Analysis of Multivariate Disease Classification Data in the Presence of Partially Missing Disease Traits

Jingang Miao¹, Samiran Sinha¹*, Suojin Wang¹, W Ryan Diver² and Susan M Gapstur²

¹Department of Statistics, Texas A&M University, College Station, TX 77843, USA
²Epidemiology Research Program, American Cancer Society, Inc., Atlanta, GA 30303, USA

Abstract

In modern cancer epidemiology, diseases are classified based on pathologic and molecular traits, and different combinations of these traits give rise to many disease subtypes. The effect of predictor variables can be measured by fitting a polytomous logistic model to such data. The differences (heterogeneity) among the relative risk parameters associated with subtypes are of great interest to better understand disease etiology. Due to the heterogeneity of the relative risk parameters, when a risk factor is changed, the prevalence of one subtype may change more than that of another subtype does. Estimation of the heterogeneity parameters is difficult when disease trait information is only partially observed and the number of disease subtypes is large. We consider a robust semiparametric approach based on the pseudo-conditional likelihood for estimating these heterogeneity parameters. Through simulation studies, we compare the robustness and efficiency of our approach with that of the maximum likelihood approach. The method is then applied to analyze the associations of weight gain with risk of breast cancer subtypes using data from the American Cancer Society Cancer Prevention Study II Nutrition Cohort.

Keywords: Dimension reduction; Disease classification; Estimating equations; Etiologic heterogeneity; Missing disease trait; Polytomous logistic model

Introduction

While disease trait information has been used in understanding survival of patients, relatively less research has been done on incorporating disease trait information into etiologic investigations. In this paper, we propose a new pseudo-conditional likelihood approach that can handle partially missing disease traits and use it to analyze data from the American Cancer Society’s Cancer Prevention Study (CPS) II Nutrition Cohort [1]. The goal of the data analysis is to investigate whether the association between weight gain and risk of breast cancer varies among different disease trait subtypes in women not using postmenopausal hormones, adjusting for important risk factors. If the association of a predictor variable varies across the subtypes, we examined how much of this variation is due to each of the disease traits. Understanding etiologic heterogeneity of a risk factor sheds light on the pathogenesis of disease [2]. In the CPS-II Nutrition Cohort, there are 5 tumor characteristics, including stage (2 levels), histology (3 levels), estrogen receptor (2 levels), progesterone receptor (2 levels), and grade (3 levels), leading to 72 (i.e., $2 \times 3 \times 2 \times 2 \times 3$) different disease subtypes.

To examine the effect of risk factors on different disease subtypes, we consider the polytomous logistic regression, which is commonly used for handling multinomial data [3-5]. There are two variants of the model: one for nominal and one for ordinal scale outcomes [6], and this paper focuses on modeling nominal outcomes. Hence, for each disease subtype, we have a set of disease-predictor association/regression parameters and a set of nuisance intercept parameters. The etiologic heterogeneity will be measured via differences among the regression parameters across subtypes. The number of regression parameters is large due to several disease characteristics (traits) while each trait has multiple levels. In this context, a second-stage model was proposed to reduce the dimension of the heterogeneity parameters when all disease traits are observed [7]. In the CPS-II Nutrition Cohort data, the missingness percentages for the five traits are 23.2%, 21.2%, 0.0%, 30.0%, and 33.6%, respectively. In particular, among the cases, approximately 45.5% had at least one missing trait.

While estimation of the heterogeneity parameters was considered in the Cox regression model in the presence of partially missing disease traits [8], the same issue has not been considered before in the context of polytomous logistic model, which will be considered in this paper. We propose to estimate the heterogeneity parameters using a pseudo-conditional likelihood. We would like to point out the distinction between Chatterjee [7] and our approach. Here we adopt the second-stage model in a polytomous logistic regression setup in the presence of partially missing disease traits and develop a robust method of inference. In particular, Chatterjee [7] did not consider the missing data issue. As a result, his pseudo-conditional likelihood function was free of the nuisance intercept parameters. In contrast, we deal with partially missing disease trait data, and consequently our pseudo-conditional likelihood involves the nuisance intercept as well as the main log-odds ratio parameters. For estimating these nuisance parameters, we use a different type of pseudo-conditional likelihood. For handling the large dimension of the nuisance parameters, we adopt another second-stage model, and estimate them from another objective function. The idea of using two objective functions, one for the main parameters of interest and the other for the nuisance parameters, was inspired by Goetghebeur and Ryan [9]. Consequently the related theory is not a straightforward extension of the theory presented in [7].

Alternative to the proposed approach, one could consider a maximum likelihood based inference for the heterogeneity parameters using the full likelihood of the data. However, misspecification of the model for the intercepts will have less bearing on our inference than on
the full likelihood based approach. Simulation studies clearly indicate this robustness property of our approach. Our inference is based on an artificially constructed pseudo-conditional likelihood function. To show its validity, we derive the large sample properties of the resulting estimator.

A brief outline of the remainder of the article is as follows. Section 2 contains the model and assumptions. In Section 3, we describe the proposed estimation methodology. The results of some simulation studies are described in Section 4. As an illustration, our method is applied to analyze the CPS-II Nutrition Cohort data in Section 5. Some concluding remarks are given in Section 6.

The Appendix contains the general methodology, the asymptotic properties, and the details of the simulation designs.

**Model and Notation**

For each subject in a cohort of \( n \) subjects, when no missingness occurs we observe \((D,Y,X)\), where \( D \) takes on one or zero according to whether the subject is diagnosed with the disease or not during the follow-up period. For the sake of simplicity and easy understanding, we shall consider only two disease traits (i.e., \( K=2 \)) and assume that \( X \) is a scalar covariate (i.e., \( P=1 \)) in Sections 2 and 3. The general case of \( K \geq 2 \) and \( P \geq 1 \) is described in Appendix A. Thus, \( Y=(Y_1,Y_2)' \) carries information on 2 disease traits. For a disease-free subject, we have \( D=0 \) and \( Y=(0,0)' \). If the \( k \)-th trait has \( M_i \) levels, then there are a total of \( M=M_1 \times M_2 \) disease subtypes. Our model is

\[
\begin{align*}
\Pr(Y\mid X) &= \exp(\alpha(Y_1)+\beta(X)Y_1) \sum_{(Y_2,Y_{m_2})} \exp(\alpha(Y_2)+\beta(X)Y_2)
\end{align*}
\]

for \( i=1,...,n \), where \( \beta(Y_{m_2}) \) denotes the log-odds ratio parameter of the disease subtype \((Y_1,Y_2)\) for the covariate, \( \alpha \) denotes the nuisance intercept parameter, and \( \sum_{(y_{m_2})} \) means summing over all \( M \) subtypes of the disease.

For a scalar continuous covariate scenario, there are \( M \) main regression (log-odds ratio) parameters of interest along with \( M \) intercept parameters, which are not the main interest here. Etiologic heterogeneity is measured via the differences among the regression parameters for a given covariate, and our focus is on estimation of the heterogeneity parameters.

**Second-stage model**

To measure heterogeneity and reduce the dimension of regression parameters of interest along with \( M \) intercept parameters, which are not the main interest here. Etiologic heterogeneity is measured via the differences among the regression parameters for a given covariate, and our focus is on estimation of the heterogeneity parameters.

\[
\begin{align*}
\beta_{(y_{m_2})} &= \alpha(Y_1)+\beta(X)Y_1 \sum_{(Y_2,Y_{m_2})} \exp(\alpha(Y_2)+\beta(X)Y_2)
\end{align*}
\]

where \( \beta_{(y_{m_2})} \) is the regression coefficient corresponding to the reference subtype of the disease, and the first-order and second-order parameter contrasts are respectively represented by \( \beta_{(y_{m_2})}^{(1)} \) and \( \beta_{(y_{m_2})}^{(2)} \), \( k=1,2 \). By assuming certain contrasts to be zero, we can reduce the number of parameters. In addition, these assumptions can be tested. Assuming the second- and higher-order contrasts are equal to zero, which we call a second-stage additive model, \( \beta_{(y_{m_2})}^{(1)} \) tells us the degree of etiologic heterogeneity with respect to the first trait, regardless of the levels of other traits. For identifiability, we set \( \beta_{(y_{m_2})}^{(1)}=\beta_{(y_{m_2})}^{(2)}=0 \) and \( \beta_{(y_{m_2})}^{(2)}=\beta_{(y_{m_2})}^{(2)} = 0 \).

More elaborately, the heterogeneity of the log-odds ratio parameters due to the first trait can be measured via the contrasts \( \beta_{(y_{m_2})}^{(1)} \) and \( \beta_{(y_{m_2})}^{(2)} \).

By assuming the second-order contrast parameters to be zero [7], we reduce the dimension of regression parameters from \( M \times M \) to \( 1+M_1-M_1+M_2-1=M_1+M_2-1 \).

Importantly, the assumption regarding the contrast parameters are testable, provided data contain enough information regarding those parameters.

To simplify the notation in the second-stage model, we use a design matrix \( \mathbf{B} \) to relate the coefficient \( \beta \) that contains all the \( \beta_{(y_{m_2})} \) parameters of the unrestricted polynmomial model to the parameters \( \theta \) of the log-linear model (2) as \( \beta=B \theta \). In particular, \( \beta_{(y_{m_2})} = B_{(y_{m_2})} \theta \), where \( B_{(y_{m_2})} \) denotes the row of \( \mathbf{B} \) corresponding to disease subtype \((y_{m_2})\).

Note that the use of the second-stage model for the regression parameters is not just for dimension reduction. More importantly, these second-stage model parameters are our main interest. As mentioned previously, these parameters directly measure the heterogeneity in the log-odds ratio parameters due to each of the disease trait. For the purpose of dimension reduction we set second and higher-order contrasts to be zero. However, this is not the only way of reducing dimension. For instance, one may keep all the second-stage model parameters, and then adopt the LASSO technique [12] to choose the important second-stage model parameters.

**Missingness mechanism**

We introduce non-missing value indicator variables, \( R=(R_1,R_2)' \), where \( R_i=1 \) (\( k=1,2 \)) if the \( k \)-th trait is observed for diseased subject \( i \) and 0 otherwise. Since for a non-diseased subject there is no relevance of disease, for all non-diseased subjects we set \( R=(1,1)' \) for convenience. Note that there are at most \( 2^2 \) types of missing data patterns: \((0,0),(0,1),(1,0)\), and \((1,1)\). For example, \((1,0)\) represents the case when the first trait is observed but not the second one. We assume that the probability of observing missingness pattern \( r \) is \( \Pr(R=r\mid Y,X)=\pi(r\mid X) \), does not depend on the disease traits. However, we not only allow the missingness probabilities to depend on \( X \) (a case of missing at random, MAR, [13,14]) but also allow the missingness indicators of different traits, \( R_1 \) and \( R_2 \), to be dependent on each other.

We introduce some additional notations to be used in the next sections. For the \( i \)-th subject, whose missing data pattern is \( r \), we partition its vector of disease traits into the observed traits \( y^{o} \) and the missing traits \( y^{m} \). Similarly, we will use \( \Sigma_{y^{m}} \) to sum over all the possible values of \( y^{m} \). For example, if \( Y=y^{o} \) but \( Y^{m} \) is missing, then \( y^{o}=y^{o}_{1} \) if \( y^{m}=y^{m}_{2} \), whose value is missing, and \( \Sigma_{y^{m}} \) means summing over all the terms corresponding to \((Y=y^{o}_{1},Y^{m}=y^{m}_{2}) \), \((Y=y^{o}_{1},Y^{m}=y^{m}_{2}) \), \((Y=y^{o}_{2},Y^{m}=y^{m}_{2}) \), \((Y=y^{o}_{1},Y^{m}=y^{m}_{2}) \), \((Y=y^{o}_{1},Y^{m}=y^{m}_{2}) \). When both traits are observed, \( \Sigma_{y^{m}} \) just uses the term corresponding to \((Y=y^{o}_{1},Y^{m}=y^{m}_{2}) \).
Estimation Methodology

Maximum likelihood method in the context of missing data

To estimate \( \theta \), one can use the maximum likelihood estimator (MLE), which is obtained by maximizing the full likelihood

\[
L = \prod_{i=1}^{n} \left[ \frac{1}{1 + \sum_{j \neq i} \exp(A_i(y_{ij}) \theta + X_i D_i \theta)} \right]^{1/2} \times \prod_{i} \sum_{y_j \in \Omega(y_i)} \exp \left( \frac{A_i^T(y_{ij}) \theta + X_i D_i \theta}{\Omega(y_i)} \right)^{1/2},
\]

where

\[
\Omega(y_i) = \sum_{y_j \in \Omega(y_i)} \exp \left( \frac{A_i^T(y_{ij}) \theta + X_i D_i \theta}{\Omega(y_i)} \right)^{1/2}.
\]

The resulting score functions for \( \theta \) and \( \xi \) can be compactly written as

\[
S_\theta = \frac{\partial \log(L)}{\partial \theta} = \sum_{i} \left[ D_i X_i \sum_{j} I(R_i = r) \sum_{y_j \in \Omega(y_i)} \exp \left( A_i^T(y_{ij}) \theta + X_i D_i \theta \right) \right],
\]

\[
S_\xi = \frac{\partial \log(L)}{\partial \xi} = \sum_{i} \left[ D_i X_i \sum_{j} I(R_i = r) \sum_{y_j \in \Omega(y_i)} \exp \left( A_i^T(y_{ij}) \theta + X_i D_i \theta \right) \right] - \sum_{i} A_i \exp \left( A_i^T(y_{i1}) \theta + X_i D_i \theta \right),
\]

where

\[
\omega_i(y_{ij}, y_{i1}, x_i, \xi) = \sum_{y_j \in \Omega(y_i)} \exp \left( A_i^T(y_{ij}) \theta + X_i D_i \theta \right) \cdot \frac{\exp \left( A_i^T(y_{ij}) \theta + X_i D_i \theta \right)}{1 + \sum_{y_j \in \Omega(y_i)} \exp \left( A_i^T(y_{ij}) \theta + X_i D_i \theta \right)}.
\]

If the model assumptions (see Appendix B) hold, then under standard regularity conditions given in Theorem 5.41 of [15], the MLE \( \hat{\eta} \) asymptotically follows a normal distribution with mean \( \eta = (\theta^T, \xi^T)^T \), and the asymptotic variance can be consistently estimated by \(-\mathbb{E} \left( \frac{\partial^2 \log(L)}{\partial \theta^2} \right)^{-1} \) where \( \mathbb{E} \eta = \left( \mathbb{E} \eta_\theta, \mathbb{E} \eta_\xi \right)^T \).

As evident from the above discussion, the inference of the heterogeneity parameters, \( \theta \), depends on the intercept parameters \( \alpha \) and their model \( \alpha = \mathcal{A} \xi \). Next we discuss an alternative inference for the heterogeneity parameters, which is more robust against the misspecification of the second-stage model for \( \alpha \).

Pseudo-conditional likelihood in the context of missing data

In order to form pseudo-conditional likelihoods (PCL), for every subject with disease, we define a matched set \( S \) consisting of the subject itself and all subjects without the disease. Thus, if \( D_i = 1 \), then \( S_i = \{i \} \cup \{j : D_j = 0 \} \). If there are \( n_{0} \) controls, then the cardinality of \( S_i \) is \( n_i + 1 \). We form the pseudo-conditional likelihood \( \mathcal{L}_{PCL} \) such that the \( i \)-th subject has a disease of subtype \( (y_i^0, y_{i1}^0) \) given that there is only one subject with disease \( (y_i^*, y_{i1}^*) \) in the set \( S_i \). Then the pseudo-conditional likelihood is defined as the product of \( \mathcal{L}_{PCL} \) over \( i \), i.e.,

\[
\mathcal{L}_{PCL} = \prod_{i} \mathcal{L}_{PCL, i},
\]

and the estimating functions are defined as the derivatives of \( \log(\mathcal{L}_{PCL}) \) with respect to \( \theta \):

\[
S_{EE, \theta} = \frac{\partial \log(\mathcal{L}_{PCL})}{\partial \theta} = \sum_{i} \left[ D_i X_i \sum_{j} I(R_i = r) \sum_{y_j \in \Omega(y_i)} \exp \left( A_i^T(y_{ij}) \theta + X_i D_i \theta \right) \right] - \sum_{i} A_i \exp \left( A_i^T(y_{i1}) \theta + X_i D_i \theta \right),
\]

where

\[
\mathcal{L}_{PCL, i} = \frac{1}{n_{0}} \sum_{x_{i1} \in \Omega(x_{i1})} \frac{P(y_{i} \mid x_{i1}) \mathcal{L}_{PCL, i}(x_{i1})}{f_{X_{i1}}(x_{i1})},
\]

and

\[
\mathcal{L}_{PCL, i}(x_{i1}) = \frac{1}{n_{0}} \sum_{x_{i1} \in \Omega(x_{i1})} P(y_{i} \mid x_{i1}) \mathcal{L}_{PCL, i}(x_{i1}),
\]

Hence, by defining \( \mathcal{L}_{PCL} = \prod_{i} \mathcal{L}_{PCL, i} \), the estimating equations for \( \xi \) are

\[
\frac{\partial S_{EE, \xi}}{\partial \xi} = \sum_{i} \left[ D_i X_i \sum_{j} I(R_i = r) \sum_{y_j \in \Omega(y_i)} \exp \left( A_i^T(y_{ij}) \theta + X_i D_i \theta \right) \right] - \sum_{i} A_i \exp \left( A_i^T(y_{i1}) \theta + X_i D_i \theta \right),
\]

where

\[
\mathcal{L}_{PCL, i} = \frac{1}{n_{0}} \sum_{x_{i1} \in \Omega(x_{i1})} \frac{P(y_{i} \mid x_{i1}) \mathcal{L}_{PCL, i}(x_{i1})}{f_{X_{i1}}(x_{i1})},
\]

and

\[
\mathcal{L}_{PCL, i}(x_{i1}) = \frac{1}{n_{0}} \sum_{x_{i1} \in \Omega(x_{i1})} P(y_{i} \mid x_{i1}) \mathcal{L}_{PCL, i}(x_{i1}),
\]

Hence, by defining \( \mathcal{L}_{PCL} = \prod_{i} \mathcal{L}_{PCL, i} \), the estimating equations for \( \xi \) are
\[
S_{EE} = \frac{\partial \log L_{EE}}{\partial \xi} = \sum_{j=1}^{q} D_j \sum_{r} I(B_r = r) \sum_i \sum_{y_i} \frac{A_i | \gamma_i, y_i, y_i' \gamma_i'}{a_i | \gamma_i, y_i, y_i' \gamma_i'}
\]

\[
- \sum_{j=1}^{q} \sum_{y_i} \sum_{y_i'} \exp(\alpha_i | \gamma_i, y_i, y_i' \gamma_i') + \frac{X_j}{\beta_j} \frac{A_j | \gamma_j, y_j, y_j' \gamma_j'}{a_j | \gamma_j, y_j, y_j' \gamma_j'} = 0.
\]

We estimate \( \theta \) and \( \xi \) by solving \( S_{EE} = 0 \) and \( S_{EE} = 0 \) simultaneously. The estimating equations are asymptotically unbiased, as is shown in Appendix B. The asymptotic covariance of \( \hat{\eta} \) is estimated by a sandwich estimator. The middle component of the sandwich estimator is obtained via a linearization technique applied to the estimating equations.

**Simulation Studies**

**Simulation design**

One of the main goals of this numerical investigation was to show how robust our method is towards a misspecification of the intercept model in the presence of partially missing disease traits. We simulated cohort data of size \( n=5,000 \) by simulating \( X, Y, D \). The scalar covariate \( X \) was simulated from the Normal(0,1) distribution. We considered two scenarios each with 3 traits. First with \( 8=(2\times2\times2) \) disease subtypes, and second with \( 23=(2\times3\times3) \) disease subtypes. For each scenario we considered a correctly specified (denoted by a) second-stage model and a misspecified one (denoted by b) for the intercepts. We created missing values in each trait where missingness probabilities depended on \( X \). Two mechanisms were used: \( M_1 \) the missingness probabilities were dependent on \( X \) but the missingness of different traits was independent; and \( M_2 \) the missingness probabilities were dependent on \( X \) and the missingness of different traits was dependent. Overall disease probability lies between 6% and 9%.

For scenario 1, we considered three disease characteristics each with two levels, resulting in \( 2\times2\times2=8 \) disease subtypes. Assuming that the second- and higher-order contrasts for the relative risk parameters are negligible, we write

\[
\beta = \Phi \theta, \eta = \Phi \theta
\]

and we chose \( \theta = (0.35, 0.15, 0, 0.5)^T \). Thus the disease subtypes were generated using the model

\[
pr\{R=(1,1,1) \mid X\} = \frac{1}{1+\exp(-1.5+0.5 \gamma X)}
\]

We created missing values in the disease traits using two mechanisms. For \( M_1 \), the missingness probabilities were allowed to depend on \( X \) through the logistic function \( \exp(-1.5+0.5X) \). For \( M_2 \), 3 traits had \( 2^3=8 \) possible missingness patterns. For each case subject these patterns were generated from a multinomial distribution with the following probabilities

\[
pr\{R=(1,1,1) \mid X\} = \frac{1}{1+\exp(\gamma X)}
\]

We would be happy to provide these omitted results upon request.

Finally, we created missing values in the disease traits. For mechanism one, the missingness probabilities were allowed to depend on \( X \) through the logistic function \( \exp(\gamma X) \). For mechanism two, we allowed the missingness probabilities of around 0.2, 0.2, and 0.3 for the three disease traits, respectively. For mechanism two, we allowed the missingness probabilities to depend on each other in a similar pattern as in scenario 1.

**Method of analysis**

Each of the simulated datasets was analyzed by the maximum likelihood approach (MLE) and by the pseudo-conditional likelihood method (PCL). Furthermore, we analyzed the data considering only the subjects without any missing disease traits using the maximum likelihood approach, and we refer to it as the complete-case maximum likelihood estimator (CMLE). In all these analyses, we adopted the second-stage additive models for the regression and intercept parameters, \( \beta = \Phi \theta \) and \( \alpha = \Phi \xi \). We present mean, median, median absolute deviation (MAD), empirical standard errors (Emp. SE), estimated standard errors (Est. SE), 95% coverage probabilities, and root mean square errors (RMSE) of all the methods based on 2,000 replications. To assess asymptotic bias, we present \( B.score = \sqrt{\text{mean estimate - truth}} / \text{Emp. SE} \).

**Results**

To save space, in both scenarios we omit the results for missingness mechanism two, which are very similar to those for mechanism one. Also, we leave out results for the correctly specified intercept model case in scenario two. The conclusions that could be drawn from the results presented here. We would be happy to provide these omitted results upon request. The results for scenario 1a (top panel of Table 1) indicate that when the intercept model is correctly specified: (1) all three methods are asymptotically unbiased; (2) the standard errors of the PCL method were slightly larger than that of the MLE method, but smaller than that of the CMLE method, which suggests that the PCL’s efficiency is close to that of the MLE method; (3) similar to the standard errors, the
θ = 0.35

For the intercepts was misspecified. The missingness probabilities depended on the covariate. This is Scenario 2b.

The results were based on 2,000 runs. There were 2×2×2 = 8 disease subtypes. The missingness probabilities depended on the covariate.

Simulation results for the complete-case MLE, the MLE, and the pseudo-conditional likelihood method. Here MAD, Emp. SE, Est. SE, Bias, B. Score, RMSE, and CP denote median absolute deviation, empirical standard error, estimated standard error, root mean squared error, bias, bias score, root mean squared error, and 95% confidence probability based on the Wald-type confidence intervals, respectively. The results were based on 2,000 replications. There were 2×2×2 = 8 disease subtypes. The missingness probabilities depended on the Wald-type confidence intervals, respectively. The results were based on 2,000 runs. There were 2×3×5 = 30 disease subtypes. The model for the intercepts was misspecified. The missingness probabilities depended on the covariate. This is Scenario 2b.

Table 1: Simulation results for the complete-case MLE, the MLE, and the pseudo-conditional likelihood method. Here MAD, Emp. SE, Est. SE, Bias, B. Score, RMSE, and CP denote median absolute deviation, empirical standard error, estimated standard error, root mean squared error, bias, bias score, root mean squared error, and 95% confidence probability based on the Wald-type confidence intervals, respectively. The results were based on 2,000 runs. There were 2×3×5 = 30 disease subtypes. The model for the intercepts was misspecified. The missingness probabilities depended on the covariate. This is Scenario 2b.

Table 2: Simulation results for the complete-case MLE, the MLE, and the pseudo-conditional likelihood method. Here MAD, Emp. SE, Est. SE, Bias, B. Score, RMSE, and CP denote median absolute deviation, empirical standard error, estimated standard error, root mean squared error, bias, bias score, root mean squared error, and 95% confidence probability based on the Wald-type confidence intervals, respectively. The results were based on 2,000 runs. There were 2×3×5 = 30 disease subtypes. The model for the intercepts was misspecified. The missingness probabilities depended on the covariate. This is Scenario 2b.
RMSEs of the PCL method were slightly larger than that of the MLE method, but smaller than that of the CMLE method; (4) the estimated standard errors of the PCL method were close to that of the empirical standard errors; and (5) all methods' coverage probabilities were close to the nominal level (95%). The trend of the results remains the same for scenario 2a.

The results for scenarios 1b (bottom panel of Table 1) and 2b (Table 2) indicate that when the intercept model is misspecified: (1) the biases of both the MLE and the CMLE methods were prominent, but the biases of the PCL method were far less serious; (2) the comparisons of the three methods in terms of standard errors, RMSEs and estimated and empirical standard errors agreement were similar to those in the model with correctly specified model for the intercepts; and (3) the coverage probabilities of the MLE and the CMLE methods deviated from the nominal level, but the coverage probabilities of the PCL stayed close to the nominal level. Finally, the PCL method was almost as efficient as the MLE method in all scenarios. The bias of the CMLE method can be attributed to model misspecification of the model for the intercepts and ignoring the subjects with missing traits. However, the main source of bias in the MLE method is due to model misspecification.

Following a referee's comment we conducted additional simulation to study the performance of the three methods in the presence of non-null second-order contrasts in the true data generating process. As in Scenario 1, we used $2 \times 2 \times 2 = 8$ disease subtypes, the misspecification probabilities were made depended on $X$, and the intercept model was misspecified. But in addition to the original $\beta_0 = (0.35, 0.15, 0, 0.5)$, the true values of the second-order contrast parameters were taken as $\beta_{1(2,2)} = 0, \beta_{2(2,2)} = -0.2, \beta_{3(2,2)} = 0.2$. We call this Scenario 1c. For Scenario 1c, we first analyzed the simulated datasets assuming a second-stage additive model, meaning second- and higher-order contrast were set to zero. Then, we analyzed the datasets adopting a second-stage model keeping all first- and second-order contrasts parameters but setting third- and higher-order contrast parameters to zero. For the misspecified additive model (top panel of Table 3), the PCL method's biases were much smaller than either MLE or CMLE for all but one parameter, and its RMSE's were smaller than CMLE and sometimes smaller than MLE. With the second-order contrasts included in the model (bottom panel of Table 3) the PCL method also performed well with the smallest biases.

### Data Example

The CPS-II Nutrition Cohort is a prospective study of cancer incidence and mortality in 86,402 men and 97,786 women and has been described in detail elsewhere [1]. Briefly, the Nutrition Cohort is a subgroup of the approximately 1.2 million participants of the CPS-II Cohort, a prospective study of cancer mortality established by the American Cancer Society in 1982 [16]. Nutrition Cohort participants resided in 21 states with population-based cancer registries, were aged 50-74 years, and completed a 10-page confidential, self-administered mailed questionnaire at enrollment in 1992 or 1993.

Excluded from this analysis were Nutrition Cohort participants who were men ($n=86,402$); women who were using hormone replacement therapy ($n=33,407$), not post-menopausal ($n=3,514$), lost to followup (i.e., alive at the first follow-up questionnaire in 1997 but did not return the 1997 or any subsequent follow-up questionnaires) ($n=2,178$), reported a personal history of cancer other than non-melanoma skin cancer in 1992 ($n=9,520$), reported a diagnosis of breast cancer on the first survey that could not be verified through medical or cancer registry records or an in situ breast cancer ($n=174$), or the subjects with missing values in any of the predictor variables or whose weight gain was more than 100 lbs ($n=7,979$). Included in the analysis were 41,014 women. There were 1,555 incident cases of breast cancer (International Classification of Disease for Oncology, Second and Third Editions site code C50) that occurred between the date of the baseline questionnaire and June 30, 2007.

The risk factor of interest in the analysis was total weight change since age 18 to 1992 (WG) as it has been shown to be related to risk of breast cancer in previous studies (e.g., [8], [17], and [18]). WG was transformed to be between 0 and 1 for numerical stability. Using $(y_1, \ldots, y_5)$ to represent levels of the five traits, stage (2 levels), histology (3 levels), estrogen receptor (2 levels), progesterone receptor (2 levels), and grade (3 levels), we can write the polychotomous logistic model and the corresponding second-stage additive model as

$$
\text{pr}(D_i = d_{i(1)}, Y_i = (y_1, \ldots, y_5)) = \frac{\exp(\alpha_{(y_1, \ldots, y_5)} + \beta_{(y_1, \ldots, y_5)} Y_i)}{1 + \sum_g \exp(\alpha_{(y_1, \ldots, y_5)} + \beta_{(y_1, \ldots, y_5)} Y_i)},
$$

$$
\text{pr}(D_i = 0 | Y_i) = \frac{1}{1 + \sum_{(y_1, \ldots, y_5)} \exp(\alpha_{(y_1, \ldots, y_5)} + \beta_{(y_1, \ldots, y_5)} Y_i)},
$$

for $i=1, \ldots, n$. Contingency tables for the disease configurations can be found in Table 4. We used the second-stage additive models for both the intercepts and regression (log-odds ratio) parameters for all three methods. For the MLE and PCL methods we used all 1,555 cases while for the CMLE approach we used 848 cases whose disease traits information was complete.

The results are presented in Table 5. Since the PCL approach is more robust towards misspecification of the intercept model, we interpret the corresponding results here. Under the PCL method, we conclude that (1) the estimate of $\theta^{(d)}$ due to weight gain is positive and statistically significant at the 5% level. The odds ratio for the incidence

---

**Table 3**: Simulation results for the complete-case MLE, the MLE, and the pseudo-conditional likelihood method. Here RMSE represents root mean squared error. The results were based on 2,000 runs. There were $2 \times 2 \times 2 \times 8$ disease subtypes. The model for the intercepts was misspecified. The misspecification probabilities depend on the covariate. This is Scenario 1c, where the true values of some of the second-order contrasts of the log-odds ratio parameters were not zero.

| Scenario 1c: Additive model (misspecified) | $\theta^{(d=0.35)}$ | $\theta^{(d=0.15)}$ | $\theta^{(d=0)}$ | $\theta^{(d=0.5)}$
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
<td>CMLE</td>
<td>-0.090</td>
<td>0.023</td>
<td>0.022</td>
</tr>
<tr>
<td></td>
<td>MLE</td>
<td>-0.080</td>
<td>0.016</td>
<td>0.020</td>
</tr>
<tr>
<td></td>
<td>PCL</td>
<td>0.015</td>
<td>-0.012</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>RMSE</td>
<td>0.224</td>
<td>0.256</td>
<td>0.250</td>
</tr>
<tr>
<td></td>
<td>MLE</td>
<td>0.191</td>
<td>0.219</td>
<td>0.217</td>
</tr>
<tr>
<td></td>
<td>PCL</td>
<td>0.211</td>
<td>0.266</td>
<td>0.272</td>
</tr>
</tbody>
</table>

---

**Table 5**: Simulation results for the complete-case MLE, the MLE, and the pseudo-conditional likelihood method. Here RMSE represents root mean squared error. The results were based on 2,000 runs. There were $2 \times 2 \times 2 \times 8$ disease subtypes. The model for the intercepts was misspecified. The misspecification probabilities depend on the covariate. This is Scenario 1c, where the true values of some of the second-order contrasts of the log-odds ratio parameters were not zero.
Table 4: Contingency tables of disease traits configurations among 1,555 cases from the CPS-II Nutrition Cohort data. Here ER, PR, and NA, stand for estrogen receptor, progesterone receptor, and not available (missing), respectively.

<table>
<thead>
<tr>
<th>METH</th>
<th>Ref.</th>
<th>Grade (Well)</th>
<th>Stage (Localized)</th>
<th>Histology</th>
<th>ER Status (ER+)</th>
<th>PR Status (PR+)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Moderate</td>
<td>Poor</td>
<td>Distant</td>
<td>Lobular</td>
<td>Other</td>
</tr>
<tr>
<td></td>
<td></td>
<td>%missing</td>
<td>Homo sapien</td>
<td>%missing</td>
<td>Homo sapien</td>
<td>%missing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>23.2</td>
<td>21.2</td>
<td>0</td>
<td>33.6</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>θ(1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>−0.160</td>
<td>−0.033</td>
<td>0.703</td>
<td>−0.652</td>
<td>0.379</td>
</tr>
<tr>
<td></td>
<td></td>
<td>−0.007</td>
<td>0.415</td>
<td>0.354</td>
<td>0.433</td>
<td>0.523</td>
</tr>
<tr>
<td></td>
<td></td>
<td>p-value</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td>0.679</td>
<td>0.937</td>
<td>0.047</td>
<td>0.142</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.949</td>
<td>0.128</td>
<td>0.810</td>
<td>0.685</td>
<td>0.368</td>
</tr>
<tr>
<td></td>
<td></td>
<td>p-value</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.002</td>
<td>0.025</td>
<td>0.128</td>
<td>0.003</td>
<td>0.030</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.904</td>
<td>0.452</td>
<td>0.003</td>
<td>0.524</td>
<td>0.535</td>
</tr>
<tr>
<td></td>
<td></td>
<td>p-value</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td>0.307</td>
<td>0.263</td>
<td>0.307</td>
<td>0.349</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.711</td>
<td>0.303</td>
<td>0.349</td>
<td>0.376</td>
<td>0.308</td>
</tr>
<tr>
<td></td>
<td></td>
<td>p-value</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td>0.317</td>
<td>0.261</td>
<td>0.303</td>
<td>0.351</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.714</td>
<td>0.348</td>
<td>0.351</td>
<td>0.439</td>
<td>0.359</td>
</tr>
<tr>
<td></td>
<td></td>
<td>p-value</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>&lt;0.001</td>
<td>0.314</td>
<td>0.261</td>
<td>0.302</td>
<td>0.294</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.711</td>
<td>0.347</td>
<td>0.352</td>
<td>0.439</td>
<td>0.359</td>
</tr>
</tbody>
</table>

Table 5: Results of the CPS-II Nutrition Cohort data analysis with five disease traits and weight gain from age 18 to 1992 as the predictor. Here EST, SE, METH, ER and PR stand for estimate, standard error, method, estrogen receptor and progesterone receptor, respectively.

of breast cancer with well differentiated grade, localized stage, histology ductal, ER status positive and PR status positive for the 3rd quartile (45 lbs, re-scaled to be 0.476) of weight gain versus 1st quartile of weight gain (15 lbs, re-scaled to be 0.190) is 1.356 (exp(0.476 − 0.190) × 1:066), 95% confidence interval (CI): 1.164–1.580); (2) the PCL method produced statistically significant estimates of θθθ(1) = (θθθ(1) θθθ(2) θθθ(3) θθθ(4) θθθ(5)), and θθθ(1) for the covariance weight gain, which can be interpreted as follows. For a women who gained 45 pounds versus one who gained 15 pounds, the test statistic is

Following a referee’s suggestion, we conducted a model assessment for the data example. There are 72 log-odds ratio parameters, and as will be discussed in the last paragraph of this section, not all of these parameters are estimable. Now, we consider a two-stage model where all third- and higher-order contrast are zero. In this setup we test $H_0$: all second-order contrasts are zero against $H_a$: at least one of the second-order contrasts is non-zero. For this purpose we fit the model with all first- and second-order contrast parameters using the proposed PCL approach. The test statistic is $T = (A \hat{θ}) (A^T A)^{-1} A^T \tilde{A}$, where $A$ is a $19 \times 27$ matrix partitioned as $A = \{θ_{19 \times 4}, I_{19 \times 19}\}$ with $I_{19 \times 19}$ being an identity matrix of order 19, and $\Sigma$ stands for the asymptotic variance covariance matrix for $\hat{θ}$. Under $H_0$, $T$ approximately follows the $\chi^2_{19}$ distribution.
indicating that some second-order contrast parameters significantly (at the 5% level) differ from zero. Please see Table 6 for the new analysis with the second-stage model containing all first- and second-order contrast parameters. Although complex due to the presence of some non-null second-order contrast parameters, the model parameters of Table 6 can be interpreted. For example, we interpret $\theta^{(3,2)}_{(2)}$ as follows:

$$\exp\{\theta^{(3,2)}_{(2)}\} = \frac{\text{p(ER +)} \times \text{p(ER -)}}{\text{p(ER -)} \times \text{p(ER +)}}$$

Where G stands for Grade. Here the numerator is the odds ratio for Grade being Poor vs. Well associated with one unit increase in weight gain when ER status is −, whereas the denominator is the same odds ratio when ER status is +. Here $\theta^{(3,2)}_{(2)}$ is non-zero, so the odds ratio varies with the change of ER status. Also, due to estimation of more parameters, the standard errors of the estimators have substantially increased (please see the standard errors of the first-order contrast parameters in the PCL method in Table 5 vs. Table 6. This entire testing procedure demonstrates one of the good features of the proposed method that we can formally test our assumptions regarding the contrast parameters.

Prompted by a reviewer’s comment, here we discuss the issue of configurations with few or no subjects. There are no subjects in 23 out of the 72 possible disease subtypes. That means that a simple polytomous logistic model cannot be fit to this data with all 72 disease subtypes. The second-stage additive model, on the other hand, can enable us to make use of the cross classification structure and thus achieve sharing information across subtypes. The proposed method with second stage additive model still works when some subtypes have no cases observed. We require some cases for every level of each trait, which is easier to have than requiring cases for each subtype. Moreover, in the data example, our method works when the second-stage model contains all first- and second order contrast parameters. In fact, one may add more higher-order contrast parameters in the second-stage model, but these additional parameters may not be estimable from the data. For example, the third- and higher-order contrast parameters involving ER−, Grade Well, and Stage Distant are not estimable as the corresponding cell frequency is zero (the third panel of Table 4).

Table 6: Results of the CPS-II Nutrition Cohort data analysis using the PCL method with five disease traits and weight gain was used as the covariate. Here EST, SE, ER and PR stand for estimate, standard error, estrogen receptor and progesterone receptor, respectively. We have included second-order contrasts in the model for the log-odds ratio parameters.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\theta^{(0)}$</th>
<th>$\theta^{(1)}_{(2)}$</th>
<th>$\theta^{(1)}_{(3)}$</th>
<th>$\theta^{(1)}_{(2,2)}$</th>
<th>$\theta^{(1)}_{(2,3)}$</th>
<th>$\theta^{(1)}_{(3,2)}$</th>
<th>$\theta^{(1)}_{(3,3)}$</th>
<th>$\theta^{(2)}_{(2)}$</th>
<th>$\theta^{(2)}_{(3,2)}$</th>
<th>$\theta^{(2)}_{(3,3)}$</th>
<th>$\theta^{(2)}_{(1,2,2)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>EST</td>
<td>1.041</td>
<td>0.029</td>
<td>0.157</td>
<td>0.316</td>
<td>-0.502</td>
<td>-0.229</td>
<td>-3.991</td>
<td>-0.118</td>
<td>0.228</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SE</td>
<td>0.347</td>
<td>0.440</td>
<td>0.493</td>
<td>0.792</td>
<td>0.807</td>
<td>0.801</td>
<td>1.589</td>
<td>0.617</td>
<td>0.882</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p-value</td>
<td>0.003</td>
<td>0.948</td>
<td>0.750</td>
<td>0.690</td>
<td>0.534</td>
<td>0.775</td>
<td>0.012</td>
<td>0.849</td>
<td>0.796</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parameter</td>
<td>$\theta^{(2)}_{(3,2)}$</td>
<td>$\theta^{(2)}_{(3,3)}$</td>
<td>$\theta^{(2)}_{(1,2,2)}$</td>
<td>$\theta^{(2)}_{(3,2)}$</td>
<td>$\theta^{(2)}_{(3,3)}$</td>
<td>$\theta^{(2)}_{(1,2,2)}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EST</td>
<td>0.061</td>
<td>-0.290</td>
<td>4.680</td>
<td>-0.840</td>
<td>1.368</td>
<td>-0.112</td>
<td>0.063</td>
<td>3.569</td>
<td>-1.743</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SE</td>
<td>0.927</td>
<td>1.228</td>
<td>1.449</td>
<td>1.004</td>
<td>0.904</td>
<td>0.940</td>
<td>1.381</td>
<td>1.298</td>
<td>0.954</td>
<td></td>
<td></td>
</tr>
<tr>
<td>p-value</td>
<td>0.947</td>
<td>0.813</td>
<td>0.001</td>
<td>0.402</td>
<td>0.130</td>
<td>0.905</td>
<td>0.964</td>
<td>0.006</td>
<td>0.608</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Discussion

The two-stage model is an efficient and flexible way to measure heterogeneity of the odds ratios. It allows a sensible way to dimension reduction. For parameter estimation of the second-stage model, one can use the MLE, PCL, or the CMLE methods. Compared with the MLE method, our method reduces the effects of the intercepts on the estimation of the regression parameters, and thus it is more robust against the misspecification of the model for the intercepts.

When the model is correct, the PCL method is asymptotically unbiased. In addition, our simulations suggest (1) when the second-stage model for the intercepts is misspecified, our bias is usually smaller than that of either the MLE method or the CMLE method, and (2) with either correctly specified or misspecified model for the intercepts, our method can usually achieve efficiency that is very close to the MLE method.

Analysis of the Cancer Prevention Study (CPS-II) Nutrition Cohort data represents the first effort that the authors are aware of to simultaneously examine the effect of multiple covariates on the outcome. We hope that it not only is a demonstration of the method but also sheds light on the etiology of breast cancer.

Acknowledgments

The authors would like to thank Nilanjan Chatterjee for his input. This research was partially supported by NSF grant SES-0956188, NIH grant R03-CA176760, and Award Number KUS-CI-016-04 made by King Abdullah University of Science and Technology (KAUST). Part of the work was carried out while Wang was visiting Australian National University supported by the Mathematical Sciences Research Visitor Program.

References


Appendix

A General Methodology

Suppose that $X$ is a vector of $P$ covariates, and $Y = (Y_1, ..., Y_k)$ carries information on disease traits, and $M$ is the total number of disease subtypes, based on all possible combinations of the various traits. We will use $M$ for $k$. Our model is

$$ p_{i,y} = \exp \left( \alpha + \sum_{p=1}^{P} \beta^{(p)} y_{i,p} \right) I \left( 1 + \sum_{y} \exp \left( \alpha + \sum_{p=1}^{P} \beta^{(p)} y_{i,p} \right) \right) $$

and

$$ \Pr(D = 0 | X_i) = 1 / \left( 1 + \sum_{y} \exp \left( \alpha + \sum_{p=1}^{P} \beta^{(p)} y_{i,p} \right) \right) $$

for $n$. For $k$ disease subtypes, we have main regression parameters of interest along with intercept parameters. The log-linear model for the log-odds ratio parameter is

$$ \beta^{(p)} = \beta_{(y_{1}, ..., y_{k})}^{(p)} = \theta^{(0)} + \sum_{k=1}^{K} \theta^{(k)}(\alpha_{(y_{1}, ..., y_{k})}) + \sum_{k=1}^{K} \sum_{k=2}^{K} \theta^{(k)}(\alpha_{(y_{1}, ..., y_{k})}) + \cdots + \theta^{(k)}(\alpha_{(y_{1}, ..., y_{k})}) $$

(1)

Suppose that is the set of log-odds ratio parameter corresponding to , then the second-stage model can be written as $\beta$. From here on, we denote by $\theta$. For each subject we introduce a vector of binary variables , where $y_{i,j} = 1$ if the $j$ trait is observed and 0 otherwise. For our convenience, we set $y_{i,k} = 0$ for a non-diseased subject. Using our methodology the estimating functions for are

$$ S_{k, \beta^{(p)}} = \frac{\partial \log (L_{k, \beta^{(p)}})}{\partial \beta^{(p)}} = \sum_{i=1}^{n} \sum_{r} I(R_i = r) \left\{ X_{i,r} \sum_{y} E_{(\gamma_{r}, ..., \gamma_{r})}^{(p)}(\theta^{(p)}) X_{i,r} B_{(\gamma_{r}, ..., \gamma_{r})}^{(p)} + M_{\gamma_{r}, ..., \gamma_{r}}^{(0)}(Q_{0,0}) \right\} $$

$$ - n_{0}^{-1} \sum_{y} \exp \left( A_{(\gamma_{r}, ..., \gamma_{r})}^{T} \xi + \sum_{p=1}^{P} X_{i,r} \sum_{y} B_{(\gamma_{r}, ..., \gamma_{r})}^{(p)}(\theta^{(p)}) X_{i,r} B_{(\gamma_{r}, ..., \gamma_{r})}^{(p)} + M_{\gamma_{r}, ..., \gamma_{r}}^{(0)}(Q_{0,0}) \right) $$

where for

$$ M_{\gamma_{r}, ..., \gamma_{r}}^{(k)}(Q_{0,0}) = n_{0}^{-1} \sum_{j \in \gamma_{r}} \exp \left( A_{(\gamma_{r}, ..., \gamma_{r})}^{T} \xi + \sum_{p=1}^{P} X_{i,r} \sum_{y} B_{(\gamma_{r}, ..., \gamma_{r})}^{(p)}(\theta^{(p)}) X_{i,r} B_{(\gamma_{r}, ..., \gamma_{r})}^{(p)} \right) $$

$$ = \int \sum_{y} \exp \left( A_{(\gamma_{r}, ..., \gamma_{r})}^{T} \xi + \sum_{p=1}^{P} X_{i,r} \sum_{y} B_{(\gamma_{r}, ..., \gamma_{r})}^{(p)}(\theta^{(p)}) X_{i,r} B_{(\gamma_{r}, ..., \gamma_{r})}^{(p)} \right) dQ_{0,0}(X) $$

and $Q_{0,0}(x) = n_{0}^{-1} \sum_{i=0}^{n} I(D_i = 0, X_i = x)$ denotes the empirical distribution function of among the controls which converges in probability to the true distribution of among the controls denoted by $Q_{0,0}$. Here for 2, respectively. We want to clarify that in signifies a summation over all possible values of the indicator vector . If there are three traits, then the possible values of are
The estimating functions for \( P \) are
\[
S_{E,\theta,\xi} = \frac{\partial \log \left( L_{N_{CL}} \right)}{\partial \xi} = \sum_{i=1}^{n} D_i \left[ \sum_j I(R_i = r) \sum_{y_{ij}, r_{ij}} A_{Y_{ij}, r_{ij}}^T \omega_{Y_{ij}, r_{ij}} \right]
\]
\[
- \frac{n_0}{\sum_j} \exp \left( A_{Y_{ij}, r_{ij}}^T \xi + \sum_{p=1}^{P} X_{y_{ij}, p} B_{Y_y}^{(p)} \theta^{(p)} \right) + N^{(i)}(Q_{0n})
\]
where for 1,
\[
N^{(i)}(Q_{0n}) = \frac{n_0}{\sum_j} \sum_{y_{ij}, r_{ij}} \exp \left( A_{Y_{ij}, r_{ij}}^T \xi + \sum_{p=1}^{P} X_{y_{ij}, p} B_{Y_y}^{(p)} \theta^{(p)} \right) A_{Y_{ij}, r_{ij}}^{\alpha_k}
\]
\[
= \sum_{y_{ij}, r_{ij}} \exp \left( A_{Y_{ij}, r_{ij}}^T \xi + \sum_{p=1}^{P} X_{y_{ij}, p} B_{Y_y}^{(p)} \theta^{(p)} \right) A_{Y_{ij}, r_{ij}}^{\alpha_k} dQ_{0n}(X).
\]
We estimate \( \theta^T, \xi \) by solving 0, \( P \), simultaneously. Denote the resulting estimator as \( \hat{\theta}^T, \hat{\xi} \).

B Asymptotic Properties
In this section, we discuss the large sample properties of \( \hat{\eta} \). We show that \( n^{-1} S_{E,\theta,\xi} \rightarrow 0 \) in probability, i.e., the estimating equations are asymptotically unbiased.

Regularity conditions:

Let \( S_0(\eta) = n^{-1}(S_{E,\theta,\xi}^T, \ldots, S_{E,\theta,\xi}^T)^T \).

1. The parameter space for \( \theta, \xi \) is a compact subset of an Euclidean space.

2. \( 0 < \exp \left( \sum_{p=1}^{P} X_{y_{ij}, p} B_{Y_y}^{(p)} \theta^{(p)} \right) < \infty \) for all \( \theta \) and \( y \).

3. \( 0 < \exp \left( A_{Y_{ij}, r_{ij}}^T \theta^{(p)} \right) < \infty \) for all \( \xi \) and \( y \).

4. The elements of the second-stage design matrices \( B \) and \( \xi \) remain uniformly bounded in absolute value by constants, say \( C_{B} \) and \( \xi \), respectively.

5. The information matrix \( H \) is positive definite.

6. The deterministic equation has only one root in the neighborhood of the true parameters.

Conditions C1-C4 are required for uniform convergence, i.e., \( n^{-1} S_{E,\theta,\xi} \rightarrow 0 \). Condition C5, C6 (identifiability) and the asymptotic unbiasedness of \( \hat{\eta} \) for zero (to be proved) together imply convergence of the estimator in probability towards the true value (Theorem 5.9 of [15]).

Asymptotic Unbiasedness:

Here we first show that as \( n \rightarrow \infty \) at the true parameter value. Due to the law of large numbers, converges to its expectation. In order to calculate this expectation, we shall use the conditional
probability that the $i$-th subject has disease of type $\theta$ given that there is one diseased subject in the matched set with this disease type. Hence,

$$E \left( \frac{S_{\theta,d}(\theta)}{n} \right) = E \left[ \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \exp \left( A_{(\gamma_j,\gamma_k)}^{T} \xi + \sum_{p=1}^{\rho} X_{j,p} B_{(\gamma_j,\gamma_k)}^{T} \theta^p \right) \pi(r, X_i) \right] \pi(r, X_i)$$

$$\times \left\{ \frac{\sum_{i=1}^{n} \exp \left( A_{(\gamma_i,\gamma_k)}^{T} \xi + \sum_{p=1}^{\rho} X_{i,p} B_{(\gamma_i,\gamma_k)}^{T} \theta^p \right) X_{i,r} B_{(\gamma_i,\gamma_k)}^{T} \theta^p \right\} - \frac{\mathcal{M}_{(\gamma_j,\gamma_k)}^{(1)}(Q_0)}{\mathcal{M}_{(\gamma_j,\gamma_k)}^{(0)}(Q_0)} \right\} d\mu(Y_{\gamma_i}) + o(1).$$

Now, the first term on the right hand side above is

$$E \left[ \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \exp \left( A_{(\gamma_j,\gamma_k)}^{T} \xi + \sum_{p=1}^{\rho} X_{j,p} B_{(\gamma_j,\gamma_k)}^{T} \theta^p \right) \pi(r, X_i) \right] \pi(r, X_i)$$

$$\times \left\{ \frac{\sum_{i=1}^{n} \exp \left( A_{(\gamma_i,\gamma_k)}^{T} \xi + \sum_{p=1}^{\rho} X_{i,p} B_{(\gamma_i,\gamma_k)}^{T} \theta^p \right) X_{i,r} B_{(\gamma_i,\gamma_k)}^{T} \theta^p \right\} d\mu(Y_{\gamma_i}) \right]\,$$

$$= E \left[ \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \exp \left( A_{(\gamma_j,\gamma_k)}^{T} \xi + \sum_{p=1}^{\rho} X_{j,p} B_{(\gamma_j,\gamma_k)}^{T} \theta^p \right) \pi(r, X_i) \right] \pi(r, X_i)$$

$$\times \left\{ \frac{\mathcal{M}_{(\gamma_j,\gamma_k)}^{(1)}(Q_0)}{\mathcal{M}_{(\gamma_j,\gamma_k)}^{(0)}(Q_0)} \right\} d\mu(Y_{\gamma_i}) \right].$$

and the second term is

$$E \left[ \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \exp \left( A_{(\gamma_j,\gamma_k)}^{T} \xi + \sum_{p=1}^{\rho} X_{j,p} B_{(\gamma_j,\gamma_k)}^{T} \theta^p \right) \pi(r, X_i) \right] \pi(r, X_i)$$

$$\times \left\{ \frac{\mathcal{M}_{(\gamma_j,\gamma_k)}^{(1)}(Q_0)}{\mathcal{M}_{(\gamma_j,\gamma_k)}^{(0)}(Q_0)} \right\} d\mu(Y_{\gamma_i}) \right].$$

The difference between the two terms is easily seen to be asymptotically the expected weighted conditional covariance between

$$\exp \left( A_{(\gamma_j,\gamma_k)}^{T} \xi + \sum_{p=1}^{\rho} X_{j,p} B_{(\gamma_j,\gamma_k)}^{T} \theta^p \right) \pi(r, X_i) \right] \pi(r, X_i)$$

$$\times \left\{ \frac{\mathcal{M}_{(\gamma_j,\gamma_k)}^{(1)}(Q_0)}{\mathcal{M}_{(\gamma_j,\gamma_k)}^{(0)}(Q_0)} \right\} d\mu(Y_{\gamma_i}) \right].$$

and weight

$$\exp \left( A_{(\gamma_j,\gamma_k)}^{T} \xi + \sum_{p=1}^{\rho} X_{j,p} B_{(\gamma_j,\gamma_k)}^{T} \theta^p \right) \pi(r, X_i) \right] \pi(r, X_i)$$

$$\times \left\{ \frac{\mathcal{M}_{(\gamma_j,\gamma_k)}^{(1)}(Q_0)}{\mathcal{M}_{(\gamma_j,\gamma_k)}^{(0)}(Q_0)} \right\} d\mu(Y_{\gamma_i}) \right].$$
Let
\[
\text{cov}^{\pi}\left\{ \pi(r, X), X \big| E_{(y^1, y^2)} \right\} = \sum_{j \in S} \sum_{r \in S} \text{exp} \left( A_T^{r, (y^1, y^2)} \xi + \sum_{p=1}^P X_{j,p} B_T^{r, (y^1, y^2)} \theta^{(p)} \right)
\]
\[
\times \frac{\text{exp} \left( A_T^{r, (y^1, y^2)} \xi + \sum_{p=1}^P X_{j,p} B_T^{r, (y^1, y^2)} \theta^{(p)} \right) \pi(r, X)}{\sum_{j \in S} \sum_{r \in S} \text{exp} \left( A_T^{r, (y^1, y^2)} \xi + \sum_{p=1}^P X_{j,p} B_T^{r, (y^1, y^2)} \theta^{(p)} \right) \pi(r, X)}
\]
\[
\times \left\{ X_{j,p} E^{r, (y^1, y^2)} - \frac{N^{(i)}(y_i^p, \theta^{(p)})}{N^{(0)}(y_i^p, \theta^{(p)})} \right\},
\]
where
\[
\bar{\pi}(r, X) = \sum_{j \in S} \sum_{r \in S} \frac{\text{exp} \left( A_T^{r, (y^1, y^2)} \xi + \sum_{p=1}^P X_{j,p} B_T^{r, (y^1, y^2)} \theta^{(p)} \right) \pi(r, X)}{\sum_{j \in S} \sum_{r \in S} \text{exp} \left( A_T^{r, (y^1, y^2)} \xi + \sum_{p=1}^P X_{j,p} B_T^{r, (y^1, y^2)} \theta^{(p)} \right) \pi(r, X)}
\]
Then, we can write
\[
E \left( \frac{S_{EE, \theta^{(p)}}}{n} \right) = E \left[ \sum_{k \in S} \text{cov}^{\pi} \left\{ \pi(r, X), X \big| E_{(y^1, y^2)} \right\} d\mu(y_i^p) \right] + o(1) = o(1),
\]
where the last equality follows due to the fact that \( \sum_{\pi} \pi(r, X) = 1 \).

Similarly, due to the law of large numbers, \( n^{-1}S_{EE, \theta} \) converges to its expectation. In order to calculate this expectation, we shall use the conditional probability that the i-th subject has disease of type \( y = (y_i^{\theta^1}, y_i^{\theta^2}) \) given that there is one diseased subject in the matched set \( S_i \) but without specifying any disease subtype information. Hence,
\[
E \left( \frac{S_{EE, \theta}}{n} \right) = E \left[ \sum_{k \in S} \sum_{r \in S} \pi(r, X_k) \frac{\sum_{i \in S} \text{exp} \left( A_T^{r, (y^1, y^2)} \xi + \sum_{p=1}^P X_{k,p} B_T^{r, (y^1, y^2)} \theta^{(p)} \right) A_{i, (y^1, y^2)}}{\sum_{i \in S} \text{exp} \left( A_T^{r, (y^1, y^2)} \xi + \sum_{p=1}^P X_{k,p} B_T^{r, (y^1, y^2)} \theta^{(p)} \right) A_{i, (y^1, y^2)}} \right] d\mu(y_i^p) + o(1)
\]
\[
= E \left[ \sum_{k \in S} \pi(r, X_k) \sum_{i \in S} \text{exp} \left( A_T^{r, (y^1, y^2)} \xi + \sum_{p=1}^P X_{k,p} B_T^{r, (y^1, y^2)} \theta^{(p)} \right) A_{i, (y^1, y^2)} \right] d\mu(y_i^p)
\]
\[
\sum_{i \in S} \sum_{r \in S} \pi (r, X_i) = \int \sum_{i \in S} \sum_{r \in S} \exp \left( A (r, X_i) + \sum_{i=1}^{P} X_{i, p} B_i \theta (p) \right) d \mu (X_i) \times \mathcal{N}^{(1)} (Q_0) / \mathcal{N}^{(0)} (Q_0) + o(1)
\]

\[
= \mathbb{E} \left[ \sum_{i \in S} \sum_{r \in S} \pi (r, X_i) \sum_{r \in S} \exp \left( A (r, X_i) + \sum_{i=1}^{P} X_{i, p} B_i \theta (p) \right) A_i \right]
\]

\[
- \sum_{i \in S} \sum_{r \in S} \sum_{r \in S} \exp \left( A (r, X_i) + \sum_{i=1}^{P} X_{i, p} B_i \theta (p) \right) \times \mathcal{N}^{(1)} (Q_0) / \mathcal{N}^{(0)} (Q_0) \right] + o(1).
\]

Now using the facts that \( \sum_{r \in S} \pi (r, X_i) = 1 \) and

\[
\sum_{i \in S} \sum_{r \in S} \exp \left( A (r, X_i) + \sum_{i=1}^{P} X_{i, p} B_i \theta (p) \right) A_i \times \mathcal{N}^{(1)} (Q_0) / \mathcal{N}^{(0)} (Q_0),
\]

we obtain that \( n^{-1} S_{EE, p} = 0 \).

**Asymptotic Normality:**

Note that for large \( n \)

\[
\frac{1}{\sqrt{n}} S_{EE, p} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_i \sum_{r \in S} I (R_i = r) \left\{ X_i, \sum_{r \in S} B_i \right\} - \frac{\mathcal{M}^{(1)} (Q_0)}{\mathcal{M}^{(0)} (Q_0)} + o(1)
\]

\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_i \sum_{r \in S} I (R_i = r) \left\{ X_i, \sum_{r \in S} B_i \right\} - \frac{\mathcal{M}^{(1)} (Q_0)}{\mathcal{M}^{(0)} (Q_0)} + o(1)
\]

(B1)

Let \( \hat{a} = \mathcal{M}^{(1)} (Q_0) \) and \( \hat{b} = \mathcal{M}^{(0)} (Q_0) \). Then using the fact that

\[
\frac{\hat{a}}{\hat{b}} - a - \frac{a}{\hat{b}} \left( \hat{b} - b \right) = o \left( n^{-1/2} \right)
\]

the summand of the second term of (B1) is

\[
\frac{\mathcal{M}^{(1)} (Q_0)}{\mathcal{M}^{(0)} (Q_0)} - \frac{\mathcal{M}^{(1)} (Q_0)}{\mathcal{M}^{(0)} (Q_0)} = \frac{1}{n \mathcal{M}^{(0)} (Q_0)} \sum_{i=1}^{n} (1 - D_i) \sum_{r \in S} \sum_{r \in S} \exp \left( A (r, X_i) + \sum_{i=1}^{P} X_{i, p} B_i \theta (p) \right)
\]
\[
\frac{1}{\sqrt{n}} S_{\text{MLE}, \phi}^2 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{p} I(R_i = r) D_j \left\{ \sum_{\gamma_{i,j} \gamma_{i,j',x_i}} \alpha_{\gamma_{i,j} \gamma_{i,j'} x_i} A_{\gamma_{i,j} \gamma_{i,j'}} - \frac{\mathcal{N}^{(0)}(Q_0)}{\mathcal{N}^{(0)}(Q_0)} \right\} + o_p(1). 
\]

Plugging (B2) into (B1) and changing the order of the two summations in the second term, we have

\[
\frac{1}{\sqrt{n}} S_{\text{MLE}, \phi}^2 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{p} I(R_i = r) D_j \left\{ \sum_{\gamma_{i,j} \gamma_{i,j',x_i}} \alpha_{\gamma_{i,j} \gamma_{i,j'} x_i} A_{\gamma_{i,j} \gamma_{i,j'}} - \frac{\mathcal{N}^{(0)}(Q_0)}{\mathcal{N}^{(0)}(Q_0)} \right\} 
\]

Finally, applying the strong law of large numbers and the Slutsky's Theorem, we obtain

\[
n^{-1/2} S_{\text{MLE}, \phi} \overset{d}{=} n^{-1/2} \sum_{i=1}^{n} \Phi_{\phi}(\theta, \xi) \text{ asymptotically. Similarly, for large } n,
\]

\[
\frac{1}{\sqrt{n}} S_{\text{MLE}, \xi}^2 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{p} I(R_i = r) D_j \left\{ \sum_{\gamma_{i,j} \gamma_{i,j',x_i}} \alpha_{\gamma_{i,j} \gamma_{i,j'} x_i} A_{\gamma_{i,j} \gamma_{i,j'}} - \frac{\mathcal{N}^{(0)}(Q_0)}{\mathcal{N}^{(0)}(Q_0)} \right\} + o(1).
\]

Employing the same technique as that used in (B2), we can write

\[
\frac{\mathcal{N}^{(1)}(Q_0)}{\mathcal{N}^{(0)}(Q_0)} - \frac{\mathcal{N}^{(0)}(Q_0)}{\mathcal{N}^{(0)}(Q_0)} = \frac{1}{n \mathcal{N}^{(0)}(Q_0)} \sum_{i=1}^{n} (1-D_i) \sum_{\gamma} \exp \left\{ A_{\gamma} \xi + \sum_{j=1}^{p} X_{i,j} B_{\gamma} \theta^{(p)} \right\} 
\]

\[
\times \left\{ \mathcal{A}_{\gamma} \frac{\mathcal{N}^{(1)}(Q_0)}{\mathcal{N}^{(0)}(Q_0)} \right\} + o_p(n^{-1/2}).
\]

Plugging (B4) into (B3) and changing the order of the two summations in the second term, we have

\[
\frac{1}{\sqrt{n}} S_{\text{MLE}, \xi}^2 = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{p} I(R_i = r) D_j \left\{ \sum_{\gamma_{i,j} \gamma_{i,j',x_i}} \alpha_{\gamma_{i,j} \gamma_{i,j'} x_i} A_{\gamma_{i,j} \gamma_{i,j'}} - \frac{\mathcal{N}^{(0)}(Q_0)}{\mathcal{N}^{(0)}(Q_0)} \right\} - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (1-D_i) 
\]

\[
\times \sum_{i=1}^{n} \sum_{j=1}^{p} I(R_i = r) D_j \frac{1}{n \mathcal{N}^{(0)}(Q_0)} \sum_{\gamma} \exp \left\{ A_{\gamma} \xi + \sum_{j=1}^{p} X_{i,j} B_{\gamma} \theta^{(p)} \right\} \left\{ \mathcal{A}_{\gamma} \frac{\mathcal{N}^{(0)}(Q_0)}{\mathcal{N}^{(0)}(Q_0)} \right\} + o_p(1)
\]

\[
= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{j=1}^{p} I(R_i = r) D_j \left\{ \sum_{\gamma_{i,j} \gamma_{i,j',x_i}} \alpha_{\gamma_{i,j} \gamma_{i,j'} x_i} A_{\gamma_{i,j} \gamma_{i,j'}} - \frac{\mathcal{N}^{(0)}(Q_0)}{\mathcal{N}^{(0)}(Q_0)} \right\} - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (1-D_i)
\]
\[
\mathbb{E}\left[ \sum_j I(R_j = r) D_j \frac{1}{N^{(0)}(Q_0)} \sum_p \exp \left( A_j^T \xi + \sum_{\nu=1}^p X_{i,p} B_{\nu,\nu}^T \theta^{(p)} \right) \left\{ A_i - \frac{N^{(i)}(Q_0)}{N^{(0)}(Q_0)} \right\} \right] + o_p(1).
\]

The last equality follows due to the application of the strong law of large numbers and Slutsky’s Theorem. Thus we have shown that \( S_{EE,\theta} \) and \( S_{EE,\xi} \) are approximately a sum of asymptotically independent random variables whose means are zero. Now,

\[
\sqrt{n} \left[ \hat{\theta} - \theta \right] = \frac{H^{-1}}{\sqrt{n}} S_n(\eta) + o_p(1) = H^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \left[ \Phi_{i,\theta}(\theta, \xi) \right] + o_p(1),
\]

where \( \theta = (\theta^{(1)^T}, \ldots, \theta^{(p)^T})^T \),

\[
H = \begin{bmatrix}
H_{11} & H_{12} & \cdots & H_{1p}

\vdots

H_{p1} & H_{p2} & \cdots & H_{pp}

H_{11} & H_{12} & \cdots & H_{1p}
\end{bmatrix}
\]

and the asymptotically independent terms are

\[
\Phi_{i,\theta}(\theta, \xi) = (\Phi_{i,\theta}^{(1)}(\theta, \xi), \ldots, \Phi_{i,\theta}^{(p)}(\theta, \xi), \Phi_{i,\xi}(\theta, \xi))^T.
\]

Also,

\[
\Phi_{i,\theta}(\theta, \xi) = \sum_j I(R_j = r) D_j \left\{ \sum_{\nu=1}^{\nu^{(0)}} \omega_{\nu,\nu} \sum_{i,j} X_{i,\nu} B_{\nu,\nu} \frac{M_{\nu,\nu}^{(1)}(Q_0)}{M_{\nu,\nu}^{(0)}(Q_0)} \right. \bigg| X_i \bigg\}
\]

\[
+ \sum_{\nu=1}^{\nu^{(p)}} \sum_{\nu'=1}^{\nu^{(0)}} X_{i,\nu} B_{\nu,\nu'}^{(p)} \frac{M_{\nu,\nu'}^{(1)}(Q_0)}{M_{\nu,\nu'}^{(0)}(Q_0)} \bigg| X_i \bigg\}
\]

\[
\Phi_{i,\xi}(\theta, \xi) = \sum_j I(R_j = r) D_j \left\{ \sum_{\nu=1}^{\nu^{(0)}} \omega_{\nu,\nu} \sum_{i,j} A_{i,\nu} \frac{N_{\nu,\nu}^{(1)}(Q_0)}{N_{\nu,\nu}^{(0)}(Q_0)} \right. \bigg| X_i \bigg\}
\]

\[
- (1 - D_j) \left[ \sum_j I(R_j = r) \frac{D}{N^{(0)}(Q_0)} \sum_{\nu} \exp \left( A_{i,\nu}^T \xi \right) \right
\]

\[
+ \sum_{\nu=1}^{\nu^{(p)}} X_{i,\nu} B_{\nu,\nu}^{(p)} \left( \theta^{(p)} \right) \frac{N_{\nu,\nu}^{(1)}(Q_0)}{N_{\nu,\nu}^{(0)}(Q_0)} \bigg| X_i \bigg\},
\]

where \( Q_0 \) represents the true distribution of \( X \) among the controls.

Therefore, the asymptotic covariance of \( \hat{\eta} \) can be consistently estimated by
\[ H_n^{-1} \sum_{i=1}^{n} \begin{bmatrix} \tilde{\Phi}_{i,\theta}(\hat{\theta}, \hat{\xi}) \\ \tilde{\Phi}_{i,\xi}(\hat{\theta}, \hat{\xi}) \end{bmatrix}^{\otimes 2} H_n^{-T}, \]

where \( \tilde{\Phi}_{i,\theta}(\hat{\theta}, \hat{\xi}) \) and \( \tilde{\Phi}_{i,\xi}(\hat{\theta}, \hat{\xi}) \) are obtained by replacing the expectations by the empirical averages, \( Q_0 \) by \( Q_{0n} \), and the true parameters by their consistent estimators.