm{T}})\Gamma^*, \quad \alpha_y^* \textit{ is the component of } \alpha^* \textit{ corresponding to } y.$$

The asymptotic variance (4) can be consistently estimated by replacing the population quantities in Eq. (4) with their empirical counterparts in the validation sample.

From the proof of the theorem (Appendix B) we can see that the component Σ_{NV} of Ω is the variance of $S_{\mathrm{NV}}(\beta^*)$ if α^* were known, and Σ_α arises from the variation due to the estimation of α^* through the validation sample. These two components will become the dominant terms of Ω as the validation fraction decreases because they are multiplied by factors $(1-\rho)$ and $(1-\rho)^2/\rho$, respectively. Since $\hat{\alpha}_\beta$ is a function of y_i, $i \in V$, $S_V(\beta^*)$ and $S_{\mathrm{NV}}(\beta^*)$ are correlated and the component Σ_{C} is the covariance between them.

Example (*Continued*). Let $\beta^*, \gamma^*, \sigma^{*2}$ and λ^{*2} be the true values for the corresponding parameters. In view of Eq. (4), the asymptotic variance for $\hat{\beta}$ is

$$\Omega = \frac{\sigma^{*2}}{F^*} + \frac{\sigma^{*2}}{\sigma^{*2} + \beta^{*2}\lambda^{*2}}\frac{(1-\rho)}{\rho}\frac{\beta^{*2}\lambda^{*2}\gamma^{*2}E(z^2)}{F^{*2}},$$

where $F^* = \rho E(x^2) + (1-\rho)\gamma^{*2}E(z^2) = \rho\lambda^{*2} + \gamma^{*2}E(z^2)$. On the other hand, based on the estimating function in Eq. (1), with γ estimated from the validation sample, Gourieroux and Monfort (1981) proposed an estimator $\tilde{\beta}$ for β, which can also be viewed as the estimator obtained in the same way as $\hat{\beta}$ but with $\boldsymbol{g}$ redefined as $\boldsymbol{g} \equiv \boldsymbol{h}$.

The asymptotic variance $\tilde{\Omega}$ for $\tilde{\beta}$ can thus be derived from Eq. (4) by replacing $\boldsymbol{g}$ with $\boldsymbol{h} = z$, and setting $\Sigma_{\mathrm{C}} \equiv 0$. Accordingly, we have

$$\tilde{\Omega} = \frac{\sigma^{*2}}{F^*} + \frac{(1-\rho)}{\rho} \frac{\beta^{*2}\lambda^{*2}\gamma^{*2}E(z^2)}{F^{*2}}.$$

It thus can be seen that $\Omega \leqslant \tilde{\Omega}$, with the equality holds when either $\beta^* = 0$ or $\gamma^* = 0$.

4. Simulation studies

In this section, through simulation studies we wish to get insight of the following issues of great interest: (1) the finite sample performance of $\hat{\beta}$; (2) the effect of the choice of the basis $\boldsymbol{h}$ on the efficiency of $\hat{\beta}$; (3) the loss in efficiency for $\hat{\beta}$ compared to the maximum likelihood estimator $\hat{\beta}^{\mathrm{ML}}$ when the model regarding $\boldsymbol{x}$ given $\boldsymbol{z}$ is correctly specified.

The simulation studies reported in Section 4.1 address problems (1) and (2), and in Section 4.2 the third problem is investigated under a simple scenario. The results presented are based on 500 replications.

4.1. Small sample properties and the choice of the basis

A linear regression model and a logistic regression model are considered in the present simulation study. In both cases, the covariate variable z is generated using the uniform distribution $U(-2,2)$, and x is generated through $x = 0.5z^2 + \xi$, where $\xi \sim \mathrm{N}(0,\sigma^2)$ which is independent of z. In the linear model case, the outcome y is generated according to $y = \eta + \varepsilon$, where $\eta = \beta_0^* + \beta_1^* z + \beta_2^* x$ and $\varepsilon \sim \mathrm{N}(0,1)$ which is independent of (z,x). In the logistic model case y is binary such that $\Pr(y=1|z,x) = \exp(\eta)/\{1+\exp(\eta)\}$, and $\eta = \beta_0^* + \beta_1^* z + \beta_2^* x$. In both cases $(\beta_0^*, \beta_1^*, \beta_2^*) = (-2,1,1)$. The validation sample is randomly selected from the total $n = 300$ observations with probability $\rho = 0.2$, and for the remaining observations x is missing.

For two choice of the basis $\boldsymbol{h}$, the 'wrong' choice $\boldsymbol{h}^{(1)} = (1,z)^{\mathrm{T}}$ and the 'correct' choice $\boldsymbol{h}^{(2)} = (1,z,z^2)^{\mathrm{T}}$, Table 1 demonstrates the simulation results, including the estimated asymptotic relative efficiency (ARE) of $\hat{\beta}$ versus the complete-case-only estimate $\hat{\beta}^V$, assessed via $\hat{\mathrm{V}}\mathrm{ar}(\hat{\beta}^V)\,/\hat{\mathrm{V}}\mathrm{ar}(\hat{\beta})$, here $\hat{\mathrm{V}}\mathrm{ar}(\cdot)$ denotes the average of the estimated asymptotic variances. As a comparison, we also present the results for $\tilde{\beta}$ obtained as $\hat{\beta}$ but with $\boldsymbol{g}$ redefined as $\boldsymbol{g} \equiv \boldsymbol{h}$.

Several conclusions can be made from this empirical study:

1. Even for the 'wrong' choice $\boldsymbol{h} = (1,z)^{\mathrm{T}}$, the bias of $\hat{\beta}$ is negligible. The variance estimate is adequate, too.
2. In most cases, the 'correct' choice $\boldsymbol{h} = (1,z,z^2)^{\mathrm{T}}$ provides more efficient $\hat{\beta}$ than the 'wrong' choice of $\boldsymbol{h} = (1,z)^{\mathrm{T}}$ does. However, such improvement seems to be slight when $\sigma^2 = 1$, that is, x and z are weakly correlated.
3. In most cases, as expected, $\hat{\beta}$ achieves higher efficiency than $\tilde{\beta}$ does.

Table 1
Results of $\tilde{\beta}$ and $\hat{\beta}$: bias, variance (Var), estimated variance ($\hat{\mathrm{V}}$ar) and asymptotic relative efficiency (ARE) to complete-case-only estimate β^V; $\eta=\beta_0^*+\beta_1^*z+\beta_2^*x, x=0.5z^2+\xi, \xi\sim \mathrm{N}(0,\sigma^2); n=300, \rho=0.2$. $\boldsymbol{h}^{(1)}=(1,z)^{\mathrm{T}}, \boldsymbol{h}^{(2)}=(1,z,z^2)^{\mathrm{T}}$

σ^2	$\boldsymbol{h}$	bias($\times 10^3$)		Var($\times 10^3$)		$\hat{\mathrm{V}}$ar($\times 10^3$)		ARE	
		$\tilde{\beta}$	$\hat{\beta}$	$\tilde{\beta}$	$\hat{\beta}$	$\tilde{\beta}$	$\hat{\beta}$	$\tilde{\beta}$	$\hat{\beta}$
(a) *Linear model*									
					$\beta_1^*=1$				
1	$\boldsymbol{h}^{(1)}$	1.99	0.72	18.76	9.72	18.25	8.55	0.73	1.56
1	$\boldsymbol{h}^{(2)}$	1.46	1.27	14.02	8.27	12.68	7.51	1.05	1.78
0.5	$\boldsymbol{h}^{(1)}$	0.79	0.04	13.38	8.54	13.39	7.70	1.00	1.74
0.5	$\boldsymbol{h}^{(2)}$	0.33	0.23	8.36	6.29	7.58	5.85	1.76	2.28
					$\beta_2^*=1$				
1	$\boldsymbol{h}^{(1)}$	1.61	1.61	12.61	12.61	13.28	13.28	1.00	1.00
1	$\boldsymbol{h}^{(2)}$	5.55	4.79	21.54	14.61	22.24	14.39	0.60	0.92
0.5	$\boldsymbol{h}^{(1)}$	2.97	2.97	20.12	20.12	21.06	21.06	1.00	1.00
0.5	$\boldsymbol{h}^{(2)}$	7.14	6.08	20.18	16.23	19.81	15.69	1.09	1.34
(b) *Logistic model*									
					$\beta_1^*=1$				
1	$\boldsymbol{h}^{(1)}$	39.04	48.77	64.56	67.95	70.75	64.59	2.23	2.44
1	$\boldsymbol{h}^{(2)}$	23.56	27.60	45.44	48.40	49.18	47.27	3.19	3.32
0.5	$\boldsymbol{h}^{(1)}$	44.34	52.66	55.24	56.85	60.49	55.96	2.65	2.86
0.5	$\boldsymbol{h}^{(2)}$	26.04	28.05	33.04	34.33	33.21	33.79	4.84	4.75
					$\beta_2^*=1$				
1	$\boldsymbol{h}^{(1)}$	44.41	66.52	17.76	19.81	17.13	17.76	1.13	1.09
1	$\boldsymbol{h}^{(2)}$	25.33	38.49	13.09	13.73	13.36	13.12	1.44	1.47
0.5	$\boldsymbol{h}^{(1)}$	67.79	82.05	25.21	26.90	24.06	24.58	1.12	1.10
0.5	$\boldsymbol{h}^{(2)}$	35.57	41.06	11.77	11.83	12.64	12.56	2.13	2.14

4.2. Comparison of efficiency

To gauge the efficiency property of the proposed method, now we consider a simple configuration for the measurement error problem. A linear model $y=\beta x+\varepsilon$ is considered where $x\sim \mathrm{N}(0,\tau^2)$ and $\varepsilon\sim \mathrm{N}(0,\omega^2)$ which is independent of x. The covariate x is subject to measurement error hence $z=x+\delta$ is observed for the total $n=100$ subjects, where $\delta\sim \mathrm{N}(0,\sigma^2)$ which is independent of x and ε. Only in the validation sample x is ascertained. The maximum likelihood estimate $\hat{\beta}^{\mathrm{ML}}$ of β is based on the likelihood

$$L=\prod_{i\in V} P_{\mathrm{N}}(y_i|x_i;\beta x_i,\omega^2)P_{\mathrm{N}}(x_i|z_i;\gamma z_i,\lambda^2)\prod_{i\in \mathrm{NV}} P_{\mathrm{N}}(y_i|z_i;\beta\gamma z_i,\beta^2\lambda^2+\omega^2),$$

where $\gamma=\tau^2/(\tau^2+\sigma^2)$, $\lambda^2=\tau^2\sigma^2/(\tau^2+\sigma^2)$, and $P_{\mathrm{N}}(\cdot;u,v)$ denotes the normal density function with mean u and variance v. Parameters $(\beta,\omega^2,\gamma,\lambda^2)$ are estimated simultaneously. True values for β,ω^2, and τ^2 are set to be 1, 1, and 4. Choose $\boldsymbol{h}=z$. For various values of ρ and σ^2, Table 2 demonstrates the estimated asymptotic relative efficiencies (ARE) of $\hat{\beta},\tilde{\beta}$, and $\hat{\beta}^V$ to $\hat{\beta}^{\mathrm{ML}}$.

Table 2
Asymptotic relative efficiency of $\hat{\beta}, \tilde{\beta}$, and the complete-case-only estimate $\hat{\beta}^V$ to maximum likelihood estimate $\hat{\beta}^{\mathrm{ML}}$: $y = \beta x + \varepsilon, z = x + \delta, \varepsilon \sim \mathrm{N}(0,1), \delta \sim \mathrm{N}(0,\sigma^2), x \sim \mathrm{N}(0,4), \beta^* = 1$. $\boldsymbol{h} = z, \boldsymbol{g} = (y,z)^{\mathrm{T}}$

σ^2	$\rho = 0.2$			$\rho = 0.5$		
	$\hat{\beta}$	$\tilde{\beta}$	$\hat{\beta}^V$	$\hat{\beta}$	$\tilde{\beta}$	$\hat{\beta}^V$
10	0.48	0.18	0.69	0.78	0.45	0.82
2	0.76	0.34	0.61	0.87	0.59	0.77
1	0.91	0.59	0.52	0.92	0.73	0.71
0.5	1.00	0.81	0.42	0.96	0.86	0.65
0.1	1.00	1.00	0.24	1.00	1.00	0.54

Here we make some remarks on Table 2:

1. $\hat{\beta}$ is highly efficient except when σ^2 is fairly large and ρ is small. On the other hand, $\tilde{\beta}$ is satisfactory only when the measurement error σ^2 is very small relative to ω^2.
2. Discarding the nonvalidation data would cause a tremendous loss in efficiency when both ρ and σ^2 are small. Nevertheless, for the case that σ^2 is very large, that is, z is a bad surrogate for x, $\hat{\beta}^V$ is more efficient than both $\hat{\beta}$ and $\tilde{\beta}$. In such case extracting information from the nonvalidation sample might be fairly difficult.

5. An illustrative example

The data for this illustration is the Husbands and Wives Data obtained from Hand et al. (1994). A random sample of 195 married women and their husbands was selected and data on their heights and ages was recorded. The focus was on the relationship between husband's and wife's heights, adjusting for wife's age. A linear regression model was to be fitted, where the outcome variable was wife's height (WHeight), and the covariates included husband's height (HHeight) and wife's age (WAge). There were 27 women with missing values in their ages. Since the data set also included data on husband's age (HAge), which was completely observed, we used it as a surrogate for wife's age, and chose $\boldsymbol{h} = (1, \mathrm{HHeight}, \mathrm{HAge})^{\mathrm{T}}$. As an empirical check for the surrogate condition, that is, controlling for WAge, HAge did not appear in the true model relating HHeight to WHeight, a linear regression model which further included HAge as one covariate was fitted based on the validation data. The p-value for the significance test of HAge was 0.29 hence the surrogate condition might be reasonable. The results are presented in Table 3.

Although the validation sample in this study was fairly large, composing 86% of the whole sample, by incorporating the nonvalidation sample we still gain about 10% of the estimation efficiency for the coefficient of HHeight, and about 5% for the coefficient of WAge, compared with the analysis using the validation data only.

Table 3
Parameter estimates and standard errors (SE) for Husbands and Wives Data

Covariate	Coefficient (SE)	
	Proposed method	Complete-case-only
HHeight (mm)	0.305 (0.062)	0.260 (0.069)
WAge (year)	−0.829 (0.386)	−0.894 (0.404)
Constant	1106.573 (111.750)	1188.998 (123.662)

6. Final discussions

In the analysis of incomplete covariates data, the association between $\boldsymbol{x}$ and $\boldsymbol{z}$ must be taken into account so that the information contained in the nonvalidation sample can be retrieved. The method proposed in this paper uses a parametric approach to obtain a 'working' association between $\boldsymbol{x}$ and $\boldsymbol{z}$, by which an estimating function for β can be constructed. The unbiasedness of the proposed estimating function has no bearing on the 'true' association between $\boldsymbol{x}$ and $\boldsymbol{z}$, hence the proposed method is robust to the specification of the model regarding $\boldsymbol{x}$ given $\boldsymbol{z}$. To enhance the efficiency of the proposed estimator, the data analysts may choose a better model relating $\boldsymbol{z}$ to $\boldsymbol{x}$ by use of the usual model-selection strategies.

The proposed estimating procedure can be extended to the case of the nonnatural link in the following way. Let $\tilde{\boldsymbol{w}} = \boldsymbol{w}v'(\eta)$. Note that in the nonnatural link case the score function for the validation sample is of the form

$$S_V(\beta) = \sum_{i \in V} \tilde{\boldsymbol{w}}_i(y_i - \mu_i).$$

For a chosen basis $\boldsymbol{h}$, redefine $\hat{\Gamma}_\beta$ as before but with $\boldsymbol{w}$ replaced by $\tilde{\boldsymbol{w}}$. Then the estimating function $S(\beta) = S_V(\beta) + S_{\mathrm{NV}}(\beta)$ is asymptotically unbiased at $\beta = \beta^*$, and the theorem in Section 3 still holds with $\boldsymbol{w}$ replaced by $\tilde{\boldsymbol{w}}$.

The limitation of the proposed approach is that the validation sample must be a simple random subsample of the whole sample. In the context of the missing data problem, the proposed estimator is valid when the missing mechanism is missing completely at random (MCAR) as described in Rubin (1976). To extend the proposed method to cases of more general missing mechanism, it may require that the 'selection process' that identifies subjects as members of the validation sample must be known up to an additional set of unknown parameters.

Acknowledgements

The authors are grateful to the referee for helpful comments.

Appendix A. Derivation of $\boldsymbol{F}(\beta)$

Let $\boldsymbol{W}=(\boldsymbol{w}_1,\dots,\boldsymbol{w}_m)^{\mathrm{T}}$, $\boldsymbol{H}=(\boldsymbol{h}_1,\dots,\boldsymbol{h}_m)^{\mathrm{T}}$, $\bar{\boldsymbol{H}}=(\boldsymbol{h}_{m+1},\dots,\boldsymbol{h}_n)^{\mathrm{T}}$, $\boldsymbol{D}_\beta=\mathrm{diag}(d_1,\dots,d_m)$, $\boldsymbol{G}=(\boldsymbol{g}_1,\dots,\boldsymbol{g}_m)^{\mathrm{T}}$, $\bar{\boldsymbol{G}}=(\boldsymbol{g}_{m+1},\dots,\boldsymbol{g}_n)^{\mathrm{T}}$, $\boldsymbol{U}=(\mu_1,\dots,\mu_m)^{\mathrm{T}}$.

Write $-(\partial/\partial\beta)S(\beta)=n\hat{\boldsymbol{F}}_n(\beta)-n\hat{\boldsymbol{R}}_n(\beta)$, where

$$\begin{aligned}\hat{\boldsymbol{F}}_n(\beta)&=n^{-1}\{\boldsymbol{W}^{\mathrm{T}}+\hat{\Gamma}_\beta^{\mathrm{T}}\bar{\boldsymbol{H}}^{\mathrm{T}}\bar{\boldsymbol{G}}(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})^{-1}\boldsymbol{G}^{\mathrm{T}}\}(\partial\boldsymbol{U}/\partial\beta)\\&=n^{-1}\{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{W}+\boldsymbol{W}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{H}(\boldsymbol{H}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{H})^{-1}\bar{\boldsymbol{H}}^{\mathrm{T}}\bar{\boldsymbol{G}}(\boldsymbol{G}^{\mathrm{T}}\boldsymbol{G})^{-1}\boldsymbol{G}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{W}\}\end{aligned}$$

and

$$\hat{\boldsymbol{R}}_n(\beta)=n^{-1}(\partial\hat{\Gamma}_\beta^{\mathrm{T}}/\partial\beta)\bar{\boldsymbol{H}}^{\mathrm{T}}(\bar{\boldsymbol{y}}-\bar{\boldsymbol{G}}\hat{\alpha}_\beta)$$

with $\bar{\boldsymbol{y}}=(y_{m+1},\dots,y_n)^{\mathrm{T}}$. For $\beta\in B$, a neighborhood of β^*, under some regularity conditions,

$$\begin{aligned}\hat{\boldsymbol{F}}_n(\beta)\xrightarrow{p}\ &\rho E(d\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})\\&+(1-\rho)E(d\boldsymbol{w}\boldsymbol{h}^{\mathrm{T}})\{E(d\boldsymbol{h}\boldsymbol{h}^{\mathrm{T}})\}^{-1}E(\boldsymbol{h}\boldsymbol{g}^{\mathrm{T}})\{E(\boldsymbol{g}\boldsymbol{g}^{\mathrm{T}})\}^{-1}E(d\boldsymbol{g}\boldsymbol{w}^{\mathrm{T}})\\&=\rho E(d\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})+(1-\rho)E(d\boldsymbol{w}\boldsymbol{h}^{\mathrm{T}})\{E(d\boldsymbol{h}\boldsymbol{h}^{\mathrm{T}})\}^{-1}E(d\boldsymbol{h}\boldsymbol{w}^{\mathrm{T}}),\end{aligned}$$

since $E(\boldsymbol{h}\boldsymbol{g}^{\mathrm{T}})\{E(\boldsymbol{g}\boldsymbol{g}^{\mathrm{T}})\}^{-1}=(I_r,0)$, where I_r is the identity matrix of size $r=\dim(\boldsymbol{h})$, and 0 denotes the r-vector of 0's. Also, $n^{-1}\bar{\boldsymbol{H}}^{\mathrm{T}}(\bar{\boldsymbol{y}}-\bar{\boldsymbol{G}}\hat{\alpha}_\beta)\xrightarrow{p}(1-\rho)\{E(\boldsymbol{h}y)-E(\boldsymbol{h}\mu)\}=0$. It then follows that, for $\beta\in B$,

$$\begin{aligned}\boldsymbol{F}(\beta)&\equiv-\lim_n n^{-1}(\partial/\partial\beta)S(\beta)\\&=\rho E(d\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})+(1-\rho)E(d\boldsymbol{w}\boldsymbol{h}^{\mathrm{T}})\{E(d\boldsymbol{h}\boldsymbol{h}^{\mathrm{T}})\}^{-1}E(d\boldsymbol{h}\boldsymbol{w}^{\mathrm{T}}).\end{aligned}$$

Appendix B. Consistency and asymptotic normality of $\hat{\beta}$

First we state the regularity conditions needed in the proof of Theorem 1.

(R1) $\beta\in\Theta$ which is a convex compact subset in R^q; the true parameter β^* lies in the interior of Θ;

(R2) $(y_i,z_i,\boldsymbol{x}_i)$, $i=1,\dots,n$, are independently and identically distributed;

(R3) μ is twice differentiable in β for each $\boldsymbol{w}$;

(R4) the matrix $\boldsymbol{F}^*\equiv\boldsymbol{F}(\beta^*)$ exists and is positive definite;

(R5) $E(\sup_{\beta\in B}\|d\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}}\|)$, $E(\sup_{\beta\in B}\|d\boldsymbol{w}\boldsymbol{h}^{\mathrm{T}}\|)$, $E(\sup_{\beta\in B}\|(d\boldsymbol{h}\boldsymbol{h}^{\mathrm{T}})^{-1}\|)$, and $E(\sup_{\beta\in B}\|(\partial\hat{\Gamma}_\beta^{\mathrm{T}}/\partial\beta)\boldsymbol{h}^{\mathrm{T}}(y-\mu)\|)$ are all finite for some neighborhood B of β^*. Here $\|M\|=(\sum_{i,j}m_{ij}^2)^{1/2}$ where $M=(m_{ij})$.

I. (*Consistency*) Following Foutz (1977), it suffices to check the following conditions to assure the consistency of $\hat{\beta}$: (a) the elements of $(\partial/\partial\beta)S(\beta)$ exist and are continuous in Θ; (b) the matrix $n^{-1}(\partial/\partial\beta)S(\beta)$ evaluated at β^* is negative definite with probability going to 1 as $n\to\infty$; (c) $n^{-1}(\partial/\partial\beta)S(\beta)$ converges to $\boldsymbol{F}(\beta)$ in probability uniformly for $\beta\in B$; (d) $n^{-1}S(\beta^*)=o_p(1)$ as $n\to\infty$.

Condition (a) follows from (R3), (b) holds by (R4) and the result in Appendix A. According to (R5), we can apply the strong law of large numbers for Banach space valued random variables to obtain uniform convergence of $n^{-1}(\partial/\partial\beta)S(\beta)$ (Fahrmeir and Kaufmann, 1985), thus (c) holds. Finally, statement (d) follows by the arguments in Section 2.

II. (*Normality*) Let $\bar{m} \equiv n-m$ be the sample size of the nonvalidation sample. Then

$$\begin{aligned} n^{-1/2}S_{\mathrm{NV}}(\beta^*) &= n^{-1/2}\hat{\Gamma}^{\mathrm{T}}_{\beta^*}\sum_{i\in\mathrm{NV}} \boldsymbol{h}_i(y_i - \boldsymbol{g}_i^{\mathrm{T}}\hat{\alpha}_{\beta^*}) \\ &= (\bar{m}/n)^{1/2}\left\{\bar{m}^{-1/2}\sum_{i\in\mathrm{NV}}\Gamma^{*\mathrm{T}}\boldsymbol{h}_i(y_i-\boldsymbol{g}_i^{\mathrm{T}}\alpha^*)\right\} \\ &\quad -(m/n)^{-1/2}(\bar{m}/n)\Gamma^{*\mathrm{T}}\left(\sum_{i\in\mathrm{NV}}\boldsymbol{h}_i\boldsymbol{g}_i^{\mathrm{T}}/\bar{m}\right)\{m^{1/2}(\hat{\alpha}_{\beta^*}-\alpha^*)\}+\mathrm{o_p}(1) \\ &\equiv (1-\rho)^{1/2}\boldsymbol{A} - \rho^{-1/2}(1-\rho)\Gamma^{*\mathrm{T}}(\boldsymbol{B}\cdot\boldsymbol{C})+\mathrm{o_p}(1), \end{aligned}$$

where $\boldsymbol{A}=\bar{m}^{-1/2}\sum_{i\in\mathrm{NV}}\boldsymbol{a}_i$ with $\boldsymbol{a}_i=\Gamma^{*\mathrm{T}}\boldsymbol{h}_i(y_i-\boldsymbol{g}_i^{\mathrm{T}}\alpha^*)$, $i\in\mathrm{NV}$. Recall that $E\{\boldsymbol{h}(y-\boldsymbol{g}^{\mathrm{T}}\alpha^*)\}=0$, so that $\boldsymbol{a}_i'$s are independent and identically distributed random vectors with mean zero and variance $\Sigma_{\mathrm{NV}}=\Gamma^{*\mathrm{T}}E\{\boldsymbol{h}(y-\boldsymbol{g}^{\mathrm{T}}\alpha^*)^2\boldsymbol{h}^{\mathrm{T}}\}\Gamma^*$. Hence $\boldsymbol{A}$ is asymptotically normal with mean zero and variance Σ_{NV}. Also, $\boldsymbol{B}\equiv\sum_{i\in\mathrm{NV}}\boldsymbol{h}_i\boldsymbol{g}_i^{\mathrm{T}}/\bar{m}\xrightarrow{p}E(\boldsymbol{h}\boldsymbol{g}^{\mathrm{T}})$ by assumption (ii), and $\boldsymbol{C}\equiv m^{1/2}(\hat{\alpha}_{\beta^*}-\alpha^*)$ is asymptotically normal with mean zero and variance $\Lambda_\alpha=\boldsymbol{R}^{-1}E\{\boldsymbol{g}(\mu^*-\boldsymbol{g}^{\mathrm{T}}\alpha^*)^2\boldsymbol{g}^{\mathrm{T}}\}\boldsymbol{R}^{-1}$, with $\boldsymbol{R}=E(\boldsymbol{g}\boldsymbol{g}^{\mathrm{T}})$. It is easy to see that $\{\boldsymbol{a}_i; i\in\mathrm{NV}\}$ and $\boldsymbol{B}\cdot\boldsymbol{C}$ are asymptotically uncorrelated.

It is known that under the regularity conditions $n^{-1/2}S_V(\beta)$ is asymptotically normal with mean zero and variance $\rho\Sigma_V$ at $\beta=\beta^*$, with $\Sigma_V=\phi E(d^*\boldsymbol{w}\boldsymbol{w}^{\mathrm{T}})$. Note that $n^{-1/2}S_V(\beta^*)$ and $n^{-1/2}S_{\mathrm{NV}}(\beta^*)$ are correlated through the term $\boldsymbol{C}$. The covariance between them is given by $(1-\rho)\Sigma_C$ where

$$\begin{aligned} \Sigma_{\mathrm{C}} &= E\{\boldsymbol{w}(y-\mu^*)(\boldsymbol{g}^{\mathrm{T}}\alpha^*-\mu^*)\boldsymbol{g}^{\mathrm{T}}\}\{E(\boldsymbol{g}\boldsymbol{g}^{\mathrm{T}})\}^{-1}E(\boldsymbol{g}\boldsymbol{h}^{\mathrm{T}})\Gamma^* \\ &= E\{\boldsymbol{w}(y-\mu^*)(\boldsymbol{g}^{\mathrm{T}}\alpha^*-\mu^*)\boldsymbol{h}^{\mathrm{T}}\}\Gamma^* \\ &= \alpha_y^*\phi E(d^*\boldsymbol{w}\boldsymbol{h}^{\mathrm{T}})\Gamma^*, \end{aligned}$$

the last equality holds from the result that $\boldsymbol{g}^{\mathrm{T}}\alpha^*$, the population least-squares regression $\boldsymbol{g}^{\mathrm{T}}\alpha^*$ of μ^* on $\boldsymbol{g}=(y,\boldsymbol{h}^{\mathrm{T}})^{\mathrm{T}}$, can be written as $\boldsymbol{g}^{\mathrm{T}}\alpha^*=\alpha_y^*y+(1-\alpha_y^*)\boldsymbol{h}^{\mathrm{T}}\theta^*$, where $\boldsymbol{h}^{\mathrm{T}}\theta^*$ is the population least-squares regression of μ^* on $\boldsymbol{h}$ with $\theta^*=E(\boldsymbol{h}\boldsymbol{h}^{\mathrm{T}})^{-1}E(\boldsymbol{h}\mu^*)$, and

$$\alpha_y^*=\frac{E(\mu^*-\boldsymbol{h}^{\mathrm{T}}\theta^*)^2}{E(y-\mu^*)^2+E(\mu^*-\boldsymbol{h}^{\mathrm{T}}\theta^*)^2}.$$

As a result, as $n\to\infty$, $n^{-1/2}S(\beta^*)$ is asymptotically normal with mean zero and variance

$$\rho\Sigma_V+(1-\rho)\Sigma_{\mathrm{NV}}+\frac{(1-\rho)^2}{\rho}\Sigma_\alpha+(1-\rho)(\Sigma_C+\Sigma_C^{\mathrm{T}}),$$

where $\Sigma_\alpha=\Gamma^{*\mathrm{T}}E(\boldsymbol{h}\boldsymbol{g}^{\mathrm{T}})\Lambda_\alpha E(\boldsymbol{g}\boldsymbol{h}^{\mathrm{T}})\Gamma^*=\Gamma^{*\mathrm{T}}E\{\boldsymbol{h}(\mu^*-\boldsymbol{g}^{\mathrm{T}}\alpha^*)^2\boldsymbol{h}^{\mathrm{T}}\}\Gamma^*$. Recall that $n^{-1}(\partial/\partial\beta)$ $S(\beta)\xrightarrow{p}-\boldsymbol{F}^*$ at $\beta=\beta^*$. By the regularity conditions and the Taylor expansion we thus conclude the result of theorem.

References

Carroll, R.J., Wand, M.P., 1991. Semiparametric estimation in logistic measurement error models. J. Roy. Statist. Soc. Ser. B 53, 573–587.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via EM algorithm (with discussion). J. Roy. Statist. Soc. Ser. B 39, 1–38.

Fahrmeir, F., Kaufmann, H., 1985. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. Ann. Statist. 13, 342–368.

Foutz, R., 1977. On the unique consistent solution to the likelihood equations. J. Amer. Statist. Assoc. 72, 147–148.

Gourieroux, C., Monfort, A., 1981. On the problem of missing data in linear models. Review of Economic Studies XLVIII, 579–586.

Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J., Ostrowski, E., 1994. A Handbook of Small Data Sets. Chapman & Hall, London.

Heyde, C.C., 1997. Quasi-Likelihood and Its Applications: A General Approach to Optimal Parameter Estimation. Springer, New York.

Nelder, J.A., Wedderburn, R.W.M., 1972. Generalized linear models. J. Roy. Statist. Soc. Ser. A 135, 370–384.

Pepe, M.S., Fleming, T.R., 1991. A non-parametric method for dealing with mismeasured covariate data. J. Amer. Statist. Assoc. 86, 108–113.

Robins, J.M., Hsieh, F., Newey, W., 1995. Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. J. Roy. Statist. Soc. Ser. B 57, 409–424.