ABSTRACT

JIANYUN, YIN. Selected Topics in Binary Classification and Missing Data Problems. (Under the direction of Howard Bondell and Daowen Zhang.)

Statistical analysis has been widely applied in different areas. In order to apply a statistical methodology, certain assumptions should be satisfied in the collected dataset. For example, we may assume the data follow the normality assumption, or we assume the data is complete. However, the data we collect may not be so standard in real cases. In this thesis, we investigate binary classification methods when data do not follow the normal distribution assumption and hypothesis tests when certain part of the dataset is subject to missing.

In chapter 2, we investigate the binary classification using the area under the ROC curve from a Bayesian perspective. Instead of assuming a certain joint distribution of predictors, we assume that the linear combination of predictors follows the normal distribution, which is more likely to hold in the real cases. In this project, we propose a Bayesian approach via empirical likelihood under the binormal assumption and choose priors by utilizing an empirical Bayes method. This Bayesian procedure outputs the Bayesian mode as the estimates of the coefficients, which optimizes the binormal AUCROC. Numerical studies are conducted to evaluate the performance of the proposed estimator comparing to the original MME estimator and other state-of-art regularized methods.

In Chapter 3, we propose the generalized score tests when data is subject to missing under the missing at random (MAR) assumption. Score statistics assume everything under the null, which can be very useful when part of the data is missing. We developed different kinds of generalized score tests for different missing patterns. Hence this work can be applied to various real cases. An extended simulation study is conducted to evaluate the proposed generalized score tests in different missing patterns. In the real world, it is hard to specify the required model correctly. Hence we also study the robustness of the proposed generalized score tests under different scenarios.
Selected Topics in Binary Classification and Missing Data Problems

by

Yin Jiajun

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Statistics

Raleigh, North Carolina

2020

APPROVED BY:

Dennis Boos

Brian Reich

Howard Bondell
Co-chair of Advisory Committee

Daowen Zhang
Co-chair of Advisory Committee
ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisors, Dr. Howard Bondell and Dr. Daowen Zhang, for their support and guidance during these years. They managed to meet with me every week even though we are apart physically. They always come up with great ideas and navigate me to the right direction. Furthermore, they taught me how to conduct research independently from coming up ideas to forming a thesis. Their expertise in statistics influenced me not only in my academic study but also in my future career development. Without their help, I can not go thus far. I’m really fortunate to have them as my advisors.

Also, I would like to thank a lot to my committee members, Dr. Dennis Boos, Dr. Brian Reich and Dr. Ernest Stitzinger, for their valuable time and suggestions. They helped me to improve my projects a lot.

I also owe great thanks to the Department of Statistics at North Carolina State University, which provides excellent academic training and support. I appreciate the help and support from all the faculty and staff in our department.
# TABLE OF CONTENTS

LIST OF TABLES ........................................................................................................... iv  
LIST OF FIGURES ........................................................................................................... v  

Chapter 1 INTRODUCTION ............................................................................................ 1  

Chapter 2 Bayesian Binormal ROC Classification via Empirical Likelihood ............... 3  
2.1 Introduction .............................................................................................................. 3  
2.2 Background ............................................................................................................ 5  
2.2.1 AUC under the binormal assumption ................................................................ 5  
2.2.2 Empirical likelihood method ............................................................................ 6  
2.3 Bayesian binormal classification method ............................................................... 8  
2.3.1 The prior specification ...................................................................................... 8  
2.3.2 Computational details ...................................................................................... 11  
2.4 Simulation ............................................................................................................. 12  
2.4.1 Simulation setting I .......................................................................................... 12  
2.4.2 Simulation setting II ......................................................................................... 14  
2.5 Conclusion ............................................................................................................ 14  

Chapter 3 Generalized Score Tests in Linear Regression Models with Missing Data .... 17  
3.1 Introduction ............................................................................................................ 17  
3.2 Case I: Only response variable is subject to missing ............................................ 20  
3.2.1 Likelihood based score test ............................................................................. 20  
3.2.2 Two Stage score test ....................................................................................... 22  
3.2.3 Inverse probability weighted score test ......................................................... 24  
3.2.4 Augmented inverse probability weighted score test ...................................... 26  
3.3 Case II: Only covariates are subject to missing .................................................. 27  
3.4 Case III: Both response variable and covariates are subject to missing ............... 30  
3.5 Simulation ............................................................................................................ 32  
3.5.1 Response variable Y is subject to missing ...................................................... 35  
3.5.2 covariates X are subject to missing ............................................................... 36  
3.5.3 Both response and covariates are subject to missing ..................................... 37  
3.6 Application ......................................................................................................... 37  
3.7 Discussion ............................................................................................................ 46  

BIBLIOGRAPHY ............................................................................................................ 47  

APPENDICES ............................................................................................................... 50  
Appendix A Appendix for project 1 ............................................................................ 51  
A.1 Proof of Proposition 1 .......................................................................................... 51  
A.2 Details for optimizing the posterior distribution ............................................... 52  
Appendix B Appendix for project 2 ............................................................................ 55  
B.1 Asymptotic Variance of the two stage likelihood score ..................................... 55
**LIST OF TABLES**

Table 2.1  Summary of the Euclidean distance from estimated $\beta$ to true $\beta$ that generate data with various dimensions $p$, correlations $\rho$ and both balanced data ($n_0 = 50, n_1 = 50$) and unbalanced data ($n_0 = 25, n_1 = 75$). The mean and standard error are calculated based on 1000 replications. All the numbers are multiplied by 100. .......................................................... 15

Table 2.2  Summary of the test AUC achieved by various methods when $p = 5, 20$ and $\rho = 0.5$ for both balanced data ($n_0 = 50, n_1 = 50$) and unbalanced data ($n_0 = 25, n_1 = 75$) in setting I. The mean and standard error are calculated based on 1000 replications. The AUC is represented as percentage. ............................... 15

Table 2.3  Summary of the test AUC achieved by various methods when $p = 5, 20$ and $\rho = 0.5$ for both balanced data ($\beta_0 = 0$) and unbalanced data ($\beta_0 = 1.2$) in setting II. The mean and standard error are calculated based on 1000 replications. The AUC is represented as percentage. .......................................................... 15

Table 3.1  Estimated type I error probability and power of each test when the response $Y$ is subject to missing in different situations. Both type I error probability and power are estimated based on 10000 replications. The standard error is between 0.001 to 0.005 approximately. ................................. 38

Table 3.2  Estimated type I error probability and power of each test when the whole vector $X$ is subject to missing in different situations. Both type I error probability and power are estimated based on 10000 replications. The standard error is between 0.001 to 0.005 approximately. ................................. 38

Table 3.3  Estimated type I error probability and power of each test when covariates $X$ follow general missing data patterns in different situations. Both type I error probability and power are estimated based on 10000 replications. The standard error is between 0.001 to 0.005 approximately. ................................. 42

Table 3.4  Estimated type I error probability and power of each test when both response $Y$ and covariates $X$ are subject to missing in different situations. Both type I error probability and power are estimated based on 10000 replications. The standard error is between 0.001 to 0.005 approximately. ................................. 42

Table 3.5  The generalized score statistics and corresponding P-value testing whether the interferon-α treatment has a significant effect on the lines of vision improvement rate. ......................................................................... 45
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 2.1</td>
<td>The prior density of $\eta$ for $a = 10$, $b = 1$.</td>
<td>9</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Power function of proposed score tests when response $Y$ is subject to missing.</td>
<td>39</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Distribution of proposed score statistics when response $Y$ is subject to missing and $p = 1$.</td>
<td>40</td>
</tr>
<tr>
<td>Figure 3.3</td>
<td>Power function of proposed score tests when the whole vector of $X$ is subject to missing.</td>
<td>41</td>
</tr>
<tr>
<td>Figure 3.4</td>
<td>Power function of proposed score tests when covariates $X$ follow general missing data patterns.</td>
<td>43</td>
</tr>
<tr>
<td>Figure 3.5</td>
<td>Power function of proposed score tests when both $Y$ and $X$ are subject to missing.</td>
<td>44</td>
</tr>
</tbody>
</table>
Statistical theory and methods allow analysts to understand the situation better and drive business decisions. As statistical analysis has been applied to various areas, the data we collect may not be regular dataset which is clean and follow classical assumptions. Hence, the reliability of statistical methods in different scenarios is critical in application. In this thesis, we investigated how to do binary classification using the area under the ROC curve and how to conduct generalized score tests when part of the dataset is subject to missing.

Chapter 2 proposes a Bayesian classification method using AUCROC under the binormal assumption. Classification can be applied to many areas such as fraudulent detection, decision making and image segmentation. The receiver operating characteristic (ROC) curve can provide a more complete picture of a binary classifier. In order to evaluate the classifier quantitatively, we apply the area under the curve (AUC) which varies from 0 to 1 as the evaluation tool for the classifier. Researchers also explored methodologies to optimize AUCROC directly. We adapt the binormal assumptions, which only requires the linear combination of multivariate predictors to follow a normal distribution. Then we review the derivation of AUCROC under the binormal assumption. We propose a Bayesian
classification approach optimizing the AUCROC under the binormal assumption. The difficulty of this problem is that we have no knowledge about the joint distribution of predictors, and it is hard to specify the likelihood of data. The empirical likelihood is employed to form the likelihood of data without specifying the parametric distribution of predictors. We impose a Beta shape prior on the binormal AUCROC and discuss how we choose the hyperparameters of the priors. This Bayesian procedure outputs the Bayesian mode as the estimates of the coefficients, which optimizes the binormal AUCROC. We conduct a simulation study to compare the performance of the proposed method with the method of moments estimator (MME). We also compare with other state of the art regularized methods in terms of test AUC in different scenarios.

In Chapter 3, we develop generalized score tests for testing the association between a response variable and covariates of interest in linear regression models with different missing data patterns. We often collect data that contains missing information in areas like clinical study, genetic study and social survey, etc. The missing data problem should be treated seriously; otherwise it may interfere with statistical estimation, inference, and testing. There is a lot of work focusing on reducing bias and increasing the efficiency of the estimators in missing data problems. However, only a few people worked on improving the testing problem when information is partially observed. We assume the missing data mechanism follows missing at random (MAR) assumption when some auxiliary variables are also introduced. Previous research concentrated on testing the partial association between the response variable and covariates conditioned on auxiliary variables. Instead, we focus on testing the direct association by leveraging the re-parameterization techniques or inverse weighting methods. We consider following three increasingly more general missing data patterns where (1) only response variable is subject to missing; (2) only covariates are subject to missing; and (3) both response and covariates are subject to missing. We construct generalized scores based on the likelihood approach and inverse weighting methods (Robins et al., 1994). By using a score test in inverse weighting methods, we incorporate more information comparing to a Wald test when covariates follow general missing patterns. The proposed generalized score tests can also be used in two sample t-test and ANOVA problems where covariates of interest are discrete. An extended simulation study and real data analysis are presented to compare different testing procedures.
2.1 Introduction

Consider a binary classification problem with $Y = 0, 1$ and a $p$-dimensional covariate $X = (x_1, x_2, \ldots, x_p)$. Denote $g$ as the group indicator which takes value in $\{0, 1\}$, and $n_g$ as the number of samples with $Y = g$. Sample size $n = n_0 + n_1$. We predict the response variable $Y$ using the linear risk score function, $f(X) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p = \beta^T X$, where $\beta \in \mathbb{R}^p$ is the parameter of interest. For any threshold $c \in (-\infty, \infty)$, we predict $\widetilde{Y} = 1$ if $f(X) > c$, and $\widetilde{Y} = 0$ otherwise. The components of $X$ can also be functions of raw data, such as transformation and interactions, which generalize the linear risk function to a non-linear function (Pepe et al., 2006).

For any given threshold $c$, we can calculate the true positive rate (TPR) and false positive rate
(FPR) as,

\[
TPR = \frac{TP}{TP + FN},
\]

\[
FPR = \frac{FP}{FP + TN}.
\]

Here True Positive count and False Positive count are denoted as TP and FP; and True Negative count and False Negative count are denoted as TN and FN. The ROC curve is a very informative graph plotting the true positive rate against the false positive rate as the classification threshold \( c \) varies from negative infinity to positive infinity (Krzanowski and Hand, 2009). We can assess the overall performance of a classifier using the ROC curve. The area under the ROC curve (AUCROC), whose value varies from 0 to 1, is often used as summary index for the ROC curve. We do not consider the classifier that is worse than the random guess when the AUCROC is smaller than 0.5. The area under ROC curve can be interpreted as the probability that the linear risk score of a randomly chosen example from group 1 is larger than that of a randomly chosen example from group 0 (Bamber, 1975),

\[
P(\beta^T X > \beta^T X^*),
\]

where we denote \( X \) is from group 1 and \( X^* \) is from group 0. A larger AUCROC indicates a better performance of a classifier.

Many researchers treat AUCROC as an objective function and optimize it directly. Pepe and Thompson (2000) proposed a method to find the best linear classifier and experiment it on \( p = 2 \) case. Pepe et al. (2006) further investigated the difference between methods using likelihood as an objective function and methods using empirical AUC as an objective function. Ma and Huang (2005) applied the TGDR algorithm to optimize a sigmoid approximation of empirical AUCROC, which can handle large dimensional problems. In real application, the importance of TPR and FPR are different, and empirical AUCROC allows us to treat them differently. Komori and Eguchi (2010) used boosting methods to optimize the Partial AUC (pAUC). As the pAUC method is threshold-dependent and difficult to inference, Yu et al. (2014) proposed to optimize the modified AUC (mAUC), which is threshold independent. Ma et al. (2006) demonstrated that compared with the empirical AUCROC, binormal AUCROC is more stable when the sample size is small and computationally less expensive when the sample size is large. The MSauc method developed by Ma et al. (2006) optimized Binormal
AUCROC using the TGDR algorithms, and the AUCPR method developed by Yu and Park (2014) transforms the problem to a penalized regression problem.

In this article, we adopt a new binormal AUCROC classification method from a Bayesian perspective. Bayesian procedure is able to consider the prior information from different aspects (Gelman et al., 2013). The proposed procedure leverages the power of Bayesian analysis by putting a non-informative flat prior on mean and covariance matrix, and a Beta-shape prior on the function of the maximized AUCROC directly. The Beta-shape prior, which can be interpreted as a regularization condition, is obtained by an empirical Bayes procedure. However, there is no parametric likelihood form for the ROC-classification method under the binormal assumption, making it difficult for Bayesian analysis. In this chapter, we use the empirical likelihood to model the likelihood of data, which avoids additional assumption on the covariate $X$. We show that the proposed Bayesian procedure can be preferable to the method of moments estimates (MME) in a simulation study. We also evaluate the proposed estimators comparing with other state of art regularization methods.

The rest of the chapter is organized as follows. Section 2.2.1 provides the details of the binormal assumption and the calculation of the AUCROC under the binormal assumption. Section 2.2.2 provides the details of the empirical likelihood method, which is used as the likelihood function in our Bayesian procedure. We develop a Bayesian binomial AUCROC classification method via empirical likelihood under the binormal assumption in section 2.3. An extended simulation study is presented in section 2.4. Finally, we give a conclusion in section 2.5.

## 2.2 Background

### 2.2.1 AUC under the binormal assumption

We can estimate the theoretical AUCROC $P(\beta^T X > \beta^T X^*)$ using two different methods. One is from empirical approach, and the other one is based on Binormal assumption. The empirical AUCROC given $\beta$ is estimated by

$$AU C(\beta) = \frac{1}{n_0 n_1} \sum_{Y_i=1; Y_j=0} I(\beta^T X_i - \beta^T X_j^* > 0).$$
The other method is based on the binormal assumption (Dorfman and Alf Jr, 1969) which assumes the linear combination $\beta^T X$ follows two independent normal distributions based on the group indicator $Y$,

$$\beta^T X|Y = 0 \sim N(v_0, \sigma^2_0),$$
$$\beta^T X|Y = 1 \sim N(v_1, \sigma^2_1).$$

This assumption is more likely to hold than the assumption that $X$ follows two multivariate normal distributions especially in high dimensional case. Based on the binormal assumption, we can rearrange (2.1) as

$$AUC(\beta) = P(\beta^T(X - X^*) > 0) = P \left( \frac{\beta^T(X - X^*) - (v_1 - v_0)}{\sqrt{\sigma^2_0 + \sigma^2_1}} > \frac{-(v_1 - v_0)}{\sqrt{\sigma^2_0 + \sigma^2_1}} \right) = \Phi \left( \frac{v_1 - v_0}{\sqrt{\sigma^2_0 + \sigma^2_1}} \right),$$

where $\Phi$ is the CDF of normal distribution. Denote the mean structure and covariance structure of covariates $X$ in each group as $E(X|Y = g) = \mu_g$ and $Cov(X|Y = g) = \Sigma_g$ for $g = 0, 1$. Note that $\Sigma_0 + \Sigma_1$ is non-singular. As we have $\beta^T \mu_g = v_g$ and $\beta^T \Sigma_g \beta = \sigma^2_g$ for $g = 0, 1$, above equation can be written as

$$AUCROC(\beta) = \Phi \left( \frac{\beta^T(\mu_1 - \mu_0)}{\sqrt{\beta^T(\Sigma_0 + \Sigma_1)\beta}} \right).$$

The linear classifier $\beta$ that optimizes (2.3) and subject to $\beta^T(\mu_1 - \mu_0) > 0$ can be written as

$$\beta^* \propto (\Sigma_0 + \Sigma_1)^{-1}(\mu_1 - \mu_0).$$

2.2.2 Empirical likelihood method

The major difficulty of using Bayesian analysis in binormal classification is that we can not write out the explicit likelihood form of this problem with respect to parameter $\beta$ since we only know the conditional distribution of $\beta^T X$ as $P(\beta^T X|Y)$, rather than the conditional distribution of $X$ as $P(X|Y; \beta)$ based on the binormal assumption. In this case, we apply empirical likelihood to replace the traditional likelihood for Bayesian Analysis.

Within the empirical likelihood framework, statisticians can apply the likelihood function avoiding additional parametric distribution assumptions on the data (Owen, 2001). In our case, the distribution of the data can not be clearly specified, and the empirical likelihood techniques can
help us overcome it.

Given data \((X_i, Y_i), i = 1, 2, \ldots, n\), the form of an empirical likelihood of a set of parameters \(\theta\) can be written as,

\[
\max_{(p_1, \ldots, p_n)} \prod_{i=1}^{n} p_i \quad \text{subject to:} \quad \sum_{i=1}^{n} p_i = 1, \quad \sum_{i=1}^{n} p_i h(X_i, Y_i, \theta) = 0, \quad p_i > 0. \tag{2.5}
\]

Here \(p_i\) is the probability weight to each data point \((X_i, Y_i)\). The weights are constrained by the moment condition \(\sum_{i=1}^{n} p_i h(X_i, Y_i, \theta) = 0\).

Following Owen (2001), we can reduce this \(n\)-dimensional optimization problem to a \(\dim(\theta)\)-dimensional problem by Lagrange multipliers, the weights \(p_i(\theta)\) in Equation (2.5) can be represent as

\[
p_i(\theta) = \frac{1}{n - \lambda^T(\theta)h(X_i, Y_i, \theta)}, \quad i = 1, 2, \ldots, n. \tag{2.6}
\]

The vector of Lagrange multiplier in the above expression satisfies

\[
\sum_{i=1}^{n} \frac{h(X_i, Y_i, \theta)}{n - \lambda^T(\theta)h(X_i, Y_i, \theta)} = 0.
\]

Built on the theory and methods about likelihood evaluation for Bayesian analysis proposed by Monahan and Boos (1992), Lazar (2003) suggested that the empirical likelihood is valid to be applied in Bayesian analysis. Schennach (2005) considered Bayesian exponential tilted empirical likelihood (BETEL), which has a well-defined probability interpretation. In practice, empirical likelihood Bayesian analysis has been widely used in the problem setting in which an explicit likelihood function does not exist. For example, Yang et al. (2012) applied the Bayesian empirical likelihood (BEL) for quantile regression that allows us to put prior information on the parameters in quantile regression, and it is very useful on the regions where the data is sparse. In our work, the empirical likelihood is served as the likelihood in the Bayesian procedure in order to avoid further distributional assumption of covariates \(X\).
2.3 Bayesian binormal classification method

In addition to the parameters $\theta = (\Sigma_0, \Sigma_1, \mu_0, \mu_1)$, we introduce $\eta = \eta(\theta)$ which is the function of the maximized AUCROC attained at $\beta_{\text{ROC}}$,

$$
\eta(\theta) = \frac{\beta^T (\mu_1 - \mu_0)(\mu_1 - \mu_0)^T \beta^*}{\beta^T (\Sigma_0 + \Sigma_1) \beta^*} = (\mu_1 - \mu_0)^T (\Sigma_0 + \Sigma_1)^{-1} (\mu_1 - \mu_0). \tag{2.7}
$$

Consider the parameter space $\Theta = \{ (\theta, \eta(\theta)) : \eta = \eta(\theta) \}$. As we only have binormal assumptions on $\beta^T X | Y$ and do not impose additional assumptions on the multi-dimensional $X$, we consider the empirical likelihood, represented as $\prod_{i=1}^n p_i(\theta)$, where the probability vector $(p_1(\theta), ..., p_n(\theta))$ is the solution of (2.5). Following Yang et al. (2012), with certain prior $f(\theta, \eta(\theta))$, we can form the posterior of $(\theta, \eta(\theta))$ as,

$$
f(\theta, \eta(\theta)|X, Y) \propto f(\theta, \eta(\theta)) \prod_{i=1}^n p_i(\theta) 
= f(\theta|\eta(\theta)) f(\eta(\theta)) \prod_{i=1}^n p_i(\theta), \tag{2.8}
$$

where $f(\theta, \eta(\theta))$ is a given prior which can be decomposed to a conditional prior on $\theta|\eta(\theta)$ and a prior on $\eta(\theta)$, and the expression of probability vector $(p_1(\theta), ..., p_n(\theta))$ is given in (2.6).

In section 2.3.1, we discuss how we choose the prior for $(\theta, \eta(\theta))$. Section 2.3.2 details how we specify the moment function $h(x, y, \theta)$ and compute our estimator based on the posterior distribution (2.8).

2.3.1 The prior specification

Imposing a prior on $\eta(\theta)$, abbreviated as $\eta$ in the future, is equivalent to impose a prior on the maximized binormal AUCROC given the parameters $\theta$. If we construct a beta distribution shape prior on the AUCROC

$$
\Phi\left( \frac{\beta^T (\mu_1 - \mu_0)}{\sqrt{\beta^T (\Sigma_0 + \Sigma_1) \beta^*}} \right) = \Phi(\sqrt{\eta}) = 0.5 + 0.5 \times \text{Beta}(a, b), \quad a > 0, \quad b > 0,
$$
the density function of $\eta$ can be represented as

$$f(\eta|a, b) = \frac{1}{B(a, b)} \phi(\sqrt{\eta}) \frac{1}{\sqrt{\eta}} \left(2\Phi(\sqrt{\eta}) - 1\right)^{a-1}(2 - 2\Phi(\sqrt{\eta}))^{b-1}. \tag{2.9}$$

**Proposition 1.** When $a = 1$, $f(\eta|a, b)$ is a monotone decreasing function in $\eta > 0$ for all $b > 0$.

**Proof.** See Appendix A.1.

**Remark.** When $b = 1$, there exists a such that $f(\eta|a, b)$ is not monotone decreasing in $\eta > 0$ as shown in Figure 2.1.

![Figure 2.1](image.png)

The choice of hyperparameters $a, b$ is very important in the posterior calculation. We adopt an approximate empirical bayes approach to determine hyperparameters. The empirical bayes approach maximizes the marginal distribution of data over all possible $a$ and $b$. However, this data-driven approach makes $a$ and $b$ converge to infinity, which put a point mass on the maximum
likelihood estimator of $\eta$. Hence we fix one of the hyperparameters and optimize another to get a reasonable prior. In order to give regularization to our estimators, we choose the constraints that make the prior of $\eta$ monotone decreasing. Based on Proposition 1, fixing $a = 1$ and tuning $b$ will always shrink the $\eta$.

On the other hand, the choice of $b$ is estimated from the empirical bayes procedure as follows. Consider the random variable

$$Z_i = (X_{i1} - X_{i0})^T (\Sigma_0 + \Sigma_1)^{-1} (X_{i1} - X_{i0}),$$

where $X_{ig}$ is from group $Y = g$. The marginal distribution of $Z_i$ follows $\chi^2_p(\eta)$ which is fully determined by $\eta$. Note that the parameterization of the non-central chi-square makes the expected value of $Z_i$ to be $p + \eta$. Hence the marginal distribution of $Z_i$ conditional on $a$ and $b$ is

$$f(z_i|a, b) = \int_0^\infty f(z_i|\eta) f(\eta|a, b) d\eta.$$

We randomly sample pairs $(X_{i0}, X_{i1})$, $i = 1, 2, ..., N$, where $N \leq n_0n_1$ in order to simplify the computation. The value of $Z$ is computed using the MLE estimates of $\Sigma_0, \Sigma_1$. Ignoring the correlation in $Z = \{Z_i, i = 1, ..., N\}$, the likelihood of $Z$ given $a, b$ is approximated as the product of marginal distribution,

$$f(z|a, b) = \prod_{i=1}^N f(z_i|a, b). \quad (2.10)$$

Together with $a = 1$, the restricted maximum likelihood estimator $\hat{b} = \arg\max_b f(z|1, b)$ of the pseudo likelihood (2.10) is chosen to be the parameters of our Beta-shape prior.

For any given $\eta$, there are infinite number of $\theta$ that satisfies (2.7). We put a prior that is uniform on all the parameter set $\theta$ that satisfies (2.7) given certain value of $\eta$, i.e.,

$$f(\theta|\eta) \propto I(\eta(\theta) = \eta).$$

This non-informative prior can be viewed as an indicator function that indicates whether certain parameter set $\theta$ satisfies the constraint $\eta(\theta) = \eta$.  

10
2.3.2 Computational details

In this section, we discuss computational details to get our estimators based on the posterior distribution. We first introduce the moment function $h(x, y, \theta)$. Then the algorithm to compute the estimator for $(\theta, \eta(\theta))$ is given. Finally, we show how to calculate the posterior estimator of $\beta$ based on the posterior estimator of $(\theta, \eta(\theta))$.

The moment function $h(x, y, \theta)$ is specified as

$$
\begin{bmatrix}
y(x - \mu_1) \\
(1-y)(x - \mu_0) \\
y(x_j - \mu_{1j})(x_k - \mu_{1k}) - \sigma_{1jk} \\
(1-y)(x_j - \mu_{0j})(x_k - \mu_{0k}) - \sigma_{0jk}
\end{bmatrix},
$$

(2.11)

where $j = 1, \ldots, p$ and $k = j, \ldots, p$; $\sigma_{gjk}$ is the $(j, k)$-th element of $\Sigma_g$, $g \in \{0, 1\}$. Here the first and second constraints are moment constraints for parameters $\mu_1$ and $\mu_0$; and the third and fourth constraints are moment constraints for $\Sigma_1$ and $\Sigma_0$. For simplicity we denote $\mu = \mu_1 - \mu_0$ and $\Sigma = \Sigma_0 + \Sigma_1$.

We use the posterior mode as the estimator of $(\theta, \eta(\theta))$. In order to calculate the posterior mode of (2.8), we first fix the $\eta$ and maximize (2.8) over $\theta$, 

$$
\max_{\theta \in \Theta_\eta} \prod_{i=1}^{n} p_i(\theta, \eta)f(\theta|\eta) f(\eta) = \max_{\theta \in \Theta_\eta} \prod_{i=1}^{n} p_i(\theta, \eta)f(\eta),
$$

(2.12)

where $\Theta_\eta = \{\theta : \eta(\theta) = \eta\}$ is the restricted parameter space of $\theta$ given $\eta$. The second equality holds as $f(\theta|\eta)$ puts uniform weights on the restricted parameter space $\Theta_\eta$ for any given $\eta$. We then maximize the equation (2.12) over $\eta$ to get the posterior mode. Details of the maximization are given in Appendix A.2.

We find $\tilde{\eta}$ and corresponding $\tilde{\theta} = (\tilde{\Sigma}_0, \tilde{\Sigma}_1, \tilde{\mu}_0, \tilde{\mu}_1)$ which attain the maximum posterior as our posterior mode estimators. Since the coefficient $\beta$ is also defined as a deterministic function of $\theta$ as
in (2.4), the corresponding posterior mode estimator (named as BaROC) of $\beta$ can be written as,

$$\tilde{\beta} = (\tilde{\Sigma}_0 + \tilde{\Sigma}_1)^{-1}(\tilde{\mu}_1 - \tilde{\mu}_0).$$

2.4 Simulation

2.4.1 Simulation setting I

In this section, we conduct a simulation study to evaluate the performance of our Bayesian posterior mode estimates (BaROC) when covariates $X$ in each group is normally generated. We generate predictors $X|Y = j$ from $N(\mu_j, \Sigma_j)$. Let $\mu_0 = 0_p$ and $\mu_1 \sim N(0_p, s I)$, where $0_p$ represents a $p$ dimensional vector of 0s. The magnitude of $s$ controls the separability between two groups. The covariance matrix $\Sigma_0$ and $\Sigma_1$ are equal to $\Sigma \in \mathbb{R}^{p \times p}$, which has the AR(1) structure, i.e. the $j,k$-th element of covariance matrix $\Sigma_{j,k} = \sigma^2 \rho^{|j-k|}$. Here $\rho$ is the correlation parameter and $\sigma^2$ is the variance. The standard deviation $\sigma$ is set to be 1. The sample size is 100 for 2 groups together and each setting is repeated 1000 times.

We first compare the Euclidean distances of estimated parameters to true parameters with the method of moments estimates (MME). Define $\hat{\theta} = (\hat{\Sigma}_0, \hat{\Sigma}_1, \hat{\mu}_0, \hat{\mu}_1)$ as the method of moments estimator of $\theta$. The corresponding MME estimator of $\beta$ can be written as

$$\hat{\beta} = (\hat{\Sigma}_0 + \hat{\Sigma}_1)^{-1}(\hat{\mu}_1 - \hat{\mu}_0).$$

In each simulation setting, we consider both balanced case ($n_0 = n_1 = 50$) and unbalanced case ($n_0 = 25$, $n_1 = 75$). On the other hand, a larger $\rho$ indicates a higher correlation between the components of the predictor $X$, which leads to a relatively worse MME estimator. We will consider both $\rho = 0.5$ representing the case with moderately correlated predictors and $\rho = 0.9$ representing the case with highly correlated predictors. The separability parameter $s$ is set to be 0.1 when $\rho = 0.5$ and 0.02 when $\rho = 0.9$.

We report the Euclidean distances of the estimated $\beta$ to the true $\beta = (\Sigma_0 + \Sigma_1)^{-1}(\mu_1 - \mu_0)$. From
Table 2.1, we observe that our proposed posterior mode estimators have shorter distances to the corresponding true parameters comparing with MME estimates in all cases. For example, when \( \rho = 0.5 \) and \( p = 5 \), the distance of \( \hat{\beta} \) to the true \( \beta \) is 5-6 standard deviation smaller than its MME counterpart. When \( \rho = 0.9 \), the BaROC estimators are also much closer to the true value.

Then we compare with other state of the art regularized methods in terms of test AUC. Two benchmark methods are used to compare with the proposed method. They are MSauc proposed by Ma et al. (2006), and the penalized regression method proposed by Yu and Park (2014). The first method optimizes

\[
\Phi \left( \frac{\beta^T(\mu_1 - \mu_0)}{\sqrt{\beta^T(\Sigma_0 + \Sigma_1)\beta}} \right)
\]

using TGDR algorithm which conducts regularization and variable selection during gradient descent step. There are two tuning parameters, \( \tau \in [0, 1] \) controls the sparsity of \( \beta \), and \( k \) is the iteration steps (Ma et al., 2006). The penalized regression method (Yu and Park, 2014) transforms \( \mu = \Sigma \beta \) into an elastic net problem

\[
\beta = \arg \min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{p} \left( \mu_i - \sum_{j=1}^{p} \sigma_{ij} \beta_j \right)^2 + \lambda \left( \alpha \sum_{i=1}^{p} |\beta_i| + \frac{1-\alpha}{2} \sum_{i=1}^{p} \beta_i^2 \right),
\]

where \( \alpha \) controls the sparsity and \( \lambda \) represents the strength of the regularization. In practice, we use the MME estimates of \( \mu \) and \( \Sigma \). The hyperparameter \( k \) in the first method and \( \lambda \) in the second method are chosen by K-fold cross-validation using cross-validation set AUC as the score. Denote MSauc as the MSauc algorithm with \( \tau = 0 \) which does not produce sparse solution; also denote Ridge as the the penalized regression method with \( \alpha = 0 \) which does not produce sparse solution, and Elastic Net as the the penalized regression method with \( \alpha = 0.5 \). The code for above two methods are available at http://bibs.snu.ac.kr/software/aucpr (Yu and Park, 2014). In addition, we compare the performance of the logistic regression method. After getting the estimator of \( \beta \) in the training set, we generate a 100k size testing set and evaluate the test AUC for each of the algorithms.

The data is generated in the same way as before with \( \rho = 0.5 \). We consider both low dimensional case \( p = 5 \) and high dimensional case \( p = 20 \); also, both balanced design and unbalanced design are considered. The result is presented in Table 2.2. The proposed BaROC always has higher test AUC than MSauc and the penalized regression method when \( p = 5 \) or \( p = 20 \) (except comparing with the
penalized ridge regression in \( p = 20 \) balanced setting). On the other hand, BaROC achieves slightly lower test AUC than logistic regression when the dimension is low \( (p = 5) \), while it has higher test AUC when the dimension is high \( (p = 20) \).

### 2.4.2 Simulation setting II

In this section, we first generate \( X \) from independent \( N(0_p, \Sigma) \) and true \( \beta_1 = (\beta_0, \beta) \) where \( \beta \sim N(0_p, 0.1 \times I) \). Here we use \( \beta_0 \) to tune the ratio between \( n_0 \) and \( n_1 \). The group indicator \( Y \) is generated from the Bernoulli distribution with probability

\[
P(Y = 1|X) = \frac{1}{1 + \exp(X^T \beta + \beta_0)}.
\]

In this case, we do not have the same interpretation of \( \mu \) and \( \Sigma \) as before, and only evaluate the performance based on the area under the curve.

We consider both low dimensional case \( p = 5 \) and high dimensional case \( p = 20 \); also both balanced design \( (\beta_0 = 0) \) and unbalanced design \( (\beta_0 = 1.2) \). The correlation parameter \( \rho \) is set to be 0.5. The result is presented in Table 2.3. The proposed BaROC always has higher test AUC than MSauc when \( p = 5 \) or \( p = 20 \). The penalized ridge regression method is more preferable than BaROC when \( p = 20 \) with paired t-test P-value < 0.01. One possible reason is that BaROC chooses hyperparameters based on the empirical Bayes procedure, while penalized regression method tunes hyperparameters based on the cross-validation AUC which is highly related to our evaluation criteria. However, it is still possible to tune hyper-parameters in BaROC by the cross-validation AUC. On the other hand, although the mean of the test AUC achieved by BaROC is within one standard deviation comparing with the mean of the test AUC achieved by the logistic regression method, the paired t-test shows that BaROC performs significantly better than logistic regression when \( p = 20 \) (P-value < 0.01).

### 2.5 Conclusion

In this chapter, we proposed a Bayesian binormal classification method leveraging the power of empirical likelihood. We use the binormal assumption to calculate the binormal AUCROC. By putting a Beta shape prior on the summary index, a function of the maximized binormal AUCROC, and
Table 2.1 Summary of the Euclidean distance from estimated $\beta$ to true $\beta$ that generate data with various dimensions $p$, correlations $\rho$ and both balanced data ($n_0 = 50, n_1 = 50$) and unbalanced data ($n_0 = 25, n_1 = 75$). The mean and standard error are calculated based on 1000 replications. All the numbers are multiplied by 100.

<table>
<thead>
<tr>
<th></th>
<th>BaROC $n_0 = n_1 = 50$</th>
<th>MME $n_0 = n_1 = 50$</th>
<th>BaROC $n_0 = 25, n_1 = 75$</th>
<th>MME $n_0 = 25, n_1 = 75$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\rho = 0.5$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 5$</td>
<td>29.3 (0.4)</td>
<td>31.8 (0.4)</td>
<td>35.1 (0.5)</td>
<td>38.4 (0.5)</td>
</tr>
<tr>
<td>$p = 10$</td>
<td>54.5 (0.6)</td>
<td>58.1 (0.6)</td>
<td>64.5 (0.7)</td>
<td>69.2 (0.8)</td>
</tr>
<tr>
<td>$p = 20$</td>
<td>109.4 (1.1)</td>
<td>114.9 (1.1)</td>
<td>137.1 (1.4)</td>
<td>143.7 (1.5)</td>
</tr>
<tr>
<td>$\rho = 0.9$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 5$</td>
<td>66.3 (0.9)</td>
<td>71.6 (1)</td>
<td>79 (1.2)</td>
<td>86 (1.3)</td>
</tr>
<tr>
<td>$p = 10$</td>
<td>125.8 (1.4)</td>
<td>134.1 (1.5)</td>
<td>154.4 (1.8)</td>
<td>165.4 (2)</td>
</tr>
<tr>
<td>$p = 20$</td>
<td>267.6 (2.8)</td>
<td>281.6 (2.9)</td>
<td>339.1 (3.6)</td>
<td>355.5 (3.7)</td>
</tr>
</tbody>
</table>

Table 2.2 Summary of the test AUC achieved by various methods when $p = 5, 20$ and $\rho = 0.5$ for both balanced data ($n_0 = 50, n_1 = 50$) and unbalanced data ($n_0 = 25, n_1 = 75$) in setting I. The mean and standard error are calculated based on 1000 replications. The AUC is represented as percentage.

<table>
<thead>
<tr>
<th></th>
<th>BaROC</th>
<th>MSauc</th>
<th>Ridge</th>
<th>Elastic Net</th>
<th>Logistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_0 = 50, n_1 = 50$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 5$</td>
<td>69.4 (0.2)</td>
<td>63.8 (0.3)</td>
<td>68.9 (0.2)</td>
<td>68.6 (0.3)</td>
<td>69.4 (0.2)</td>
</tr>
<tr>
<td>$n_0 = 25, n_1 = 75$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 5$</td>
<td>68.6 (0.3)</td>
<td>63.2 (0.3)</td>
<td>68 (0.3)</td>
<td>67.7 (0.3)</td>
<td>68.7 (0.2)</td>
</tr>
</tbody>
</table>

Table 2.3 Summary of the test AUC achieved by various methods when $p = 5, 20$ and $\rho = 0.5$ for both balanced data ($\beta_0 = 0$) and unbalanced data ($\beta_0 = 1.2$) in setting II. The mean and standard error are calculated based on 1000 replications. The AUC is represented as percentage.

<table>
<thead>
<tr>
<th></th>
<th>BaROC</th>
<th>MSauc</th>
<th>Ridge</th>
<th>Elastic Net</th>
<th>Logistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_0 = 50, n_1 = 50$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 5$</td>
<td>63.8 (0.2)</td>
<td>61.2 (0.2)</td>
<td>63.9 (0.2)</td>
<td>63.8 (0.2)</td>
<td>63.8 (0.2)</td>
</tr>
<tr>
<td>$n_0 = 25, n_1 = 75$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 5$</td>
<td>63.1 (0.2)</td>
<td>60.7 (0.2)</td>
<td>63.3 (0.2)</td>
<td>63 (0.2)</td>
<td>63.2 (0.2)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>BaROC</th>
<th>MSauc</th>
<th>Ridge</th>
<th>Elastic Net</th>
<th>Logistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_0 = 50, n_1 = 50$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 20$</td>
<td>72 (0.2)</td>
<td>69.5 (0.2)</td>
<td>72.7 (0.2)</td>
<td>71.8 (0.2)</td>
<td>71.8 (0.2)</td>
</tr>
<tr>
<td>$n_0 = 25, n_1 = 75$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$p = 20$</td>
<td>71.3 (0.2)</td>
<td>68.8 (0.2)</td>
<td>72.2 (0.2)</td>
<td>71.3 (0.2)</td>
<td>71.1 (0.2)</td>
</tr>
</tbody>
</table>
uniform priors on all $\theta | \eta(\theta)$ satisfying (2.7), we avoid putting the prior on a $p$ dimensional linear combination coefficient $\beta$. In order to avoid further assumption on the distribution of covariates $X$, empirical likelihood is applied to replace the traditional likelihood. Simulation studies show that the proposed Bayesian procedure (BaROC) outperforms the frequentist MME in all cases since the Euclidean distance between the decision boundary of BaROC and optimal decision boundary is much smaller. Hence estimators of BaROC are closer to the true value comparing with the MME counterpart. Comparing with other regularized algorithms that optimize the binormal ROC directly, BaROC has relatively higher test AUC in many cases.
CHAPTER

3

GENERALIZED SCORE TESTS IN LINEAR REGRESSION MODELS WITH MISSING DATA

3.1 Introduction

Consider a regression problem

\[ Y = \beta_0 + \beta_1^T X + e_1, \]  

(3.1)

where a continuous response variable \( Y \) and \( p \)-dimensional covariates \( X = (X_1, X_2, \ldots, X_p)^T \) are subject to missing. The vector \( \beta_1 = (\beta_1, \beta_2, \ldots, \beta_p)^T \) is a \( p \)-dimensional vector of regression coefficients, and \( e_1 \) is the error term with expectation \( \mathbb{E}(e_1) = 0 \). We are interested in testing the null hypothesis \( H_0: \beta_1 = 0_p \). However, data may be subject to missing in many real world applications. For example, people may not complete certain questions in surveys, or expensive variables are not collected for
some subjects. Define

\[ R \equiv (R_0, R_1, \ldots, R_p), \]

where \( R_0 \) is the indicator that \( Y \) is observed and \( R_j \) is the indicator that \( X_j \) is observed for \( j = 1, 2, \ldots, p \). The observed data can be written as \((R_i, R_{0i}Y_i, R_{1i}X_{1i}, \ldots, R_{pi}X_{pi})\) for \( i = 1, \ldots, n \).

There are three most commonly used test procedures, namely score test, Wald test and likelihood ratio test, which are asymptotically equivalent. However, the score test assumes everything under the null hypothesis which simplifies the estimating procedure (Wong et al., 2019). Moreover, a score test considers more information than a Wald test and a likelihood ratio test when we apply inverse weighting methods to problems that covariates \( X \) follow general missing data patterns. When data is incomplete, most software reports results based on a complete case analysis. The pseudo-score with complete data can be written as

\[ SCC(\beta_0) = \sum_{i=1}^{n} R_{0i} R_X (Y_i - \beta_0)X_i, \]

where \( R_X = \prod_{j=1}^{p} R_j \) is the indicator that \( X \) is fully observed. The complete case method loses a lot of information in the incomplete observations, and furthermore, may lead to incorrect result in some scenarios. If the data is missing at random (MAR), we can conduct a score test correctly using the observed likelihood function. However, MAR with observed \( Y \) and \( X \) is a strong assumption which may not hold, and it is more likely to hold with extra fully observed auxiliary variables denoted as \( W \). Therefore, we finally collect data \((R_i, R_{0i}Y_i, R_{1i}X_{1i}, \ldots, R_{pi}X_{pi}, W_i)\) for \( i = 1, \ldots, n \).

One example of the above problem setting is the two-phase study where the covariates of interest \( X \) is difficult or expensive to collect. A common situation is that we have observations of \( Y \) and \( W \) for all the individuals in phase 1, and collect the expensive variable \( X \) from the selected sub-group in phase 2. We assume the sampling scheme does not depend on \( X \) itself, i.e. 

\[ P(R_X = 1|Y, X, W) = P(R_X = 1|Y, W), \]

which follows MAR condition.

Most of the research work focused on the estimation in missing data problems. Ibrahim (1990), Ibrahim et al. (1999) developed an EM algorithm for inference in generalized linear models with missing discrete covariates and missing continuous covariates correspondingly. Ibrahim et al. (2005) further gave a comprehensive review for above methods together with other widely-used approaches.
including multiple imputation, fully Bayesian and weighted estimating equations for inference in generalized linear models with missing covariates. Chen et al. (2008) discussed the inference method when both response and covariates are subject to missing in regression models. When the covariates are missing at random, a model for \( X \) is often required in likelihood methods. In practice, the distribution function of \( X \) can be hard to model. Some researchers considered the semi-parametric method which does not specify a parametric model for the marginal distribution of \( X \). Suppose the missingness depends on finite and fully observed strata of data, Lawless et al. (1999) provided a semi-parametric likelihood framework that can handle certain missing data problems and response-selective sampling problems. Since the parametric model of missing covariates can be misspecified, Zhang and Rockette (2005) proposed a semiparametric maximum likelihood method for the case that the missing covariates are discrete.

Recently, some researchers start to work on score tests for two-phase studies. Derkach et al. (2015) proposed a likelihood based score test for generalized linear models which is valid for different sampling schemes. Lawless (2018) extended the work of Derkach et al. (2015) in order to test the relationship between failure time and the covariates of interest \( X \). Wong et al. (2019) developed imputation based robust score statistics in order to test the association of the response variable and expensive covariates in the two phase study.

However, all of the above research work actually tested the hypothesis of partial independence of the response and covariates \( X \), which in our problem setting corresponds to \( H_0 : \gamma_1 = 0_p \) under the following model

\[
Y = \gamma_0 + \gamma_1^T X + \gamma_2^T W + e_2, \tag{3.2}
\]

where \( \gamma_2 \in \mathbb{R}^{q \times 1} \) is the vector of coefficients for auxiliary variable \( W \) and \( e_2 \) is the error term with \( \mathbb{E}(e_2) = 0 \). We are more interested in the direct association between the response variable \( Y \) and covariates \( X \) in many real applications. When the extra variable \( W \) is a necessary condition for MAR assumption to hold, testing the direct association in model (3.1) can be difficult. We can either reparameterize model (3.2) or use the inverse weighting methods based on model (3.1) while incorporate the information of \( W \) in the missing mechanism.

In this chapter, we do a comprehensive research on testing the association between the response variable \( Y \) and covariates of interest \( X \) in model (3.1) for different missing data patterns. In section
3.2, we develop two likelihood based score tests and two inverse weighted score tests for the situation that only the response variable $Y$ is subject to missing. In section 3.3, we provide the inverse weighted score tests when covariates $X$ follow the general missing pattern, and based on the good property of score test, the proposed generalized score tests achieve higher power. These approaches can also be applied to the two phase study problem which is a special scenario. Inverse weighting methods for the case that both $Y$ and $X$ are subject to missing is discussed in section 3.4. An extended simulation for different missing scenarios is presented in section 3.5. We present a real data example in section 3.6 and have a brief discussion in section 3.7.

3.2 Case I: Only response variable is subject to missing

In this section, we consider the case where covariates $X$ are completely observed and only the response $Y$ is subject to missing. In this case, $R_j = 1$ for $j = 1, 2, ..., p$ and $R_0$ is the only random observed data indicator. We assume the observed data probability,

$$ P(R_0 = 1|Y, X) = P(R_0 = 1|X), \quad (3.3) $$

follows the MAR assumption.

3.2.1 Likelihood based score test

Our interest is to test $H_0 : \beta_1 = 0_p$ for model (3.1). If the observed probability follows the MAR assumption, $P(R_0 = 1|Y, X) = P(R_0 = 1|X)$, the complete analysis will give a correct result. However, if the observed probability also depends on $Y$, namely missing not at random (MNAR), we can not conduct correct hypothesis tests with observed $Y$ and $X$ based on model (3.1) by the complete analysis or the likelihood approach. Assuming the MAR assumption holds after introducing $W$ as (3.3), we can get consistent estimators of parameters in model (3.2), and furthermore, conduct valid score tests with the observed data $(R_0, Y_i, X_{1i}, \ldots, X_{pi}, W_i)$.

The conditional expectation of model (3.2) given covariates $X$ is

$$ E(Y|X) = \gamma_0 + \gamma_1^T X + \gamma_2^T E(W|X). \quad (3.4) $$
Hence model (3.2) implies model (3.1) if variables $W$ and $X$ follow the linear relationship

$$W = \tau^T \tilde{X} + e_3,$$

(3.5)

where $\tau = (\tau_1, \ldots, \tau_q) \in \mathbb{R}^{(p+1) \times q}$ and $e_3$ is the vector of residuals with mean structural $E(e_3) = 0_q$.

For the rest of the chapter, we denote a tilde sign on the top of the variable when the intercept term is incorporated, i.e. $\tilde{X} = (1, X)^T$. We further denote $W(\tau) = W - \tau^T \tilde{X}$. Decomposing $\tau^T = (\tau'_0, \tau'_1)$ where $\tau'_0 \in \mathbb{R}^{q \times 1}$ corresponding to the intercept and $\tau'_1 \in \mathbb{R}^{q \times p}$ corresponding to the covariates $X$,

equation (3.4) can be written as

$$E(Y|X) = \gamma_0 + \gamma_1^T X + \gamma_2^T \tau^T \tilde{X}$$

$$= (\gamma_0 + \gamma_2^T \tau'_0) + (\gamma_1^T + \gamma_2^T \tau'_1)X$$

$$= \beta_0 + \beta_1^T X,$$

(3.6)

where $\beta_0 = \gamma_0 + \gamma_2^T \tau'_0$ and $\beta_1 = \gamma_1^T + \gamma_2^T \tau'_1$. Hence model (3.2) can be reparameterized as

$$Y = \beta_0 + \beta_1^T X + \gamma_2^T (W - \tau^T \tilde{X}) + e_2.$$  

(3.7)

Therefore, testing $H_0 : \beta_1 = 0_p$ in model (3.1) is equivalent to testing $H_0 : \beta_1 = 0_p$ in model (3.7).

Define the model of interest as $P(Y|X, W; \theta)$, where $\theta$ includes $\beta_0, \beta_1, \gamma_2, \tau$ and other parameters. Assuming the separability condition holds, the observed likelihood can be written as

$$L(\theta) = \prod_{i=1}^n P(Y_i, X_i, W_i; \theta)^{R_{oi}} \times P(X_i, W_i; \theta)^{1-R_{oi}}$$

$$= \prod_{i=1}^n P(Y_i|X_i, W_i; \theta)^{R_{oi}} \times P(X_i, W_i; \theta).$$

(3.8)

If we further assume that residuals $e_{2i} \overset{i.i.d.}{\sim} N(0, \sigma_1^2)$ and $e_{3i} \overset{i.i.d.}{\sim} N(0, \Sigma)$, the log-likelihood formula of the observed data from (3.8) can be given as

$$l(\theta) = \sum_{i=1}^n \left[ R_{oi} \left( -\frac{1}{2} \log \sigma_1^2 - \frac{1}{2\sigma_1^2} (Y_i - \beta_0 - \beta_1^T X_i - \gamma_2^T W_i(\tau)) \right) + \right.$$

$$\left. -\frac{1}{2} \log((2\pi)^k | \Sigma|) - \frac{1}{2} W_i(\tau)^T \Sigma^{-1} W_i(\tau) \right].$$

(3.9)
Under the error assumptions described above, \( \theta = (\beta^T, \beta_0, \gamma_2^T, \tau_1^T, \tau_2^T, \ldots, \tau_q^T, \sigma_1^2, \text{vec}(\Sigma))^T \) is a collection of all the parameters where \( \theta_1 = \beta_1 \) is the parameter of interest and \( \theta_2 = (\beta_0, \gamma_2^T, \tau_1^T, \tau_2^T, \ldots, \tau_q^T, \sigma_1^2, \text{vec}(\Sigma))^T \) is a collection of nuisance parameters. Define \( \hat{\theta} = (0_p, \hat{\beta}_0^T, \hat{\gamma}_2^T, \hat{\tau}_1^T, \hat{\tau}_2^T, \ldots, \hat{\tau}_q^T, \hat{\sigma}_1^2, \text{vec}(\hat{\Sigma})) \) as the MLE of (3.9) under the null hypothesis \( H_0 : \beta_1 = 0_p \).

The score under the null can be written as

\[
S_L(\theta_2) = \frac{\partial l}{\partial \beta_1} \bigg|_{\beta_1 = 0_p} = \sum_{i=1}^{n} \frac{R_{0i}}{\sigma_1^2} \left\{ Y_i - \beta_0 - \gamma_2^T W_i(\tau) \right\} X_i,
\]

and the estimated score can be written as

\[
S_L(\hat{\theta}_2) = \sum_{i=1}^{n} \frac{R_{0i}}{\hat{\sigma}_1^2} \left\{ Y_i - \hat{\beta}_0 - \hat{\gamma}_2^T W_i(\hat{\tau}) \right\} X_i.
\]

Denote \( I(\hat{\theta}_2) \) as the estimated observed information matrix under the null hypothesis and \( \{ I^{-1}(\hat{\theta}_2) \}_{1:p,1:p} \) as the \( \beta_1 \) component of \( I^{-1}(\hat{\theta}_2) \). The score test statistics can be written as

\[
\chi^2_L = S_L(\hat{\theta}_2)^T \{ I^{-1}(\hat{\theta}) \}_{1:p,1:p} S_L(\hat{\theta}_2)/n,
\]

which converges in distribution to \( \chi^2_p \) as \( n \to \infty \) under the null hypothesis. We will reject the null hypothesis \( H_0 : \beta_1 = 0_p \) at significant level \( \alpha \) if

\[
\chi^2_L > \chi^2_{p, \alpha}.
\]

### 3.2.2 Two Stage score test

The likelihood based score test is the most efficient test when the model specification is correct. However, in real cases, the normality assumption of \( W \) and \( Y \) may be too strong. Actually, the individual component of \( W \) may follow different continuous distributions or even discrete distributions. For example, in two phase studies, the auxiliary variables often include age, gender, etc. In this section, we develop a two stage score test extending the previous likelihood based score test. The term "two stage" means that, using the estimating equations, we first estimate the parameters in model (3.5), and then construct the score test from model (3.7) by using the estimated covariate \( W(\hat{\tau}) \). While the previous likelihood based approach estimates the parameters of both models in
(3.9) by updating the parameters iteratively.

Without normality assumptions, we still assume the response variable $Y$ has the mean structure

$$E(Y|W, X) = \gamma_0 + \gamma_1^T X + \gamma_2^T W,$$  \hspace{1cm} (3.11)

and a linear relationship between $W$ and $X$,

$$E(W|X) = \tau^T \tilde{X}.$$  \hspace{1cm} (3.12)

We propose a pseudo-score by dropping the variance term $\sigma_1^2$ in (3.10) as

$$S_{TS}(\theta_2) = \sum_{i=1}^n R_0 i \{ Y_i - \beta_0 - \gamma_2^T W_i(\tau) \} X_i.$$  \hspace{1cm} (3.13)

Here the parameter $\theta = (\beta_1^T, \theta_2^T)^T$ where nuisance parameters $\theta_2 = (\beta_0, \gamma_2^T, \tau_1^T, \tau_2^T, \ldots, \tau_q^T)^T$. At the true parameter values, each component of this pseudo-score has expectation 0 under the null hypothesis since

$$E[R_0 \{ Y - \beta_0 - \gamma_2^T W(\tau) \} X] = E[E[R_0 (Y - \beta_0 - \gamma_2^T W(\tau))X|Y, X, W]]$$
$$= E[E[R_0|Y, X, W](Y - \beta_0 - \gamma_2^T W(\tau))X]$$
$$= E[E[\pi(X, W)(Y - \beta_0 - \gamma_2^T W(\tau))X|X, W]]$$
$$= E[\pi(X, W) \{ E(Y|X, W) - \beta_0 - \gamma_2^T W(\tau) \} X]$$
$$= 0,$$

where $E(R_0|Y, X, W) = E(R_0|X, W) = \pi(X, W)$ follows the MAR assumption.

The M-estimator of nuisance parameter $\theta_2$ satisfies

$$\sum_{i=1}^n \psi_1(Y_i, X_i, W_i, R_{0i}, \hat{\theta}_2) = 0,$$
where

\[
\psi_1(Y_i, X_i, W_i, R_{0i}, \theta_2) = \begin{pmatrix}
R_{0i} \tilde{W}_i(\tau) \{ Y_i - \beta_0 - \gamma_2^T W_i(\tau) \} \\
\tilde{X}_i(W_{li} - \tau_i^T \tilde{X}_i) \\
\tilde{X}_i(W_{2i} - \tau_2^T \tilde{X}_i) \\
\vdots \\
\tilde{X}_i(W_{qi} - \tau_q^T \tilde{X}_i)
\end{pmatrix}.
\]

Hence, the estimated pesudo-score function can be written as

\[
S_{TS}(\tilde{\theta}_2) = \sum_{i=1}^{n} R_{0i} \{ Y_i - \tilde{\beta}_0 - \tilde{\gamma}_2^T \tilde{W}_i(\tilde{\tau}) \} X_i.
\]

The estimating equation also provides us the asymptotic distribution of \( \tilde{\theta}_2 \), and we can estimate the asymptotic distribution of \( S_{TS}(\tilde{\theta}_2) \) (see Appendix B). The two-stage score test statistics can be written as

\[
\chi^2_{TS} = S_{TS}(\tilde{\theta}_2)^T \text{Var}(S_{TS}(\tilde{\theta}_2))^{-1} S_{TS}(\tilde{\theta}_2),
\]

which converges in distribution to \( \chi^2_p \) as \( n \to \infty \) under the null hypothesis.

### 3.2.3 Inverse probability weighted score test

Previous two methods rely on certain assumptions of \( Y|X, W \) and \( W|X \). If these assumptions do not hold, we may obtain incorrect test results. On the other side, the complete case score (after dropping \( \sigma^2_1 \)) derived from model (3.1),

\[
S_{CC}(\beta_0) = \sum_{i=1}^{n} R_{0i} (Y_i - \beta_0) X_i,
\]  

(3.14)

does not rely on both assumptions. However, the complete case score puts the same weight, actually 1, to all subjects with completely observed data. At the true value of parameters, the expectation of each component of this complete case score under the null hypothesis and MAR assumption given \( X \) and \( W \) is

\[
E \{ R_0(Y - \beta_0)X \} = E \{ E \{ R_0(Y - \beta_0)X|Y, X, W \} \} = E \{ E(R_0|X, W)(Y - \beta_0)X \}.
\]
which is not equal to zero generally.

We correct this bias leveraging the inverse probability weighting method which is first proposed
by Horvitz and Thompson (1952) in order to deal with sampling problems. Later, by re-weighting the
complete data points, Robins et al. (1994) introduced a class of estimators without the full parametric
specification of the distribution of data under the MAR assumption. By positing a parametric model
for \( E(R_0|X, W) = P(R_0 = 1|V) \) where \( V = (X^T, W^T)^T \) is fully observed and \( \pi(V; \phi) > 0 \)
with probability 1, the inverse probability weighted score for testing \( H_0: \beta_1 = 0 \) is given as

\[
S^{(1)}_{iW}(\theta_2) = \sum_{i=1}^{n} \frac{R_{0i}}{\pi(V_i; \phi)} (Y_i - \beta_0) X_i.
\]

Here the parameter \( \theta = (\beta_1^T, \beta_0, \phi^T)^T \) can be decomposed as the parameter of interest \( \theta_1 = \beta_1 \) and the nuisance parameter \( \theta_2 = (\beta_0, \phi^T)^T \). At the true parameter values, each component of this score
has expectation 0 under the null hypothesis since

\[
E\left\{ \frac{R_0}{\pi(V; \phi)} (Y - \beta_0) X \right\} = E\left[ E\left( \frac{R_0}{\pi(V; \phi)} (Y - \beta_0) X | Y, X, W \right) \right] = E\left( \frac{E(R_0|X, W)}{\pi(V; \phi)} (Y - \beta_0) X \right) = E\left( [Y - \beta_0] X \right) = E\left[ E(Y|X) - \beta_0 \right] X = 0.
\]

In practice, a logistic regression model is often considered for \( \pi(V; \phi) \), i.e.,

\[
\pi(V; \phi) = \frac{1}{1 + \exp(-\phi^T V)}.
\]

The M-estimator of the nuisance parameter \( \theta_2 \) satisfies

\[
\sum_{i=1}^{n} \psi_2(Y_i, X_i, W_i, R_{0i}, \hat{\theta}_2) = 0,
\]

where

\[
\psi_2(Y_i, X_i, W_i, R_{0i}, \theta_2) = \begin{pmatrix}
\frac{R_{0i}}{\pi(V_i; \phi)} (Y_i - \beta_0) \\
|R_{0i} - \pi(V_i; \phi)| V
\end{pmatrix}.
\]
Finally we can calculate the asymptotic variance of \( S_{1W}^{(1)}(\hat{\theta}_2) \) and construct a test statistic using the same techniques as Appendix B.

### 3.2.4 Augmented inverse probability weighted score test

The inverse probability weighted score test described above can lead to incorrect results when we mis-specify the observed data probability model \( \pi(V; \phi) \). On the other hand, the likelihood based score test and the two-stage score test do not require to model \( \pi(V; \phi) \), but will be invalid if certain assumptions of \( Y|X, W \) and \( W|X \) do not hold. In practice, we do not know which model is correct. Here we introduce an augmented inverse probability weighted score test which yields valid hypothesis testing results when the MAR assumption holds and either of the observed data probability model or the mean model of \( Y|X, W \) (3.11) holds. The augmented propensity score for testing \( H_0: \beta_1 = 0 \) can be written as

\[
S_{AIW}^{(1)}(\theta_2) = \sum_{i=1}^{n} \left\{ \frac{R_{0i}}{\pi(V_i; \phi)}(Y_i - \beta_0)X_i - \frac{R_{0i} - \pi(V_i; \phi)}{\pi(V_i; \phi)}(m(V_i; \gamma) - \beta_0)X_i \right\},
\]

where the augmented term is \( m(V; \gamma) = \mathbb{E}(Y|X, W) = \gamma^T \tilde{V} \). Here the coefficients \( \gamma = (\gamma_0, \gamma_1^T, \gamma_2^T)^T \in \mathbb{R}^{(p+q+1) \times 1} \) where \( \gamma_0, \gamma_1, \gamma_2 \) corresponds to intercept, \( X \) and \( W \). Given the MAR assumption (3.3), this augmented term can be fitted using complete data since \( \mathbb{E}(Y|X, W) = \mathbb{E}(Y|X, W, R_0 = 1) \). In this work, we demonstrate the case that \( m(V; \gamma) \) is a linear function of \( V \). However, it can be extended to more general settings. The parameter \( \theta = (\beta_1^T, \beta_0, \phi^T, \gamma^T)^T \) can be decomposed to the parameter of interest \( \theta_1 = \beta_1 \) and the nuisance parameter \( \theta_2 = (\beta_0, \phi^T, \gamma^T)^T \). The M-estimator of the nuisance parameter \( \theta_2 \) satisfies

\[
\sum_{i=1}^{n} \psi_3(Y_i, X_i, W_i, R_{0i}, \hat{\theta}_2) = 0,
\]

where

\[
\psi_3(Y_i, X_i, W_i, R_{0i}, \theta_2) = \begin{pmatrix}
\frac{R_{0i}}{\pi(V_i; \phi)}(Y_i - \beta_0) - \frac{R_{0i} - \pi(V_i; \phi)}{\pi(V_i; \phi)}(m(V_i; \gamma) - \beta_0) \\
[R_{0i} - \pi(V_i; \phi)]\tilde{V}_i \\
R_{0i}(Y_i - \gamma^T \tilde{V}_i)\tilde{V}_i
\end{pmatrix}.
\]

The augmented inverse probability weighted score (3.16) has the double robustness property.
At the true parameter values, each component of $S_{AIW}^{(1)}(\theta_2)$ has expectation

$$E \left\{ \frac{R_0}{\pi(V; \phi)} (Y - \beta_0)X - \frac{R_0 - \pi(V; \phi)}{\pi(V; \phi)} (m(V; \gamma) - \beta_0)X \right\}$$

$$= E \left\{ (Y - \beta_0)X + \frac{R_0 - \pi(V; \phi)}{\pi(V; \phi)} (Y - m(V; \gamma))X \right\}$$

The first term will equal to $0_p$ if $H_0 : \beta_1 = 0_p$ holds. Conditioning on the random variable $V$, the second term can be written as

$$E \left[ E \left\{ \frac{R_0 - \pi(V; \phi)}{\pi(V; \phi)} (Y - m(V; \gamma))X \mid V \right\} \right]$$

$$= E \left[ E \left\{ \frac{R_0 - \pi(V; \phi)}{\pi(V; \phi)} \mid V \right\} E\{ (Y - m(V; \gamma))X \mid V \},$$

where the equation is given by the MAR assumption. If we correctly specify either the observed data probability model or the mean model (3.11), this term will equal to $0_p$. Therefore, each component of (3.16) has expectation $0_p$ under the null hypothesis.

Finally, we can calculate the asymptotic variance of $S_{AIW}^{(1)}(\hat{\theta}_2)$ and construct a test statistic using the same techniques as Appendix B.

### 3.3 Case II: Only covariates are subject to missing

In this section, we consider the case where covariates $X$ are subject to missing while the response $Y$ is fully observed. In this case, $R_0 = 1$ and $R_j$ is the random observed data indicator for $j = 1, 2, \ldots, p$. We assume the observed data probability, $P(R_j = 1|Y, X, W) = P(R_j = 1|Y, W)$ for $j = 1, 2, \ldots, p$, follows the MAR assumption.

Traditional inverse weighting methods estimate parameters by positing a model of $R_X$ as

$$P(R_X = 1|V) = \pi(V; \phi) = \frac{1}{1 + \exp(-\phi^T \tilde{V})},$$

where $R_X$ indicates that whether the whole vector $X$ is observed. The estimator of $\theta = (\beta_1^T, \beta_0, \phi^T)^T$
is based on
\[
\sum_{i=1}^{n} \left( \frac{R_{Xi}}{\pi(V_i; \phi)} [Y_i - \beta_0 - \beta_1^T X_i] \bar{X}_i \right) = 0.
\]
Hence we can construct the score
\[
S_{IW}(\theta_2) = \sum_{i=1}^{n} \frac{R_{Xi}}{\pi(V_i; \phi)} (Y_i - \beta_0) X_i, \tag{3.17}
\]
where the parameter \( \theta = (\beta_1^T, \beta_0, \phi_1^T)^T \) can be decomposed to the parameter of interest \( \theta_1 = \beta_1 \) and the nuisance parameter \( \theta_2 = (\beta_0, \phi)^T \).

However, the drawback of above method is that it only re-weights the cases that the whole \( X \) is observed. By leveraging the good property of score test that every estimation step is under the null hypothesis, we incorporate the cases that \( X \) is partially missing. Assuming the MAR assumption holds as
\[
P(R_j = 1|Y, X, W) = P(R_j = 1|Y, W) \quad j = 1, 2, \ldots, p,
\]
we posit logistic regression models for \( P(R_j = 1|V) \) as \( \pi_j(V; \phi_j) \) for \( j = 1, 2, \ldots, p \) separately where \( V = (Y, W^T)^T \) is fully observed.

The proposed inverse weighted score for testing \( H_0 : \beta_1 = 0_p \) can be written as
\[
S_{IW}^{(2)}(\theta_2) = \sum_{i=1}^{n} \left( Y_i - \beta_0 \right) \begin{bmatrix} \frac{R_{1i}}{\pi_1(V_i; \phi_1)} X_{1i} \\ \vdots \\ \frac{R_{pi}}{\pi_p(V_i; \phi_p)} X_{pi} \end{bmatrix}, \tag{3.18}
\]
where the parameter \( \theta = (\beta_1^T, \beta_0, \phi_1^T, \phi_2^T, \ldots, \phi_p^T)^T \) can be decomposed to the parameter of interest \( \theta_1 = \beta_1 \) and the nuisance parameter \( \theta_2 = (\beta_0, \phi_1^T, \phi_2^T, \ldots, \phi_p^T)^T \). The M-estimator of the nuisance parameter \( \theta_2 \) satisfies
\[
\sum_{i=1}^{n} \psi_0(Y_i, X_i, W_i, R_i, \hat{\theta}_2) = 0,
\]
where

\[
\psi_6(Y_i, X_i, W_i, R_i, \theta_2) = \begin{pmatrix}
Y_i - \beta_0 \\
[R_{1i} - \pi_1(V_i; \phi_1)]V_i \\
\vdots \\
[R_{pi} - \pi_p(V_i; \phi_p)]V_i
\end{pmatrix}.
\]

Both (3.17) and (3.18) have expectation 0 when the null hypothesis is true and missing mechanism is correct. However, as (3.18) incorporates the information where \(X\) is partially missing, it has higher power when the null hypothesis does not hold.

Similarly, the augmented inverse weighted score for testing \(H_0 : \beta_1 = 0_p\) can be written as

\[
S^{(2)}_{AIW}(\theta_2) = \sum_{i=1}^n \left( Y_i - \beta_0 \right) \begin{bmatrix}
\frac{R_{1i}}{\pi_1(V_i; \phi_1)} X_{1i} - \frac{R_{1i} - \pi_1(V_i; \phi_1)}{\pi_1(V_i; \phi_1)} m_1(V_i; \gamma_1) \\
\vdots \\
\frac{R_{pi}}{\pi_p(V_i; \phi_p)} X_{pi} - \frac{R_{pi} - \pi_p(V_i; \phi_p)}{\pi_p(V_i; \phi_p)} m_p(V_i; \gamma_p)
\end{bmatrix},
\]

where we denote \(\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_p)\) as a \((2 + q) \times p\) matrix and the augmented term \(m(V; \gamma) = E(X|Y, W) = \gamma^T \tilde{V}\). The M-estimator of the nuisance parameter \(\theta_2 = (\beta_0, \phi^T, \gamma_1^T, \gamma_2^T, \ldots, \gamma_p^T)^T\) satisfies

\[
\sum_{i=1}^n \psi_7(Y_i, X_i, W_i, R_i, \tilde{\theta}_2) = 0,
\]

where

\[
\psi_7(Y_i, X_i, W_i, R_i, \theta_2) = \begin{pmatrix}
Y_i - \beta_0 \\
[R_{1i} - \pi_1(V_i; \phi_1)]V_i \\
\vdots \\
[R_{pi} - \pi_p(V_i; \phi_p)]V_i
\end{pmatrix}.
\]

Finally we can calculate the asymptotic variance of the proposed estimated scores and construct test statistics using the same techniques as Appendix B.

When \(R_j = R_X\) for \(j = 1, 2, \ldots, p\) where covariates are either totally missing or totally observed, and MAR assumption \(P(R_X = 1|Y, X, W) = P(R_X = 1|X, W)\) holds, the problem reduces to the two
3.4 Case III: Both response variable and covariates are subject to missing

In this section, we consider the case where both response variable $Y$ and covariates $X$ are subject to missing. We assume the observed probabilities follow the MAR assumption

$$P(R_0 = 1|Y, X, W) = P(R_0 = 1|W),$$

$$P(R_j = 1|Y, X, W) = P(R_j = 1|W) \quad j = 1, 2, \ldots, p,$$

and further assume that the conditional independence assumption between $R_0$ and $R_j$ for $j = 1, 2, \ldots, p$ holds, i.e.,

$$R_0 \perp R_j | W \quad j = 1, 2, \ldots, p.$$

We posit logistic regression models for $P(R_0 = 1|W)$ as $\pi_0(W; \phi_0)$ and $P(R_j = 1|W)$ as $\pi_j(W; \phi_j)$, $j = 1, \ldots, p$, separately. The inverse weighted score for testing $H_0 : \beta_1 = 0_p$ can be written as

$$S^{(3)}_{IW}(\theta_2) = \sum_{i=1}^{n} \frac{R_{0i}}{\pi_0(W_i; \phi_0)} (Y_i - \beta_0) \left[ \begin{array}{c} \frac{R_1}{\pi_1(W_i; \phi_1)} X_{1i} \\ \vdots \\ \frac{R_p}{\pi_p(W_i; \phi_p)} X_{pi} \end{array} \right],$$

where the parameter $\theta = (\beta_1^T, \beta_0, \phi_0^T, \phi_1^T, \ldots, \phi_p^T)^T$ can be decomposed to the parameter of interest $\theta_1 = \beta_1$ and the nuisance parameter $\theta_2 = (\beta_0, \phi_0^T, \phi_1^T, \ldots, \phi_p^T)^T$. The M-estimator of the nuisance parameter $\theta_2$ satisfies

$$\sum_{i=1}^{n} \psi_0(Y_i, X_i, W_i, R_i, \hat{\theta}_2) = 0,$$
where

$$
\psi_0(Y_i, X_i, W_i, R_i, \theta_2) = \begin{pmatrix}
\frac{R_{0i}}{\pi_0(W_i; \phi_0)}(Y_i - \beta_0) \\
[R_{0i} - \pi_0(W_i; \phi_0)]W_i \\
[R_{1i} - \pi_1(W_i; \phi_1)]W_i \\
\vdots \\
[R_{pi} - \pi_p(W_i; \phi_p)]W_i
\end{pmatrix}.
$$

Similarly, the augmented inverse weighted score for testing $H_0 : \beta_1 = 0_p$ can be written as

$$
S_{A1W}^{(3)}(\theta_2) = \sum_{i=1}^{n} \begin{pmatrix}
\frac{R_{0i}R_{1i}}{\pi_0(W_i; \phi_0)\pi_1(W_i; \phi_1)}(Y_i - \beta_0)X_{1i} - \frac{R_{0i}R_{1i} - \pi_0(W_i; \phi_0)\pi_1(W_i; \phi_1)}{\pi_0(W_i; \phi_0)\pi_1(W_i; \phi_1)}m_1(W_i; \gamma_1, f_1) \\
\vdots \\
\frac{R_{0i}R_{pi}}{\pi_0(W_i; \phi_0)\pi_p(W_i; \phi_p)}(Y_i - \beta_0)X_{pi} - \frac{R_{0i}R_{pi} - \pi_0(W_i; \phi_0)\pi_p(W_i; \phi_p)}{\pi_0(W_i; \phi_0)\pi_p(W_i; \phi_p)}m_p(W_i; \gamma_p, f_p)
\end{pmatrix},
$$

where the augmented term $m(W_i; \gamma, f) = (m_1(W_i; \gamma_1, f_1), m_2(W_i; \gamma_2, f_2), \ldots, m_p(W_i; \gamma_p, f_p))^T = E((Y - \beta_0)X|W) = E(YX|W) - \beta_0E(X|W)$. If we assume a linear relationship between $Y$ and $W$, the linear model between $X$ and $W$,

$$
E(X|W) = \gamma^TW, \quad \gamma = (\gamma_1, \gamma_2, \ldots, \gamma_p) \in \mathbb{R}^{q \times p},
$$

suggests the linear model between $YX$ and $W^*$, the polynomials of degree two over $W$ with dimension $q^*$,

$$
E(YX|W) = f^TW^*, \quad f = (f_1, f_2, \ldots, f_p) \in \mathbb{R}^{q^* \times p}.
$$

The M-estimator of the nuisance parameter $\theta_2 = (\beta_0, \phi_0^T, \ldots, \phi_p^T, \gamma_1^T, \ldots, \gamma_p^T, f_1^T, \ldots, f_p^T)^T$ satisfies

$$
\sum_{i=1}^{n} \psi_0(Y_i, X_i, W_i, R_i, \hat{\theta}_2) = 0,
$$
where

\[
\psi_0(Y_i, X_i, W_i, R_{ii}, C_i, \theta_2) = \begin{pmatrix}
\frac{R_{0i}}{\pi_0(W_i; \phi_0)}(Y_i - \beta_0) \\
[R_{0i} - \pi_0(W_i; \phi_0)]W_i \\
[R_{1i} - \pi_1(W_i; \phi_1)]W_i \\
\vdots \\
[R_{pi} - \pi_p(W_i; \phi_p)]W_i \\
R_{0i}R_{1i}(X_{1i} - \gamma_1^T \tilde{W}_i)W_i \\
\vdots \\
R_{0i}R_{pi}(Y X_{pi} - f_p^T \tilde{W}_i)W_i \\
\end{pmatrix}. 
\]

Note that if covariates \(X\) are fully observed, the problem reduces section 2, and if the response variable \(Y\) is fully observed, the problem reduces to section 3.

### 3.5 Simulation

In this section, we would like to evaluate the behavior of the proposed test statistics in terms of type I error probability and power through simulation. For simplicity, we consider \(p = q = 2\). We first generate \(X = (X_1, X_2)^T\) where both \(X_1\) and \(X_2\) follow independent Normal(0,1). Then \(W = (W_1, W_2)^T\) is generated from

\[W = \tau^T \tilde{X} + e_3,\]

where \(\tau = (\tau_1, \tau_2) \in \mathbb{R}^{3 \times 2}\) and \(e_3 = (e_{31}, e_{32})^T\). We further denote \(\tau_j = (\tau_{j0}, \tau_{j1}, \tau_{j2})\) for \(j = 1, 2\). The response variable is generated from a linear model with interaction terms of \(X\) and \(W\) as

\[Y = c_0 + c_1 X_1 + c_2 X_2 + c_3 W_1 + c_4 W_2 + c_5 X_1 W_1 + c_6 X_1 W_2 + c_7 X_2 W_1 + c_8 X_2 W_2 + e_2,\] (3.19)

where \(e_2\) is independent Normal(0,1). Denote \(c = (c_1, c_2, \ldots, c_8)^T\) for simplicity. We want to test \(\beta_1 = 0_p\) in \(E(Y|X) = \beta_0 + \beta_1^T X\). Generalized score tests like two stage method and AIW method may
require to model \( Y|X,W \) or the augmented term. Here we only model the main effects and will misspecify it when any \( c_k \neq 0 \) for \( k \in \{5,6,7,8\} \).

On the other side, the observed indicator \( R_j \) for certain \( j \in \{0,1,\ldots,p\} \) is generated from \( P(R_j = 1|X,W,Y) = \frac{1}{1 + \exp(-r_j(X,W,Y,\xi_j,\omega_j))} \) where

\[
r_j(X,W,Y,\xi_j,\omega_j) = \xi_{j0} + \xi_{j1}X_1 + \xi_{j2}X_2 + \xi_{j3}W_1 + \xi_{j4}W_2 + \xi_{j5}Y + \omega_j,
\]

where coefficients \( \xi_j = (\xi_{j0},\xi_{j1},\xi_{j2},\xi_{j3},\xi_{j4},\xi_{j5})^T \) and \( \omega_j \) is misspecification term. Generalized score tests like IW method and AIW method require to model the missing mechanism. Here we only model the main effects and will misspecify it when the misspecification term is not zero.

When only response \( Y \) is subject to missing, we generate the error term \( e_{31} \) from independent Normal(0,1) and \( e_{32} \) from independent Uniform(-1,1) to demonstrate the difference between the likelihood based score test and the two stage score test. We generate \( r_0 \) using linear model of \( X \) and \( W \) plus the misspecification term \( \omega_0 \). When covariates \( X \) is subject to missing, both \( e_{31} \) and \( e_{32} \) are generated from Normal(0,1). Hence \( X \) is also linear in \( W \). We use the linear function to model the augmented term which should be correct if \( Y|X,W \) is linear. We also generate \( r_j \) using linear model of \( Y \) and \( W \) plus the misspecification term \( \omega_j \) for \( j \in \{1,2\} \). When both response and covariates are subject to missing, \( e_{31} \) and \( e_{32} \) are also generated from standard normal distribution which suggests a linear model between \( X|W \). If the mean model \( Y|W,X \) is also linear, we can build a linear model for \( X|W \) and quadratic model for \( Y|W \) and \( X|W \) which form the augmented term together. We generate \( r_j \) using linear model of \( W \) plus the misspecification term \( \omega_j \) for \( j \in \{0,1,2\} \).

Although the model can be more complicated in the real world, we only consider the main effects for the model (3.19) and (3.20) in this simulation. When the true model includes interaction terms (\( c_k \neq 0 \) for \( k \in \{5,6,7,8\} \) or \( \omega_j \neq 0 \) for \( j \in \{0,1,2\} \)), we may misspecify the model. For each missing pattern case, we evaluate generalized score tests in three scenarios: (1) We correctly specify the augmented term and the missing mechanism. (2) We correctly specify the augmented term while misspecify the missing mechanism. In this case, the inverse probability weighted score test is not valid anymore, while the augmented inverse probability weighted score test provides robustness. (3) We correctly specify the missing mechanism while misspecify the augmented term.

We evaluate the test assuming that the limiting distribution of the generalized score statistics
proposed converge to $\chi^2$ as $n \to \infty$ under the null hypothesis, and calculate the size and power of different tests empirically. We use the chi-square cutoff at certain threshold $\alpha$ as our rejection criteria and check whether the test attains the correct size. The estimated type I error probability

$$\frac{1}{S} \sum_{i=1}^{S} I(T_i > \chi^2_{2,\alpha} | H_0)$$

is calculated based on $\alpha = 0.01, 0.05$ and $0.1$ over $S$ replications. We also report the estimated power at the significance level $\alpha = 0.05$, i.e.,

$$\frac{1}{S} \sum_{i=1}^{S} I(T > \chi^2_{2,\alpha} | H_a),$$

where $H_a$ is the alternative hypothesis that $\beta_1 \neq 0_p$. The power function corresponding to different $\beta$ in model (3.1) is also plotted, i.e.,

$$\frac{1}{S} \sum_{i=1}^{S} I(T > \chi^2_{2,\alpha} | \beta_1 j = \Delta, j = 1, \ldots, p),$$

where $\Delta$ is the value of difference between components of $\beta$ and 0. We set $\Delta = 0.2$ when we calculate the power in (3.22) and $\Delta \in \{0, 0.2, 0.4, 0.8\}$ to plot the power function.

We have $c = (1, 1, -1, 1, 1, 0, 0, 0, 0)^T$, $\tau_1 = (2, -0.5, 0.75)^T$ and $\tau_2 = (2, -0.5, 0.25)^T$ in scenario (1) and (2). However, in scenario (3), we have $c = (1, -0.5, -0.25, 1, 1, 0.5, 0.5, 0.5)$ with $\tau_1 = \tau_2 = (0.5, 0, 0)^T$. Although there are interaction terms, the conditional expectation $E(Y|X)$ still has a linear form

$$E(Y|X_1, X_2) = c_0 + c_1 X_1 + c_2 X_2 + c_3 E(W_1|X_1, X_2) + c_4 E(W_2|X_1, X_2) + c_5 X_1 E(W_1|X_1, X_2)$$

$$+ c_6 X_1 E(W_2|X_1, X_2) + c_7 X_2 E(W_1|X_1, X_2) + c_8 X_2 E(W_2|X_1, X_2)$$

$$= (c_0 + c_3 E(W_1|X_1, X_2) + c_4 E(W_2|X_1, X_2)) + (c_1 + c_5 E(W_1|X_1, X_2) + c_6 E(W_2|X_1, X_2))X_1$$

$$+ (c_2 + c_7 E(W_1|X_1, X_2) + c_8 E(W_2|X_1, X_2))X_2.$$ 

Hence our interest parameter $\beta_1$ is equal to

$$\begin{pmatrix}
  c_1 + c_5 \tau_{10} + c_6 \tau_{20} \\
  c_2 + c_7 \tau_{10} + c_8 \tau_{20}
\end{pmatrix}.$$
as $\tau_{j1} = \tau_{j2} = 0$ for $j = 1, 2$.

When we calculate the power with $\alpha = 0.05$, parameter $c$ is set to be $(1, 1 + \Delta, -1 + \Delta, 1, 1, 0, 0, 0, 0)^T$ in scenario (1) and (2), and $(1, -0.5 + \Delta, -0.25 + \Delta, 1, 1, 0.5, 0.5, 0.5)^T$ in scenario (3). Although there are interaction terms in scenario (3), the conditional expectation $E(Y|X)$ still have a linear form. In this case, the variance of $Y$ depends on $X$. However, the estimating equation based methods IW and AIW should be immune to heteroskedasticity.

For each missing mechanism, we compare different testing procedures. The 'CC' method conducts the traditional score test only based on the complete data, which should be incorrect due to MAR assumptions. The inverse probability weighted score test and the augmented inverse probability weighted score test are denoted as 'IW' and 'AIW' for simplicity. When only the response variable $Y$ is subject to missing, we also have the likelihood based score test denoted as 'Likelihood', and the two stage score test denoted as 'TS'.

We set the sample size $N = 1000$ when we estimate the Type 1 error of test and $N = 200$ when we evaluate the power. The type 1 error/power is estimated based on $S = 10000$ replications. The power plot is based on $S = 1000$ replications.

### 3.5.1 Response variable $Y$ is subject to missing

We generate $R_0$ from (3.20) with $\xi_0 = (1, 0.25, -0.1, -0.25, 0.1, 0)^T$. The interaction term $\omega_0$ is set to be 0 in scenario (1) and (3), where we correctly specify the relationship. However, in scenario (2), we have $\omega_0 = 0.2X_1W_1 + 0.1X_1W_2 - 0.1X_2W_1 + 0.05X_2W_2$ while misspecify the observed data probability by modeling only main effects of $X$ and $W$.

The simulation result in three scenarios is shown in Table 3.1, and the plot of the power function is represented in Figure 3.1. In scenario (1) and (2), the likelihood based score test does not attain the correct size while the two stage score test is relatively closer to the true size since the latter does not require to specify the distribution of $e_2$ and $e_3$ correctly. When $R_0|X, W$ is misspecified, IW obtains the incorrect size while AIW is still robust. When $Y|X, W$ is misspecified, both IW and AIW obtain the correct size since IW does not rely on the mean structure (3.11) and is derived from the original model (3.1), and on the other hand, AIW has the double robustness property. The AIW approach gives the correct size and reasonable power for all three scenarios and is the most robust approach.

When the problem reduces to $p = 1$ and $q = 1$, we also ran a simulation in Figure 3.2 to demon-
strate that the square root of the proposed score statistics follow the Normal distribution when the corresponding required model are correctly specified.

### 3.5.2 covariates X are subject to missing

#### 3.5.2.1 Missing data in two-phase study

We generate $R_1 = R_2 = R_X$ from (3.20) with $\xi = (-0.2, 0, 0, -0.4, 0.2, 0.2)^T$. The interaction term $\omega$ is set to be 0 in scenario (1) and (3), where the missing mechanism is correctly specified; and $\omega = -0.1 Y W_1 + 0.1 Y W_2$ in scenario (2), where we misspecify the observed data probability by modeling main effects of $Y$ and $W$.

The simulation result is presented in Table 3.2, and the power plot is presented in Figure 3.3. The AIW method is more robust than the IW method, which has incorrect estimated type I error probability when $R_X | Y, W$ is misspecified.

#### 3.5.2.2 Each component of X is subject to missing

We generate $R_j$ from (3.20) for $j = 1, 2$ with $\xi_1 = (0.1, 0, 0, 0.1, 0.1, 0.2)^T$, $\xi_2 = (0.2, 0, 0, 0.25, 0.1, 0.2)^T$. The interaction term $\omega_1 = \omega_2 = 0$ in scenario (1) and (3). However, we have $\omega_1 = 0.2 Y W_1 - 0.05 Y W_2$ and $\omega_2 = -0.1 Y W_2$ and misspecify the observed data probability by modeling only main effects of $Y$ and $W$ in scenario (2).

The simulation result is presented in Table 3.3 and the power plot is presented in Figure 3.4. We compare the proposed methods "IW covariate" and "AIW covariate" that consider the information of $X$ that is partially missing to methods "IW whole" and "AIW whole" which solve two-phase study problems. In the first scenario, all inverse weighting methods attain the correct size while the methods considering partial missing covariates attain higher power than their counterparts. In the second scenario where the missing mechanism is misspecified, both "AIW covariate" and "AIW whole" have the correct size due to the double robustness property, and "AIW covariate" attains higher power than "AIW whole" as the former method applies more information. In the third scenario, both "IW covariate" and "AIW covariate" attain the correct size while "IW whole" and "AIW whole" can not attain the correct size. Theoretically, the methods solving two-phase study problems should also attain correct size but lower power if we correctly model $R_X | Y, W$. However, we model...
the $R_X$ using logistic regression with the main effects of $Y$ and $W$, which is incorrect since there is no logistic regression model can fit $R_X = R_1 R_2$ perfectly given that $R_1$ and $R_2$ are generated from the logistic regression model. We notice that the 'IW whole' attains the correct size in the first scenario because the main effects of $Y$ and $W$ are sufficient to approximate the model for $R_X$ in this setting.

### 3.5.3 Both response and covariates are subject to missing

We generate $R_j$ from (3.20) for $j = \{0, 1, 2\}$ with $\xi_0 = (0.6, 0, 0, -0.1, 0.1, 0)^T$, $\xi_1 = (0.8, 0, 0.25, 0.1, 0)^T$, $\xi_2 = (0.8, 0, 0.1, 0.1, 0)^T$. The interaction term $\omega_0 = \omega_1 = \omega_2 = 0$ in scenario (1) and (3). However, we have $\omega_0 = 0.4W_1^2 - 0.1W_2^2 + 0.2W_1 W_2$, $\omega_1 = -0.1W_1^2 + 0.1W_2^2 - 0.2W_1 W_2$ and $\omega_2 = 0.1W_2^2 - 0.1W_1 W_2$, while misspecify the observed data probability by modeling main effects of $W$ in scenario (2).

The simulation result is presented in Table 3.4 and the power plot is presented in Figure 3.5. When all models are correctly specified, the complete case analysis has the correct estimated type I error probability. However, when $R_0|W$ and $R_j|W$ for $j = 1, 2$ are misspecified, the complete case method and IW method no longer obtain the correct estimated type I error probability while AIW still performs robust.

### 3.6 Application

In this section, we apply the proposed generalized score tests on the age-related macular degeneration (AMD) dataset (Molenberghs and Kenward, 2007). This clinical trial is interested in comparing the effect of interferon-α treatment and placebo for patients with AMD (Molenberghs and Kenward, 2007). There are 240 patients participating in this trial, and they are divided equally into each treatment group. The visual acuity and lines of vision are assessed over four consecutive visits. We concentrate on the lines of vision at the last visit. Since the patient may miss the visit, there are about 17% missing observations. We also define the treatment group as our variable of interest $X$ labeled as 1 if the patient is assigned to the active treatment group, and 0 if the patient is assigned to the placebo group. Before the trial start, researchers collect the baseline lines of vision and lesion grade, which scales from 1 to 4. Lesion grade, served as auxiliary variable $W_1$ in our study, is collected for all the patients except a patient with id 21, so we can safely delete this observation. Define the
Table 3.1 Estimated type I error probability and power of each test when the response $Y$ is subject to missing in different situations. Both type I error probability and power are estimated based on 10000 replications. The standard error is between 0.001 to 0.005 approximately.

<table>
<thead>
<tr>
<th></th>
<th>$\alpha = 0.01$</th>
<th>$\alpha = 0.05$</th>
<th>$\alpha = 0.1$</th>
<th>power($\alpha = 0.05$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>models are correctly specified</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>0.011</td>
<td>0.058</td>
<td>0.112</td>
<td>0.455</td>
</tr>
<tr>
<td>Likelihood</td>
<td>0.014</td>
<td>0.068</td>
<td>0.120</td>
<td>0.611</td>
</tr>
<tr>
<td>TS</td>
<td>0.010</td>
<td>0.051</td>
<td>0.101</td>
<td>0.518</td>
</tr>
<tr>
<td>IW</td>
<td>0.010</td>
<td>0.052</td>
<td>0.102</td>
<td>0.449</td>
</tr>
<tr>
<td>AIW</td>
<td>0.010</td>
<td>0.052</td>
<td>0.102</td>
<td>0.543</td>
</tr>
<tr>
<td></td>
<td>$R_0</td>
<td>X, W$ is misspecified</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>0.114</td>
<td>0.279</td>
<td>0.395</td>
<td>0.540</td>
</tr>
<tr>
<td>Likelihood</td>
<td>0.015</td>
<td>0.064</td>
<td>0.122</td>
<td>0.586</td>
</tr>
<tr>
<td>TS</td>
<td>0.010</td>
<td>0.050</td>
<td>0.099</td>
<td>0.485</td>
</tr>
<tr>
<td>IW</td>
<td>0.070</td>
<td>0.197</td>
<td>0.304</td>
<td>0.502</td>
</tr>
<tr>
<td>AIW</td>
<td>0.012</td>
<td>0.053</td>
<td>0.100</td>
<td>0.508</td>
</tr>
<tr>
<td></td>
<td>$Y</td>
<td>X, W$ is misspecified</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>0.020</td>
<td>0.085</td>
<td>0.147</td>
<td>0.390</td>
</tr>
<tr>
<td>Likelihood</td>
<td>0.043</td>
<td>0.125</td>
<td>0.200</td>
<td>0.504</td>
</tr>
<tr>
<td>TS</td>
<td>0.016</td>
<td>0.067</td>
<td>0.125</td>
<td>0.357</td>
</tr>
<tr>
<td>IW</td>
<td>0.011</td>
<td>0.052</td>
<td>0.104</td>
<td>0.378</td>
</tr>
<tr>
<td>AIW</td>
<td>0.010</td>
<td>0.052</td>
<td>0.101</td>
<td>0.415</td>
</tr>
</tbody>
</table>

Table 3.2 Estimated type I error probability and power of each test when the whole vector $X$ is subject to missing in different situations. Both type I error probability and power are estimated based on 10000 replications. The standard error is between 0.001 to 0.005 approximately.

<table>
<thead>
<tr>
<th></th>
<th>$\alpha = 0.01$</th>
<th>$\alpha = 0.05$</th>
<th>$\alpha = 0.1$</th>
<th>power($\alpha = 0.05$)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>models are correctly specified</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>0.013</td>
<td>0.060</td>
<td>0.113</td>
<td>0.381</td>
</tr>
<tr>
<td>IW</td>
<td>0.012</td>
<td>0.051</td>
<td>0.103</td>
<td>0.329</td>
</tr>
<tr>
<td>AIW</td>
<td>0.010</td>
<td>0.056</td>
<td>0.105</td>
<td>0.349</td>
</tr>
<tr>
<td></td>
<td>$R_X</td>
<td>Y, W$ is misspecified</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>0.012</td>
<td>0.054</td>
<td>0.105</td>
<td>0.324</td>
</tr>
<tr>
<td>IW</td>
<td>0.017</td>
<td>0.073</td>
<td>0.140</td>
<td>0.237</td>
</tr>
<tr>
<td>AIW</td>
<td>0.011</td>
<td>0.057</td>
<td>0.112</td>
<td>0.310</td>
</tr>
<tr>
<td></td>
<td>$X</td>
<td>Y, W$ is misspecified</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>0.297</td>
<td>0.494</td>
<td>0.606</td>
<td>0.604</td>
</tr>
<tr>
<td>IW</td>
<td>0.013</td>
<td>0.058</td>
<td>0.111</td>
<td>0.288</td>
</tr>
<tr>
<td>AIW</td>
<td>0.016</td>
<td>0.064</td>
<td>0.121</td>
<td>0.299</td>
</tr>
</tbody>
</table>
(a) $\alpha = 0.01$
(b) $\alpha = 0.05$
(c) $\alpha = 0.1$

Both models are correctly specified

(d) $\alpha = 0.01$
(e) $\alpha = 0.05$
(f) $\alpha = 0.1$

$R_0 \mid X, W$ is misspecified

(g) $\alpha = 0.01$
(h) $\alpha = 0.05$
(i) $\alpha = 0.1$

$Y \mid X, W$ models is misspecified

Figure 3.1 Power function of proposed score tests when response $Y$ is subject to missing.
Figure 3.2 Distribution of proposed score statistics when response Y is subject to missing and $p = 1$. 

(a) two stage method  
(b) iw method  
(c) aiw method  
Both models are correctly specified  

(d) two stage method  
(e) iw method  
(f) aiw method  
$R_0|X, W$ is misspecified  

(g) two stage method  
(h) iw method  
(i) aiw method  
$Y|X, W$ models is misspecified
Figure 3.3 Power function of proposed score tests when the whole vector of $X$ is subject to missing.
Table 3.3  Estimated type I error probability and power of each test when covariates $X$ follow general missing data patterns in different situations. Both type I error probability and power are estimated based on 10000 replications. The standard error is between 0.001 to 0.005 approximately.

<table>
<thead>
<tr>
<th></th>
<th>$\alpha = 0.01$</th>
<th>$\alpha = 0.05$</th>
<th>$\alpha = 0.1$</th>
<th>$\text{power}(\alpha = 0.05)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>models are correctly specified</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>0.025</td>
<td>0.086</td>
<td>0.153</td>
<td>0.409</td>
</tr>
<tr>
<td>IW covariate</td>
<td>0.010</td>
<td>0.053</td>
<td>0.102</td>
<td>0.431</td>
</tr>
<tr>
<td>IW whole</td>
<td>0.012</td>
<td>0.053</td>
<td>0.105</td>
<td>0.355</td>
</tr>
<tr>
<td>AIW covariate</td>
<td>0.011</td>
<td>0.051</td>
<td>0.103</td>
<td>0.471</td>
</tr>
<tr>
<td>AIW whole</td>
<td>0.012</td>
<td>0.053</td>
<td>0.104</td>
<td>0.370</td>
</tr>
<tr>
<td>**$R_1</td>
<td>Y, W$ and $R_2</td>
<td>Y, W$ are misspecified**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>0.089</td>
<td>0.230</td>
<td>0.338</td>
<td>0.436</td>
</tr>
<tr>
<td>IW covariate</td>
<td>0.017</td>
<td>0.071</td>
<td>0.136</td>
<td>0.337</td>
</tr>
<tr>
<td>IW whole</td>
<td>0.076</td>
<td>0.208</td>
<td>0.315</td>
<td>0.392</td>
</tr>
<tr>
<td>AIW covariate</td>
<td>0.011</td>
<td>0.053</td>
<td>0.104</td>
<td>0.440</td>
</tr>
<tr>
<td>AIW whole</td>
<td>0.013</td>
<td>0.053</td>
<td>0.104</td>
<td>0.323</td>
</tr>
<tr>
<td>**$X</td>
<td>Y, W$ is misspecified**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>complete case</td>
<td>0.801</td>
<td>0.918</td>
<td>0.950</td>
<td>0.771</td>
</tr>
<tr>
<td>IW covariate</td>
<td>0.014</td>
<td>0.057</td>
<td>0.109</td>
<td>0.317</td>
</tr>
<tr>
<td>IW whole</td>
<td>0.030</td>
<td>0.095</td>
<td>0.158</td>
<td>0.342</td>
</tr>
<tr>
<td>AIW covariate</td>
<td>0.014</td>
<td>0.061</td>
<td>0.114</td>
<td>0.315</td>
</tr>
<tr>
<td>AIW whole</td>
<td>0.042</td>
<td>0.120</td>
<td>0.191</td>
<td>0.371</td>
</tr>
</tbody>
</table>

Table 3.4  Estimated type I error probability and power of each test when both response $Y$ and covariates $X$ are subject to missing in different situations. Both type I error probability and power are estimated based on 10000 replications. The standard error is between 0.001 to 0.005 approximately.

<table>
<thead>
<tr>
<th></th>
<th>$\alpha = 0.01$</th>
<th>$\alpha = 0.05$</th>
<th>$\alpha = 0.1$</th>
<th>$\text{power}(\alpha = 0.05)$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>models are correctly specified</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>0.013</td>
<td>0.059</td>
<td>0.110</td>
<td>0.288</td>
</tr>
<tr>
<td>IW</td>
<td>0.010</td>
<td>0.050</td>
<td>0.102</td>
<td>0.289</td>
</tr>
<tr>
<td>AIW</td>
<td>0.011</td>
<td>0.049</td>
<td>0.098</td>
<td>0.238</td>
</tr>
<tr>
<td>**$R_0</td>
<td>W$ and $R_1, R_2</td>
<td>W$ are misspecified**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>0.188</td>
<td>0.380</td>
<td>0.502</td>
<td>0.329</td>
</tr>
<tr>
<td>IW</td>
<td>0.030</td>
<td>0.113</td>
<td>0.200</td>
<td>0.438</td>
</tr>
<tr>
<td>AIW</td>
<td>0.009</td>
<td>0.049</td>
<td>0.098</td>
<td>0.234</td>
</tr>
<tr>
<td>**$X Y</td>
<td>W^+$ is misspecified**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CC</td>
<td>0.062</td>
<td>0.157</td>
<td>0.243</td>
<td>0.352</td>
</tr>
<tr>
<td>IW</td>
<td>0.008</td>
<td>0.049</td>
<td>0.099</td>
<td>0.214</td>
</tr>
<tr>
<td>AIW</td>
<td>0.009</td>
<td>0.051</td>
<td>0.104</td>
<td>0.211</td>
</tr>
</tbody>
</table>
(a) $\alpha = 0.01$
(b) $\alpha = 0.05$
(c) $\alpha = 0.1$

Both models are correctly specified

(d) $\alpha = 0.01$
(e) $\alpha = 0.05$
(f) $\alpha = 0.1$

$R_1|Y, W$ and $R_2|Y, W$ are misspecified

(g) $\alpha = 0.01$
(h) $\alpha = 0.05$
(i) $\alpha = 0.1$

$Y|X, W$ model is misspecified

Figure 3.4 Power function of proposed score tests when covariates $X$ follow general missing data patterns.
Both models are correctly specified

\( R_0 | W \) and \( R_1, R_2 | W \) are misspecified

\( Y | X, W \) is misspecified

Figure 3.5 Power function of proposed score tests when both \( Y \) and \( X \) are subject to missing.
The generalized score statistics and corresponding P-value testing whether the interferon-α treatment has a significant effect on the lines of vision improvement rate.

<table>
<thead>
<tr>
<th></th>
<th>score</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CC</td>
<td>4.282</td>
<td>0.039</td>
</tr>
<tr>
<td>Likelihood</td>
<td>4.452</td>
<td>0.035</td>
</tr>
<tr>
<td>TS</td>
<td>4.519</td>
<td>0.034</td>
</tr>
<tr>
<td>IW</td>
<td>3.904</td>
<td>0.048</td>
</tr>
<tr>
<td>AIW</td>
<td>4.094</td>
<td>0.043</td>
</tr>
</tbody>
</table>

response variable $Y$

$$Y = \frac{\text{lines of vision at fourth visit} - \text{baseline lines of vision}}{\text{baseline lines of vision}},$$

as the lines of vision improvement rate which is subject to missing. We are interested in testing the association between lines of vision improvement rate and treatment group. This is the missing data problem where the response variable $Y$ is subject to missing. We build a logistic regression model for the observed indicator $R_0$ using the auxiliary variable $W = (W_1, W_1^2)$ and the variable of interest $X$,

$$P(R_0 = 1|X, W) = \frac{1}{1 + \exp(\xi_0 + \xi_1 X + \xi_2 W_1 + \xi_3 W_1^2)}.$$ 

We see all the predictors have significant effects, and the model AUCROC is 0.6859, which suggests that the missing completely at random (MCAR) assumption is not appropriate, and we may assume MAR holds. Under MAR assumption, the complete case score test may not attain the correct size and can lead to misleading results. We conduct proposed generalized score tests on this dataset. Here the likelihood based score test and two stage score test is fitted using $W = (W_1)$, and inverse weighted score tests are fitted using $W = (W_1, W_1^2)$ as we need to fit the observed indicator. The test result is presented in Table 3.5. We see that while the complete case score test rejects the null hypothesis aggressively, the proposed generalized score tests, including IW and AIW, are more conservative. The likelihood based score test and two stage score test is more aggressive than the complete case analysis. I prefer the result of AIW as we have a good model for the missing mechanism and the AIW method has double robustness property.
3.7 Discussion

In this chapter, we have provided several score test procedures for testing the partial association between a response variable $Y$ and covariates $X$ given certain auxiliary variables when there are missing observations. When the distribution of data is known, the most efficient way is to form a likelihood function and derive the score function from it. If the distribution function is unknown or hard to model, we can still use the two stage score test. However, both of the above methods require to correctly model (3.2). On the other side, inverse weighting methods are based on the model between the response variable and covariates directly, using auxiliary information to re-balance the data based on MAR assumptions. The drawback of inverse weighting methods is that they only consider the cases where the response variable and covariates variable are all completely observed, which loses a lot of information on those covariates that are partially observed. By using the score statistics, which assumes everything under the null, we make the inverse weighted score statistics incorporate all the information even part of covariates are missing, and hence improve the power. The corresponding augmented inverse weighted score test share the same advantage as the inverse weighted score test while provides robustness when the missing mechanism is misspecified. The proposed score tests can also be applied to the two sample $t$ test, such as the real data example presented in section 3.6, and ANOVA problem where covariates $X$ are discrete.
BIBLIOGRAPHY


A.1 Proof of Proposition 1

The first order derivative of (2.9) is

\[
\frac{dp(\eta|1,b)}{d\eta} = \frac{1}{B(1,b)} \phi'(\sqrt{\eta}) \frac{1}{\sqrt{\eta}}(2-2\Phi(\sqrt{\eta}))^{b-1} \\
+ \frac{1}{B(1,b)} \phi(\sqrt{\eta}) \left( -\frac{1}{2} \eta^{-3/2} \right) (2-2\Phi(\sqrt{\eta}))^{b-1} \\
+ \frac{1}{B(1,b)} \phi(\sqrt{\eta}) \frac{1}{\sqrt{\eta}} (b-1) (2-2\Phi(\sqrt{\eta}))^{b-2} \left( -2\phi(\sqrt{\eta}) \frac{1}{2\sqrt{\eta}} \right) \\
= \frac{1}{B(1,b)} \phi(\sqrt{\eta}) \eta^{-3/2} (2-2\Phi(\sqrt{\eta}))^{b-2} \left[ -\Phi(\sqrt{\eta}) + \sqrt{\eta}(1-b)\phi(\sqrt{\eta}) \right] \\
< \frac{1}{B(1,b)} \phi(\sqrt{\eta}) \eta^{-3/2} (2-2\Phi(\sqrt{\eta}))^{b-2} \left[ -\Phi(\sqrt{\eta}) + \sqrt{\eta}(1-b)\phi(\sqrt{\eta}) \right].
\]

Let \( u = \sqrt{\eta} \), it suffices to show that \( h(u) = -(u^2 + 1)(1 - \Phi(u)) + u\phi(u) < 0 \), \( \forall u \in (0, \infty) \). The first order derivative of \( h \) is

\[
h'(u) = -2(u(1 - \Phi(u)) - \phi(u)) < 0.
\]
Hence $h$ is a monotone increasing function for $u \in (0, \infty)$, and $h(u) \to \infty$ as $u \to \infty$. Therefore, $h$ is always negative for $u \in (0, \infty)$, in other word, $\frac{dp(\eta|1, b)}{d\eta} < 0$ for $\eta \in (0, \infty)$. The prior function $p(\eta|1, b)$ is a monotone decreasing function.

### A.2 Details for optimizing the posterior distribution

In order to solve

$$\max_{\theta \in \Theta} \prod_{i=1}^{n} p_i(\theta, \eta), \quad (A.1)$$

we can plug in the expression of $p_i(\theta, \eta)$ based on equation (2.6) and optimize it over $\theta \in \Theta_\eta$. However, the optimization can be difficult as $\lambda$ is implicitly defined. Instead, we step back to the constrained optimization method used to solve (2.5) and extend it to a new algorithm which can be used to solve (A.1).

We form the Lagrange multiplier

$$L = -\sum_{i=1}^{n} \log p_i - \lambda_1 (\sum_{i=1}^{n} p_i - 1) - \lambda_2 (\mu^T \Sigma^{-1} \mu - \eta) - \lambda_3 \sum_{i=1}^{n} p_i [y_i (x_i - \mu_1)]$$

$$- \lambda_4 \sum_{i=1}^{n} p_i [(1 - y_i)(x_i - \mu_0)] - \lambda_5^T m_1 - \lambda_6^T m_0, \quad (A.2)$$

where $m_1$ is the vectorized version of the third moment function in $h(x_i, y_i, \theta)$, and $m_0$ is the vectorized version of the forth moment function in $h(x_i, y_i, \theta)$. The second constraint is from the constraint over parameter space $\theta \in \Theta_\eta$, and remaining moment constraints are from $h(x_i, y_i, \theta)$.

Based on (2.6), for any given $\theta$, we have

$$p_i = \frac{1}{n - \lambda_3^T [y_i (x_i - \mu_1)] - \lambda_4^T [(1 - y_i)(x_i - \mu_0)] - \lambda_5^T \frac{\partial m_1}{\partial p_i} - \lambda_6^T \frac{\partial m_0}{\partial p_i}}.$$
Setting derivatives with respect to other parameters to zero results in the equations

\[
\begin{align*}
\mu^T \Sigma^{-1} \mu - \eta &= 0 \\
\sum_{i=1}^{n} p_i [y_i (x_i - \mu_1)] &= 0, \\
\sum_{i=1}^{n} p_i [(1 - y_i) (x_i - \mu_0)] &= 0, \\
\sum_{i=1}^{n} p_i y_i [(x_{ij} - \mu_{1j}) (x_{ik} - \mu_{1k}) - \sigma_{1jk}] &= 0, \\
\sum_{i=1}^{n} p_i [1 - y_i] [(x_{ij} - \mu_{0j}) (x_{ik} - \mu_{0k}) - \sigma_{0jk}] &= 0, \\
\left[-2 \lambda_2 \Sigma^{-1} (\mu_1 - \mu_0) + \sum_{i=1}^{n} p_i y_i \lambda_3 \right]^T \frac{\partial \mu_1}{\partial \mu_1} - \lambda_5 \frac{\partial m_1}{\partial \mu_1} &= 0, \\
\left[-2 \lambda_2 \Sigma^{-1} (\mu_0 - \mu_1) + \sum_{i=1}^{n} p_i (1 - y_i) \lambda_4 \right]^T \frac{\partial \mu_0}{\partial \mu_0} - \lambda_6 \frac{\partial m_0}{\partial \mu_0} &= 0,
\end{align*}
\]

Here \(\mu_{gj}\) is the \(j\)-th element of \(\mu_g\); \(\lambda_{6jk}\) is the element of the lagrange multiplier corresponding to \(\sigma_{0jk}\), and \(\lambda_{5jk}\) is the element of the lagrange multiplier corresponding to \(\sigma_{1jk}\).

When there is no extra constraint of \(\eta\), or say \(\lambda_2 = 0\), above procedure reduces to finding the maximum empirical likelihood where we have moment constraints on mean and covariance. A non-zero \(\lambda_2\) drives system (A.3) into a new stable point. For any given \(\lambda_2\), we compute the parameters iteratively based on above equations, and the algorithm converges to certain parameter set \(\theta\) and corresponding weights \(\{p_i(\theta, \eta), i = 1, \ldots, n\}\), where \(\eta = \eta(\theta)\) can be calculated from the equation (2.7). This \(\theta\) is the parameter set that achieves \(\max_{\theta \in \Theta} \prod_{i=1}^{n} p_i(\theta, \eta)\). By varying \(\lambda_2\) in certain range, the algorithm converges to certain \(\theta\) and \(\eta\), and meanwhile provides the quantity \(\max_{\theta \in \Theta} \prod_{i=1}^{n} p_i(\theta, \eta)\).

After getting a discretized approximation of \(\max_{\theta \in \Theta} \prod_{i=1}^{n} p_i(\theta, \eta)\) over \(\eta\), we optimize (2.12) over \(\eta\) on the grid. Finally, we find the posterior mode of \(\theta\) and \(\eta\) as
\[ \theta, \eta = \arg\max_{\theta, \eta} \left\{ \prod_{i=1}^{n} p_i(\theta, \eta) f(\eta) \right\}. \]
APPENDIX FOR PROJECT 2

B.1 Asymptotic Variance of the two stage likelihood score

Define $A(\theta_2)$ as

$$A(\theta_2) = -E[\psi'_1(Y_i, X_i, W_i, R_{0i}, \theta_2)]$$

$$= E \begin{pmatrix} R_0 \hat{W}(\tau) \hat{W}^T(\tau) & -R_0 \gamma_{21} \hat{W}(\tau) \hat{X}^T & \ldots & -R_0 \gamma_{2q} \hat{W}(\tau) \hat{X}^T \\ 0 & \hat{X} \hat{X}^T & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \hat{X} \hat{X}^T \end{pmatrix},$$

where the second equation holds when the reparameterized model is correctly specified.
From the book of Boos and Stefanski (2013), under suitable regularity conditions,

\[ \hat{\theta}_2 - \theta_2 = \frac{1}{n} \sum_{i=1}^{n} A(\theta_2)^{-1} \psi(Y_i, X_i, W_i, R_{0i}, \theta_2) + R_n \]

\[ = \frac{1}{n} \sum_{i=1}^{n} h(Y_i, X_i, W_i, R_{0i}) + R_n, \]

where \( \sqrt{n} R_n \overset{p}{\to} 0_{1+q+q(p+1)} \) as \( n \to \infty \).

Further denote \( S_i(\theta_2) = R_{0i}[Y_i - \beta_0 - \gamma T W_i(\tau)X_i] \) as the contribute of \( i \)-th data point to the score under the null hypothesis. And hence \( S(\hat{\theta}) \) can be written as \( \sum_{i=1}^{n} S_i(\hat{\theta}_2) \). We have

\[ \sum_{i=1}^{n} S_i(\hat{\theta}_2) - nE\{S_i(\theta_2)\} = h_T(Y_i, X_i, W_i, R_{0i}) + nR_n', \]

where

\[ h_T(Y_i, X_i, W_i, R_{0i}) = S_i(\theta_2) - E\{S_i(\theta_2)\} + \left[ E\{S_i'(\theta_2)\}\right] h(Y_i, X_i, W_i, R_{0i}). \]

Hence the variance of (3.13) can be written as

\[ \text{Var}(S(\hat{\theta})) = \text{Var}(h_T(Y_i, X_i, W_i, R_{0i})), \]

which can be estimated from the data.