Using Proxy Information in Linear Regression Model

Rong Huang, K. C. Carrière, University of Alberta
Rong Huang, 632 CAB, University of Alberta, Edmonton, T6G 2G1, Canada
huang@stat.ualberta.ca

Key Words : longitudinal data; missing data; crossover design.

1 Introduction

Proxy information is often collected for those incomplete variables in investigations to avoid missing data situation. For example, in a survey of elderly people or in cancer clinical trials, some participants may be too sick to respond and the researcher may take approximate information from their care providers. Then the issue is how to treat these proxy information. However, most of the related work todate are limited to the situation that the incomplete information occurred on covariates. And the solutions suggested debated between omitting the incomplete covariates or using the proxy variables instead. McCallum (1972) and Wickens (1972) demonstrated that using even a poor proxy is better than using only the observable regressor in terms of asymptotic bias for the case of one unobservable regressor. Barnow (1976) showed that when there is more than one unobservable regressor, deleting the unobservable regressor may be a better choice. Aigner (1974) found that proxies are preferable in most empirical situations in terms of scalar-valued mean squared error. But, Maddala (1977) showed that including the proxy variable may results with nonignorable bias. Frost (1979) found that using proxy information indiscriminately may be very risky from the view point of the squared bias and the variance of the estiamtor. Dhrymes (1978), and Trenkler and Stahlecker (1996) considered the situations where the estimator without considering the proxy dominates the estimator with the proxy with respect to MSE-matrix criterion.

The main advantage of utilizing proxy information lies in that it avoids the situation of missing data and thus standard statistical analysis can be performed. Further, it could lead to great cost saving when the procedure of obtaining actual data is expensive if not impossible. The situation we consider is when the dependent data are measured with uncertainty as not all data are actual but proxy. The question is; in this case, how does the estimator utilizing proxy information behave compared with that using complete actual data, in terms of efficiency.

In this paper, we also extend to multivariate repeated measures data. We propose an estimator utilizing proxy information, which is unbiased for the parameter of interest. We give the conditions under which estimators using proxy information will be nearly efficient as they are from complete actual data. For several two and three-period two-treatment designs, we illustrate our findings. Specifically, we give combinations of proxy and actual data that are allowed to attain desirable efficiency.

The first author is funded in part by a Studentship Award from the Alberta Heritage Foundation for Medical Research, Alberta, Canada. The second author is supported in part by a grant from Natural Sciences and Engineering Research Council of Canada, and is a National Health Research Scholar (6607- 2120-48) with the National Health Research and Development Program and a Heritage Senior Scholar with the Alberta Heritage Foundation for Medical Research.

2 The Model and Estimators

Consider the linear regression model

$$\mathbf{z} = \mathbf{X}\,\boldsymbol{\beta} + \tilde{\mathbf{X}}\,\boldsymbol{\alpha} + \boldsymbol{\epsilon} \tag{2.1}$$

with $\mathbf{z} = (\mathbf{z}_1^T, \ldots, \mathbf{z}_s^T)^T$, $\mathbf{z}_k = (\mathbf{z}_{1k}^T, \ldots, \mathbf{z}_{n_k k}^T)^T$, $\mathbf{z}_{jk} = (z_{1jk}, \ldots, z_{pjk})^T$, $z_{ijk} = y_{ijk}$ if y_{ijk} is observed, $z_{ijk} = p_{ijk}$ if y_{ijk} is missing, where p_{ijk} is the proxy for y_{ijk}, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_l)^T$, $\mathbf{X} = ((\mathbf{1}_{N1} \bigotimes \mathbf{X}_1)^T, \ldots, (\mathbf{1}_{Ns} \bigotimes \mathbf{X}_s)^T)^T$ is the matrix relating to $\boldsymbol{\beta}$, where $\mathbf{X}_k = (\mathbf{x}_{1k}^T, \ldots, \mathbf{x}_{pk}^T)^T$ for $\mathbf{x}_{ik} = (x_{ik1}, \ldots, x_{ikl})^T$. Here, $\tilde{\mathbf{X}} = (\mathbf{X}_1^T(\mathbf{M}_{11}^T, \ldots, \mathbf{M}_{N_1 1}^T), \ldots, \mathbf{X}_s^T(\mathbf{M}_{1s}^T, \ldots, \mathbf{M}_{N_s s}^T))^T$ is the matrix corresponding to proxy data, so that $\mathbf{M}_{jk} = diag(\delta_{1jk}, \ldots, \delta_{pjk})$ with $\delta_{ijk} = 1$ if y_{ijk} is missing, and 0, otherwise. And, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_l)^T$ measures possible bias due to proxy. The $\boldsymbol{\epsilon}$ has mean $\mathbf{0}$ with covariance matrix $\mathbf{V}$, where $\mathbf{V} = \mathbf{I}_n \bigotimes \boldsymbol{\Sigma}$ for a $p \times p$ covariance matrix $\boldsymbol{\Sigma}$ for $\mathbf{y}_{jk}$.

Let $\boldsymbol{\xi} = (\boldsymbol{\beta}^T, \boldsymbol{\alpha}^T)^T$, the estimator for $\boldsymbol{\xi}$ based on generalized least squares method is

$$\hat{\boldsymbol{\xi}} = \mathbf{I}^{-1} \begin{pmatrix} \sum_{jk} \mathbf{X}_k^T \Sigma^{-1} \mathbf{z}_{jk} \\ \sum_{jk} \mathbf{X}_k^T \mathbf{M}_{jk} \Sigma^{-1} \mathbf{z}_{jk} \end{pmatrix} \tag{2.2}$$

which is unbiased. The information matrix for $\hat{\xi}$ is $\mathbf{I}$

$$\mathbf{I} = \begin{bmatrix} \mathbf{I}_{11} & \mathbf{I}_{12} \\ \mathbf{I}_{21} & \mathbf{I}_{22} \end{bmatrix} \tag{2.3}$$

where $\mathbf{I}_{11} = \sum_k N_k \mathbf{X}_k^T \Sigma^{-1} \mathbf{X}_k$, $\mathbf{I}_{12} = \sum_k \mathbf{X}_k^T \Sigma^{-1} \mathbf{M}_{.k} \mathbf{X}_k$, $\mathbf{I}_{21} = \mathbf{I}_{12}^T$, $\mathbf{I}_{22} = \sum_k \mathbf{X}_k^T (\sum_j \mathbf{M}_{jk} \Sigma^{-1} \mathbf{M}_{jk}) \mathbf{X}_k$, $\mathbf{M}_{.k} = \sum_j \mathbf{M}_{jk}$.

By (2.2) and (2.3), the $l \times 1$ subvector of $\hat{\xi}$, is $\hat{\beta}$, obtained as

$$\begin{aligned} \hat{\beta} &= (\mathbf{I}_{11} - \mathbf{I}_{12}\mathbf{I}_{22}^{-}\mathbf{I}_{21})^{-1} \sum_{jk} \mathbf{X}_k^T \Sigma^{-1} \mathbf{z}_{jk} \\ &\quad -\mathbf{I}_{11}^{-1}\mathbf{I}_{12}(\mathbf{I}_{22} - \mathbf{I}_{21}\mathbf{I}_{11}^{-1}\mathbf{I}_{12})^{-1} \\ &\quad \sum_{jk} \mathbf{X}_k^T \mathbf{M}_{jk} \Sigma^{-1} \mathbf{z}_{jk} \end{aligned} \tag{2.4}$$

which is unbiased for β, the parameter of interest. The covariance matrix for $\hat{\beta}$ is

$$\begin{aligned} Cov(\hat{\beta}) &= \mathbf{I}(\hat{\beta})^{-1} \\ &= (\mathbf{I}_{11} - \mathbf{I}_{12}\mathbf{I}_{22}^{-}\mathbf{I}_{21})^{-1} \end{aligned} \tag{2.5}$$

which depends on the assumed error structure and also on the number of missing observations $\delta_{i.k} (= \sum_j \delta_{ijk})$, in period i of sequence k.

For complete data with no missing value, the $\hat{\beta}$ reduces to $\hat{\beta}^f = \mathbf{I}_{11}^{-1} \sum_k \mathbf{X}_k^T \mathbf{\Sigma}^{-1} \mathbf{y}_{.k}$, where $\mathbf{y}_{.k} = \sum_j y_{jk}$, and the covariance matrix for $\hat{\beta}^f$ is $\mathbf{I}_{11}^{-1}$, which only depends on the error structure.

For complete subset data, the estimator $\hat{\beta}^c$ and its covariance matrix are obtained by removing the incomplete pairs and are similar in form as $\hat{\beta}^f$ and $Cov(\hat{\beta}^f)$.

Obviously, for model (2.1), the estimator (2.4) utilizing proxy information is more efficient than that using only the complete subset data.

3 Efficiency of Estimators

Comparing the efficiency of the estimator utilizing proxy information to that using complete actual data, we obtain the conditions under which the former is approximately as efficient as the latter considering both dependent and independent measure error structures.

Without loss of generality, we assume β_1, the first element of β, is the parameter of interest.

Result 3.1: For $\Sigma = \sigma_\varepsilon^2 I_p$, estimator $\hat{\beta}_1$ for β_1 using proxy information is at least $c \times 100\%$ efficient as that of complete actual data, if the number of missing observations, $\delta_{i.k}, i = 1, \ldots, p, k = 1, \ldots, s$ are the solutions to inequality $g^l(\delta_{i.k}) \geq 0$ subject to $N_k \geq \delta_{i.k} \geq 0$ for $i = 1, \ldots p, k = 1, \ldots s$. Here, $g^l(\delta_{i.k})$ is a l^{th} order function of $\delta_{i.k}$.

$$g^l(\delta_{i.k}) = \sigma_\varepsilon^2(|\mathbf{I}(\hat{\beta})| - c|\mathbf{I}(\hat{\beta}^f)||\mathbf{I}(\hat{\beta})_r|/|\mathbf{I}(\hat{\beta}^f)_r|) \tag{3.1}$$

where $\mathbf{I}(\hat{\beta})_r$ and $\mathbf{I}(\hat{\beta}^f)_r$ are the submatrix of $\mathbf{I}(\hat{\beta})$ and $\mathbf{I}(\hat{\beta}^f)$ respectively by deleting the first row and the first column.

Proof: Note that, when $\Sigma = \sigma_\varepsilon^2 I_p$, we have $\mathbf{I}_{12} = \mathbf{I}_{21} = \mathbf{I}_{22}$. The information matrix for $\hat{\beta}$ and $\hat{\beta}^f$ are simplified to

$$\begin{aligned} \mathbf{I}(\hat{\beta}) &= \mathbf{I}_{11} - \mathbf{I}_{12} \\ &= \frac{1}{\sigma_\varepsilon^2} \begin{pmatrix} \sum_{ik}(N_k - \delta_{i.k})x_{ik1} & \cdots \\ \vdots & \vdots \\ \sum_{ik}(N_k - \delta_{i.k})x_{ik1}x_{ikl} & \cdots \\ \sum_{ik}(N_k - \delta_{i.k})x_{ik1}x_{ikl} & \\ \vdots & \\ \sum_{ik}(N_k - \delta_{i.k})x_{ikl}^2 & \end{pmatrix} \end{aligned} \tag{3.2}$$

$$\begin{aligned} \mathbf{I}(\hat{\beta}^f) &= \mathbf{I}_{11} \\ &= \frac{1}{\sigma_\varepsilon^2} \begin{pmatrix} \sum_{ik} N_k x_{ik1}^2 & \cdots \\ \vdots & \vdots \\ \sum_{ik} N_k x_{ik1}x_{ikl} & \cdots \\ \sum_{ik} N_k x_{ik1}x_{ikl} & \\ \vdots & \\ \sum_{ik} N_k x_{ikl}^2 & \end{pmatrix} \end{aligned} \tag{3.3}$$

$$\tag{3.4}$$

The determinant $|\mathbf{I}(\hat{\beta})|$ is a l^{th} order funtion of $\delta_{i.k}$, $i = 1, \ldots, p, k = 1, \ldots, s$, and $|\mathbf{I}(\hat{\beta})_r|$ is a $(l-1)^{th}$ order function of $\delta_{i.k}$, $i = 1, \ldots, p, k = 1, \ldots, s$. Thus, the conditions for $\delta_{i.k}, i = 1, \ldots, p, k = 1, \ldots, s$, such that $var(\hat{\beta}_1^f)/var(\hat{\beta}_1) \geq c$, are given as in Result 3.1.

Remark 3.1 If $N_k = n$ and $\delta_{i.k} = \delta$ for any i and k, (3.2) and (3.3) are simply

$$\mathbf{I}(\hat{\beta}) = (n - \delta)\mathbf{I}^* \tag{3.5}$$

and

$$\mathbf{I}(\hat{\beta}^f) = n\mathbf{I}^* \tag{3.6}$$

where

$$\mathbf{I}^* = \frac{1}{\sigma_\varepsilon^2} \begin{pmatrix} \sum_{ik} x_{ik1}^2 & \cdots & \sum_{ik} x_{ik1}x_{ikl} \\ \vdots & \vdots & \vdots \\ \sum_{ik} x_{ik1}x_{ikl} & \cdots & \sum_{ik} x_{ikl}^2 \end{pmatrix} \tag{3.7}$$

Note also that in complete subset data analyses, the information matrix becomes

$$\mathbf{I}(\hat{\beta}^f) = (n-\delta)\mathbf{I}^{**} \tag{3.8}$$

where I^{**} is the same as in (3.6) except that the missing period i in sequence k is eliminated in computing (3.6).

Result 3.2: For $\Sigma = \sigma_\varepsilon^2 I_p + \sigma_\xi^2 \mathbf{1}\mathbf{1}^T$, estimator $\hat{\beta}_1$ for β_1 using proxy information is at least $c \times 100\%$ efficient as that utilizing complete actual data, if the number of missing observations, $\delta_{i.k}, i = 1,\ldots,p, k = 1,\ldots,s$ are the solutions to inequality $g^{l(l+1)}(\delta_{i.k}) \geq 0$ subject to $N_k \geq \delta_{i.k} \geq 0$ for $i = 1,\ldots p, k = 1,\ldots s$. Here, $g^{l(l+1)}(\delta_{i.k})$ is a $l(l+1)^{th}$ order function of $\delta_{i.k}$.

$$g^{l(l+1)}(\delta_{i.k}) = f^{l(l+1)}(\delta_{i.k}) - \tilde{c}h^{(l-1)(l+1)}(\delta_{i.k})\tilde{f}^l(\delta_{i.k}) \tag{3.9}$$

where

$$\begin{aligned} |\mathbf{I}(\hat{\beta})| &= |\mathbf{I}_{11} - \mathbf{I}_{12}\mathbf{I}_{22}^{-}\mathbf{I}_{21}| \qquad (3.10)\\ &= \frac{f^{l(l+1)}(\delta_{i.k})}{\sigma_\varepsilon^2(\tilde{f}^l(\delta_{i.k}))^l} \end{aligned}$$

$$|\mathbf{I}(\hat{\beta})_{11}| = \frac{h^{(l-1)(l+1)}(\delta_{i.k})}{\sigma_\varepsilon^2(\tilde{f}^l(\delta_{i.k}))^{(l-1)}} \tag{3.11}$$

$$\tilde{c} = c|\mathbf{I}(\hat{\beta}^f)|/|\mathbf{I}(\hat{\beta}^f)_{11}| \tag{3.12}$$

Proof: The proof of the result is straightforward and omitted here.

Remark 3.2 If all data are actual, we see that (3.1) $g^l(\delta_{i.k}) = 0$ and (3.8) $g^{l(l+1)}(\delta_{i.k}) = 0$, as $c = 1$, $\delta_{i.k} = 0$ for any i and k.

4 Crossover Designs for Two-treatment Trials with proxy information

For two-treatment trials, a design is called *dual balanced*, if the design allocates an equal number of subjects to sequence k and its dual k^*, where the treatments assigned for k^{th} sequence are the opposite of those for the k^{th} sequence for each period.

The parameter β in (2.1) can be partitioned as $(\beta_{(1)}^T, \beta_{(2)}^T)^T$, where $\beta_{(1)} = (\mu, \pi)^T$ and $\beta_{(2)} = (\tau, \gamma)^T$, μ is the overall mean, π is the period contrast, τ is the treatment effect contrast and γ is the carryover effect contrast. The additional parameter α is partitioned similarly. Matrix $\mathbf{X}_k$ is a design matrix relating the responses from subjects in k^{th} sequence to β. It is partitioned as $(\mathbf{X}_{k(1)}^T, \mathbf{X}_{k(2)}^T)^T$, where $\mathbf{X}_{k(1)}$ is the design matrix relating to $\beta_{(1)}$ and $\mathbf{X}_{k(2)}$ relating to $\beta_{(2)}$. Define $d(i,k)$ to be the treatment given in period i to subjects in sequence k. Define $a_{ik} = 1$ if $d(i,k) = A$ and -1 if $d(i,k) = B$ so that $\mathbf{X}_{k(2)} = ((a_{1k},\ldots,a_{pk})^T, (0, a_{1k},\ldots,a_{p-1,k})^T)^T$.

Assuming $\delta_{i.k} = \delta_{i.k^*}$, we have

$$\sum_k N_k \mathbf{X}_{k(1)}^T \Sigma^{-1} \mathbf{X}_{k(2)} = 0$$

$$\sum_k \mathbf{X}_{k(1)}^T \Sigma^{-1} \mathbf{M}_{.k} \mathbf{X}_{k(2)} = 0$$

$$\sum_k \mathbf{X}_{k(1)}^T \mathbf{M}_{.k} \Sigma^{-1} \mathbf{X}_{k(2)} = 0$$

$$\sum_k \mathbf{X}_{k(1)}^T (\sum_j \mathbf{M}_{jk} \Sigma^{-1} \mathbf{M}_{jk}) \mathbf{X}_{k(2)} = 0$$

In which case, the information matrix of $(\hat{\beta}_{(2)}^T, \hat{\alpha}_{(2)}^T)^T$ does not depend on those of $(\hat{\beta}_{(1)}^T, \hat{\alpha}_{(1)}^T)^T$, and its expression is the same as (2.3) with $\mathbf{X}_k$ replaced by $\mathbf{X}_{k(2)}$. The covariance matrix for $\hat{\beta}_{(2)}$ is the same as (2.5) with $\mathbf{X}_k$ replaced by $\mathbf{X}_{k(2)}$.

Applying results in section 3 on $\hat{\beta}_{(2)}$, we can obtain conditions that $\delta_{i.k}$, $i = 1,\ldots,p, k = 1,\ldots,s$ have to satisfy to generate at least $c \times 100\%$ efficient estimator using proxy information. In sections 4.1 and 4.2, we assume that the pattern of missing data (which were filled in by their proxy) is monotonic and the data in the first period is available for all subjects. The case of $\delta_{i.k} = N_k$ for any $i > 1$ can be considered as the comparison of completely randomized design to the pseudo repeated measures design by creating an extra period(s) using proxy information. Therefore, we will limit our investigation to $\delta_{i.k} < N_K$ for $i > 1$.

4.1 Two-period Two-treatment Designs

For two-period designs, there are four possible sequences to consider: AA, AB, BA, and BB. We consider the following designs:

Design I: AB and BA

Design II: AB, BA, AA, and BB

Design III: AA and BB

Design I is the optimal design under no carryover effect (Grizzle 1965), Design II is universally optimal (Carriere and Reinsel 1992), and Design III is the parallel group two-period two-treatment design.

Design I and Design III: The estimator for tratment effect contrast τ using proxy information is as efficient as the estimator using complete actual data regardless of the fraction of proxy data and the measurement error structure.

Remark 4.1 This is due to the fact that in these 2 periods designs, the estimator for τ only uses the first period data under the model with unequal residual effects. Hence gathering proxy information to fill in the missing second period data does not result in any gain (or loss) in efficiency.

Design II: Let $r_1 = \delta_{2.1}/(N/4) = \delta_{2.2}/(N/4)$ and $r_2 = \delta_{2.3}/(N/4) = \delta_{2.4}/(N/4)$ be the proportions of data filled in by the proxy in second period of sequence AB(BA) and AA(BB) respectively. Results 3.1 and 3.2 lead to

$$r_2 \leq \frac{4(c-1)}{(2c-3)}, r_1 \leq \frac{4(c-1)-(2c-3)r_2}{2r_2+2c-3} \quad (4.1)$$

when $\Sigma = \sigma_\epsilon^2 I_p$, while

$$r_2 \leq \frac{-b2}{b1}, r_1 \leq \frac{-b_1 r_2 - b_2}{r_2 + b_1} \quad (4.2)$$

when $\Sigma = \sigma_\epsilon^2 I_p + \sigma_\xi^2 \mathbf{1}\mathbf{1}^T$, where $b_1 = [-2b^2 + 6b - 3 + (b^2 - 4b + 2)c]/[2(1-b)^2]$ and $b_2 = [(b^2 - 4b + 2)(1-c)]/(1-b)^2$, with $b = \rho/(1+\rho)$, such that the estimator for tratment effect contrast τ using proxy information is at least as $c \times 100\%$ efficient as the estimator using complete actual data.

Table 1 reports the possible proportions of proxy information permitted to generate an approximately efficient estimator for τ for Design II. The proportions of proxy information permitted decrease with increasing efficiency coefficient c, and so with increasing correlation coefficient ρ. The proportion of proxy information permitted in the sequence AA(BB) increases when that in the sequence AB(BA) decreases. Because sequence AA(BB) are less favorable (Carriere and Reinsel, 1992), we might allocate more resources to collect actual information in sequences AB(BA). For example, when error structure is independent, to obtain 80% efficient estimator for τ, 50% of the data in second period of sequences AA(BB) and 25% of those in sequence AB(BA) can be collected as proxy data. Another example is, when errors are correlated with correlation coefficient $\rho = 0.6$, we could have 20% of the data in second period of sequence AA(BB) and 12% in sequence AB(BA) collected as proxy information, with less than 10% loss of efficiency for estimating τ.

4.2 Three_period Two_treatment Designs

For three-period designs, there are eight possible sequences to consider: AAA, AAB, ABA, ABB and their duals. We consider the following designs:

Design IV: ABB and BAA

Design V: AAA and BBB

Design VI: ABB, BAA, AAB, and BBA

Design VII: ABB, BAA, ABA, and BAB

Under an equicorrelated covariance structure, the dual-balanced Design IV with sequences ABB and BAA is known to be the universally optimal design within the class of three-period designs(Kershner 1986, Laska, Meisner and Kushner 1983). Balancing statistical optimality and clinical suitability, Carriere (1994) recommended Design VI as a 'nearly' optimal three-period design, which performed very competitively under various models. Ebbutt (1984) considered Design VII seriously, and Design V is the parallel group three-period two-treatment design.

To simplify the presentation, we assume that missing data occurs only in the third (last) period.

Design IV: Let $r_1 = \delta_{3.1}/(N/2) = \delta_{3.2}/(N/2)$ be the proportion of data in the third period of sequence ABB(BAA) filled in by the proxy. Results 3.1 and 3.2 lead to

$$r_1 \leq \frac{6(1-c)}{5-3c} \quad (4.3)$$

when $\Sigma = \sigma_\epsilon^2 I_p$, while

$$r_1 \leq \frac{2b^2 - 8b + 6 + 2c(3-b)(b-1)}{2b^2 - 5b + 5 - c(3-b)} \quad (4.4)$$

when $\Sigma = \sigma_\epsilon^2 I_p + \sigma_\xi^2 \mathbf{1}\mathbf{1}^T$, where $b = \rho/(1+2\rho)$, such that the estimator for treatment effect contrast τ using proxy information is at least as $c \times 100\%$ efficient as the estiamtor using complete actual data.

For Design IV, Table 2 shows that the proportion of proxy information permitted in the third period is decreasing with increasing correlation coefficient ρ and increasing c.

Design V: The estimator for treatment effect contrast τ using proxy information is as efficient as the estiamtor using complete actual data regardless of the fraction of proxy data in the third period of sequence AAA(BBB) and the measurement error structures, with similar reasons laid out in cases with Design I and Design III.

Design VI: Let $r_1 = \delta_{3.1}/(N/4) = \delta_{3.2}/(N/4)$ and $r_2 = \delta_{3.3}/(N/4) = \delta_{3.4}/(N/4)$ be the proportions of data filled in by proxy in the third period of sequence ABB(BAA) and AAB(BBA) respectively. Results 3.1 and 3.2 lead to

$$r_2 \leq \frac{6(1-c)}{5-3c}, r_1 \leq \frac{-24(1-c)+(10-6c)r_2}{4r_2-(10-6c)} \tag{4.5}$$

when $\Sigma = \sigma_\varepsilon^2 I_p$, while

$$r_2 \leq min\{\frac{-a_1}{a_0}, \frac{-a_3}{a_2}\}, r_1 \leq \frac{-a_3-a_2r_2}{a_0r_2+a_1} \tag{4.6}$$

when $\Sigma = \sigma_\varepsilon^2 I_p + \sigma_\xi^2 \mathbf{1}\mathbf{1}^T$, where $a_0 = -2(2b^4-4b^3+6b^2-4b+2), a_1 = -2(-2b^4+6b^3-9b^2+10b-5)-c(b^2-8b+6), a_2 = -2(-2b^4+18b^3-29b^2+18b-5)-c(b^2-8b+6)(4b^2-4b+1), a_3 = -2(2b^4-20b^3+46b^2-40b+12)-c(b^2-8b+6)(-4b^2+8b-4)$, $b = \rho/(1+2\rho)$, such that the estimator for treatment effect contrast τ using proxy information is at least as $c \times 100\%$ effeicient as the estiamtor using complete actual data.

The trend for Design VI is similar to Table 1 for Design II. If sequence ABB(BAA) is preferred to sequence AAB(BBA), we could allocate more resources to collect actual data in ABB(BAA). For example, when $\Sigma = \sigma_\varepsilon^2 I_p$, we could have 90% and 7% of the data in the third periods of sequences AAB(BBA) and ABB(BAA) respectively to be filled in by proxy, with less than 20% loss of efficiency.

Design VII: Let $r_1 = \delta_{3.1}/(N/4) = \delta_{3.2}/(N/4)$ and $r_2 = \delta_{3.3}/(N/4) = \delta_{3.4}/(N/4)$ be the proportions of the data filled in by proxy in the third period of sequence ABB(BAA) and ABA(BAB) respectively. Results 3.1 and 3.2 lead to

$$r_1 \leq \frac{(6-5c)r_2-20(1-c)}{4r_2-(14-5c)} \tag{4.7}$$

with $0 \leq r_2 \leq 1$, when $\Sigma = \sigma_\varepsilon^2 I_p$, while

$$r_1 \leq \frac{-a_3-a_2r_2}{a_0r_2+a_1}, r_2 \leq min\{\frac{-a3}{a2}, 1\} \tag{4.8}$$

when $\Sigma = \sigma_\varepsilon^2 I_p + \sigma_\xi^2 \mathbf{1}\mathbf{1}^T$, where $a_0 = 4-4b, a_1 = -2(7+2b^2-7b)-c(2b-5), a_2 = -2(3-3b+2b^2)-c(2b-5), a_3 = -2(-10+14b-4b^2)-c(2b-5)(-4+4b)$, $b = \rho/(1+2\rho)$, such that the estimator for treatment effect contrast τ using proxy information is at least as $c \times 100\%$ effeicient as the estiamtor using complete actual data.

The summarization of the proportions (r_1, r_2) of proxy information permitted in the third period of sequences ABB(BAA) and ABA(BAB), shows that r_1 is almost stationary as r_2 decreases from 1 to 0, except for $\rho = 0$. When errors are independent, r_1 increases from .09 to .21 for $c = .9$ and from .33 to .39 for $c = .8$.

5 Discussion

We have investigated the role of proxy information in statistical analyses. Proxy information can be available in diverse form. Its quality also vary in great deal. For example, these can be rather heterogeneous measure on neighborhood affluence in describing one's socio-economic status. In clinical trial context, they can contain relevant characteristics to the respondents, when the sick respondents' care-providers act as the respondents in providing the pertinent information.

Gathering proxy information in an effort to maximize the data collection strategy is particularly useful in statistical analyses, as the usual software can then be used to analyze the data. In essence, this approach is similar to the single imputation strategy, where missing data are filled in with some values generated via various methods (Little and Rubin, 1987). Common software gives the users an option as to what methods to adopt. Most frequent imputation may be to use the mean of the sub-group by considering particular covariate pattern shared by the subject with missing data. The usual analytic methods then proceed assuming that the imputed values are actual data. With proxy data, such approach has also been the usual practice.

We have employed a slightly more general model attempting to capture the possible bias, as the proxy information is not really the actual responses. We anticipate that if the proxy information is indeed alighned with the actual data, there will be no real loss in efficiency.

Indeed even when over 90% of the data are replaced by its proxy information, we find that the efficiency compared to the ideal situation of complete data can be over 90%. As the within-subject correlation increases, the proportion of proxy information allowed to attain desired efficiency decreases.

Efficiency comparison to other existing techniques can be seen in Remark 3.1. For example, the equation (3.5) can be regarded as the information matrix being assumed in a single imputation strategy. By comparing (3.4) to (3.5), we can see that our approach reduce the degrees of freedom by the same amount as the missing data while accommodating the bias that the single imputation strategy does not. Similarly comparing (3.4) to (3.7), we can appreciate that the information using proxy is always

much larger than that of the complete subset data analysis.

The assumptions we used and therefore the limitation of our approach is that we have presumed the proxy to behave similarly as the actual data, although we allowed the room for possible bias. That is, we assumed that the covariance matrix for each subject is identical. This may not hold in some situations. We are currently investigating the extension of our approach to the situations with heterogeneous covariance matrix. Depending upon the degree of heterogeneity, our findings can give practical guidelines as to what level of heterogeniety is acceptable to attain desired efficiency. Our approach might be a simpler alternative to such current missing data techniques as (Patel, 1985, 1991; Carriere 1994, 1999; Jennrich and Schluchter 1986) that often requires iterative techniques or special software. Power comparisons against these missing data techniques are under investigation.

Reference

1. Aigner, D. J. (1974), MSE dominace of least squares with errors of observation, Journal of Econometrics, 2, 365-372

2. Barnow, B. S. (1976), The use of proxy variables when one or two independent variables are measured with error, The American Statistician, 30, 119-121

3. Carriere, K.C. and Reinsel, G. (1992), Investigation of dual-balanced crossover designs for two treatments. Biometrics, 48, 1157-1164.

4. Carriere, K.C. (1994), Crossover designs for clinical trials. Statistics in Medicine, 13, 1063-1069.

5. Carriere, K.C. (1999), Methods for repeated measures data analysis with missing values, Journal of Statistical Planning and Inference 77, 221-236.

6. Dhrymes, P. J. (1978), Introductory Econometrics, Springer, New York.

7. Ebbutt, A.F. (1984), Three-period crossover designs for two treatments. Biometrics, 40, 219-224

8. Frost, P. A. (1979), Proxy variables and specification bias, The review of economics and Statistics, 61, 323-325

9. Grizzle, J.E. (1965), The two-period chang-over design and its use in clinical trials. Biometrics, 21, 467-480

10. Jennrich, R.I. and Schluchter, M.D. (1986), Unbalanced repeated-measures models with structured covariance matrices, Biometrics, 805-820.

11. Kershner, R. P. (1986). Optimal 3-period 2-treatment crossover designs with and without baseline measurements. Proceedings of the Biopharmaceutical Section of the American Statistical Association, 152-156.

12. Laska, E. M., Meisner, M, and Kushner, H. B. (1983). Optimal crossover designs in the presence of carryover effects. Biometrics, 39, 1087-1091.

13. Maddala, G. S. (1977), Econometrics, McGraw-Hill, New York.

14. McCallum, B. T. (1972), Relative asymptotic bias from errors of omission and measurement, Econometrica, 40, 757-758

15. Patel, H.I. (1985), Analysis of incomplete data in a two-period crossover design with reference to clinical trials, Biometrika, 72, 411-418.

16. Patel, H.I. (1991), Analysis of incomplete data from a clinical trial woth repeated measurements, Biometrika, 78, 609-619.

17. Trenkler, G. and Stahlecker, P.(1996), Dropping variables versus use of proxy variables in linear regression, Journal of Statistical Planning and Inference, 50, 65-75

18. Wickens, M. R. (1972), A note on the use of proxy variables, Econometrika, 40, 756-761

Table 1. Proxy information permitted to generate $c \times 100\%$ efficient estimator $\hat{\tau}$ for Design II

c	r_2	ρ					
		0.0	0.2	0.4	0.6	0.8	1.0
.8	.50	.25	.22	.14			
	.40	.40	.38	.33	.23	.07	
	.30	.47	.46	.42	.35	.24	.06
	.20	.52	.51	.48	.42	.33	.20
	.10	.55	.54	.51	.46	.38	.28
	.00	.57	.56	.53	.49	.42	.33
.9	.30	.06	.05	.02			
	.20	.20	.19	.16	.12	.06	
	.10	.28	.27	.25	.21	.16	.10
	.00	.33	.32	.31	.28	.23	.18

Note: r_2 is the proportion of proxy information in the second period of sequence AA(BB). The entries r_1s, the proportions of proxy information in the second period of sequence AB(BA) were divided according to $r_2 > r_1$ or $r_2 < r_1$.

Table 2. Proxy information permitted to generate $c \times 100\%$ efficient estimator $\hat{\tau}$ for Design IV

c	ρ					
	0.0	0.2	0.4	0.6	0.8	1.0
.8	.46	.48	.48	.49	.49	.50
.9	.26	.27	.29	.29	.30	.30

Note: The entries r_1s, are the proportions of proxy information in the third period of sequence ABB(BAA).