Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

Logistic regression in meta-analysis using aggregate data

Bei-Hung Chang , Stuart Lipsitz & Christine Waternaux

To cite this article: Bei-Hung Chang , Stuart Lipsitz & Christine Waternaux (2000) Logistic regression in meta-analysis using aggregate data, Journal of Applied Statistics, 27:4, 411-424, DOI: 10.1080/02664760050003605

To link to this article: https://doi.org/10.1080/02664760050003605

Published online: 02 Aug 2010.

Journal of Applied Statistics, Vol. 27, No. 4, 2000, 411–424

Logistic regression in meta-analysis using aggregate data

BEI-HUNG CHANG[1], **STUART LIPSITZ**[2] **& CHRISTINE WATERNAUX**[3], [1]*Center for Health Quality, Outcomes, and Economic Research, Bedford VA Medical Center, USA;* [2]*Department of Biostatistics, Harvard School of Public Health, Boston, USA and* [3]*Division of Biostatistics, New York State Psychiatric Institute, New York, USA*

ABSTRACT *We derived two methods to estimate the logistic regression coefficients in a meta-analysis when only the 'aggregate' data (mean values) from each study are available. The estimators we proposed are the discriminant function estimator and the reverse Taylor series approximation. These two methods of estimation gave similar estimators using an example of individual data. However, when aggregate data were used, the discriminant function estimators were quite different from the other two estimators. A simulation study was then performed to evaluate the performance of these two estimators as well as the estimator obtained from the model that simply uses the aggregate data in a logistic regression model. The simulation study showed that all three estimators are biased. The bias increases as the variance of the covariate increases. The distribution type of the covariates also affects the bias. In general, the estimator from the logistic regression using the aggregate data has less bias and better coverage probabilities than the other two estimators. We concluded that analysts should be cautious in using aggregate data to estimate the parameters of the logistic regression model for the underlying individual data.*

1 Introduction

Meta-analysis concerns pooling data from various studies. For example, Hegarty *et al.* (1994) conducted a meta-analysis of the international literature on outcome in schizophrenia, and identified more than 300 outcome studies (both controlled and naturalistic) that examined the improvement of schizophrenic patients. Each

Correspondence: Bei-Hung Chang, Center for Health Quality, Outcomes, and Economic Research, Bedford VA Medical Center, 200 Springs Road (152), Bedford, MA 01730, USA. E-mail: bhchang@bu.edu.

ISSN 0266-4763 print; 1360-0532 online/00/040411-14

of these outcome studies examined the effect of a single treatment on the binary outcome, the improvement of schizophrenia (1 = improved, 0 = not improved). In this paper, we consider meta-analysis methods for analysing aggregate data gathered from such studies that have binary outcomes. In particular, we have an outcome measurement y_{ij} that is a binary variable for the jth subject in study i. For each subject we also have a covariate vector x_{ij}. However, since each study used only one treatment, the treatment covariate in x_{ij} is the same for each subject in a study. There are other covariates of interest, such as follow-up years in the schizophrenia meta-analysis data, which vary for subjects within a study.

A logistic regression model is commonly used to model the relationship between y_{ij} and covariate vector x_{ij},

$$E(y_{ij}|\mathbf{x}_{ij}) = \Pr(y_{ij} = 1|\mathbf{x}_{ij}) = \frac{e^{\mathbf{x}_{ij}^t\beta}}{1 + e^{\mathbf{x}_{ij}^t\beta}} \quad \text{or} \quad \text{logit}(p(y_{ij} = 1|\mathbf{x}_{ij})) = \mathbf{x}_{ij}^t\boldsymbol{\beta} \tag{1}$$

where y_{ij} = binary outcome measurement for observation j in study i,
$\mathbf{x}_{ij}$ = the covariate vector for observation j in study $i, j = 1, \ldots, n_i; i = 1, \ldots, k;$
n_i = the number of observations in study i, k = the number of studies.

When the analyst has the individual y_{ij}s and $\mathbf{x}_{ij}$s, the maximum likelihood can be used to estimate the logistic regression vector $\boldsymbol{\beta}$ which consists of an intercept β_0 and vector $\boldsymbol{\beta}_1$. However, in meta-analysis, it often happens that the available data is ‘aggregate’ data of the form:

$$\bar{y}_{i+} = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} \quad \text{and/or} \quad \bar{x}_{i+} = \frac{\sum_{j=1}^{n_i} \mathbf{x}_{ij}}{n_i}$$

With a binary response y_{ij}, knowing $\bar{y}_{i+}$ and n_i also gives us knowledge of how many y_{ij}s equal 0 and how many equal 1. The most popular method for estimating the parameters of a logistic regression model with aggregate data is to fit the model (Fridstrøm & Ingebrigtsen, 1991):

$$E(y_{ij}|i\text{th study}) = \frac{e^{\bar{\mathbf{x}}_{i+}^t\beta^\star}}{1 + e^{\bar{\mathbf{x}}_{i+}^t\beta^\star}} \tag{2}$$

We call this model a mean logistic regression model based on the fact that the covariate vector $\mathbf{x}_{ij}$ in model (2) has a value equal to the study mean vector $\bar{\mathbf{x}}_{i+}$ for all observations within the same study. It is noticeable that since the relationship between x and y is non-linear, $\boldsymbol{\beta}^\star$ in (2) does not necessarily equal $\boldsymbol{\beta}$ in (1) unless each subject within a study has the same value $\mathbf{x}_{ij}$, i.e. $\mathbf{x}_{ij} = \bar{\mathbf{x}}_{i+}$. Therefore, using maximum likelihood with (2) will not give consistent estimates of $\boldsymbol{\beta}$ in model (1). In this paper, we developed two other methods to estimate $\boldsymbol{\beta}$ in model (1) using the aggregate data $\bar{y}_{i+}$ and $\bar{\mathbf{x}}_{i+}$.

The two other methods we propose to estimate $\boldsymbol{\beta}$ are based on a discriminant function estimator (Efron, 1975; Press & Wilson, 1978) and a reverse Taylor series approximation (Press & Wilson, 1978; Nerlove & Press, 1973). These two estimators were originally developed to provide the starting values in an iterative maximum likelihood estimation procedure for a logistic regression model before the high speed computer became available. The derivations of these two estimators using individual data are discussed in Section 2. The derivations of these two estimators, as well as the estimators using the mean logistic regression model using aggregate data, are discussed in Section 3. As we will show, the mean logistic regression model can be obtained by a Taylor series approximation to model equation (1). In Section 4, these two estimators as well as the mean logistic regression were applied to the above example of the schizophrenia meta-analysis data. Section 5 describes a simulation study that was conducted to evaluate the performance of these two estimators as well as the estimator obtained from the mean logistic regression model. The performance of the estimators was evaluated for conditions which vary by three parameters: numbers of studies in a meta-analysis, sample sizes in each study, and distribution types of the covariates. The results of the simulation study are presented in Section 6.

2 Estimators

This section describes the derivation of the discriminant function estimators and the reverse Taylor series approximation estimators using individual data. As mentioned above, in the past, these two estimators have been used as an approximate of a logistic regression coefficient estimator.

2.1 Discriminant function estimators

By applying Bayes' theorem, $\text{logit}(p(y_{ij} = 1 \mid \mathbf{x}_{ij}))$ can be expressed as:

$$\log\left[\frac{p(y_{ij}=1)}{1-p(y_{ij}=1)}\right] + [\log f(\mathbf{x}_{ij} \mid y_{ij}=1) - \log f(\mathbf{x}_{ij} \mid y_{ij}=0)]$$

where $f(\mathbf{x}_{ij} \mid y_{ij})$ is the probability density function for $\mathbf{x}_{ij}$ given y_{ij}. Suppose that the covariate vector $\mathbf{x}_{ij}$ given $y_{ij} = y$ is a random sample from the p-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}_y$ and covariance matrix $\boldsymbol{\Sigma}$: $(\mathbf{x}_{ij} \mid y_{ij} = y) \sim N(\boldsymbol{\mu}_0 + \Delta y, \boldsymbol{\Sigma})$, where $\Delta = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$, and $y = 0, 1$. Then

$$\text{logit}(p(y_{ij}=1 \mid x_{ij})) = \log\left[\frac{p(y_{ij}=1)}{1-p(y_{ij}=1)}\right] - \frac{1}{2}\mu_1^t \boldsymbol{\Sigma}^{-1}\mu_1 + \frac{1}{2}\mu_0^t \boldsymbol{\Sigma}^{-1}\mu_0 + x_{ij}^t \boldsymbol{\Sigma}^{-1}(\mu_1 - \mu_0)$$

Therefore, the regression coefficient vector $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1)$ in model (1) is:

$$\begin{aligned} \boldsymbol{\beta}_0 &= \log\left[\frac{p(y_{ij}=1)}{1-p(y_{ij}=1)}\right] - \frac{1}{2}\mu_1^t \boldsymbol{\Sigma}^{-1}\mu_1 + \frac{1}{2}\mu_0^t \boldsymbol{\Sigma}^{-1}\mu_0 \\ \beta_1 &= \boldsymbol{\Sigma}^{-1}(\mu_1 - \mu_0) \end{aligned} \tag{3}$$

The estimators of $(\beta_0, \boldsymbol{\beta}_1)$ calculated from the above equations are called discriminant function estimators. To calculate the discriminant function estimator, it is necessary to estimate vectors $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}$, and $p(y_{ij} = 1)$. In Section 3.1, we show how to estimate these parameters by applying a multivariate linear model to the aggregate data $\bar{y}_{i+}$ and $\bar{\mathbf{x}}_{i+}$. Note that one important assumption required to derive the discriminant function estimator is the normality of the distribution of covariate vector $\mathbf{x}$.

2.2 A reverse Taylor series approximation

A reverse Taylor series approximation is obtained by expanding the logistic function (model 1) about $\mathbf{x}_{ij} = \bar{\mathbf{x}}$ (the grand mean $\bar{\mathbf{x}} = [\Sigma n_i \bar{\mathbf{x}}_{i+}]/\Sigma n_i$) in a Taylor series.

$$
\begin{aligned}
E(y_{ij}|\mathbf{x}_{ij}) &= \frac{e^{\beta_0 + \mathbf{x}_{ij}^t \beta_1}}{1 + e^{\beta_0 + \mathbf{x}_{ij}^t \beta_1}} \\
&= \left[\frac{e^{\beta_0 + \bar{\mathbf{x}}^t \beta_1}}{1 + e^{\beta_0 + \bar{\mathbf{x}}^t \beta_1}} - \frac{e^{\beta_0 + \bar{\mathbf{x}}^t \beta_1}}{(1 + e^{\beta_0 + \bar{\mathbf{x}}^t \beta_1})^2} \boldsymbol{\beta}_1^t \bar{\mathbf{x}} \right] + \left[\frac{e^{\beta_0 + \bar{\mathbf{x}}^t \beta_1}}{(1 + e^{\beta_0 + \bar{\mathbf{x}}^t \beta_1})^2} \boldsymbol{\beta}_1^t \right] \mathbf{x}_{ij} + R(\mathbf{x}_{ij}) \\
&= \alpha_0 + \boldsymbol{\alpha}_1^t \mathbf{x}_{ij} + R(\mathbf{x}_{ij})
\end{aligned} \tag{4}
$$

where

$$
\alpha_0 = \frac{e^{\beta_0 + \bar{\mathbf{x}}^t \beta_1}}{1 + e^{\beta_0 + \bar{\mathbf{x}}^t \beta_1}} - \frac{e^{\beta_0 + \bar{\mathbf{x}}^t \beta_1}}{(1 + e^{\beta_0 + \bar{\mathbf{x}}^t \beta_1})^2} \boldsymbol{\beta}_1^t \bar{\mathbf{x}}, \qquad \boldsymbol{\alpha}_1 = \frac{e^{\beta_0 + \bar{\mathbf{x}}^t \beta_1}}{(1 + e^{\beta_0 + \bar{\mathbf{x}}^t \beta_1})^2} \boldsymbol{\beta}_1
$$

$$
R(\mathbf{x}_{ij}) = \left(\frac{e^{\beta_0 + \bar{\mathbf{x}}_{ij}^t \beta_1}}{1 + e^{\beta_0 + \bar{\mathbf{x}}_{ij}^t \beta_1}} - \frac{e^{\beta_0 + \bar{\mathbf{x}}^t \beta_1}}{1 + e^{\beta_0 + \bar{\mathbf{x}}^t \beta_1}} \right) + \left(\frac{e^{\beta_0 + \bar{\mathbf{x}}^t \beta_1}}{(1 + e^{\beta_0 + \bar{\mathbf{x}}^t \beta_1})^2} \boldsymbol{\beta}_1^t \right) (\mathbf{x}_{ij} - \bar{\mathbf{x}})
$$

$R(\mathbf{x}_{ij})$ is the 'remainder' term.

Neglecting the remainder term, a 'linear' model for y_{ij} given $\mathbf{x}_{ij}$ is obtained;

$$
E(y_{ij}|\mathbf{x}_{ij}) \approx \alpha_0 + \boldsymbol{\alpha}_1^t \mathbf{x}_{ij}
$$

β_0 and $\boldsymbol{\beta}_1$ can then be expressed as a function of α_0, $\boldsymbol{\alpha}_1$ and $\bar{\mathbf{x}}$ from solving these equations in a reverse direction,

$$
\begin{aligned}
\boldsymbol{\beta}_1 &= \frac{\boldsymbol{\alpha}_1}{(\alpha_0 + \bar{\mathbf{x}}^t \boldsymbol{\alpha}_1)\,(1 - \alpha_0 - \bar{\mathbf{x}}^t \boldsymbol{\alpha}_1)} \\
\beta_0 &= -\bar{\mathbf{x}}^t \boldsymbol{\beta}_1 - \log\left(\frac{1}{\alpha_0 + \bar{\mathbf{x}}^t \boldsymbol{\alpha}_1} - 1 \right)
\end{aligned} \tag{5}
$$

Therefore, the logistic regression coefficients, βs, can be estimated once the linear regression coefficients, αs, are estimated. In Section 3.2, we show how to estimate αs using aggregate data $\bar{y}_{i+}$ and $\bar{\mathbf{x}}_{i+}$. The reverse Taylor series approximation will be close to the MLE of $(\beta_0, \boldsymbol{\beta}_1)$ obtained from a logistic regression (with the individual y_{ij} and $\mathbf{x}_{ij}$) when the variance of the $\mathbf{x}_{ij}$s is small, i.e. $R(\mathbf{x}_{ij})$ is close to 0.

3 Deriving estimators from aggregate data

This section describes the derivation of the discriminate function estimator, the Reverse Taylor series approximation estimator, and the mean logistic regression estimator using aggregate data. An extension of the model used in this paper (i.e. fixed effects and single treatment in each study) to a random effects model and studies with two treatments, as in a clinical trial, is also discussed in this section.

3.1 Discriminant function estimator

As described in Section 2.1, to obtain the discriminant function estimator of $\boldsymbol{\beta}$, we need to estimate $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, $\boldsymbol{\Sigma}$, and $p(y_{ij}=1)$. Since y_{ij} is a Bernoulli variable, an estimator of $\mathrm{pr}(y_{ij}=1)$ is the grand mean $\bar{y}$

$$\bar{y} = \frac{\sum_{i=1}^{k} n_i \bar{y}_{i+}}{\sum_{i=1}^{k} n_i}$$

To estimate $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, and $\boldsymbol{\Sigma}$ using aggregate data, we can use the multivariate linear model applied to $\bar{y}_{i+}$ and $\bar{\mathbf{x}}_{i+}$. Although not always reasonable, to derive the discriminant function estimator, we must assume that $(\mathbf{x}_{ij}|y_{ij}=y) \sim N_{\mathrm{p}}(\boldsymbol{\mu}_0+\Delta y, \boldsymbol{\Sigma})$, i.e. $E(\mathbf{x}_{ij}|y_{ij}=y)=\boldsymbol{\mu}_0+\Delta y$ and $\mathrm{VAR}(\mathbf{x}_{ij}|y_{ij}=y)=\boldsymbol{\Sigma}$, for $y=0,1$. This implies that $\mathrm{E}(\sqrt{n_i}\bar{\mathbf{x}}_{i+}|\sqrt{n_i}\bar{y}_{i+})=\sqrt{n_i}\boldsymbol{\mu}_0+\Delta\sqrt{n_i}\bar{y}_{i+}$, and $\mathrm{VAR}(\sqrt{n_i}\bar{\mathbf{x}}_{i+}|\sqrt{n_i}\bar{y}_i)=\boldsymbol{\Sigma}$. Based on these two functions (mean and variance), the ordinary least squares (OLS) estimators of $\boldsymbol{\mu}_0$, Δ(i.e. $\boldsymbol{\mu}_1-\boldsymbol{\mu}_0$), and $\boldsymbol{\Sigma}$ can be obtained from the following multivariate linear regression model

$$E\begin{bmatrix}\sqrt{n_1}\bar{\mathbf{x}}_{1+}\\ \cdot\\ \cdot\\ \cdot\\ \cdot\\ \sqrt{n_k}\bar{\mathbf{x}}_{k+}\end{bmatrix}=\begin{bmatrix}\sqrt{n_1} & \sqrt{n_1}\bar{y}_{1+}\\ \cdot & \cdot\\ \cdot & \cdot\\ \cdot & \cdot\\ \cdot & \cdot\\ \cdot & \cdot\\ \sqrt{n_k} & \sqrt{n_k}\bar{y}_{k+}\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}_0\\ \Delta\end{bmatrix}$$

The OLS estimators of μ_0, Δ and $\boldsymbol{\Sigma}$ can then be used to calculate the logistic regression coefficients, βs, using equation (3). The variance of β_0 and $\boldsymbol{\beta}_1$ can also be calculated from $\boldsymbol{\Sigma}$ using the Delta method (Appendix A).

3.2 Reverse Taylor series approximation

As described in Section 2.2, the reverse Taylor series approximation gave $E(y_{ij}|\mathbf{x}_{ij}) \approx \alpha_0+\boldsymbol{\alpha}_1^t\mathbf{x}_{ij}$. This implies that $E(\bar{y}_{i+}|\bar{\mathbf{x}}_{i+}) \approx \alpha_0+\boldsymbol{\alpha}_1^t\bar{\mathbf{x}}_{i+}$. Since the variance of $(\bar{y}_{i+}|\bar{\mathbf{x}}_{i+})$ is proportional to n_i, a weighted OLS regression (with weights n_i) should be used to estimate α_0, and $\boldsymbol{\alpha}_1$ in this model. The estimators of α_0 and $\boldsymbol{\alpha}_1$ can then be used to calculate the logistic regression coefficients using equation (5). For the same reason that the variance of $(\bar{y}_{i+}|\bar{\mathbf{x}}_{i+})$ is not the same across studies, a robust jack-knife variance estimator proposed by Wu (1986) was used to estimate the variance of $\hat{\alpha}_0$ and $\hat{\boldsymbol{\alpha}}_1$ (see Appendix B). The estimate of the variance of $\hat{\beta}_0$ and $\hat{\boldsymbol{\beta}}_1$ was then calculated by the Delta method (Appendix B).

3.3 Reverse Taylor series approximation with one categorical covariate

Using a simple example, we show that when the covariates are indicator variables, which are derived from one categorical variable, the reverse Taylor series approximation without an intercept gives consistent estimates of βs in model 1. For example, suppose we want to predict the probability that drivers get injured in motor vehicle crashes (y_{ij}) by the type of restraint they used (x_{ij}s). If there are four types of restraints used: none, seat belt only, airbag only, and both seat belt and airbag, we can create four indicator variables: x_1, x_2, x_3, x_4. Suppose the following logistic regression model is the true model:

$$\operatorname{logit}(\operatorname{pr}(y_{ij} = 1 \mid x_{1ij}, x_{2ij}, x_{3ij}, x_{4ij})) = \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} \tag{6}$$

where $y_{ij} = 1$ if the jth driver in the ith city is injured and $y_{ij} = 0$ otherwise, $x_{kij} = 1$ if the jth driver in the ith city used the kth type restraint and $x_{kij} = 0$ otherwise, ($k = 1, 2, 3, 4$). There is an exact relationship between the βs in the logistic regression model and the αs in the linear model:

$$E(y_{ij} \mid x_{kij}) = \operatorname{pr}(y_{ij} = 1 \mid x_{kij}) = \alpha_1 x_{1ij} + \alpha_2 x_{2ij} + \alpha_3 x_{3ij} + \alpha_4 x_{4ij}$$

Note, $\operatorname{logit}(\operatorname{pr}(y_{ij} = 1 \mid x_{1ij}, x_{2ij}, x_{3ij}, x_{4ij})) = \operatorname{logit}(\alpha_1 x_{1ij} + \alpha_2 x_{2ij} + \alpha_3 x_{3ij} + \alpha_4 x_{4ij})$, so that β_k in the true logistic model in (6) is equal to $\operatorname{logit}(\alpha_k)$, $k = 1, 2, 3, 4$. If the only available data are the proportion of injured drivers and proportion of each type of restraint used in a city, then we can obtain consistent estimates of αs in the following linear model without an intercept:

$$E(\bar{y}_{i+} \mid \bar{x}_{ki+}) = \alpha_1 \bar{x}_{1i+} + \alpha_2 \bar{x}_{2i+} + \alpha_3 \bar{x}_{3i+} + \alpha_4 \bar{x}_{4i+}$$

and then obtain a consistent estimate of $\beta_k = \operatorname{logit}(\alpha_k)$ from the estimate of α_k. In this case, the mean logistic regression in (2) does not give consistent estimates of β_ks.

3.4 *Mean logistic regression*

If we expand the logistic function in model (1) about $\bar{\mathbf{x}}_{i+}$ in a Taylor series we obtain:

$$\begin{aligned} E(y_{ij} \mid \mathbf{x}_{ij}) &= \frac{e^{\beta_0 + \mathbf{x}_{ij}^t \beta_1}}{1 + e^{\beta_0 + \mathbf{x}_{ij}^t \beta_1}} \\ &= \left[\frac{e^{\beta_0 + \bar{\mathbf{x}}_{i+}^t \beta_1}}{1 + e^{\beta_0 + \bar{\mathbf{x}}_{i+}^t \beta_1}} + \frac{e^{\beta_0 + \bar{\mathbf{x}}_{i+}^t \beta_1}}{(1 + e^{\beta_0 + \bar{\mathbf{x}}_{i+}^t \beta_1})^2} \boldsymbol{\beta}_1^t (\mathbf{x}_{ij} - \bar{\mathbf{x}}_{i+}) \right] + R^\star(\mathbf{x}_{ij}) \end{aligned}$$

where $R^\star(\mathbf{x}_{ij})$ denotes the remainder term. By summing over the observations in study i, then dividing the sum by study sample size n_i, we obtain the expected value of $\bar{y}_{i+}$ for study i as:

$$E(\bar{y}_{i+} \mid i\text{th study}) = \frac{e^{\bar{\mathbf{x}}_{i+}^t \beta}}{1 + e^{\bar{\mathbf{x}}_{i+}^t \beta}} + \frac{\Sigma(R_{ij}^\star)}{n_i}$$

Neglecting the remainder term, this is the same as in model (2).

3.5 *Fixed versus random effects*

In model (1), we assume that a fixed effects model without a study effect is the correct model for the data. However, if the random effects model

$$E(y_{ij} \mid \mathbf{x}_{ij}, \delta_i) = \frac{e^{\mathbf{x}_{ij}^t \beta}}{1 + e^{\mathbf{x}_{ij}^t \beta}} + \delta_i \tag{7}$$

where δ_i is a random study effect with mean 0, is a more appropriate model for the data, the biases of the three estimates discussed will be the same as the fixed effects model in (1), but the variance will not be the same as the estimates when a fixed effects model holds. This is because, in the random effects model, the expected value of y_{ij} is the same as in the fixed effect model, but the variance of y_{ij} includes two components: the within and between study variance. In our method, we use a robust estimator to estimate the variance of $\hat{\beta}$, therefore, even if we misspecify the variance of y_{ij}, the variance estimator is still consistent. In the random effects model, observations in the same study are correlated; to simulate correlated data, the higher moments of y_{ij} are required. Since we are mainly interested in bias, and simulating correlated data from a random effects model is much more complicated, we chose to simulate data from a fixed effects model with no study effects.

In this paper, we have focused on studies with only one treatment per study, as in a pilot study. For future research we can derive methods to combine the study results of randomized trials with two treatments per study. In this type of trial, we obtain two aggregate proportions from each study, one from each of the treatment groups. The model (7) will then need to be modified by adding a subscript to indicate treatment ($k = 1, 2$).

$$E(y_{ijk} | \mathbf{x}_{ijk}, \delta_i) = \frac{e^{\mathbf{x}^t_{ijk}\beta}}{1 + e^{\mathbf{x}^t_{ijk}\beta}} + \delta_i$$

where we assume β already contains a treatment effect, and δ_i has mean 0. We can actually still use a fixed effects model to estimate β, treating the two proportions from the same studies as independent. However, the basic sampling unit is the study, and not the treatment group within the study. We have to calculate the variance of $\hat{\beta}$ by a robust jack-knife method (Wu, 1986) using the study as the basic sampling unit.

4 An application to an example

The example used is the earlier mentioned meta-analysis of the international literature on outcome in schizophrenia (Hegarty *et al.*, 1994). Only 63 studies, which adopt a narrow diagnostic system and used either drug or non-specific treatment, were included in this example. Each of these outcome studies examined the effect of a single treatment (drug or non-specific) on the binary outcome, improvement of schizophrenia (1 = improved, 0 = not improved). In addition to the treatment type, follow-up years was also included as the independent variable. However, only the mean follow-up years in each study is available, but not the follow-up years of each individual within the study.

All three estimation methods were first applied to the individual data. Since individual data were not available in this meta-analysis dataset, the individual data were derived from the aggregate data by assuming a sample size of one in each study, and rounding the proportion of improvement in each study to be either 0 or 1. The estimators of both the regression coefficients β and their standard errors obtained from three methods are fairly close to each other using individual data (Table 1). When the three methods were applied to the aggregate data, the reverse Taylor series approximation gives similar estimators of β as the mean logistic regression but much larger standard errors. The discriminant function estimators are, however, quite different from the other two estimators (Table 1). These differences could be due to the non-normality of the independent variables. To evaluate better the performance of these three estimators under various conditions, a simulation study was conducted.

TABLE 1. Estimators of β (and S.E.) of the schizophrenia outcome studies using three estimation methods

	Individual data			Aggregate data		
Covariates	Logistic	Reverse Taylor	Discriminant function	Mean logistic	Reverse Taylor	Discriminant function
Intercept	−1.917	−1.760	−1.920	−0.795	−0.775	−1.070
	(0.851)	(0.890)	(0.707)	(0.054)	(0.185)	(0.014)
Follow-up years	−0.037	−0.038	−0.038	−0.003	−0.003	−0.0003
	(0.066)	(0.046)	(0.064)	(0.003)	(0.010)	(0.001)
Treatment	−1.232	−1.190	−1.180	−0.429	−0.441	−0.047
	(1.202)	(0.960)	(1.180)	(0.052)	(0.182)	(0.022)

5 Simulation study

5.1 Simulation design

In our simulation study, we evaluated the performance of the three estimators of the logistic regression coefficients of two correlated covariates; one is continuous (x_{1ij}), and the other is dichotomous (x_{2ij}) with value 1 or 0. The two covariates are correlated because we let the mean and variance of x_{1ij} depend on the value of x_{2ij}. Within each study, the value of x_{1ij} varies among individuals, while the value of x_{2ij} does not and it is equal to x_{2i} for $j = 1, \ldots, n_i$. Two distribution types were considered for x_{1ij}: normal and exponential.

We simulated three random variables y_{ij}, x_{1ij}, x_{2i} with the joint density:

$$f(y_{ij}, x_{1ij}, x_{2i}) = f(y_{ij} | x_{1ij}, x_{2i}) f(x_{1ij} | x_{2i}) f(x_{2i})$$

where $f(x_{2i})$ is the Bernoulli density function with parameter p, $f(x_{1ij} | x_{2i})$ is either a normal or an exponential density, and $f(y_{ij} | x_{1ij}, x_{2i})$ is the Bernoulli density function with parameter p_y which is a logistic function of x_{1ij}, and x_{2i} with coefficients β_1, β_2 and an intercept β_0. Because the distribution of x_{1ij} might be different for each study, we draw the study parameters for each study from a normal distribution, then fix them for the simulation of the individual data within the study, i.e. the parameters of the distribution of $x_{1ij} | x_{2i}$ were different for each study, but were fixed within each simulation.

5.2 Simulation parameters

The parameters used in the simulation were based on the schizophrenia outcome studies meta-analysis data set mentioned above. Let y_{ij} indicate the improvement (binary outcome 1/0: yes/no), x_{1ij} indicate the follow up years (continuous variable) and x_{2ij} indicate the treatment type (dichotomous variable 1, 0: non-specific versus drug) for individual j in study i. The available data are the mean values of both the outcome variable, $\bar{y}_{i+}$ and the covariates, $\bar{x}_{1i+}$, $\bar{x}_{2i+}$ and the number of patients in each study, n_i. In this dataset $x_{2ij} = \bar{x}_{2i+}$ for $j = 1, \ldots, n_i$, i.e. the values of x_2 are all the same within each study. Also, the sample mean and variance of x_{1ij} are equal to 12 and 130 when $x_{2ij} = 1$, and are equal to 6.5 and 21 when $x_{2ij} = 0$. We used these values as the true value of the parameters in the simulation, i.e. $\mu_1 = 12$, $\sigma_1^2 = 130$, $\mu_0 = 6.5$, $\sigma_0^2 = 21$, from which to draw x_{1ij}.

Two sets of variances of the normal distribution $f(x_1 | x_2)$ were used, $(\sigma_1^2, \sigma_0^2) = (130, 21)$ and $(\sigma_1^2, \sigma_0^2) = (260, 42)$, where the first set is obtained from the

TABLE 2. Simulation design

Number of studies $k = 30, 60$
Sample size in each study $n = 30, 70$
Distribution of the dichotomous covariate x_2: Bernoulli (0.5)
Conditional distribution of the continuous covariate x_1:
• $(x_{1ij} \mid x_{2i} = 1, \mu_{1i}, \sigma) \sim N(\mu_{1i}, \sigma)$, where $\mu_{1i} \sim N(\mu = 12, \sigma = 130)$, or $N(12, 260)$
• $(x_{1ij} \mid x_{2i} = 0, \mu_{0i}, \sigma) \sim N(\mu_{0i}, \sigma)$, where $\mu_{0i} \sim N(\mu = 6.5, \sigma = 21)$, or $N(6.5, 42)$
or
• $(x_{1ij} \mid x_{2i} = 1, \lambda_{1i}) \sim \text{Exponential}(\lambda_{1i} = 12^{-1})$, or $\text{Exponential}(\lambda_{1i} = 17^{-1})$ where $\log(\lambda_{1i}) \sim N(-\log 12, 1/N_1)$ or $N(-\log 17, 1/N_1)$; N_1 is the number of studies with $x_{2i} = 1$
• $(x_{1ij} \mid x_{2i} = 0, \lambda_{0i}) \sim \text{Exponential}(\lambda_{0i} = 6.5^{-1})$, or $\text{Exponential}(\lambda_{0i} = 8.5^{-1})$ where $\log(\lambda_{0i}) \sim N(-\log 6.5, 1/N_0)$ or $N(-\log 8.5, 1/N_0)$; N_0 is the number of studies with $x_{2i} = 0$
Logistic regression coefficients β_0, β_1 and β_2: $(-0.90, -0.05, -0.35)$

schizophrenia dataset, the second set is twice the first set. To have similar values of mean and variance as the normal covariates, two sets of means of the exponential $f(x_1 \mid x_2)$ were chosen to be $(\mu_1, \mu_0) = (12, 6.5)$ and $(\mu_1, \mu_0) = (17, 8.5)$, which yield the variances $(144, 42)$ and $(289, 72)$. The number of studies k was chosen to be 30 and 60. For simplicity, the sample size of each study, n, was set to be the same for each study, which was either 30 or 70. The true regression coefficients were chosen to be $(\beta_0, \beta_1, \beta_2) = (-0.90, -0.05, -0.35)$ which were close to the mean logistic regression estimators obtained from this schizophrenia meta-analysis dataset. A summary of the parameters used in the simulation is listed in Table 2.

5.3 Simulation procedure

The simulation was carried out in a sequential order. We first generated k of x_{2i}, each of which was generated from a Bernoulli distribution with probability 0.5. Next, μ_{1i} (or μ_{0i}) was generated from a normal distribution with mean equal to μ_1 (or μ_0) and variance equal to σ_1^2 (or σ_0^2). Then, for each study, x_{1ij} was generated from $N(\mu_{1i}, \sigma_1^2)$ if $x_{2i} = 1$, and from $N(\mu_{0i}, \sigma_0^2)$ if $x_{2i} = 0$. Only one value was generated for each of the two covariates x_{2i} and x_{1ij}. The generated values of x_{2i}, and x_{1ij} were then used to generate y_{ij}. y_{ij} was generated 1000 times using the same generated x_{2i}, x_{1ij}, and the true logistic regression coefficients. In the case when x_{1ij} is exponentially distributed with mean equal to $\lambda_{1i}^{-1} = \mu_{1i}$ (or $\lambda_{0i}^{-1} = \mu_{0i}$), $\log(\lambda_{1i})$ (or $\log(\lambda_{0i})$) was generated from a normal distribution with mean equal to $-\log(\mu_1)$ (or $-\log(\mu_0)$) and variance equal to $1/N_1$ (or $1/N_0$), where N_1 (or N_0) is the number of the study with generated x_{2i} equal to 1 (or 0). Then, for each study, x_{1ij} was generated from $\text{Exponential}(\lambda_{1i})$ if $x_{2i} = 1$, and from $\text{Exponential}(\lambda_{0i})$ if $x_{2i} = 0$. Following the above simulation procedure, the two covariates x_1 and x_2 are correlated. The continuous covariate x_1 varies among individuals in any study. For the dichotomous covariate, x_2, its value does not vary among individuals within each study.

Once the individual data were generated, $\bar{y}_{i+}$, $\bar{x}_{1+}$ and $\bar{x}_{2+}$ of each study were calculated and then used to calculate the discriminant function estimators and the reverse Taylor series approximation. The mean values were also fitted by the mean logistic regression model. The bias and the 95% probability coverage of the estimated regression coefficients β_1 and β_2 were calculated.

TABLE 3. The standardized bias of $\hat{\beta}$

$x_1\|x_2$	k	n	Variance of $x_1\|x_2$		β_1 Mean logistic	β_1 Reverse Taylor	β_1 Discriminant function	β_2 Mean logistic	β_2 Reverse Taylor	β_2 Discriminant function
Normal	30	30	130	21	3.44	8.8	2.45	10.6	11.4	6.5
			260	42	6.27	22.1	9.89	21.4	22.9	15.4
		70	130	21	1.82	16.2	3.41	19.4	12.9	2.8
			260	42	3.58	11.0	−3.52	40.1	37.1	31.5
	60	30	130	21	4.82	12.3	6.73	12.7	9.7	4.0
			260	42	5.08	13.3	4.67	31.7	32.2	24.1
		70	130	21	5.33	14.3	8.74	19.9	15.9	11.5
			260	42	8.69	13.9	5.51	38.1	36.7	32.3
Exponential	30	30	144	42	14.1	18.2	14.2	−1.18	−0.39	−5.8
			289	72	22.9	31.4	27.5	−1.94	−1.19	−7.1
		70	144	42	15.9	17.2	11.0	1.13	3.75	−7.9
			289	72	35.1	56.5	53.1	3.85	−0.67	−10.6
	60	30	144	42	15.1	27.4	22.0	0.14	−0.93	−5.1
			289	72	23.0	33.7	27.9	−2.71	−4.35	−8.4
		70	144	42	23.1	31.9	29.3	−0.73	−8.66	−17.7
			289	72	33.2	45.6	46.4	−3.90	−8.33	−15.8

6 Simulation results

6.1 Standardized bias of $\hat{\beta}$: $Z_{\hat{\beta}} = (\hat{\beta} - \beta)/\text{s.e.}(\hat{\beta})$ (Table 3)

To take into account the variance of the estimates, the bias of $\hat{\beta}$ was standardized by subtracting the true value of β from the mean of 1000 values of $\hat{\beta}$ then dividing by the standard error of the mean of $\hat{\beta}$ and denoting it as $Z_{\hat{\beta}}$. By the central limit theorem, $Z_{\hat{\beta}}$ is distributed as $N(0, 1)$ if $\hat{\beta}$ is asymptotically unbiased for β. Unfortunately, the $Z_{\hat{\beta}}$ obtained from the three methods of estimation are significantly greater than zero in almost all simulation conditions. This indicates all three types of estimators are biased and usually overestimate the true β. In most cases, the $Z_{\hat{\beta}}$ of the mean logistic regression estimators are either the lowest or not substantially larger than the other two.

The simulation results also show that the magnitude of $Z_{\hat{\beta}}$ depends on the distribution type and variance of $x_1|x_2$. The standardized biases of β_1 are consistently larger when $x_1|x_2$ is exponential as compared to normal distribution. Although this is not restricted to the discriminant function estimator, which is based on normal assumption, the same was also observed in the other two types of estimators. For all three type of estimators, the $Z_{\hat{\beta}}$ of the regression coefficient of the dichotomous covariate, β_2, are larger than those of β_1 when $x_1|x_2$ is normal. On the contrary, the $Z_{\hat{\beta}}$ of β_2 are much smaller than those of β_1 when $x_1|x_2$ is exponential. As discussed in Section 2.2, the reverse Taylor series approximation will be less biased if the variance of the covariate is small. This is confirmed by the simulation. In most cases, a larger variance of x_1 is associated with a larger bias of not only β_1 but also β_2. This association was also observed for the other types of estimators. As for the effects of study sample size (n) and number of studies (k) on the bias, no clear and consistent association with $Z_{\hat{\beta}}$ was observed.

TABLE 4. The 95% CI coverage of $\hat{\beta}$

					β_1			β_2	
$x_1 \mid x_2$	k	n	Variance of $x_1 \mid x_2$	Mean logistic	Reverse Taylor	Discriminant function	Mean logistic	Reverse Taylor	Discriminant function
Normal	30	30	130 21	0.963	0.925	0.944	0.935	0.930	0.916
			260 42	0.954	0.899	0.943	0.915	0.904	0.919
		70	130 21	0.960	0.906	0.853	0.930	0.942	0.852
			260 42	0.956	0.952	0.905	0.798	0.813	0.758
	60	30	130 21	0.952	0.917	0.948	0.926	0.933	0.934
			260 42	0.962	0.940	0.971	0.846	0.833	0.886
		70	130 21	0.949	0.926	0.895	0.895	0.913	0.876
			260 42	0.951	0.925	0.896	0.786	0.780	0.779
Exponential	30	30	144 42	0.931	0.895	0.932	0.971	0.953	0.957
			289 72	0.907	0.841	0.854	0.966	0.950	0.947
		70	144 42	0.931	0.917	0.882	0.959	0.944	0.883
			289 72	0.826	0.596	0.527	0.968	0.950	0.866
	60	30	144 42	0.924	0.870	0.914	0.957	0.951	0.959
			289 72	0.902	0.825	0.858	0.947	0.947	0.942
		70	144 42	0.896	0.839	0.817	0.956	0.948	0.863
			289 72	0.834	0.696	0.631	0.949	0.936	0.857

6.2 Coverage of 95% CI for β (Table 4)

The coverage probability was calculated as the percentage of times in the 1000 repetitions that the 95% confidence interval, calculated using $\hat{\beta} \pm 1.96$(s.e. of $\hat{\beta}$), included the true value of β. With 1000 repetitions, if the true coverage is 95%, the proportion of 95% CI containing the true value should be in the interval $(0.936, 0.964)$.

When $x_1 \mid x_2$ is normal, the coverage of the 95% CI for β_1 using the mean logistic regression estimator is within the range 0.936 and 0.964. The coverage of the 95% CI for β_2 using the mean logistic regression estimator is, however, smaller than the lower limit of 0.936. The coverage for both β_1 and β_2 using the other two types of estimators is smaller than 0.936, i.e. the coverages are significantly less than 0.95 at the 0.05 level.

When $x_1 \mid x_2$ is exponential, the 95% CI coverages for $\hat{\beta}_1$ are smaller than the lower limit of 0.936 in all simulation conditions for all three types of estimators. While the 95% CI coverage for $\hat{\beta}_2$ is either in 95% CI range (0.936 to 0.964) or close to this range when it was calculated, based on the mean logistic regression estimator or the reverse Taylor estimators, most of the coverages calculated based on the discriminant function estimator are out of the 95% CI range.

7 Discussion

In this paper, we derived two methods for estimating the logistic regression coefficients using aggregate data: the reverse Taylor series estimator and the Discriminant function estimator. These two methods give similar estimators as the maximum likelihood estimators for the logistic regression coefficients when the individual data are used. However, these two methods can be biased when the aggregated data are used. Theoretically, when the variance of the covariate is small, the reverse Taylor series estimator will give a consistent estimator of β, and when the distribution of the covariate is normal over all studies, the Discriminant function estimator will give a consistent estimator. In all the conditions of the simulation study we performed, these two methods and the logistic regression model using the aggregate data all had a standardized bias significantly greater than zero. This indicates that the estimators are biased. The magnitudes of the biases are smaller for the mean logistic regression estimators than those of the other two types of estimators.

In the simulation study, the dichotomous covariate x_2 has the same value within each study; however, due to its correlation with the continuous covariate x_1, its regression coefficient was not correctly estimated using the aggregate data model. The conditional distribution type of $x_1|x_2$ affects the bias of both x_1 and x_2. When $x_1|x_2$ is normal, the bias of $\hat{\beta}_1$ is smaller than that of $\hat{\beta}_2$, when $x_1|x_2$ is exponential, the bias of $\hat{\beta}_1$ increases substantially, especially for the discriminant function estimator, and the bias of $\hat{\beta}_2$ is relatively small.

The results from the simulation study also show that when the variance of the continuous covariate increases, the standardized bias of its regression coefficient estimate increases for all three estimators. However, no clear relationship was observed between the sample size of each study and the bias and/or between the number of studies and the bias.

The performance of the three types of the estimators was also evaluated in terms of coverage probability. When $x_1|x_2$ is normal, the mean logistic regression estimator gives the best coverage of the 95% CI for β_1 and the coverage for β_2 is significantly smaller than 95% for all three types of estimators. When $x_1|x_2$ is exponential, all three types of estimators give significantly small coverage for β_1, while both the mean logistic regression and the reverse Taylor series estimators give good coverage for β_2 and the discriminant function estimator gives significantly small coverage.

Based on the simulation results, for the most part, the two estimators we derived did not perform as well as the mean logistic regression with respect to standardized bias and coverage probability. Although the mean logistic regression performed better than the other two methods, it also gives biased estimators and smaller than expected coverage of the 95% CI. Thus, analysts should be cautious in using aggregate data $\bar{y}_{i+}$ and $\bar{x}_{i+}$ to estimate the parameters of the logistic regression model for the underlying y_{ij} given the x_{ij}s, except under special conditions. If all covariates appear approximately normal, then the discriminant function estimator can be used since it will give consistent estimators. If there is only one categorical covariate, then the reverse Taylor series approximation will be the method of choice, as discussed in Section 3.3.

REFERENCES

EFRON, B. (1975) The efficiency of logistic regression compared to Normal discriminant analysis, *Journal of the American Statistical Association*, 70(352), pp. 892–898.

FRIDSTRØM, L. & INGERBRIGTSEN, S. (1991) An aggregate accident model based on polled, regional time-series data, *Accident Analysis and Prevention*, 23(5), pp. 363–378.

HEGARTY, J., BALDESSARINI, R. J., TOHEN, M., WATERNAUX, C. & OEPEN, G. (1994) One hundred years of schizophrenia: a meta-analysis of the outcome literature, *American Journal of Psychiatry*, 151(10), pp. 1409–1416.

NERLOVE, M. & PRESS, J. (1973) *Univariate and Multivariate Log-linear and Logistic Models*, R-1306, The Rand Corporation, Santa Monica, CA.

PRESS, J. & WILSON, S. (1978) Choosing between logistic regression and discriminant analysis, *Journal of the American Statistical Association*, 73(364), pp. 699–705.

WU, C. F. J. (1986) Jack-knife, bootstrap and other resampling methods in regression analysis, *The Annals of Statistics*, 14(4), pp. 1261–1295.

Appendix A

Derive the variance of discriminant function estimators using the Delta method

Let

$$Z = \begin{bmatrix} \sqrt{n_1}\bar{x}_{1+} \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \sqrt{n_k}\bar{x}_{k+} \end{bmatrix}, \quad D = \begin{bmatrix} \sqrt{n_1} & \sqrt{n_1}\bar{y}_{1+} \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ \sqrt{n_k} & \sqrt{n_k}\bar{y}_{k+} \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_0 \\ \Delta \end{bmatrix} \Rightarrow E(Z) = D\mu$$

discriminant function estimator

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

$$= \begin{bmatrix} \log\left[\dfrac{p(y_{ij}=1)}{1-p(y_{ij}=1)}\right] - \dfrac{1}{2}\mu_1^t\boldsymbol{\Sigma}^{-1}\mu_1 + \dfrac{1}{2}\mu_0^t\boldsymbol{\Sigma}^{-1}\mu_0 \\ \boldsymbol{\Sigma}^{-1}\Delta \end{bmatrix} = f(\mu)$$

By delta method

$$\mathrm{var}(\beta) = \left(\frac{\mathrm{d}\beta}{\mathrm{d}\mu}\right)\mathrm{var}(\mu)\left(\frac{\mathrm{d}\beta}{\mathrm{d}\mu}\right)^t$$

where

$$\frac{\mathrm{d}\beta}{\mathrm{d}\mu} = \begin{bmatrix} \dfrac{\mathrm{d}\beta_0}{\mathrm{d}\mu_0} & \dfrac{\mathrm{d}\beta_0}{\mathrm{d}\Delta} \\ \dfrac{\mathrm{d}\beta_1}{\mathrm{d}\mu_0} & \dfrac{\mathrm{d}\beta_1}{\mathrm{d}\Delta} \end{bmatrix} = \begin{bmatrix} \mu_0^t\boldsymbol{\Sigma}^{-1} & -\mu_1^t\boldsymbol{\Sigma}^{-1} \\ -\boldsymbol{\Sigma}^{-1} & \boldsymbol{\Sigma}^{-1} \end{bmatrix}$$

Appendix B

Derive the variance of reverse Taylor series approximation estimators using the Delta method

$$E(\bar{y}_{i+}|\bar{\mathbf{x}}_{i+}) = \alpha_0 + \boldsymbol{\alpha}_1^t\bar{\mathbf{x}}_{1+}, \qquad \text{let } \alpha = \begin{bmatrix} \alpha_0 \\ \boldsymbol{\alpha}_1 \end{bmatrix}, \quad Y = \begin{bmatrix} \bar{y}_{1+} \\ \vdots \\ \bar{y}_{k+} \end{bmatrix}, \quad X = \begin{bmatrix} \bar{\mathbf{x}}_{1+} \\ \vdots \\ \bar{\mathbf{x}}_{k+} \end{bmatrix}$$

then $\hat{\alpha} = (X^tWX)^{-1}X^tWY$ with robust variance equal to $(X^tWX)^{-1}X^tW\,\mathrm{var}(Y)WX(X^tWX)^{-1}$

where

$$W = \text{diag}\{n_i\}, \quad \text{var}(Y) = \text{diag}\{v_{ii}\} \text{ with } v_{ii} = \frac{(\bar{y}_{i+} - \hat{y})^2}{(1 - h_i)^2}$$

and h_i is the ith diagonal element of the matrix $XW(X^tWX)^{-1}WX^t$

The reverse Taylor series approximation estimator β is a function of α, i.e.

$$\beta = \begin{bmatrix} \beta_0 \\ \boldsymbol{\beta}_1 \end{bmatrix} = \begin{bmatrix} -\bar{\mathbf{x}}^t\boldsymbol{\beta}_1 - \log\left(\dfrac{1}{\alpha_0 + \bar{\mathbf{x}}^t\boldsymbol{\alpha}_1} - 1\right) \\ \dfrac{\boldsymbol{\alpha}_1}{(\alpha_0 + \bar{\mathbf{x}}^t\boldsymbol{\alpha}_1)\ (1 - \alpha_0 - \bar{\mathbf{x}}^t\boldsymbol{\alpha}_1)} \end{bmatrix} = f(\alpha)$$

By Delta Method,

$$\text{var}(\beta) = \left(\frac{d\beta}{d\alpha}\right)\text{var}(\alpha)\left(\frac{d\beta}{d\alpha}\right)^t$$

where

$$\frac{d\beta}{d\alpha} = \begin{bmatrix} \dfrac{d\beta_0}{d\alpha_0} & \dfrac{d\beta_0}{d\boldsymbol{\alpha}_1} \\ \dfrac{d\boldsymbol{\beta}_1}{d\alpha_0} & \dfrac{d\boldsymbol{\beta}_1}{d\boldsymbol{\alpha}_1} \end{bmatrix}$$

$$= \begin{bmatrix} -\bar{\mathbf{x}}^t\dfrac{d\boldsymbol{\beta}_1}{d\alpha_0} + \dfrac{1}{(1 - \alpha_0 - \bar{\mathbf{x}}^t\boldsymbol{\alpha}_1)\ (\alpha_0 + \bar{\mathbf{x}}^t\boldsymbol{\alpha}_1)} & -\bar{\mathbf{x}}^t\dfrac{d\boldsymbol{\beta}_1}{d\boldsymbol{\alpha}_1} + \dfrac{\bar{\mathbf{x}}^t}{(1 - \alpha_0 - \bar{\mathbf{x}}^t\boldsymbol{\alpha}_1)\ (\alpha_0 + \bar{\mathbf{x}}^t\boldsymbol{\alpha}_1)} \\ \dfrac{\boldsymbol{\alpha}_1(2\alpha_0 + 2\bar{\mathbf{x}}^t\boldsymbol{\alpha}_1 - 1)}{(\alpha_0 + \bar{\mathbf{x}}^t\boldsymbol{\alpha}_1)^2(1 - \alpha_0 - \bar{\mathbf{x}}^t\boldsymbol{\alpha}_1)^2} & \dfrac{(\alpha_0 + \bar{\mathbf{x}}^t\boldsymbol{\alpha}_1)\ (1 - \alpha_0 - \bar{\mathbf{x}}^t\boldsymbol{\alpha}_1) + \boldsymbol{\alpha}_1\bar{\mathbf{x}}^t(2\alpha_0 + 2\bar{\mathbf{x}}^t\boldsymbol{\alpha}_1 - 1)}{(\alpha_0 + \bar{\mathbf{x}}^t\boldsymbol{\alpha}_1)^2(1 - \alpha_0 - \bar{\mathbf{x}}^t\boldsymbol{\alpha}_1)^2} \end{bmatrix}$$