2023

Estimation in Generalized Estimating Equation measurement error models using instrumental variables/ Estimating the cardinality of latent defective edges in hypergraphs

Damin Zhu
University of Vermont

Estimation in Generalized Estimating Equation measurement error models using instrumental variables/ Estimating the cardinality of latent defective edges in hypergraphs

A Dissertation Presented

by

Damin Zhu

to

The Faculty of the Graduate College

of

The University of Vermont

In Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
Specializing in Mathematical Science

May, 2023

Defense Date: March 20th, 2023
Dissertation Examination Committee:

Jeffrey S. Buzas, Ph.D., Advisor
Yves Dubief, Ph.D., Chairperson
Greg Warrington, Ph.D.
Jean-Gabriel Young, Ph.D.
Cynthia J. Forehand, Ph.D., Dean of the Graduate College

Abstract

This dissertation consists of two studies. The first develops theory for a new method for estimating regression parameters using generalized estimating equations (GEE) with panel data prone to covariate measurement error. The focus is on logistic regression, though the method is applicable to other models. The method requires availability of instrumental variables (IV) to identify model parameters. Simulations are performed to assess the performance of the proposed estimator. The method, abbreviated GEEIV, is able to accurately estimate logistic regression parameters masked by measurement error in a variety of population configurations.

In the second study, an algorithm is proposed to estimate the number of latent defective edges in large hypergraphs. The new statistical method combines the strength of sampling strategies and an existing algorithmic method known for efficient latent edge identification for small graphs. Our statistical approach strikes a balance between computational time consumption and estimation power, with the flexibility to adapt to several assumption violations. Simulations are performed on both synthetic data and a simulator loaded with US western grid structures. The new algorithm was able give unbiased estimates using relatively little computational time for the synthetic data for a wide range of combinations of graph sizes, defective graph edges and defective edge distributions. Simulation results from US western grid data agreed with a previous study on relatively small latent edge sets. On a large edge set, previous studies were not able to provide a reasonable estimate. The new algorithm was able to give estimates and confidence intervals for the larger problem.

Table of Contents

List of Figures iv
List of Tables v

1 Generalized Estimating Equations with Instrumental Variables 4
1.1 Introduction 4
1.1.1 From linear models to generalized linear models 6
1.1.2 Robust statistics and M-estimators 10
1.1.3 Panel data and Generalized estimating equations 14
1.2 Measurement error 17
1.2.1 Notation 19
1.2.2 Attenuation effect of measurement error in the linear model . 20
1.2.3 Attenuation effect of measurement error in GLMs 22
1.2.4 Correcting measurement error induced bias using instrumental variables 25
1.2.5 Instrumental variable method for GLM correction 27
1.2.6 Standardized residual preserves covariance structure 30
1.3 Instrumental variables approach to GEE with measurement error–Logistic Regression 36
1.3.1 Model statement and notation 37
1.3.2 Variance estimation 42
1.4 Simulation study 44
1.4.1 Data generation 45
1.4.2 Simulation results 49
1.5 Conclusions 59
1.6 Appendix 60
1.6.1 Derivation of $\boldsymbol{\mathcal{A}}$ 60
1.6.2 Generating correlated binary clusters 62
1.6.3 Tables on Newton-Raphson convergence rate 74

2 Estimating the Cardinality of Latent Defective Sets 76
2.1 Introduction 76
2.2 Background 77
2.2.1 Power grid and cascading failure 77
2.2.2 Group testing and edge search problem 78
2.3 Terminology and notation 79
2.4 Sampling algorithm for estimating d 80
2.4.1 Choice of the estimation function g 83

2.4.2 Determining the optimum subgraph size subject to a computational cost constraint . . . 89
2.4.3 Determining the number of subgraphs subject to a computational cost constraint . . . 93
2.5 Simulation results . . . 94
2.5.1 Estimation performance for large d . . . 94
2.5.2 Estimation performance for small d . . . 95
2.5.3 Summary of simulation results . . . 95
2.5.4 Assessing the performance of the optimal m and W . . . 99
2.6 Western US power grid test case . . . 101
2.7 Conclusion . . . 104
2.8 Appendix . . . 106
2.8.1 Hypergeometric distribution and maximum likelihood estimator 106
2.8.2 A heuristic explanation of why there are very few d-edges in an optimal m-graph . . . 111
2.8.3 Brief explanation of Chen and Hwang's algorithm in our setting and the derivation of W . . . 112

List of Figures

1.1 Measurement error bending regression lines toward 0 slope. 22
1.2 The attenuation effect on the slope estimators for logistic regression. . 24
1.3 The similarity between F(1.7x) and standard normal c.d.f $\Phi(x)$. . . . 25
1.4 A comparison of estimator sample means. Normal Measurement error. Plotted values are multiplied by 1000. 52
1.5 A comparison of estimator sample standard deviations ordered by cluster size. Normal Measurement error. Values are multiplied by 1000. Note the scale of the vertical axes differ. 56
1.6 A comparison of estimator sample standard deviations ordered by total number of observations. Normal Measurement error. Values are multiplied by 1000. 57
1.7 A comparison of estimator mean absolute error ordered by cluster size. Normal Measurement error. Values are multiplied by 1000. 58
1.8 Correlation upper bound by marginal probabilities. 66
1.9 For a given α, p_j valid range is determined by maximum and minimum p_i. Intermediate p (Purple) plays no role. 69
1.10 Fixing the variance of X causes the p_i to have varied variance and skewness with respect to OR. 70
1.11 p_i generated with standard normal X is likely to have wider range than that from a normal distribution with identical mean and variance. . . 71
1.12 For a given (α, EY), σ_p is chosen such that all p's would very likely to fall into the valid region posed by $p_{(1)} = EY - 3\sigma_p$ and $p_{(n)} = EY + 3\sigma_p$. 72

2.1 A comparison of efficiency of Chen and Hwang's algorithm to edge testing all edges. Uniformly distributed d-edges of size 3 were generated in 750 graphs of order 100. The number of d-edges in each graph is shown on the x-axis. 81
2.2 A comparison of m efficiency. Standard deviation of adjacent m's are compared to the that of the mode of algorithm selected m. 101

List of Tables

1.1 Average failed attempts to generate 40 clusters of size 25, $\alpha = 0.25$. . 68
1.2 Cluster Size=25; Number of Cluster=160; Normal Measurement error with attenuation factor 0.8. Reporting number of converged estimation via Newton Rapson method. Total number of cases is 1000 for all rows. 74
1.3 Cluster Size=25; Number of Cluster=40; Normal Measurement error with attenuation factor 0.8. Reporting number of converged estimation via Newton Rapson method. Total number of cases is 1000 for all rows. 75
1.4 Cluster Size=6; Number of Cluster=20; Normal Measurement error with attenuation factor 0.8. Reporting number of converged estimation via Newton Rapson method. Total number of cases is 1000 for all rows. 75

2.1 Simulation results for d-edges following uniform distribution. The statistics on each row are calculated over 1000 random graphs. 96
2.2 Simulation result for d-edges following power law distribution. For each row 1000 random graphs were estimated. For each graph, L was calculated by the algorithm to achieve a standard deviation 5% of $\hat{d}$, after finding an optimal m. 97
2.3 Simulation results for d-edges following uniform distribution. For each row 1000 random graphs were estimated. For each graph, L was calculated by the algorithm to achieve a standard deviation 5% of $\hat{d}$, after finding an optimal m. 98
2.4 Performance of optimal m and W 100
2.5 Estimations for the Western US power grid cascade failure test case. $m = 6$, $L = 8,000,000$. 103

Overview

Estimation of parameters for probability models is a central focus of the field of Statistics. Estimation methodologies have been widely adopted in various fields such as medical science, sociology, manufacturing, electrical engineering, etc. for purposes including but not limited to prediction, assessment and decision making.

This dissertation develops estimation methods for two distinct application areas. The first study examines correcting for covariate measurement error in Generalized Linear Models (GEEs). GEE's were designed for analyzing clustered data. The focus herein is primarily on binary outcomes, though the results extend to other outcome models. Clustered binary outcomes are common in medical studies where patient status is considered binary (dead/alive, healthy/ill), and patients are grouped by hospital/region or physician (for example). GEE gives biased estimation of the odds ratio if the input covariates are measured with error. The bias typically attenuates the estimate toward zero, thereby masking or diminishing the potential impact of the covariate on the outcome. We describe two sources of bias in the estimating equations, namely the dependency between mis-measured covariates and the residual, and the non-linearity of link function. The first source is addressed by introducing instrumental variables, whose independence with most model components can be

used to eliminate the aforementioned dependency. The second issue is addressed by standardizing the estimating equation component that contains the link function. A new issue arises with the standardizing process: the working correlation matrix, which is a component in GEE to incorporate cluster correlation, needs to be adjusted according to the standardized component. We prove that, for the logistic regression model, the new working correlation matrix preserves the correlation structure in the original error-free model. Our new method, called GEEIV, is able to provide unbiased and efficient estimation.

The second topic addresses estimating the number of latent edges in a graph. At least two approaches to the problem have been studied in the past. One approach is purely algorithmic, and the other uses extrapolation [23][34] when the number of latent edges is very large. The algorithmic approach can find all latent edges, but it is applicable only for graphs with relatively small numbers of latent edges, as otherwise the computational time is prohibitive. The extrapolation approach is able to estimate the number of latent edges when there are too many to find all of them algorithmically, but the method becomes unworkable when the latent edges are very large in number. Our study is motivated by finding nodes in a national power grid that would cause catastrophic failure if the nodes themselves failed. The power grid contains hundreds of millions of such latent edges representing potential risk. Previous methods are too slow or inaccurate to analyze graphs of this magnitude. Using sampling theory, we develop a statistical algorithm whose run-time does not increase with the number of latent edges. The key idea is to extract information from small subgraphs to make inference on the full graph. A method of moments estimator is used to obtain an unbiased estimator of the latent edge number. To ensure statistical efficiency, we

developed a procedure to determine an optimum order for the subgraphs. Optimum subgraphs have maximum information "density". Users can then sample as many optimum subgraphs as needed to reach their desired precision.

Chapter 1

Generalized Estimating Equations with Instrumental Variables

1.1 Introduction

This chapter explores using instrumental variables to correct for bias induced through covariate measurement error in regression models fit using generalized estimating equations(GEE).

The term measurement error refers to the discrepancy between a quantity's measurement and its true value. It is well known that covariate measurement error in a parametric regression model causes bias in coefficient estimation, and the behavior of this bias varies with the nature of the model and the nature of the measurement error itself [21]. In this section of the dissertation, we describe and analyze generalized estimating equations when classic additive measurement error is present.

Briefly, GEE's were proposed as an extension to the generalized linear model(GLM) [6]. GEE's are intended for clustered data, where observations within clusters can be

correlated. GEE's relax GLM's independence assumption by incorporating a working correlation matrix to model intra-cluster correlation. The increased complexity of GEE's require development of extensions to established covariate measurement error correction methods for GLM.

Methods that effectively correct for covariate measurement error require additional information that allow estimation of the magnitude of the measurement error. Additional information can come in many forms, and be internal or external to the study data, see [22], Chapter 3-6. Instrumental variables (IVs) are additional measurements internal to the study that are correlated with the true, mis-measured covariate. [16] studied an approach to IV estimation for GLM models. The instrumental variable correction method in that paper can be adapted to GEEs.

This chapter is organized as follows. Section 1.1 introduces the models and parameter estimation techniques for clustered binary data in the absence of measurement error. Section 1.2 discusses the effects of covariate measurement error for linear models, and describes a consistent estimator using an instrumental variable. In Section 1.3, we combine the knowledge of the previous two sections and define a generalized estimating equation utilizing instrumental variables, abbreviated GEEIV. GEEIV consistently or nearly consistently estimates regression parameters for clustered data in the presence of covariate measurement error. Section 1.4 contains a simulation study for the logistic regression model.

1.1.1 From linear models to generalized linear models

Linear regression models are the bread and butter of statistical analysis and have been studied and employed extensively over the last 100+ years. Examples of linear models include analysis of variance models and polynomial regression, see [17]. The General Linear Model (distinct from generalized linear models) is of the form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.1}$$

where $\boldsymbol{Y}$ is a $n \times 1$ response vector, $\boldsymbol{X}$ is a $n \times k$ explanatory random variable matrix, $\boldsymbol{\beta}$ is a $k \times 1$ regression parameter vector and $\boldsymbol{\epsilon}$ is a $n \times 1$ error vector. The model posits a linear relation between the expectation of a response variable Y and one or more explanatory variables $\boldsymbol{X}$ such that the expectation is the aggregated effect of a linear combination of $\boldsymbol{X}$ and regression parameters $\boldsymbol{\beta}$. The random component ϵ represents all latent factors that are not taken into consideration in the model.

An important reason for fitting linear regression models is to estimate regression coefficients $\boldsymbol{\beta}$. Coefficients carry information about the direction and strength of the correlation and impact between covariates and response. Estimation and inference on $\boldsymbol{\beta}$ is a common way of drawing correlation-related conclusions. By far the most popular choice for coefficient estimation is least squares. Under a normality assumption on $\boldsymbol{\epsilon}$, least squares is equivalent to the method of maximum likelihood. Maximum likelihood estimators(MLE) are obtained by maximizing a likelihood function so that the observed data is most probable under the model inferred by the likelihood, a sta-

tistical analogy to "seeing is believing". Besides its intuition, least squares estimation results in the "best linear unbiased estimator"(BLUE) for linear models, in a sense that it has the least variance among all unbiased estimators.

J.A.Nelder and R.W.M. Wedderburn extended the linear models in (1.1) to data drawn from distributions in the exponential family, see [4]. Their extended models are called Generalized Linear Models (GLMs). The extension naturally incorporates (1.1) when $\boldsymbol{\epsilon}$ is normally distributed, as the normal distribution is in the exponential family. Other common exponential family distributions include the exponential, Bernoulli(binary data), Poisson(count data), and gamma. Diverse fields including epidemiology, economics, medicine, and geostatistics often result in data that is well-modeled by a distribution in the exponential family.

GLM models the expectation of a response Y as a function of $\eta = \boldsymbol{X}\boldsymbol{\beta}$. Specifically,

$$E[Y|\boldsymbol{X}] = \mu(\eta) \tag{1.2}$$

where $\mu(\cdot)$ is termed an inverse link function. In the general linear model (1.1), μ is the identity function. The inverse link function μ can be specified independently of the distribution for Y. However, there exists a natural choice called the canonical link that greatly simplifies likelihood calculation. To see this, we first define the exponential family of distributions.

Let θ and Φ denote scalar valued location and scale parameters, and $b(\cdot), C(\cdot,\cdot)$, and $\alpha(\cdot)$ functions. The exponential family of densities $f(y \mid \theta, \Phi)$ (or probability mass functions) is defined via

$$f(y|\theta,\Phi) = \exp\left(\frac{y\theta - b(\theta)}{\alpha(\Phi)} + c(y,\Phi)\right). \tag{1.3}$$

It is not difficult to show that

$$\begin{aligned} E[Y] &= b^{'}(\theta), \\ Var(Y) &= \alpha(\Phi) b^{''}(\theta). \end{aligned} \tag{1.4}$$

As noted above, in a regression setting the conditional mean of the response is modeled as $E[Y|\boldsymbol{X}] = \mu(\eta)$ where $\mu(\cdot)$ is the inverse link function. The canonical (inverse) link is defined as $\mu = b'$. The canonical link equates θ to $\eta = \boldsymbol{X}\boldsymbol{\beta}$:

$$\theta = b'^{-1}(\mu(\eta)) = \eta.$$

Example (Logistic model): A popular choice for analyzing binary responses is the logistic regression model. A Bernoulli random variable y, given expectation p, has the following probability mass function:

$$f_{y|p}(y) = p^y(1-p)^{1-y} \quad y \in \{0, 1\}$$

which can also be written in the form of (1.3):

$$\begin{aligned} p^y(1-p)^{1-y} &= \exp\left[y \log p + (1-y)\log(1-p) \right] \\ &= \exp\left[y \log \frac{p}{1-p} + \log(1-p) \right] \\ &= \exp\left[y\theta - \log(e^\theta + 1) \right] \end{aligned} \tag{1.5}$$

where $\theta = \log \frac{p}{1-p}$.

Equation (1.5) is a special case of (1.3) with $b(\theta) = \log(e^\theta + 1)$, $c(y, \Phi) = 0$ and

$\alpha(\Phi) = 1$. The canonical link for binary regression yields

$$E[Y \mid \boldsymbol{X}] = b'(\eta) = \frac{1}{1 + e^{-\eta}} = F(\eta) \tag{1.6}$$

where F is the well-known logistic distribution function. Therefore, the canonical link for a binary outcome model results in the logistic regression.

Example (Poisson regression): The probability mass function for a Poisson random variable with scalar valued parameter λ is

$$P(Y|\lambda) = \frac{\lambda^y e^{-\lambda}}{y!}$$

which can also be written as:

$$\frac{\lambda^y e^{-\lambda}}{y!} = \exp\Big(y\theta - e^{\theta} - \log(y!)\Big)$$

where $\theta = \log(\lambda)$. It is a special case of (1.3) where $b(\theta) = e^{\theta}$, $c(y, \Phi) = \log(y!)$ and $\alpha(\Phi) = 1$. The canonical link for Poisson regression yields

$$E[Y|\boldsymbol{X}] = b'(\eta) = \exp(\eta).$$

Suppose the data consist of independent pairs of observations $\{(Y_1, \boldsymbol{X_1}), (Y_2, \boldsymbol{X_2}), \ldots, (Y_n, \boldsymbol{X_n})\}$, and a GLM is postulated. Under regularity restrictions, the MLE, denoted $\hat{\boldsymbol{\theta}}$, is a solution to $\frac{\partial}{\partial \boldsymbol{\beta}} l(\boldsymbol{\theta}|\boldsymbol{Y}) = 0$, where $l(\boldsymbol{\theta}|\boldsymbol{Y}) = \sum_{i=1}^{n} l_i(\boldsymbol{\theta}|Y_i)$ is the log-likelihood function. The derivative can be further expanded

with the chain rule:

$$\sum_{i=1}^{n} \frac{\partial l_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \theta_i} \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \mathbf{0}. \tag{1.7}$$

By (1.3) and (1.4), three of the four components can be simplified:

$$\begin{aligned} \frac{\partial l_i}{\partial \theta} &\propto (y_i - \mu_i) \\ \frac{\partial \theta}{\partial \mu} &\propto \delta_i^{-1} \\ \frac{\partial \eta}{\partial \boldsymbol{\beta}} &= \boldsymbol{X}_i \end{aligned}$$

where $\delta_i = \mathrm{Var}(Y_i|\boldsymbol{X}_i)$. The MLE estimating equations simplify to

$$\sum_{i=1}^{n} \frac{\partial l_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} (Y_i - \mu_i)\delta_i^{-1} \frac{\partial}{\partial \boldsymbol{\beta}} \mu_i = \sum_{i=1}^{n} (Y_i - \mu_i) \frac{\frac{\partial}{\partial \boldsymbol{\beta}} \mu_i}{\mathrm{Var}(Y_i|\boldsymbol{X}_i)} = \mathbf{0}. \tag{1.8}$$

If the link is canonical, i.e., $\theta_i = \eta_i$, the estimating equations further simplify to

$$\sum_{i=1}^{n} \frac{\partial l_i}{\partial \theta_i} \frac{\partial \eta_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} (Y_i - \mu_i)\boldsymbol{X}_i = \mathbf{0}.$$

Note that (1.8) is a non-linear system of equations whenever μ is nonlinear. In general, a closed form solution for (1.8) cannot be obtained, and therefore the estimating equations are solved numerically.

1.1.2 Robust statistics and M-estimators

In GLMs, maximum likelihood estimators are not generally unbiased, and therefore they are not BLUE. However, they are consistent and efficient, as they asymptotically achieve the Cramer-Rao lower bound. A consistent estimator converges in probability

to the true parameter value. In other words, a consistent estimator's bias and variability goes to 0 as sample size increases, as long as the model is correctly specified.

Unfortunately, model mis-specification is difficult to avoid in applied data analysis, and MLEs can be sensitive to model violations. In the following chapters we discuss the effect of two sources of model mis-specification, ignored correlation and measurement error, which would undermine the MLE's efficiency and/or consistency. But we first introduce a general class of estimators called M-estimators. M-estimation is a generalization of maximum likelihood estimation.

Definition 1.1.1. *Suppose* $Y \sim P(\boldsymbol{\theta})$ *where* $\boldsymbol{\theta}$ *is a vector of unknown parameters. For observations* $Y_1, \ldots, Y_n$*, an M-estimator for* $\boldsymbol{\theta}$ *is defined as the solution(s) of the vector valued estimating equations*

$$\sum_{i=1}^{n} \boldsymbol{\psi}(\boldsymbol{\theta}; Y_i) = \mathbf{0}. \tag{1.9}$$

$\boldsymbol{\psi}$ *is called the estimating function.*

M-estimation allows flexibility in modeling, to potentially achieve robustness against model mis-specification. Consistency of M-estimators is considered in the following lemma.

Lemma 1.1.1. *If the estimating function* $\boldsymbol{\psi}$ *satisfies*

$$E[\boldsymbol{\psi}(\boldsymbol{\theta}_0, Y)] = \mathbf{0}$$

for some $\boldsymbol{\theta}_0 \in \mathcal{R}^k$*, then the solution to* $\sum_{i=1}^{n} \psi(\theta, Y_i) = 0$*, denoted* $\hat{\boldsymbol{\theta}}$*, converges to* $\boldsymbol{\theta_0}$*, under regularity conditions.*

Example (normal mean and variance): Suppose $Y_1, ..., Y_n$ are i.i.d. $N(\mu, \sigma^2)$ where $\boldsymbol{\theta} = (\mu, \sigma^2)$ is unknown. The maximum likelihood estimating equations for $\boldsymbol{\theta}$ are

$$\psi(\boldsymbol{\theta}; Y) = \sum_{i=1}^{n} \psi_i(\boldsymbol{\theta})$$

where

$$\psi_i(\boldsymbol{\theta}) = \begin{bmatrix} Y_i - \mu \\ (Y_i - \mu)^2 - \sigma^2 \end{bmatrix}.$$

The solution $\hat{\boldsymbol{\theta}}$ is consistently estimating $\boldsymbol{\theta}$ because

$$E[\psi_i(\boldsymbol{\theta}, Y)] = \begin{bmatrix} E[Y_i] - \mu \\ E[(Y_i - \mu)^2] - \sigma^2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Example (logistic regression): Suppose the binary variable $y \in \{0, 1\}$ has mean $E[Y \mid \boldsymbol{X}] = F(\boldsymbol{X}^T \beta)$ where $F(\cdot)$ is the logistic distribution function. The maximum likelihood estimating equations for logistic regression are:

$$\psi(\boldsymbol{\beta}; \boldsymbol{Y}, \boldsymbol{X}) = \sum_{i=1}^{n} \psi_i(\boldsymbol{\beta}) = \mathbf{0}$$

where

$$\psi_i(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} \Big[y_i \eta_i - \log(e^{\eta_i} + 1) \Big] = \Big[y_i - F(\boldsymbol{X}_i^T \beta) \Big] \boldsymbol{X}_i.$$

The estimating function is unbiased:

$$E[\psi_i(\boldsymbol{\beta}) \mid \boldsymbol{X}_i] = \Big[E(y_i \mid \boldsymbol{X_i}) - F(\boldsymbol{X}_i^T \beta) \Big] \boldsymbol{X}_i = \mathbf{0}$$

where the last equality follows from (1.6).

M-estimators are (generally) implicitly defined, rendering impossible the determination of exact sampling distribution results in finite samples. The asymptotic distribution of M-estimator's can be derived by expanding the estimating equation in a Taylor series around $\boldsymbol{\theta_0}$ and applying the Central Limit Theorem and Slutsky's Theorem. The following lemma results.

Lemma 1.1.2. *Suppose* $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ *are independent observations with* $X_i \sim P(\boldsymbol{\theta})$. *Let* $\hat{\boldsymbol{\theta}}$ *solve*

$$\sum_{i=1}^{n} \boldsymbol{\psi}(\boldsymbol{X}_i, \boldsymbol{\theta}) = \mathbf{0}. \tag{1.10}$$

Then

$$\hat{\boldsymbol{\theta}} \text{ is } AN\left(\boldsymbol{\theta}_0,\ \frac{\boldsymbol{V}(\boldsymbol{\theta}_0)}{n}\right),$$

where

$$\begin{aligned}
\boldsymbol{V}(\boldsymbol{\theta}_0) &= \boldsymbol{\mathcal{A}}(\boldsymbol{\theta}_0)^{-1}\boldsymbol{\mathcal{B}}(\boldsymbol{\theta}_0)\boldsymbol{\mathcal{A}}(\boldsymbol{\theta}_0)^{-T}, \\
\boldsymbol{\mathcal{A}}(\boldsymbol{\theta}_0) &= E\left[\boldsymbol{\psi}'(\boldsymbol{\theta}_0)\right], \\
\boldsymbol{\mathcal{B}}(\boldsymbol{\theta}_0) &= E\left[\boldsymbol{\psi}(\boldsymbol{\theta}_0)\boldsymbol{\psi}^T(\boldsymbol{\theta}_0)\right],
\end{aligned}$$

and $\boldsymbol{\psi'} = \frac{\partial}{\partial\theta}\boldsymbol{\psi}$.

The dependence of $\boldsymbol{\psi}$ on $\boldsymbol{X}$ has been suppressed. The proofs for Lemmas 1.1.1 and 1.1.2 can be found in Boos and Stefanski [27]. In practice, $\boldsymbol{\mathcal{A}}(\boldsymbol{\theta}_0)$, $\boldsymbol{\mathcal{B}}(\boldsymbol{\theta}_0)$ must

be estimated. Empirical estimators are given by

$$\boldsymbol{\mathcal{A}}_n(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}_i^{'}(\hat{\boldsymbol{\theta}})$$
$$\boldsymbol{\mathcal{B}}_n(\hat{\boldsymbol{\theta}}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\psi}_i(\hat{\boldsymbol{\theta}})\boldsymbol{\psi}_i^{T}(\hat{\boldsymbol{\theta}}).$$

1.1.3 Panel data and Generalized estimating equations

The use of GLMs requires independent observations, in which case the likelihood is comprised of a product of density functions evaluated at each observation. When the data are correlated, the likelihood in general does not have a closed form. Assuming independence when data are correlated does not typically affect consistency of the MLE. In the case of GLMs, this is easily seen by inspecting (1.8) and noting the MLE estimating equation remains unbiased regardless of the correlation structure among observations.

The penalty of mis-specifying the correlation is instead on efficiency. Ignoring the correlation structure results in an estimator with larger variance than would occur if the correlation is correctly modelled [19].

Liang and Zeger proposed Generalized estimating equations(GEE) [6] to address the efficiency loss caused by correlation in panel data. Panel data has a correlation structure where observations are grouped into clusters. Cluster members are correlated with each other while being independent from those in a different cluster. A double subscript will be used for clustered data. For example, $\boldsymbol{Y}_i = [Y_{i1}, \ldots, Y_{in_i}]^T$ denotes the n_i outcomes in the ith cluster.

GEE extends GLMs by adding a so-called "working correlation matrix" to the estimating equations. This is done as follows. First write (1.7) in an equivalent matrix form:

$$\mathbf{0} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \frac{\partial l_i}{\partial \theta} \frac{\partial \theta}{\partial \mu} \frac{\partial \mu}{\partial \eta} \frac{\partial \eta}{\partial \boldsymbol{\beta}} = \boldsymbol{D}^T \boldsymbol{B} \boldsymbol{\Delta}^{-1} \boldsymbol{S} \tag{1.11}$$

where

$$\boldsymbol{D} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ & & \vdots & \\ 1 & x_{1n} & \cdots & x_{kn} \end{pmatrix}, \qquad \boldsymbol{\Delta} = \begin{pmatrix} \delta_1 & & & \\ & \delta_2 & & \\ & & \ddots & \\ & & & \delta_n \end{pmatrix},$$

$$\boldsymbol{B} = \begin{pmatrix} \frac{\partial \mu_1}{\partial \eta_1} & & & \\ & \frac{\partial \mu_2}{\partial \eta_2} & & \\ & & \ddots & \\ & & & \frac{\partial \mu_n}{\partial \eta_n} \end{pmatrix}, \quad \boldsymbol{S} = \begin{pmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \\ \vdots \\ y_n - \mu_n \end{pmatrix}. \tag{1.12}$$

With independent data, $\boldsymbol{\Delta}$ coincides with the variance-covariance matrix of $\boldsymbol{Y}|\boldsymbol{X}$, denoted $\boldsymbol{V}$. Note that $\boldsymbol{V}$ can be partitioned as $\boldsymbol{V} = \boldsymbol{\Delta}^{\frac{1}{2}} \boldsymbol{I} \boldsymbol{\Delta}^{\frac{1}{2}}$ where $\boldsymbol{I}$ is an $n \times n$ identity matrix representing the correlation matrix for independent outcomes. Liang and Zeger consider replacing $\boldsymbol{I}$ with a working correlation matrix $\boldsymbol{R}(\boldsymbol{\alpha})$ that represents the correlation structure of $\boldsymbol{Y}|\boldsymbol{X}$. Using $\boldsymbol{V} = \boldsymbol{\Delta}^{\frac{1}{2}} \boldsymbol{R}(\boldsymbol{\alpha}) \boldsymbol{\Delta}^{\frac{1}{2}}$ results in the estimating equations

$$\boldsymbol{D}^T \boldsymbol{B} \boldsymbol{V}^{-1} \boldsymbol{S} = \mathbf{0}. \tag{1.13}$$

With clustered observations, $\boldsymbol{R}(\boldsymbol{\alpha})$ is a diagonal block matrix with each block

being the assumed cluster correlation:

$$\boldsymbol{R}(\boldsymbol{\alpha}) = \begin{pmatrix} \boldsymbol{R}_1(\boldsymbol{\alpha}) & & & \\ & \boldsymbol{R}_2(\boldsymbol{\alpha}) & & \\ & & \ddots & \\ & & & \boldsymbol{R}_n(\boldsymbol{\alpha}) \end{pmatrix}.$$

Then (1.13) results in summing over clusters:

$$\sum_i \boldsymbol{D}_i^T \boldsymbol{B}_i \boldsymbol{V}_i^{-1} \boldsymbol{S}_i = \boldsymbol{0} \tag{1.14}$$

where $\boldsymbol{D}_i$, $\boldsymbol{B}_i$, $\boldsymbol{S}_i$ and $\boldsymbol{V}_i$ are submatrices of $\boldsymbol{D}$, $\boldsymbol{B}$, $\boldsymbol{S}$ and $\boldsymbol{V}$, respectively for the ith cluster.

Note that (1.14) matches the form given by (1.10) and hence its solution $\hat{\boldsymbol{\beta}}$ is a M-estimator. The estimating function is unbiased:

$$E[\boldsymbol{D}_i^T \boldsymbol{B}_i \boldsymbol{V}_i^{-1} \boldsymbol{S}_i(\boldsymbol{\beta}_0) | \boldsymbol{X}] = \boldsymbol{D}_i^T \boldsymbol{B}_i \boldsymbol{V}_i^{-1} E[\boldsymbol{S}_i(\boldsymbol{\beta}_0) | \boldsymbol{X}] = 0.$$

By Lemma 1.1.1, $\hat{\boldsymbol{\beta}}$ converges in probability to the true parameter $\boldsymbol{\beta}_0$:

The nuisance parameter vector $\boldsymbol{\alpha}$ in the working correlation matrix $\boldsymbol{R}$ requires its own vector of estimating equations. Since the dimension of $\boldsymbol{\alpha}$ and its interpretation vary according to the correlation structure, its estimating equation has no general form. A commonly used working correlation structure is the exchangeable structure.

The exchangeable correlation matrix for a cluster is given by

$$\boldsymbol{R_i}(\alpha) = \begin{pmatrix} 1 & \alpha & \dots & \alpha \\ \alpha & 1 & \dots & \alpha \\ \vdots & & \ddots & \vdots \\ \alpha & & & 1 \end{pmatrix}.$$

For this model, α is a scaler representing the correlation between any two observations within a cluster. A common way to consistently estimate α is via the estimating equation

$$\sum_i \sum_{j<k} \left(\frac{Y_{ij} - \mu_{ij}}{\sqrt{\delta_{ij}}} \frac{Y_{ik} - \mu_{ik}}{\sqrt{\delta_{ik}}} - \alpha \right) = 0. \tag{1.15}$$

The estimating equation is easily seen to be unbiased. The complete GEEs for this case is then

$$\begin{bmatrix} \sum_i \boldsymbol{D}_i^T \boldsymbol{B}_i \boldsymbol{V}_i^{-1} \boldsymbol{S}_i \\ \sum_i \sum_{j<k} \left(\frac{Y_{ij} - \mu_{ij}}{\sqrt{\delta_{ij}}} \frac{Y_{ik} - \mu_{ik}}{\sqrt{\delta_{ik}}} - \alpha \right) \end{bmatrix} = \mathbf{0}. \tag{1.16}$$

When the working correlation matrix closely models the true correlation structure in the data, GEE estimators have smaller variance than estimators that ignore the correlation structure. The gain in efficiency increases with the strength of the correlation among observations within a cluster [6].

1.2 Measurement error

Measurement error is the error induced in the process measuring a quantity, and is another common sources of model mis-specification. Measurement error does not necessarily involve human mistakes, and is in general considered impossible to eliminate.

For example, when measuring some chemical quantity with an instrument, the reading is often a sum of a random component and the true value. Another example is measuring of blood pressure. In this case, the true value is considered as the average value of the quantity over many readings, which does not coincide with any single reading.

Measurement error can arise on either the explanatory variable or response variable in a regression analysis. The effect of measurement errors behaves distinctively when attached to different components of a model. This dissertation only considers measurement error on the explanatory variable $\boldsymbol{X}$ in a regression model. See [13][26][9] for studies and applications of other types of measurement error. It is worth noting that measurement error on $\boldsymbol{X}$ is considered a source of endogeneity. Endogeneity refers to the situation where the explanatory variable is correlated with the systematic error, with other common sources being a simultaneous system and missing variables. It is known to bias parameter estimation and reduce efficiency.

Measurement error on explanatory variables can be further classified by whether it is additive and/or homoscedastic. This dissertation focus on the classical measurement error model, that is measurement error that is additive, unbiased, and non-differential. A classical measurement error model assumes a additive relationship

$$W = X + U \tag{1.17}$$

where X and W denote the unobserved true explanatory variable and the mismeasured one. U represents stochastic measurement error that is independent of X and the outcome variable. It is the most natural way of modeling measurement error in the sense that it captures the action of measuring a quantity with an instrument-the

error is related to the precision of the instrument, not the quantity being measured.

Another type of measurement error is Berkson measurement error. The Berkson model is

$$X = W + U \tag{1.18}$$

where the measurement error U is independent from W. In this case, W is the fixed target quantity and the true quantity X varies around W. Classical measurement error is known to bias regression parameter estimators. Berkson measurement error does not introduce bias in linear models. Rather, the penalty is increased variance and loss of power.

Unbiased measurement error simply means $E[U] = 0$, implying that $E[W \mid X] = X$. Intuitively, it is assuming properly functioning instrument and no human error. Non-differential measurement error assumes that W contains no additional information about Y when X is known. Formally, measurement error is non-differential when the distribution of $Y \mid X, W$ is equivalent to the distribution of $Y \mid X$. In this case W is called a surrogate. An example of differential measurement error is given in [22] in the context of study on the relationship between diet and breast cancer where a woman's diet after cancer diagnosis is taken as a measurement of her long-term dietary intake, which is believed to be related to the development of breast cancer. However the diet afterward(W) can very possibly be altered by the diagnosis result(Y).

1.2.1 Notation

Classical measurement error (unbiased, additive and non-differential) is defined as follows.

1. $W = X + U$. $U \perp\!\!\!\perp X$.

2. $E[U|X] = 0$.

3. $f_{Y|XW} = f_{Y|X}$ (non-differential)

where f denotes either a density or probability mass function.

A note on notation. A subscript W will be added to matrices (and partial derivatives) to indicate contamination. For example,

$$\boldsymbol{D}_W = \begin{pmatrix} 1 & w_{11} & \cdots & w_{k1} \\ 1 & w_{12} & \cdots & w_{k2} \\ & & \vdots & \\ 1 & w_{1n} & \cdots & w_{kn} \end{pmatrix}$$

is the error-contaminated matrix of covariates.

1.2.2 Attenuation effect of measurement error in the linear model

Assume a simple linear model with classical measurement error:

$$Y_i = \beta_0 + \beta X_i + \epsilon_i$$

$$W_i = X_i + U_i$$

where $i = 1, ..., n$, ϵ_i is independent of X_i and is normally distributed with mean 0 and variance σ^2. The naive estimator, defined as the least squares estimator of β

ignoring measurement error, is

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{D}_W^T \boldsymbol{D}_W)^{-1} \boldsymbol{D}_W^T \boldsymbol{Y}.$$

The slope parameter consistently estimates

$$\begin{aligned} \beta_* &= \frac{\mathrm{Cov}(W, Y)}{\mathrm{Var}(W)} \\ &= \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X + U)} \\ &= \frac{\sigma_X^2}{\sigma_U^2 + \sigma_X^2} \beta. \end{aligned} \tag{1.19}$$

Note that $|\beta_*| < |\beta|$, in other words, the naive estimator is always biased toward 0 by a ratio of $\lambda = \frac{\sigma_X^2}{\sigma_U^2 + \sigma_X^2}$. This ratio is called the attenuation factor. For fixed σ_X^2, larger amounts of measurement error, quantified through σ_U^2, result in stronger attenuation.

Figure 1.1 is a visualization of the attenuation effect. We regressed sepal length(Y) on petal length(X) from Fisher's Iris flower setosa species data [2] with three levels of normal measurement error(U) added to petal length. The sample variance of X is $\sigma_X^2 = 0.03$. The three levels of measurement errors are $\sigma_U^2 \in \{0, 0.03, 0.3\}$. Simple linear regressions are performed 1000 times, each time with a different random set of measurement error. The mean slope coefficients are 0.54,0.27 and 0.05, respectively. The slope in average is attenuated by roughly 50% and 11%, which agree with (1.19).

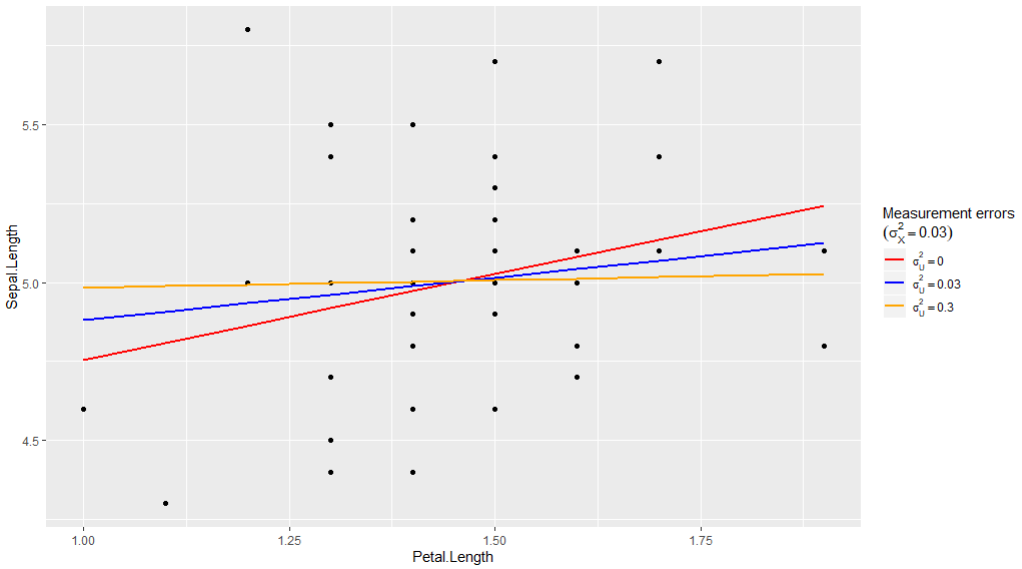

Figure 1.1: Measurement error bending regression lines toward 0 slope.

1.2.3 Attenuation effect of measurement error in GLMs

It is difficult to make general statements on the effect of measurement error on parameter estimation in non-linear regression models. Overall, for differential measurement error, there is a smoothing effect in that $E[Y \mid W] = E[E[Y \mid X] \mid W]$, understanding that expectation is a smoothing operation. Then measurement error smooths the regression between the true predictor X and the outcome Y. To gain additional insight into the effect of measurement error in non-linear regression models, this section contains a simulation study looking at measurement error in the logistic model.

Consider a logistic regression model with one predictor, where Y is a binary outcome with expectation $E[Y \mid X] = F(\beta_0 + \beta X)$, F is the logistic distribution function, that is, $F(\eta) = (1 + e^{-\eta})^{-1}$. X is normally distributed with mean μ_x and variance

σ_X^2. The measurement error U is additive, non-differential and normally distributed with mean 0 and variance σ_U^2. Let Φ and ϕ denote the standard normal c.d.f. and p.d.f.

Lemma 1.2.1. *For normally distributed measurement error, it was shown in [5] that:*

$$E[Y \mid W] \approx F\left(\frac{\beta_0 + \beta E[X \mid W]}{\sqrt{1 + \frac{\sigma^2_{X|W}}{1.7^2}\beta^2}}\right). \tag{1.20}$$

Then naive estimator of the simple logistic regression model with classical measurement error is biased toward 0, approximately by a factor of $\left(1 + \frac{\sigma_X^2(1-\lambda)}{1.7^2}\beta^2\right)^{-\frac{1}{2}}$*, where* $\lambda = \frac{\sigma_X^2}{\sigma_U^2+\sigma_X^2}$.

We finish this subchapter with a simulation study verifying (1.20). Figure 1.2 shows the distribution of regression coefficient estimators calculated with and without measurement error over 1000 datasets, each with 1000 observations. Data were generated with the following parameters.

1. $\beta_0 = 0.00596$.
2. $\beta = \log 2$.
3. $X \sim N(0, 1)$.
4. Measurement error variance $\sigma_U^2 = 0.5$.

The parameter configuration is such that there is a modest odds ratio (given by $e^\beta = 2$), and the choice of β_0 results in $E[Y] = 0.5$.

Two estimators of β were computed in the simulations. $\hat{\beta}_X$ denotes the estimator computed when there is no measurement error, i.e. using the true covariate values X.

$\hat{\beta}_W$ uses the mis-measured covariate values W. As can be seen in Figure 1.2, ignoring measurement error causes severe bias. The red vertical dashed line in Figure 1.2 marks the true coefficient $\beta = \log 2 = 0.693$. The sample means of $\hat{\beta}_X$ and $\hat{\beta}_W$ are 0.692 and 0.443, respectively. Plugging in $\sigma_x^2 = 1$ and $\sigma_U^2 = 0.5$ yields $\lambda = \frac{2}{3}$ and $\sigma_{X|W}^2 = \frac{1}{3}$. The (logistic) attenuation factor in Equation (1.20) is approximately 0.649. The expected value of the naive estimator would then be approximately $\log 2 * 0.649 = 0.450$, which is close the observed mean value 0.443. The difference is a result of the approximation $F(1.7x) \approx \Phi(x)$ (see [11] and Figure 1.3), sampling error, and the finite sample size. Note that $\hat{\beta}_W$ has smaller variance than $\hat{\beta}_X$ (0.003 v.s. 0.005).

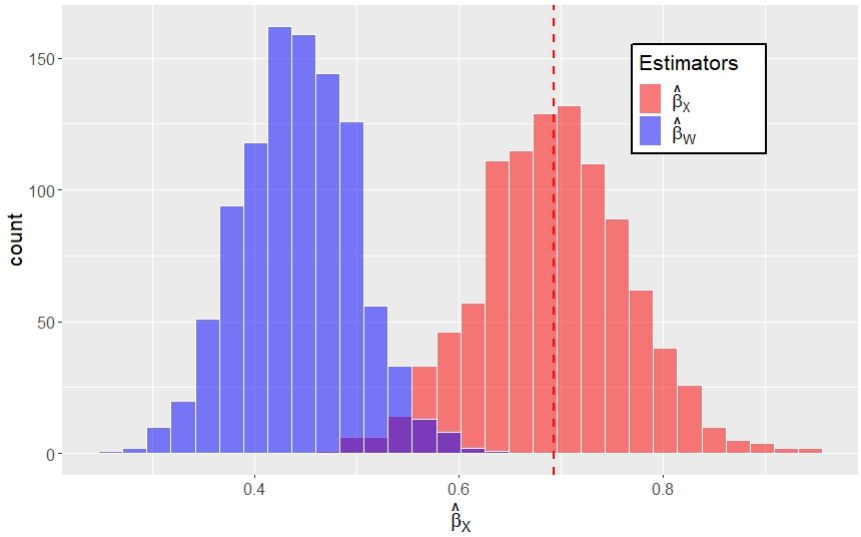

Figure 1.2: The attenuation effect on the slope estimators for logistic regression.

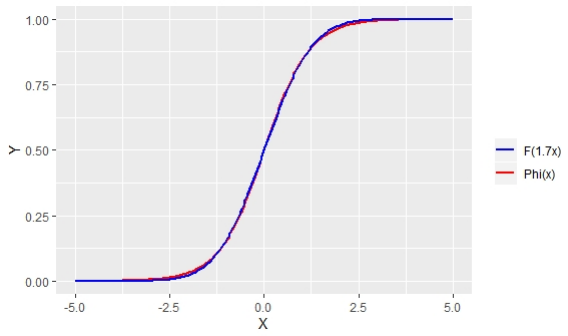

Figure 1.3: The similarity between F(1.7x) and standard normal c.d.f $\Phi(x)$.

1.2.4 Correcting measurement error induced bias using instrumental variables

The are many approaches to correcting for bias in regression parameter estimators induced by measurement error. The choice of method used is often dictated by the type of additional information that is available. As seen above, for the linear regression model, knowledge of measurement error variance, or have the ability to consistently estimate it, will yield a consistent estimator for the slope coefficient. In this case the correction is done by dividing the naive estimator by the attenuation factor.

Instrumental variables are additional data that correlate with the true value of the mis-measured explanatory variable, and are independent of both the measurement error and the random error of the model. The correlation structure makes it possible to obtain consistent estimators without direct knowledge on the measurement error variance. The method of instrumental variables has been used in econometrics to address endogeneity, which includes measurement error, simultaneity and omitted

variables as its major sources. As mentioned in the previous chapter, endogeneity refers to the situation when the explanatory variable is correlated with the error term. It is known to bias parameter estimation if left untended.

The basic method of correcting with IV is called two-stage least squares (2SLS). This method was published along with the discovery of IV [1] and later became a standard practice in econometrics [25]. The following provides a brief synopsis of the method of 2SLS for a linear regression model. We'll then explain why it can not be applied to non-linear regression models to obtain consistent parameter estimates, and one possible modification to remedy the deficiency.

Definition 1.2.1. *A measurement T is an instrument if*

1. *T is correlated with X.*
2. *T is independent of measurement error U.*
3. *T is independent of error $Y - E[Y|X]$.*

Chap 6.2 of Carroll, el at. [22] provide intuition on how IV estimation works in a measurement error model of the form

$$Y = f(X) + \epsilon$$
$$W = X + U.$$

T being independent of both measurement and systematic error implies

$$\frac{\partial W}{\partial T} = \frac{\partial X}{\partial T}$$
$$\frac{\partial Y}{\partial T} = \frac{\partial f(X)}{\partial T}.$$

Using the chain rule, we have

$$\begin{aligned}\frac{\partial f(X)}{\partial X} &= \frac{\partial f(X)}{\partial T}\frac{\partial T}{\partial X} \\ &= \frac{\partial Y}{\partial T}\Big/\frac{\partial W}{\partial T}\,.\end{aligned} \tag{1.21}$$

Equation (1.21) implies that we can understand how $f(X)$ changes with respect to X if we know how Y and W change with T. In the context of a simple linear regression, $f(X) = \beta_0 + \beta X$, and (1.21) implies that β can be estimated with two linear regressions: Y on T and W on T. To see this, let $\hat{\beta}_{Y|T}$ and $\hat{\beta}_{W|T}$ be the slope estimators from the two regressions, respectively. Then $\hat{\beta}_{Y|T}/\hat{\beta}_{W|T}$ consistently estimates β.

1.2.5 Instrumental variable method for GLM correction

Unfortunately the 2SLS method does not immediately extend to GLMs. To see this, we first make a comparison between the expectation of GLM estimating equations with and without measurement error. In what follows we assume there are not additional covariates measured without error. It is straightforward to extend the results to the case where additional covariates are present. In absence of measurement error,

$$\begin{aligned}E\left[\frac{\partial l}{\partial\theta}\frac{\partial\theta}{\partial\mu}\frac{\partial\mu}{\partial\eta}\frac{\partial\eta}{\partial\boldsymbol{\beta}}\middle| X\right] &= E\left[\frac{\partial l}{\partial\theta}\middle| X\right]\frac{\partial\theta}{\partial\mu}\frac{\partial\mu}{\partial\eta}\frac{\partial\eta}{\partial\boldsymbol{\beta}} \\ &\propto E[Y-\mu|X] \\ &= 0.\end{aligned} \tag{1.22}$$

In the presence of measurement error,

$$E\left[\left(\frac{\partial l}{\partial\theta}\frac{\partial\theta}{\partial\mu}\frac{\partial\mu}{\partial\eta}\frac{\partial\eta}{\partial\boldsymbol{\beta}}\right)_W\middle|X\right]=E\left\{[Y-\mu_W]\left(\frac{\partial\theta}{\partial\mu}\frac{\partial\mu}{\partial\eta}\frac{\partial\eta}{\partial\boldsymbol{\beta}}\right)_W\middle|X\right\}. \tag{1.23}$$

Unfortunately, the expectation on the RHS of (1.23) is not zero, that is $E[Y-\mu_W|X]\neq 0$ because $E[\mu_W|X]=E[F(\eta_W)|X]\neq\mu(\eta_X)$ for non-linear F. Furthermore, W is correlated with Y through X. Hence the last three components in (1.23) can't be factored out of the expectation.

A general method for correcting for covariate measurement error in non-linear regression models was proposed in [16]. The method was developed for uncorrelated outcomes, and therefore does not apply to data modelled through generalized estimating equations. The approach is described here, and modifications of the method for application in the GEE setting are developed in the next section.

Suppose that in the absence of measurement error, $E[Y\mid\boldsymbol{X}]=\mu(\boldsymbol{X^T\beta})\equiv\mu_{\boldsymbol{X}}$ and $\boldsymbol{\beta}$ is estimated via the estimating function

$$\psi(Y,W,\boldsymbol{\beta})=(Y-\mu_X)g(X,\boldsymbol{\beta}).$$

Often

$$g(X,\boldsymbol{\beta})=\frac{\partial}{\partial\boldsymbol{\beta}}\mu_X/V(X;\boldsymbol{\beta})=[\mu'_X/V(X;\boldsymbol{\beta})]\,\boldsymbol{X}$$

where $V(X;\boldsymbol{\beta})$ models the variance of Y conditional on $\boldsymbol{X}$ and $\mu'_{\boldsymbol{X}}=\frac{d}{dx}\mu(x)$. Let $\mu_{\boldsymbol{W}}\equiv\mu(\boldsymbol{W}^T\boldsymbol{\beta})$, $\mu_{\boldsymbol{T}}\equiv\mu(E[\boldsymbol{X}\mid\boldsymbol{T}]^T\boldsymbol{\beta})$ and define $\mu'_{\boldsymbol{W}}$ and $\mu'_{\boldsymbol{T}}$ similarly. Define

$$A=\frac{Y-\mu_{\boldsymbol{W}}}{\sqrt{\mu'_{\boldsymbol{W}}}}.$$

In [16], it was shown that

$$E[A \mid \boldsymbol{X}, \boldsymbol{T}] = o(\sigma_U^2)$$

where recall σ_U^2 is the measurement error variance. The implication is that A can be used to define an instrumental variable estimating function that will reduce measurement error induced bias. Define an IV estimating function as

$$\psi_{IV}(Y, \boldsymbol{W}, \boldsymbol{T} \mid \boldsymbol{\beta}) = \sqrt{\frac{\mu_T'}{\mu_W'}}(Y - \mu_{\boldsymbol{W}}) \frac{\mu_T'}{V(E[\boldsymbol{X} \mid \boldsymbol{T}]; \boldsymbol{\beta})} E[\boldsymbol{X} \mid \boldsymbol{T}]. \tag{1.24}$$

Then

$$\begin{aligned} E[\psi_{IV}(Y, \boldsymbol{W}, \boldsymbol{T} \mid \boldsymbol{\beta}) \mid \boldsymbol{X}, \boldsymbol{T}] &= E\left[\frac{(Y - \mu_W)}{\sqrt{\mu_W'}} \mid \boldsymbol{X}, \boldsymbol{T}\right] \frac{\mu_T'}{V(E[\boldsymbol{X} \mid \boldsymbol{T}]; \boldsymbol{\beta})} E[\boldsymbol{X} \mid \boldsymbol{T}] \mu_T' \\ &= E\left[A \mid \boldsymbol{X}, \boldsymbol{T}\right] \frac{\mu_T'}{V(E[\boldsymbol{X} \mid \boldsymbol{T}]; \boldsymbol{\beta})} E[\boldsymbol{X} \mid \boldsymbol{T}] \mu_T' \\ &= o(\sigma_U^2). \end{aligned}$$

The approach works because the IV estimating function factors as a product of A (a function of Y and W only), and a function of T. The conditional expectation then also factors, allowing us to leverage the fact that $E\left[A \mid \boldsymbol{X}, \boldsymbol{T}\right] = o(\sigma_U^2)$.

It was also shown in [16] that exactly consistent estimators result for a large class of mean models. The form of these models is defined in the following lemma.

Lemma 1.2.2. *Suppose that the distribution of U is symmetric around zero, the moment generating function for U exists, and*

$$\mu_{\boldsymbol{X}} = E[Y \mid X] = \frac{a_1 + a_2 e^{a_5 \eta_{\boldsymbol{X}}}}{a_3 + a_4 e^{a_5 \eta_{\boldsymbol{X}}}} \tag{1.25}$$

for scalar constants $a_1, \ldots, a_5$ *and* $\eta_{\boldsymbol{X}} = \boldsymbol{X}^T\boldsymbol{\beta}$. *Then*

$$E\left[\frac{Y - \mu_{\boldsymbol{W}}}{\sqrt{\mu'_{\boldsymbol{W}}}} \mid \boldsymbol{X}, \boldsymbol{W}\right] = 0.$$

The logistic regression model is of the form (1.25), as explicated in the following example.

Example (Logistic regression): For the logistic regression model, $\mu_{\boldsymbol{X}} = E[Y \mid \boldsymbol{X}] = F(\boldsymbol{X}^T\boldsymbol{\beta}) = 1/(1 + \exp\{-\boldsymbol{X}^T\boldsymbol{\beta}\})$. This is of the form (1.25) where $a_1 = a_2 = a_4 = 1$, $a_5 = -1$ and $a_2 = 0$. It is easy to show that

$$\mu'_{\boldsymbol{X}} = F(\boldsymbol{X}^T\boldsymbol{\beta})(1 - F(\boldsymbol{X}^T\boldsymbol{\beta})).$$

The instrumental variable estimating function is then

$$\psi_{IV}(Y, \boldsymbol{W}, \boldsymbol{T} \mid \boldsymbol{\beta}) = \sqrt{\frac{\mu'_{\boldsymbol{T}}}{\mu'_{\boldsymbol{W}}}}(Y - F(\boldsymbol{W}^T\boldsymbol{\beta}))\left(E[\boldsymbol{X} \mid \boldsymbol{T}].\right)$$

For the logistic model, $E[\psi_{IV}(Y, \boldsymbol{W}, \boldsymbol{T} \mid \boldsymbol{\beta}) \mid \boldsymbol{X}, \boldsymbol{T}] = 0$, and fully consistent estimators result.

1.2.6 Standardized residual preserves covariance structure

We seek to extend the above approach to IV estimation in GLIMs to GEEs.

Let

$$\boldsymbol{A_i} = \begin{pmatrix} A_{i1} \\ \vdots \\ A_{in_i} \end{pmatrix}, \qquad \boldsymbol{C_i} = \begin{pmatrix} \sqrt{\mu'_{\boldsymbol{T}_{i1}}} A_{i1} \\ \vdots \\ \sqrt{\mu'_{\boldsymbol{T}_{n1}}} A_{in_i} \end{pmatrix}$$

represent standardized difference vectors for the ith cluster and where $\mu'_{\boldsymbol{T}_{i1}} = \mu'(E[\boldsymbol{X} \mid \boldsymbol{T}_{i1}]^T \boldsymbol{\beta})$. For GLIMs, only the mean of the standardized difference A_{ij} was important. For data modelled with GEEs, the variance and covariance of $\boldsymbol{A_i}$ are required to properly model the correlation structure.

A naive proposal for an IV estimating function would be to use

$$\psi_{i,IV,\text{naive}} = \boldsymbol{D}_{\boldsymbol{i},\boldsymbol{T}}^{\boldsymbol{T}} \boldsymbol{C_i} \tag{1.26}$$

as the estimating function for the ith cluster where

$$\boldsymbol{D}_{\boldsymbol{i},\boldsymbol{T}}^{\boldsymbol{T}} = \begin{pmatrix} E[\boldsymbol{X}_1^T \mid \boldsymbol{T}_1] \\ \vdots \\ E[\boldsymbol{X}_{n_i}^T \mid \boldsymbol{T}_{n_i}] \end{pmatrix}.$$

It is straightforward to show that for mean models of the form given in (1.25), $E[\psi_{i,IV,\text{naive}} \mid \boldsymbol{X}_i, \boldsymbol{T}_i] = 0$, implying that consistent estimators of the regression parameters are obtained. However, the proposal is naive because the variance covariance matrix of $\boldsymbol{C_i}$ is not modelled, and therefore inefficient estimators result.

Ideally, the variance matrix for $\boldsymbol{C_i}$ would be of a similar form to that of the difference vector $\boldsymbol{S_i}$ used in the absence of measurement error. Unfortunately, the reality is more complex for most models. As detailed in the following two lemmas,

the covariance structure is preserved, but the variance of C_{ij} is not unless special conditions are met. Fortunately, the conditions are met for a very important model.

First some preliminaries. Suppose that $\mu_{\boldsymbol{X}} = E[Y|\boldsymbol{X}]$ is of the form given by (1.25). Let $k = 1/\sqrt{a_5(a_2a_3 - a_1a_4)}$. Note that

$$\begin{aligned}\mu'_{\boldsymbol{W}} &= \frac{e^{a_5\eta_{\boldsymbol{W}}}}{k^2(a_3 + a_4e^{a_5\eta_{\boldsymbol{W}}})^2},\\ \frac{\mu_{\boldsymbol{W}}}{\sqrt{\mu'_{\boldsymbol{W}}}} &= k(a_1 + a_2e^{a_5\eta_{\boldsymbol{W}}})e^{-a_5\eta_{\boldsymbol{W}}/2}\end{aligned}$$

where recall $\eta_{\boldsymbol{W}} = \boldsymbol{W}^T\boldsymbol{\beta}$. If the moment generating function for U, denoted $m_U(t)$, exists and is an even function (implying U has a symmetric distribution), then it is not difficult to show that

$$\begin{aligned}E\left[\frac{\mu_{\boldsymbol{W}}}{\sqrt{\mu'_{\boldsymbol{W}}}} \mid \boldsymbol{X}, \boldsymbol{T}\right] &= m_U(a_5\beta_1/2)k(a_1 + a_2e^{a_5\eta_{\boldsymbol{X}}})e^{-a_5\eta_{\boldsymbol{X}}/2}\\ &= m_U(a_5\beta_1/2)\frac{\mu_{\boldsymbol{X}}}{\sqrt{\mu'_{\boldsymbol{X}}}}\end{aligned} \tag{1.27}$$

and

$$\begin{aligned}E\left[\frac{1}{\sqrt{\mu'_{\boldsymbol{W}}}} \mid \boldsymbol{X}, \boldsymbol{T}\right] &= m_U(a_5\beta_1/2)k(a_3 + a_4e^{a_5\eta_{\boldsymbol{X}}})e^{-a_5\eta_{\boldsymbol{X}}/2}\\ &= m_U(a_5\beta_1/2)\frac{1}{\sqrt{\mu'_{\boldsymbol{X}}}}.\end{aligned} \tag{1.28}$$

The following lemma says that the covariance structure of the standardized differences A_{ij} is the same as the covariance structure in the absence of measurement error. This result is key for developing an IV approach to GEE.

Lemma 1.2.3. *Suppose that $E[Y|\boldsymbol{X}]$ is of the form given by* (1.25)*. Assume classical*

measurement error where the moment generating function for U exists and is an even function. Then,

$$\begin{aligned} Cov(A_1, A_2 \mid \boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{T}_1, \boldsymbol{T}_2) &= Cov\left(\frac{Y_1 - \mu_{\boldsymbol{W}_1}}{\sqrt{\mu'_{\boldsymbol{W}_1}}}, \frac{Y_2 - \mu_{\boldsymbol{W}_2}}{\sqrt{\mu'_{\boldsymbol{W}_2}}} \mid \boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{T}_1, \boldsymbol{T}_2\right) \\ &= m_U^2(a_5\beta_1/2)\, Cov\left(\frac{Y_1 - \mu_{\boldsymbol{X}_1}}{\sqrt{\mu'_{\boldsymbol{X}_1}}}, \frac{Y_2 - \mu_{\boldsymbol{X}_2}}{\sqrt{\mu'_{\boldsymbol{X}_2}}} \mid \boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{T}_1, \boldsymbol{T}_2\right). \end{aligned} \tag{1.29}$$

Proof.

$$\begin{aligned} \mathrm{Cov}(A_1, A_2 \mid \boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{T}_1, \boldsymbol{T}_2) &= E\left[\left(\frac{Y_1 - \mu_{\boldsymbol{W}_1}}{\sqrt{\mu'_{\boldsymbol{W}_1}}}\right)\left(\frac{Y_2 - \mu_{\boldsymbol{W}_2}}{\sqrt{\mu'_{\boldsymbol{W}_2}}}\right) \mid \boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{T}_1, \boldsymbol{T}_2\right] \\ &= E\left[\left(\frac{Y_1}{\sqrt{\mu'_{\boldsymbol{W}_1}}}\right)\left(\frac{Y_2 - \mu_{\boldsymbol{W}_2}}{\sqrt{\mu'_{\boldsymbol{W}_2}}}\right) \mid \boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{T}_1, \boldsymbol{T}_2\right] \\ &= E\left[\left(\frac{Y_1 Y_2}{\sqrt{\mu'_{\boldsymbol{W}_1}}\sqrt{\mu'_{\boldsymbol{W}_2}}}\right) - \left(\frac{Y_1 \mu_{\boldsymbol{W}_2}}{\sqrt{\mu'_{\boldsymbol{W}_1}}\sqrt{\mu'_{\boldsymbol{W}_2}}}\right) \mid \boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{T}_1, \boldsymbol{T}_2\right] \\ &= m_U^2(a_5\beta_1/2) E\left[\left(\frac{Y_1 Y_2}{\sqrt{\mu'_{\boldsymbol{X}_1}}\sqrt{\mu'_{\boldsymbol{X}_2}}}\right) - \left(\frac{Y_1 \mu_{\boldsymbol{X}_2}}{\sqrt{\mu'_{\boldsymbol{X}_1}}\sqrt{\mu'_{\boldsymbol{X}_2}}}\right) \mid \boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{T}_1, \boldsymbol{T}_2\right] \\ &= m_U^2(a_5\beta_1/2) \mathrm{Cov}\left(\frac{Y_1 - \mu_{\boldsymbol{X}_1}}{\sqrt{\mu'_{\boldsymbol{X}_1}}}, \frac{Y_2 - \mu_{\boldsymbol{X}_2}}{\sqrt{\mu'_{\boldsymbol{X}_2}}} \mid \boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{T}_1, \boldsymbol{T}_2\right). \end{aligned} \tag{1.30}$$

The second equality follows because $E[A_2 \mid \boldsymbol{X}, \boldsymbol{T}] = 0$. The penultimate equality follows from (1.27) and (1.28), and because of the mutual independence of Y_1, Y_2, U_1 and U_2.

□

The following lemma derives an expression for the variance of the standardized differences. Unlike the prior lemma, the variance structure of the A_{ij} does not match that of standardized differences in the absence of measurement error unless a specific condition is met.

Lemma 1.2.4. *Suppose that $E[Y|\boldsymbol{X}]$ is of the form given by* (1.25)*. Assume classical measurement error where the moment generating function for U exists and is an even function. Then,*

$$\mathrm{Var}(A \mid \boldsymbol{X}, \boldsymbol{T}) = m_U(a_5\beta_1)\mathrm{Var}\left(\frac{Y - \mu_{\boldsymbol{X}}}{\sqrt{\mu'_{\boldsymbol{X}}}} \mid \boldsymbol{X}, \boldsymbol{T}\right) - \xi$$

where

$$\xi = 2m_U(a_5\beta_1)k^2\Big(a_3a_4E[Y^2 \mid \boldsymbol{X}, \boldsymbol{T}] - (a_1a_4 + a_2a_3)E[Y \mid \boldsymbol{X}, \boldsymbol{T}] + a_1a_2\Big).$$

Proof.

$$\begin{aligned}\mathrm{Var}\left(\frac{Y - \mu_{\boldsymbol{W}}}{\sqrt{\mu'_{\boldsymbol{W}}}} \mid \boldsymbol{X}, \boldsymbol{T}\right) &= E\left[\left(\frac{Y - \mu_{\boldsymbol{W}}}{\sqrt{\mu'_{\boldsymbol{W}}}}\right)^2 \mid \boldsymbol{X}, \boldsymbol{T}\right] \\ &= E\left[\frac{Y^2}{\mu'_{\boldsymbol{W}}} - \frac{2Y\mu_{\boldsymbol{W}}}{\mu'_{\boldsymbol{W}}} + \frac{\mu^2_{\boldsymbol{W}}}{\mu'_{\boldsymbol{W}}} \mid \boldsymbol{X}, \boldsymbol{T}\right].\end{aligned}$$

We consider each term in turn. For the first term,

$$\begin{aligned}E\left[\frac{Y^2}{\mu'_{\boldsymbol{W}}} \mid \boldsymbol{X}, \boldsymbol{T}\right] &= k^2E\left[Y^2(a_3^2e^{-a_5\eta_{\boldsymbol{W}}} + 2a_3a_4 + a_4^2e^{a_5\eta_{\boldsymbol{W}}}) \mid \boldsymbol{X}, \boldsymbol{T}\right] \\ &= m_U(a_5\beta_1)E\left[\frac{Y^2}{\mu'_{\boldsymbol{X}}} \mid \boldsymbol{X}, \boldsymbol{T}\right] - 2m_U(a_5\beta_1)k^2a_3a_4E\left[Y^2 \mid \boldsymbol{X}, \boldsymbol{T}\right].\end{aligned}$$

Next,

$$\begin{aligned}E\left[\frac{2Y\mu_{\boldsymbol{W}}}{\mu'_{\boldsymbol{W}}} \mid \boldsymbol{X}, \boldsymbol{T}\right] &= 2k^2E\left[Y(a_1a_3e^{-a_5\eta_{\boldsymbol{W}}} + a_1a_4 + a_2a_3 + a_2a_4e^{a_5\eta_{\boldsymbol{W}}}) \mid \boldsymbol{X}, \boldsymbol{T}\right] \\ &= m_U(a_5\beta_1)E\left[\frac{2Y\mu_{\boldsymbol{X}}}{\mu'_{\boldsymbol{X}}} \mid \boldsymbol{X}, \boldsymbol{T}\right] - 2m_U(a_5\beta_1)k^2(a_2a_4 + a_2a_3)E\left[Y \mid \boldsymbol{X}, \boldsymbol{T}\right].\end{aligned}$$

Finally,

$$E\left[\frac{\mu_{\boldsymbol{W}}^2}{\mu'_{\boldsymbol{W}}} \mid \boldsymbol{X}, \boldsymbol{T}\right] = k^2 E\left[(a_1^2 e^{-a_5\eta_{\boldsymbol{W}}} + 2a_1a_2 + a_2^2 e^{a_5\eta_{\boldsymbol{W}}}) \mid \boldsymbol{X}, \boldsymbol{T}\right]$$
$$= m_U(a_5\beta_1) E\left[\frac{\mu_{\boldsymbol{X}}^2}{\mu'_{\boldsymbol{X}}} \mid \boldsymbol{X}, \boldsymbol{T}\right] - 2m_U(a_5\beta_1)k^2 a_1 a_2.$$

Combining, we have

$$\mathrm{Var}\left(\frac{Y - \mu_{\boldsymbol{W}}}{\sqrt{\mu'_{\boldsymbol{W}}}} \mid \boldsymbol{X}, \boldsymbol{T}\right) = m_U(a_5\beta_1)\mathrm{Var}\left(\frac{Y - \mu_{\boldsymbol{X}}}{\sqrt{\mu'_{\boldsymbol{X}}}} \mid \boldsymbol{X}, \boldsymbol{T}\right)$$
$$- 2m_U(a_5\beta_1)k^2\Bigg(a_3a_4 E\left[Y^2 \mid \boldsymbol{X}, \boldsymbol{T}\right]$$
$$- (a_1a_4 + a_2a_3)E\left[Y \mid \boldsymbol{X}, \boldsymbol{T}\right] + a_1a_2\Bigg).$$

□

We are particularly interested in models where $\xi = 0$, as in that case the variance structure in the absence of measurement carries over to A_{ij}. Unfortunately, $\xi = 0$ is not likely satisfied by many models. However, it is satisfied by a very important model, namely the logistic regression model.

Example (Logistic model): As noted above, the logistic model for a binary outcome Y postulates that $\mu_{\boldsymbol{X}} = E[Y|\boldsymbol{X}] = (1 + e^{-\eta_{\boldsymbol{X}}})^{-1}$, which is of the form (1.25) where $a_1 = a_3 = a_4 = 1$, $a_2 = 0$, $a_5 = -1$. Plugging these values into the expression for ξ and noting that for a binary outcome $E[Y^2 \mid \boldsymbol{X}] = E[Y \mid \boldsymbol{X}]$, it follows that $\xi = 0$ and therefore for the logistic model Lemma 1.2.4 implies

$$\mathrm{Var}(A \mid \boldsymbol{X}, \boldsymbol{T}) = m_U(\beta_1)\mathrm{Var}\left(\frac{Y - \mu_{\boldsymbol{X}}}{\sqrt{\mu'_{\boldsymbol{X}}}} \mid \boldsymbol{X}, \boldsymbol{T}\right) = m_U(\beta_1).$$

The last equality follows because the logistic model is in the exponential family with canonical link and therefore $\text{Var}\left(\frac{Y-\mu_{\boldsymbol{X}}}{\sqrt{\mu'_{\boldsymbol{X}}}} \mid \boldsymbol{X}\right) = 1$.

Next, consider the Poisson model with canonical link. The hope would be that models with canonical link would preserve the variance structure, because in this case, the variance of the standardized difference in the absence of measurement error is unity. Unfortunately, that is not the case.

Example (Poisson model): Poisson regression with canonical link is such that $E[Y|\boldsymbol{X}] = e^{\eta \boldsymbol{X}}$, which is of the form (1.25) where $a_1 = a_4 = 0$, $a_2 = a_3 = 1$, $a_5 = 1$. Then $k = 1$ and from Lemma 1.2.4 we conclude

$$\begin{aligned}\text{Var}(A \mid \boldsymbol{X}, \boldsymbol{T}) &= m_U(\beta_1)\text{Var}\left(\frac{Y-\mu_{\boldsymbol{X}}}{\sqrt{\mu'_{\boldsymbol{X}}}} \mid \boldsymbol{X}, \boldsymbol{T}\right) + 2m_U(\beta_1)e^{\eta \boldsymbol{X}} \\ &= m_U(\beta_1)(1 + 2e^{\eta \boldsymbol{X}}).\end{aligned}$$

The last equality follows because the Poisson model with canonical link is such that $\text{Var}\left(\frac{Y-\mu_{\boldsymbol{X}}}{\sqrt{\mu'_{\boldsymbol{X}}}} \mid \boldsymbol{X}\right) = 1$.

1.3 Instrumental variables approach to GEE with measurement error–Logistic Regression

In this section we leverage the results of the previous section to define an IV estimating function for clustered binary outcomes modelled with the logistic regression mean function.

1.3.1 Model statement and notation

Let Y_{ij} represent the jth binary outcome in cluster i where $i = 1, ..., n$ and $j = 1, ..., n_i$. The logistic model stipulates that $\mu_{\boldsymbol{X}_{ij}} = E[Y_{ij} \mid \boldsymbol{X}_{ij}] = F(\eta_{\boldsymbol{X}_{ij}})$ where $\eta_{\boldsymbol{X}_{ij}} = \boldsymbol{X}_{ij}^T\boldsymbol{\beta}$, and recall that $\boldsymbol{X_{ij}} = (1, X_{ij1}, X_{ij2}, \ldots, X_{ijp})^T$.

An exchangeable correlation structure with constant correlation is assumed: $\text{Corr}(Y_{i1}, Y_{i2}) = \alpha$. That is,

$$\text{Corr}(Y_{ij}, Y_{ik}) = \begin{cases} 1 & \text{if } j = k \\ \alpha & \text{else.} \end{cases}$$

In the absence of measurement error the GEE estimating function for $\boldsymbol{\beta}$ for cluster i is given by

$$\psi_\beta(Y_i, \boldsymbol{X}_i) = \boldsymbol{D_i^t \Delta_i V_i^{-1} S_i} \tag{1.31}$$

where

$$\boldsymbol{S_i} = \begin{pmatrix} Y_{i1} - \mu_{\boldsymbol{X}i1} \\ \vdots \\ Y_{in_i} - \mu_{\boldsymbol{X}in_i} \end{pmatrix}, \qquad \boldsymbol{D_i} = \begin{pmatrix} \boldsymbol{X}_{i1}^T \\ \vdots \\ \boldsymbol{X}_{in_i}^T \end{pmatrix},$$

$$\boldsymbol{\Delta_i} = \text{diag}\{\mu'_{\boldsymbol{X}i1}, \ldots, \mu'_{\boldsymbol{X}in_i}\},$$

$$\boldsymbol{V_i} = \boldsymbol{\Delta_i^{1/2} R_i}(\alpha)\boldsymbol{\Delta_i^{1/2}},$$

and

$$\boldsymbol{R_i}(\alpha) = \begin{pmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & & \\ \vdots & & \ddots & \vdots \\ \alpha & \cdots & & 1 \end{pmatrix}.$$

We assume, without loss of generality, that the first covariate in $\boldsymbol{X}$ is measured with additive, non-differential measurement error. All other covariates are measured without error. To that end, define $\boldsymbol{W} = (1, W, X_2, \ldots, X_p)$ where

$$\begin{aligned} W &= X_1 + U, \\ E[U|\boldsymbol{X}] &= 0, \\ E[Y|\boldsymbol{X}] &= E[Y \mid \boldsymbol{X}, \boldsymbol{W}]. \end{aligned}$$

The subscripts for cluster and observation within cluster have been suppressed for clarity. It is also assumed the moment generating function of the measurement error exists and is an even function. For each observation in each cluster, there exists an instrumental variable $\boldsymbol{T_{ij}}$ for the mis-measured X_{ij1}. The dimension of the instrumental variable may be greater than one, i.e. there may several instruments for the

covariate that is measured with error. Define

$$
\begin{aligned}
\boldsymbol{W_{ij}} &= (1, W_{ij}, X_{ij2}, \ldots, X_{ijp})^T \\
\mu_{\boldsymbol{W_{ij}}} &= F(\boldsymbol{W_{ij}^T}\boldsymbol{\beta}), \\
\mu'_{\boldsymbol{W_{ij}}} &= F(\boldsymbol{W_{ij}^T}\boldsymbol{\beta})(1 - F(\boldsymbol{W_{ij}^T}\boldsymbol{\beta})) \\
E[\boldsymbol{X_{ij}} \mid \boldsymbol{T_{ij}}] &= (1, E[X_{ij1} \mid \boldsymbol{T_{ij}}], X_{ij2}, \ldots, X_{ijp})^T, \\
\mu_{\boldsymbol{T_{ij}}} &= F(E[\boldsymbol{X_{ij}} \mid \boldsymbol{T_{ij}}]^T\boldsymbol{\beta}), \\
\mu'_{\boldsymbol{T_{ij}}} &= F(E[\boldsymbol{X_{ij}} \mid \boldsymbol{T_{ij}}]^T\boldsymbol{\beta})(1 - F(E[\boldsymbol{X_{ij}} \mid \boldsymbol{T_{ij}}]^T\boldsymbol{\beta})), \\
\boldsymbol{A_i} &= \begin{pmatrix} A_{i1} \\ \vdots \\ A_{in_i} \end{pmatrix} = \begin{pmatrix} \frac{Y_{i1} - \mu_{\boldsymbol{W}i1}}{\sqrt{\mu'_{\boldsymbol{W}i1}}} \\ \vdots \\ \frac{Y_{i1} - \mu_{\boldsymbol{W}in_i}}{\sqrt{\mu'_{\boldsymbol{W}in_i}}} \end{pmatrix}, \qquad \boldsymbol{C_i} = \begin{pmatrix} \sqrt{\mu'_{\boldsymbol{T}_{i1}}} A_{i1} \\ \vdots \\ \sqrt{\mu'_{\boldsymbol{T}_{in1}}} A_{in_i} \end{pmatrix} \\
\boldsymbol{D_{Ti}} &= \begin{pmatrix} E[\boldsymbol{X_{i1}^T} \mid \boldsymbol{T_{i1}}] \\ \vdots \\ E[\boldsymbol{X_{in_i}^T} \mid \boldsymbol{T_{in_i}}] \end{pmatrix}, \qquad \boldsymbol{\Delta_{Ti}} = \mathrm{diag}(\mu'_{\boldsymbol{T_{i1}}}, \ldots, \mu'_{\boldsymbol{T_{in_i}}}).
\end{aligned}
$$

The following corollary to Lemmas 1.2.3 and 1.2.4 is key to constructing an IV GEE estimating function for logistic regression. The corollary gives the form of the variance/covariance matrix for $\boldsymbol{C_i}$.

Corollary 1.3.0.1. *For the logistic regression measurement error model described above,*

$$Var(\boldsymbol{C_i}) = m_U(\beta_1)\boldsymbol{\Delta_{Ti}^{1/2}}\boldsymbol{R_i}(\alpha_C)\boldsymbol{\Delta_{Ti}^{1/2}}$$

where

$$\alpha_C = \frac{m_U^2(\beta_1/2)}{m_U(\beta_1)}\alpha.$$

Proof. It follows directly from Lemmas 1.2.3 and 1.2.4 that

$$\text{Corr}(C_{ij}, C_{ik}) = \frac{m_U^2(\beta_1/2)}{m_U(\beta_1)}\alpha,$$

and

$$\text{Var}(C_{ij}) = m_U(\beta_1)\mu'_{\boldsymbol{T_{ij}}}.$$

The corollary follows immediately. □

We are now in a position to define an IV estimating function for the logistic regression measurement error model with clustered data. The estimating function is analogous to (1.31), where $\boldsymbol{C_i}$ replaces $\boldsymbol{S_i}$. Define

$$\boldsymbol{V_{Ci}} = \boldsymbol{\Delta}_{\boldsymbol{Ti}}^{1/2}\boldsymbol{R_i}(\alpha_C)\boldsymbol{\Delta}_{\boldsymbol{Ti}}^{1/2}.$$

Note that it is not necessary to include the $m_U(\beta_1)$ term in the definition of $\boldsymbol{V_{Ci}}$ as it will factor out of the estimating equation.

The IV estimating function for $\boldsymbol{\beta}$ for cluster i is

$$\boldsymbol{\psi}_{\beta,IV}(\boldsymbol{Y_i}, \boldsymbol{W}_i, \boldsymbol{T}_i) = \boldsymbol{D}_{\boldsymbol{Ti}}^{\boldsymbol{T}}\boldsymbol{\Delta_{Ti}}\boldsymbol{V}_{\boldsymbol{Ci}}^{-\mathbf{1}}\boldsymbol{C_i}.$$

The estimating function for α_C for cluster i is

$$\psi_{\alpha_C}(\boldsymbol{Y_i}, \boldsymbol{W}_i, \boldsymbol{T}_i) = \sum_{j<k}(A_{ij}A_{ik} - \gamma\alpha_C),$$

where $\gamma = m_U(\beta_1)$. The estimating function for γ is

$$\psi_\gamma(\boldsymbol{Y_i}, \boldsymbol{W}_i, \boldsymbol{T}_i) = \sum_{j=1}^{n_i}(A_{ij}^2 - \gamma).$$

The combined estimating function is then

$$\boldsymbol{\psi}_{IV}(\boldsymbol{Y_i}, \boldsymbol{W}_i, \boldsymbol{T}_i) = \begin{bmatrix} \boldsymbol{\psi}_{\boldsymbol{\beta},IV}(\boldsymbol{Y_i}, \boldsymbol{W}_i, \boldsymbol{T}_i) \\ \psi_{\alpha_C}(\boldsymbol{Y_i}, \boldsymbol{W}_i, \boldsymbol{T}_i) \\ \psi_\gamma(\boldsymbol{Y_i}, \boldsymbol{W}_i, \boldsymbol{T}_i) \end{bmatrix}. \tag{1.32}$$

The IV estimator $\hat{\varphi} = (\hat{\boldsymbol{\beta}}^T, \hat{\alpha}_C, \hat{\gamma})^T$ solves

$$\sum_{i=1}^{n} \boldsymbol{\psi}_{IV}(\boldsymbol{Y_i}, \boldsymbol{W}_i, \boldsymbol{T}_i) = \mathbf{0}. \tag{1.33}$$

Note that $\hat{\alpha}_C$ and $\hat{\gamma}$ can be solved for explicitly as a function of $\hat{\boldsymbol{\beta}}$. From the definitions of the estimating equations, it follows that

$$\hat{\gamma} = \frac{1}{\sum_{i=1}^n n_i} \sum_{i=1}^{n} \sum_{j=1}^{n_i} A_{ij}^2,$$
$$\hat{\alpha}_C = \frac{1}{\hat{\gamma} \sum_{i=1}^n \binom{n_i}{2}} \sum_{i=1}^{n} \sum_{j<k} A_{ij} A_{ik}.$$

The following proposition states the the GEEIV estimating function is unbiased, implying that the estimating function will yield consistent estimators of the regression coefficients.

Proposition. *Under the assumed measurement error model,*

$$E\left[\boldsymbol{\psi}_{IV}(\boldsymbol{Y_i}, \boldsymbol{W}_i, \boldsymbol{T}_i)\right] = \mathbf{0}.$$

Proof. Note that $\boldsymbol{D_{Ti}^T}$, $\boldsymbol{\Delta_{Ti}}$ and $\boldsymbol{V_{Ci}}$ depend on the data only through $(\boldsymbol{X}_i, \boldsymbol{T}_i)$. We show that the expectation conditional on $(\boldsymbol{X}_i, \boldsymbol{T}_i)$ is zero, from which it follows that the unconditional expectation is zero. First consider the estimating function for $\boldsymbol{\beta}$:

$$E\left[\boldsymbol{\psi}_{\beta,IV}(\boldsymbol{Y_i}, \boldsymbol{W}_i, \boldsymbol{T}_i) \mid \boldsymbol{X}_i, \boldsymbol{T}_i\right] = \boldsymbol{D_{Ti}^T \Delta_{Ti} V_{Ci}^{-1}} E\left[\boldsymbol{C_i} \mid \boldsymbol{X}_i, \boldsymbol{T}_i\right] = \mathbf{0}$$

where the last equality follows from Lemma 1.2.2.

Next, note that Lemma 1.2.3 gives $E[A_{ij}A_{ik} \mid \boldsymbol{X}_i] = \mathrm{Cov}[A_{ij}A_{ik} \mid \boldsymbol{X}_i] = m_U(\beta_1/2)^2\alpha$, and from Lemma 1.2.4 it follows that $E[A_{ij}^2 \mid \boldsymbol{X}, \boldsymbol{T}] = m_U(\beta_1)$. Then it easily follows that

$$E\left[\psi_\gamma(\boldsymbol{Y_i}, \boldsymbol{W}_i, \boldsymbol{T}_i) \mid \boldsymbol{X}_i, \boldsymbol{T}_i\right] = 0,$$

and from the definition of α_C, it follows readily that

$$E\left[\psi_{\alpha_C}(\boldsymbol{Y_i}, \boldsymbol{W}_i, \boldsymbol{T}_i) \mid \boldsymbol{X}_i, \boldsymbol{T}_i\right] = 0.$$

□

1.3.2 Variance estimation

The IV estimator defined in (1.33) is an M-estimator, and therefore is asymptotically normal with variance matrix of the form given in Lemma 1.1.2. Details of the

construction of the variance matrix are given here.

An estimator of the variance of $\hat{\varphi} = (\hat{\boldsymbol{\beta}}^T, \hat{\alpha}_C, \hat{\gamma})^T$ is constructed as follows. Let

$$\mathcal{A} = E\left[\frac{\partial}{\partial\varphi}\boldsymbol{\psi}_{IV}(\boldsymbol{Y_i}, \boldsymbol{W}_i, \boldsymbol{T}_i)\right].$$

and

$$\mathcal{B} = E\left[\boldsymbol{\psi}_{IV}(\boldsymbol{Y_i}, \boldsymbol{W}_i, \boldsymbol{T}_i)\boldsymbol{\psi}_{IV}^{\boldsymbol{T}}(\boldsymbol{Y_i}, \boldsymbol{W}_i, \boldsymbol{T}_i)\right].$$

The sandwich estimator for the variance of the asymptotic distribution of $\hat{\varphi}$ has the form

$$\boldsymbol{\mathcal{V}} = \frac{1}{n}\boldsymbol{\mathcal{A}}^{-1}\boldsymbol{\mathcal{B}}\boldsymbol{\mathcal{A}}^{-T}.$$

In practice, empirical versions of $\mathcal{A}$ and $\mathcal{B}$ are used, as detailed in the discussion following Lemma 1.1.2.

In the absence of measurement error, the GEE estimating function for the logistic model satisfies $E\boldsymbol{\psi}\boldsymbol{\psi}^{\boldsymbol{t}} = -E\boldsymbol{\psi}'$, and the variance estimator can then be simplified. It can be shown that

$$E\left[\boldsymbol{\psi}_{IV}\boldsymbol{\psi}_{IV}^{\boldsymbol{T}}\right] \neq -E\left[\frac{\partial}{\partial\varphi}\boldsymbol{\psi}_{IV}\right],$$

and therefore the full sandwich estimate must be used for the IV estimation procedure.

In practice, $E[X_{ij1} \mid \boldsymbol{T_{ij}}]$ must be estimated. Note that $E[X_{ij1} \mid \boldsymbol{T_{ij}}] = E[W_{ij1} \mid \boldsymbol{T_{ij}}]$. Therefore, we can regress W on $\boldsymbol{T}$ to estimate $E[X_{ij1} \mid \boldsymbol{T_{ij}}]$. Typically W and $\boldsymbol{T}$ are strongly correlated and a linear regression will suffice. Estimation of the regression parameters in $E[X_{ij1} \mid \boldsymbol{T_{ij}}]$ will not affect the asymptotic variance of $\hat{\boldsymbol{\beta}}$.

1.4 Simulation study

Here we describe the results of a large simulation study designed to assess the performance of the logistic regression IV estimator in finite samples. The GEE logistic model and IV estimator are defined in Section 1.3.

Questions addressed in the simulations are as follows:

1. How well does GEEIV perform compared to naive estimator in terms of bias and variance?
2. How accurate is the asymptotic variance estimator?
3. What is the coverage rate for confidence intervals?
4. How robust is the GEEIV estimator to the assumption of symmetric measurement error?
5. How does sample size (number of clusters and cluster size), magnitude of the odds ratio, and amount of measurement error effect the above?

This section is organized as follows:

- Section 1.4.1 includes the details of data generation, including choice of parameter values, and generating algorithm, and restrictions on generating correlated binary data.
- Section 1.4.2 describes the simulation results. A comparison of effectiveness between GEEIV and naive estimators in terms of bias and variability is provided. The performance of the GEEIV asymptotic variance and confidence interval coverage are also included.

1.4.1 Data generation

There are a number of parameter choices required for the generation of correlated binary data. Parameter choices dictate the overall rate of outcome in the population (marginal positive rate), the odds ratio for a one standard deviation increase in the predictor measured with error, the amount of measurement error, and the strength of correlation between outcomes within a cluster (intra-cluster correlation). Additionally, the number of clusters and the number of observations within a cluster must be chosen. If only two values are considered for each condition, there are $2^6 = 64$ conditions. As detailed below, more than two values were examined for important parameters (odds ratio, cluster size,....). For each combination of conditions, 1000 datasets were generated and analyzed.

1.4.1.1 Marginal positive rate, odds ratio and intra-cluster correlation

The intra-cluster correlation coefficient (α), marginal positive rate (EY) and odds ratio (OR) are meaningful parameters, and together dictate the difficulty of data generation. Detailed discussion will be given later in this section.

The marginal positive rate varied between 0.5 and 0.1, representing both common and somewhat rare outcomes.

The odds ratio for a one standard deviation increase in the predictor varied between 2, 3 and 4, representing moderate to strong association between the outcome and predictor.

Two values for the within cluster correlation coefficient were considered: $\alpha = 0.25$ and $\alpha = 0.60$. Larger α implies a stronger correlation between cluster members,

and fewer effective observations as a consequence. A dataset would have an effective number of observations equal to the number of clusters if $\alpha = 1$, and an effective number of observations equal to the total number of observations if $\alpha = 0$.

Datasets with fewer effective observations yield parameter estimators with larger variability. A higher correlation coefficient also increases the penalty of an incorrect independence assumption. Liang and Zeger [6] considered $\alpha = 0.25$ too weak for GEE to yield significant improvement over GLM. Fitzmaurice [12] demonstrated that the efficiency, defined as the ratio of GEE v.s. GLM estimators asymptotic variance, declines with increasing correlation, and the decline is most notable when the correlation is greater than 0.4. We'll show that intra-cluster correlation has a larger impact on the efficiency loss when measurement error is present and that a correlation coefficient of 0.25 is no longer insignificant. We also pick a second $\alpha = 0.6$ that was considered significant in the original GEE paper. It is worth noting that a larger α greatly increases the difficulty of data generation and decreases the variability of the independent variable. This effect will be discussed in detail later in the section.

1.4.1.2 Cluster size and number

Large values for α significantly impact the ability to generate correlated binary outcomes, and generating correlated binary outcomes for large clusters becomes practically impossible (details are given in the appendix). For $\alpha = 0.25$, datasets were generated with 160 clusters of size 25, and 100 clusters of size 16 were generated for $\alpha = 0.6$.

We also considered the impact of cluster size and number of clusters by varying these quantities. We sampled generated datasets of the sizes indicated above to

construct datasets with a smaller number of clusters and/or cluster size. This was accomplished by considered the following fractions of the cluster size: $\frac{1}{2}$ and $\frac{1}{4}$, and the fractions of size $\frac{1}{2}$, $\frac{1}{4}$ and $\frac{1}{8}$ for the number of clusters. Including the full number of clusters and sizes, there are $3 \times 4 = 12$ combinations of cluster size and number of clusters in total.

1.4.1.3 Generating correlated binary clusters

Clustered binary data were generated through a modification of the algorithm given in Emrich and Piedemonte [10]. The algorithm was adapted to the regression context, and modifications were developed to increase the data generation speed.

A brief description of the algorithm is given here. For simplicity, we drop the subscript for cluster, that is, we consider generating observations within a cluster, denoted (X_i, Y_i), for $i = 1, ..., n$ where Y_i is binary random variable with individual expectation of p_i. Define $q_i = 1 - p_i$. Parentheses in subscripts are used to represent order. For example, $p_{(1)} = \min\{p_i\}$ and $p_{(n)} = \max\{p_i\}$.

1. Set values for EY, α, and the odds ratio. Note that $\beta_0 = F^{-1}(EY)$ and $\beta_1 = \log(OR)$.

2. Generate $p_i \sim N(EY, \sigma_Y^2)$ for $i = 1, ..., n$. Discard p_i's and repeat Step 1 if

$$\sqrt{p_{(1)}q_{(n)}/p_{(n)}q_{(1)}} < \alpha.$$

3. Generate $[Y_1, ..., Y_n]$ with $p_1, ..., p_n, \alpha$ using Emrich & Piedemonte's algorithm [10].

4. Define $X_i = (F^{-1}(Y_i) - \beta_0)/\beta_1$.

A detailed description about this procedure, including the choice of $\sigma^2_{Y'}$ and the reason for discarding generated expectations, is contained in Appendix 1.6.2.

1.4.1.4 Generating measurement error and instrumental variables

Measurement error was added after successful generation of $\boldsymbol{X}$ and $\boldsymbol{Y}$. The process is much simpler as measurement error are independent within a cluster. We varied both the distribution and the variance of measurement error to assess GEEIV performance.

Two types of measurement error distribution were considered: normal and standardized chi-squared. The normal measurement errors are symmetric about 0 in distribution. The standardized chi-squared is a chi-squared random variable with three degrees of freedom, shifted to have mean 0 and scaled to have the specified measurement error variance. The Chi-square measurement errors were included to assess the robustness of the method to the assumption of symmetric measurement errors. A chi-squared random variable on three degrees of freedom is quite skewed.

The data simulation algorithm is such that the variance of X is not fixed. Therefore, rather than fixing the measurement error variance, we specified two levels of attenuation factor: 0.33, 0.8. Larger attenuation corresponds to larger measurement error.

The instrumental variable $\boldsymbol{T}$ was defined as $T = X + \tilde{U}$ where $\tilde{U}$ has the same distribution as the measurement error, but with a fixed attenuation factor of 0.3. In this way, the correlation between $\boldsymbol{T}$ and $\boldsymbol{X}$ is around 0.8(0.5) for measurement error attenuation factor=0.8(0.3) across all (EY, OR, α) combination.

1.4.2 Simulation results

1.4.2.1 Rate of successful convergence

For each dataset, the naive and IV estimator were calculated by solving the nonlinear equations (1.14) and (1.32) with the Newton Raphson method. Newton Raphson is a gradient descent method for finding an approximation to the root $f(x) = 0$, where x can be either a scalar or vector. The starting value was set to be naive GLM estimator. The algorithm stops at x_0 when each component of $f(x_0)$ is less than a critical value, which in our case is 10^{-8}, or when the default maximum number of loops ($= 20$) is reached. If the estimating function was not less than the critical value before the maximum number of loops was reached, the resulting x is in general not close enough to the root and we say the algorithm did not converge.

In Appendix 1.6.3 we provide a few tables on Newton Raphson convergence rates with some parameter settings. In general a "balanced" dataset, that is, with a 0.5 marginal positive rate is more likely to result in convergence than those with 0.1 marginal positive rate. High Odds ratios and/or measurement error variance also negatively impact the convergence rate (this is common in measurement error simulation studies). Smaller cluster size and fewer clusters also negatively affect convergence rates. Perhaps this is somewhat counter intuitive, as larger cluster sizes and number of clusters might seem to make the estimating equations more complicated. However, having more and larger clusters results in less variation in the estimating function itself, rendering it more likely to have a solution.

The tables in Appendix 1.6.3 also include a comparison of success rate of the IV estimator and estimator with true covariate, that is, estimated without measurement

error. It seems that the two columns are quite similar, suggesting that measurement error is not the primary culprit for failed convergence. Running Newton's method with a different starting value on a failed dataset rarely results in a convergence. Increasing the maximum number of iteration also did not help.

When analyzing the performance of GEEIV, datasets where the algorithm did not converge are excluded.

1.4.2.2 Summary of bias elimination

Figure 1.4 shows the bias of the naive and IV estimators over a multitude of parameter configurations. A summary of the findings is as follows.

The GEEIV estimator effectively reduced measurement error induced bias across all parameter settings studied with perhaps the exception of when simultaneously the overall response rate was low, the OR is large, and the cluster size is small. None of the parameters alone seems to have a significant impact on the unbiasedness of GEEIV.

- $E[Y]$: The marginal positive rate alone has a relatively minor effect on both naive and GEEIV estimator in terms of bias, with $E[Y] = 0.1$ having slightly larger bias compared to $E[Y] = 0.5$.

- OR: the naive estimator is heavily affected by the odds ratio. The attenuation ratio, defined as 1-$\hat{\beta}_1/\beta_1$, is approximately 40%, 60%, 70% for OR 2, 3 and 4, respectively.

  GEEIV estimators performs well across all OR's, with a few exceptions.

- σ_U^2: Measurement error has a significant attenuation effect on the naive estima-

tor which gets stronger with lower attenuation factor.

On GEEIV the attenuation effect is almost negligible.

- Measurement error type: asymmetric measurement error has little effect on the bias of the GEEIV estimator for most settings.

For small datasets with $E[Y] = 0.1$ and large OR, GEEIV shows exception to the above summary which is worth their own discussion. GEEIV estimates noticeably strayed away from the true values when sample size drops below a certain point. Low attenuation factors and high correlation exacerbate the bias. Both upward and downward biases were observed. This can be a sign of the mean not being stable due to insufficient number of datasets.

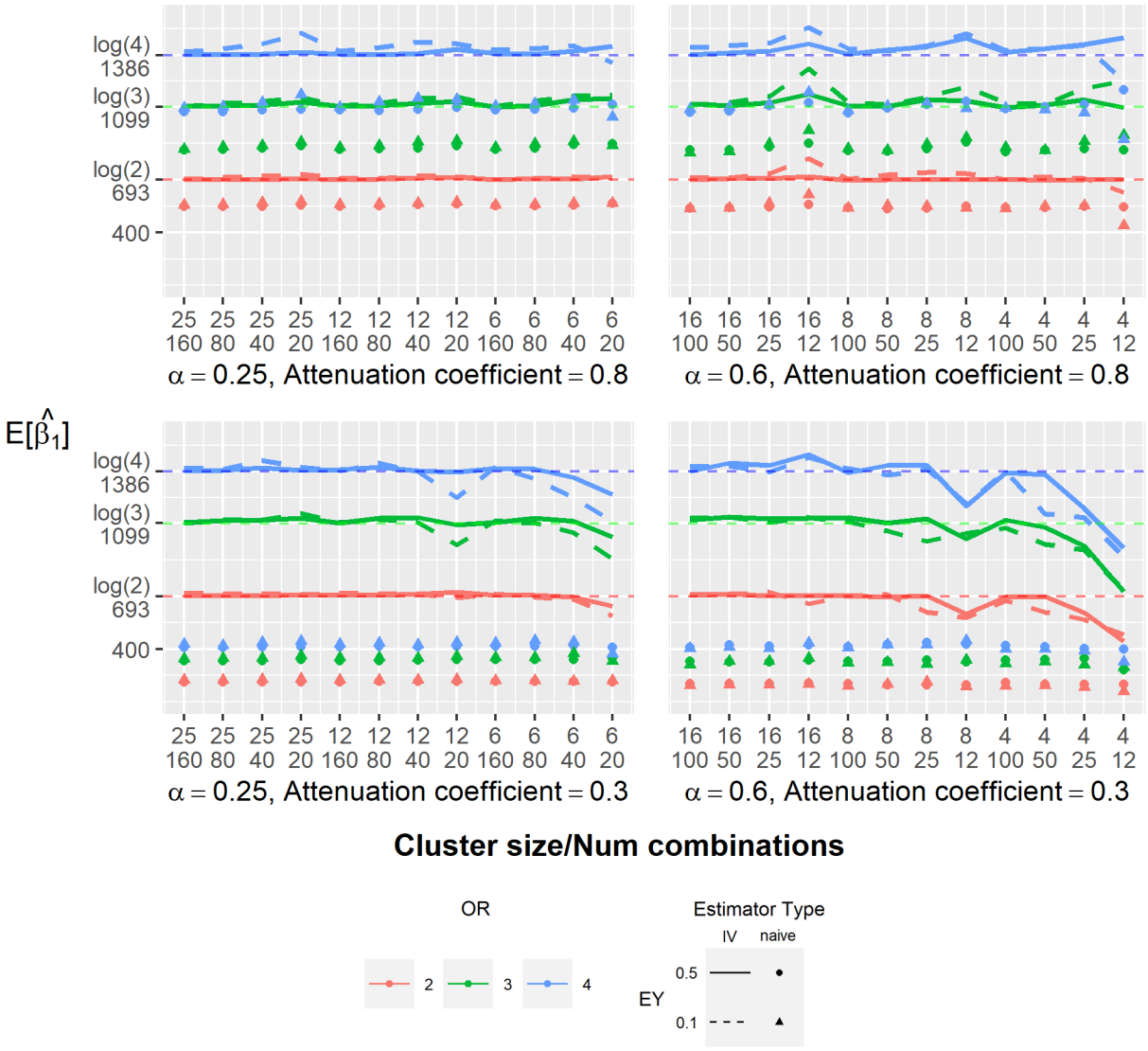

Figure 1.4: A comparison of estimator sample means. Normal Measurement error. Plotted values are multiplied by 1000.

1.4.2.3 Variability of $\hat{\beta}_1$

Here the variability of GEEIV is evaluated. Figure 1.5 illustrates that GEEIV has larger variance compared to the naive estimator. Increased variability is the price one pays for reduced bias–this bias/variance trade-off has been observed in essentially all other estimators proposed for reducing measurement error induced attenuation. An explanation of this phenomenon is that measurement error smooths the relation between Y and X, thereby induces bias but decreasing variability. It is also evident from the figure that the standard deviation of GEEIV increases with the OR and as the sample size decreases. Smaller population response rates (EY) result in increased variation.

A summary of the effects of parameter choices on estimator variability are as follows.

- $E[Y]$: Fixing all other factors, "unbalanced" datasets($E[Y] = 0.1$]) have approximately double standard deviation than "balanced" datasets.

- OR: Fixing all other factors, datasets with higher odds ratio have larger standard deviation. GEEIV is affected in a larger scale compared to the naive estimator. OR 2:3:4 has an approximate standard deviation ratio of $1 : 2 : 3$ for GEEIV and $8 : 9 : 10$ for naive estimator.

- σ_U^2: Increasing the measurement error variance had opposite effects on the variability of the GEEIV and naive estimators. Larger measurement error variance reduced the standard deviation for the naive estimator, while it is increased the standard deviation for GEEIV. This effect is expected. As mentioned above, measurement error smooths the relation between Y and X, and more measure-

ment error imparts more smoothing resulting in less variability. More measurement error also means more bias, and correcting for additional bias adds complexity, thereby increasing variability.

- Cluster number and size: The standard deviation of the estimator varied depending on the total number of observations, regardless of the cluster size/number combination. Figure 1.6 is a re-ordered version of Figure 1.5, with standard deviations sorted by total number of observations. It can be seen that standard deviations are similar when the total number of observations are the same. When the total number of observations is fixed, datasets with a small number of clusters seem to have a slight decrease in the standard deviation.

- Measurement error distribution: The effect of non-symmetric measurement error was minor, and appears to have an opposite effect on the GEEIV and naive estimator standard deviation. The difference is not significant expect for the most "extreme" settings ($E[Y] = 0.1$, $OR = 3, 4$).

It is counter-intuitive to see that cluster size has little effect on the standard deviation. In most cases, one would usually expect a negative correlation between standard deviation and number of independent observations. For example, it is well known that the variance of the sample mean in a dataset with k clusters of size n and correlation ρ is given by

$$\frac{\sigma^2}{nk}[1 + (n-1)\rho]$$

where σ^2 is the variance of a single observation [31]. The idea is that as cluster members are correlated, the number of effective independent observations in a cluster must be less than its size. Imagine in an extreme situation where ρ is 1, a dataset

with 1 cluster of size n would have a much larger standard deviation compared to a dataset with n clusters of size 1, as the former has only 1 effective independent observation and the latter has n.

However this rule does not seem to apply in our simulation, and not only for GEEIV. The naive estimator standard deviation, and even the estimator obtained from error-free regression(data not shown in Appendix) are also not affected. We do not know the nature of this phenomenon. It might be a feature of our correlated binary data generation method.

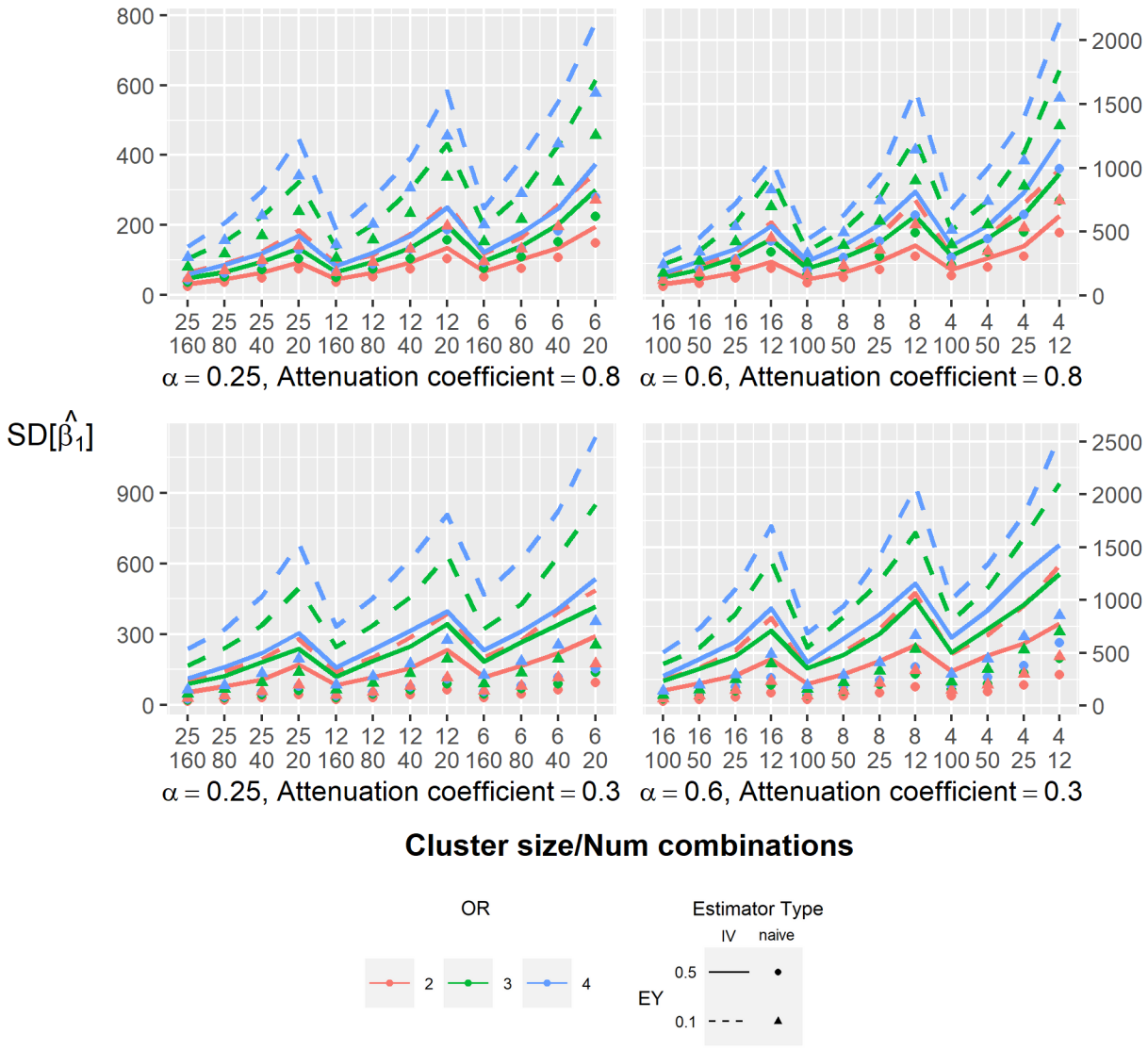

Figure 1.5: A comparison of estimator sample standard deviations ordered by cluster size. Normal Measurement error. Values are multiplied by 1000. Note the scale of the vertical axes differ.

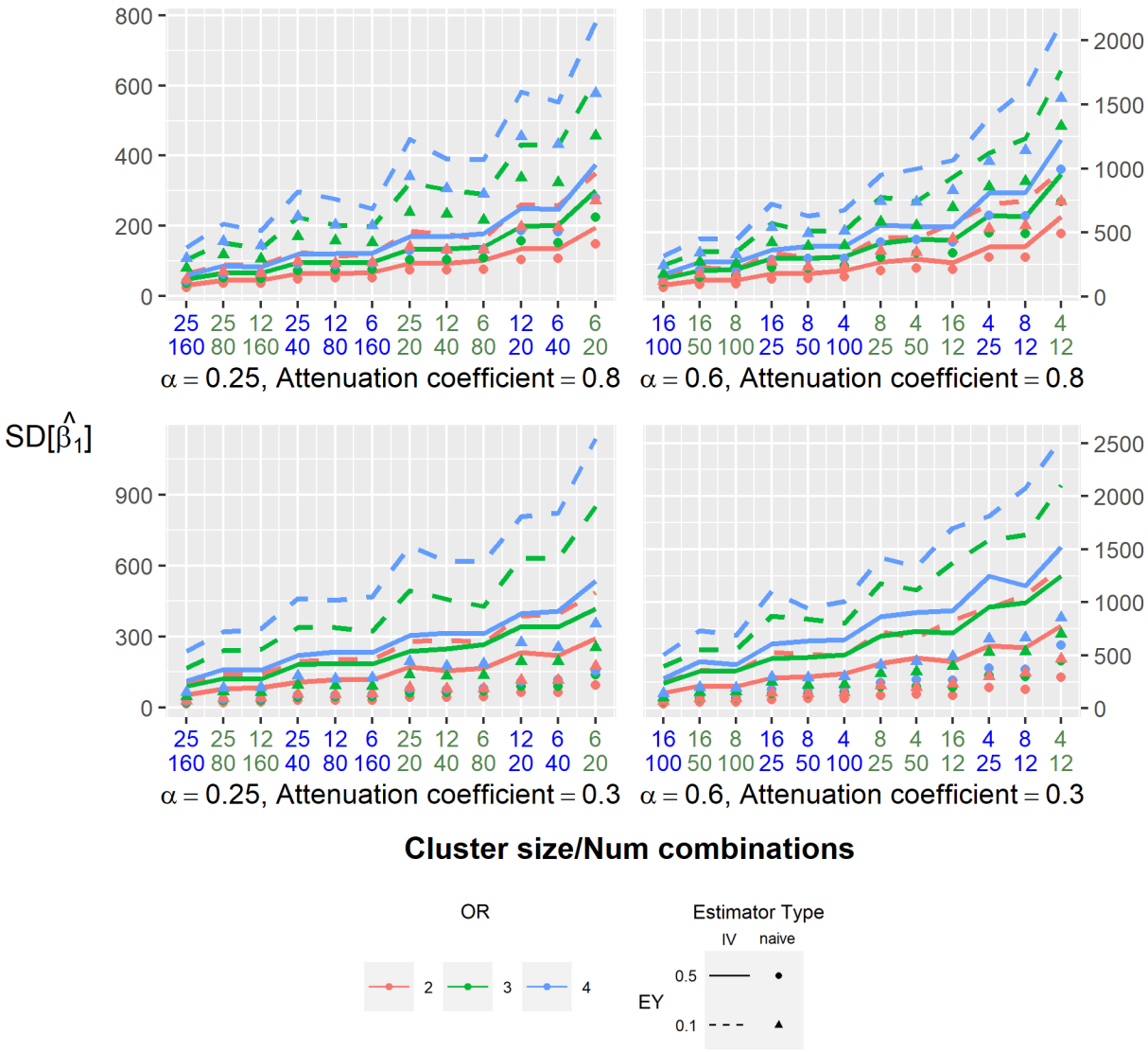

Figure 1.6: A comparison of estimator sample standard deviations ordered by total number of observations. Normal Measurement error. Values are multiplied by 1000.

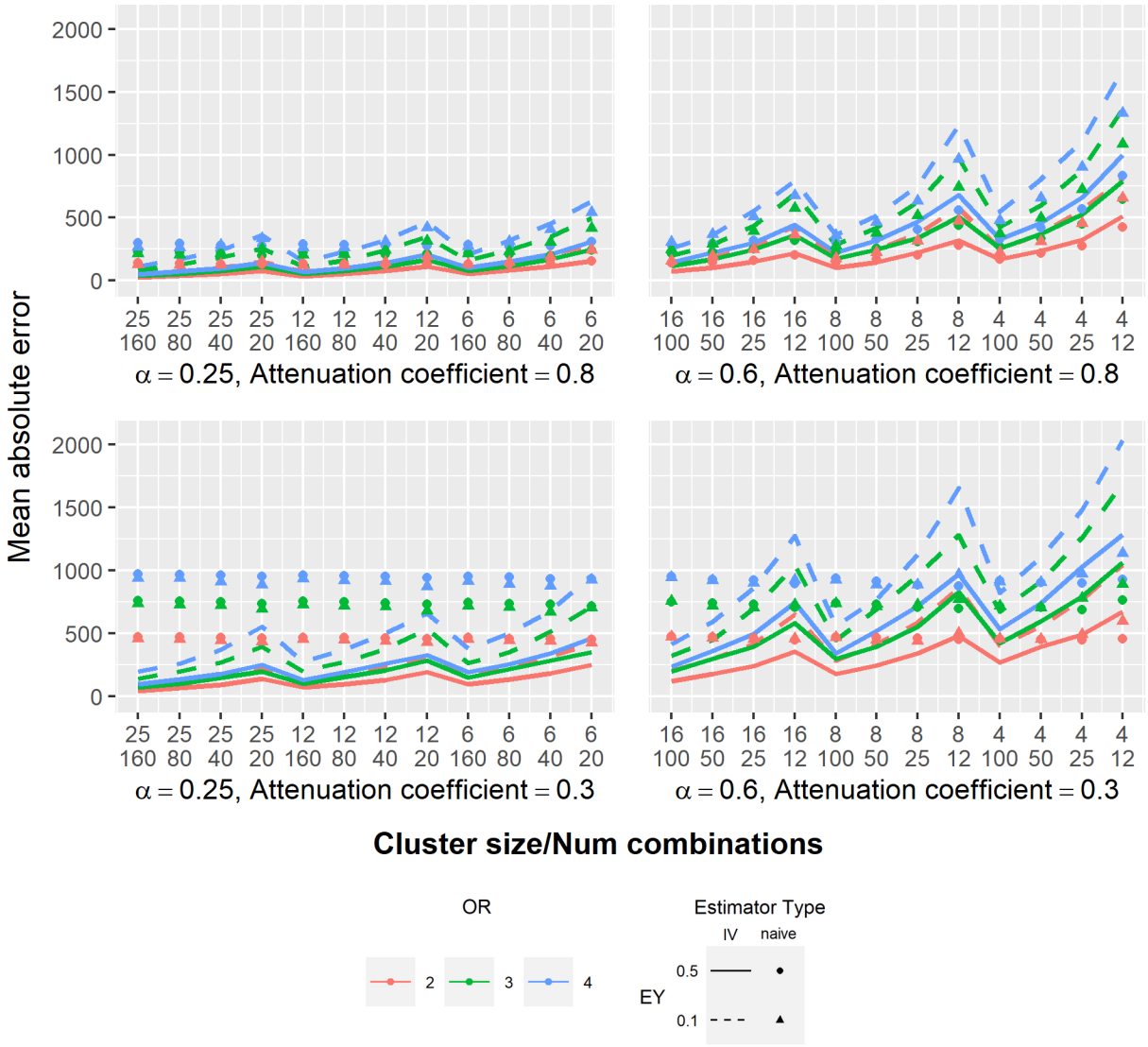

Figure 1.7: A comparison of estimator mean absolute error ordered by cluster size. Normal Measurement error. Values are multiplied by 1000.

1.4.2.4 Bias and variability of $\hat{\beta}_0$

The intercept parameter β_0 is of secondary interest, and typically intercept parameters are not greatly affected by covariate measurement error. This was the case in our simulations. Both naive and GEEIV estimators are nearly unbiased for all settings. GEEIV estimators have slightly larger standard deviation. Results are not shown for the intercept parameter.

1.5 Conclusions

In this chapter we developed unbiased estimating equations for clustered data with covariate measurement error contamination. The naive estimating equations contain two sources of bias: dependency between the residual and the covariates, and non-linearity of the link function. The first source of bias was eliminated by the introduction of instrumental variables, which by definition, are independent of measurement and systematic error. The second source of bias is removed by an appropriate standardization of the residuals. A new issue arises as the standardized residuals require a matching working correlation structure. For binary regression models, We proved that the standardized residuals preserve the original correlation structure, with correlation coefficients scaled by a constant factor. As a result, an additional estimating equation to estimate the scaled correlation coefficient is sufficient–no new modelling of the correlation structure is required.

A large simulation study was conducted to examine the effectiveness of the approach in the setting of correlated binary outcomes modelled with the logistic link. The simulations show GEE instrumental variable appraoch (GEEIV) yields an unbi-

ased estimator. Under somewhat extreme conditions, defined as a small population prevalence of outcome, a high odds ratio, and small sample size, the GEEIV estimator retains some bias, but nonetheless suffers much less bias compared to ignoring measurement error.

For future work, it is possible to extend the application to other members of the exponential family, Poisson and beta distribution, for example. The working correlation matrices will not be the same as in the absence of measurement error, and this will require additional modeling. Most importantly, the GEEIV method would still yield an asymptotically unbiased estimator as long as the standardization process is modified appropriately.

1.6 Appendix

1.6.1 Derivation of $\boldsymbol{\mathcal{A}}$

The asymptotic covariance matrix $\boldsymbol{\mathcal{V}}$ in Section 1.3.2 can be estimated computationally. In case readers are interested in analytical estimation, we provide some insight here.

The calculation of $\boldsymbol{\mathcal{B}}$ is a straightforward matrix multiplication and thus is omitted. We mainly discuss the upper left four components of $\boldsymbol{\mathcal{A}}$, which involves complex partial derivatives. The remaining components are related to the nuisance parameter γ. Their derivation may be considered in the future.

Recall the four components are:

$$\boldsymbol{\mathcal{V}}_{\boldsymbol{\beta},\alpha_C} = E\left(\begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{\psi}_{\boldsymbol{\beta},i} & \frac{\partial}{\partial \alpha_C}\boldsymbol{\psi}_{\boldsymbol{\beta},i} \\ \frac{\partial}{\partial \boldsymbol{\beta}}\psi_{\alpha_C,i} & \frac{\partial}{\partial \alpha_C}\psi_{\alpha_C,i} \end{bmatrix} \middle| \boldsymbol{X}\right).$$

The expectation of two derivatives on the right side $\frac{\partial}{\partial \alpha_C}\boldsymbol{\psi}_{\boldsymbol{\beta},i}$ and $\frac{\partial}{\partial \alpha_C}\psi_{\alpha_C,i}$ can be easily calculated and thus no estimation is needed:

$$\begin{aligned} E\left[\frac{\partial}{\partial \alpha_C}\boldsymbol{\psi}_{\boldsymbol{\beta},i}\middle|\boldsymbol{X}\right] &= E\left[\boldsymbol{D}_{Ti}^T\boldsymbol{\Delta}_{Ti}\left(\frac{\partial}{\partial \alpha_C}\boldsymbol{V}_{Ti}^{-1}\right)\boldsymbol{\Delta}_{Ti}^{\frac{1}{2}}\middle|\boldsymbol{X}\right]E[\boldsymbol{A_i}|\boldsymbol{X}] \\ &\quad + E\left[\boldsymbol{D}_{Ti}^T\boldsymbol{\Delta}_{Ti}\boldsymbol{V}_{Ti}^{-1}\boldsymbol{\Delta}_{Ti}^{\frac{1}{2}}\middle|\boldsymbol{X}\right]E\left[\frac{\partial}{\partial \alpha_C}\boldsymbol{A_i}\middle|\boldsymbol{X}\right] \\ &= \mathbf{0}; \\ E\left[\frac{\partial}{\partial \alpha_C}\psi_{\alpha_C,i}\middle|\boldsymbol{X}\right] &= \frac{\partial}{\partial \alpha_C}\sum_{j<k}(A_{ij}A_{ik}-\gamma\alpha_C) = -\binom{n_i}{2}\gamma. \end{aligned}$$

Expectation of the first component is $\mathbf{0}$ because both $E[\boldsymbol{A_i}|\boldsymbol{X}]$ and $\frac{\partial}{\partial \alpha_C}\boldsymbol{A_i}$ are 0 vectors.

The two components on the left can not be reduced to simple terms and need to be estimated with sample mean. Still, $\frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{\psi}_{\boldsymbol{\beta},i}$ can be simplified:

$$\begin{aligned} E\left[\frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{\psi}_{\boldsymbol{\beta},i}\middle|\boldsymbol{X}\right] &= E\left[\frac{\partial}{\partial \boldsymbol{\beta}}\left(\boldsymbol{D}_{Ti}^T\boldsymbol{\Delta}_{Ti}\boldsymbol{V}_{Ti}^{-1}\boldsymbol{\Delta}_{Ti}^{\frac{1}{2}}\right)|\boldsymbol{X}\right]E[\boldsymbol{A_i}|\boldsymbol{X}] + E[\boldsymbol{D}_{Ti}^T\boldsymbol{\Delta}_{Ti}\boldsymbol{V}_{Ti}^{-1}\boldsymbol{\Delta}_{Ti}^{\frac{1}{2}}\frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{A_i}|\boldsymbol{X}] \\ &= E\left[\boldsymbol{D}_{Ti}^T\boldsymbol{\Delta}_{Ti}\boldsymbol{V}_{Ti}^{-1}\boldsymbol{\Delta}_{Ti}^{\frac{1}{2}}\frac{\partial}{\partial \boldsymbol{\beta}}\boldsymbol{A_i}|\boldsymbol{X}\right], \end{aligned} \tag{1.34}$$

while $\frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{\psi}_{\alpha_C,i}$ remains nearly unchanged:

$$\begin{aligned} E\left[\frac{\partial}{\partial \boldsymbol{\beta}} \psi_{\alpha_C,i} | \boldsymbol{X}\right] &= E\left[\frac{\partial}{\partial \boldsymbol{\beta}} \sum_{j<k} (A_{ij}A_{ik} - \gamma\alpha_C) \middle| \boldsymbol{X}\right] \\ &= \sum_{j \neq k} E\left[A_{ij} \frac{\partial}{\partial \boldsymbol{\beta}} A_{ik} \middle| \boldsymbol{X}\right]. \end{aligned} \tag{1.35}$$

It remains to derive $\frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{A}_i$:

$$\begin{aligned} \frac{\partial}{\partial \boldsymbol{\beta}} \boldsymbol{A}_i &= \frac{\partial \boldsymbol{A}_i}{\partial \eta_{W,i}} \frac{\partial \eta_{W,i}}{\partial \boldsymbol{\beta}} \\ &= \frac{\partial \boldsymbol{A}_i}{\partial \eta_{W,i}} \boldsymbol{D}_{W,i}^T. \end{aligned}$$

$\frac{\partial}{\partial \eta_{W,i}} \boldsymbol{A}_i$ is a $n_i \times 1$ vector with the jth element being $\frac{\partial}{\partial \eta_{W,i}} A_{ij}$. For logistic regression:

$$\frac{\partial}{\partial \eta_{W,i}} A_{ij} = \frac{1}{2} Y_{ij} [e^{\frac{1}{2}\eta_{W,ij}} - e^{-\frac{1}{2}\eta_{W,ij}}] - \frac{1}{2} e^{\frac{1}{2}\eta_{W,ij}}.$$

1.6.2 Generating correlated binary clusters

1.6.2.1 Problem statement and existing methods

This section of the appendix serves as a description of technical details for generating correlated clusters. For simplicity the cluster subscript is omitted. Observations within a cluster are denoted as (X_i, Y_i) for $i = 1, ..., n$. Y_i is binary random variable with a randomly generated marginal expectation of p_i, such that $E[P_i] = EY$. $\text{Corr}(Y_i, Y_j) = \alpha, 1 \leq i < j \leq n$. The marginal expectation p_i is connected to the independent variable through inverse link function $p_i = F(\eta_i) = F(\beta_0 + \beta_1 X_i)$. In our simulation, F is the logistic function and therefore $\beta_1 = \log(OR)$.

There have been several proposed methods for generating correlated binary clusters. However, as far as we know, there has been no published work on generating correlated binary clusters in a regression setting. More specifically, existing methods do not satisfy our needs in that:

1. The focus of prior methods was on generating $Y_1, .., Y_n$ unconditionally on the X_i. The generation of $X_1, ..., X_n$ was not considered.

2. With one exception, all methods we encountered require the specification of $p_1, ..., p_n$, which are supposed to be randomly generated in our regression simulation. The exception allows random p_i but not a specified correlation matrix.

An exchangeable correlation structure is used in our simulation setting. Therefore, we choose to generate p_i's first, then generate Y_i with one of the existing methods.

1.6.2.2 Prentice's constraint and validity of expectations

In a correlated cluster, the correlation coefficient between Y_i and Y_j, denoted ρ_{ij}, does not range freely from -1 to 1. As shown in [8] ρ_{ij} is bounded by

$$[\max\{-(p_i p_j/q_i q_j)^{\frac{1}{2}}, -(q_i q_j/p_i p_j)^{\frac{1}{2}}\}, \min\{(p_i q_j/p_j q_i)^{\frac{1}{2}}, (p_j q_i/p_i q_j)^{\frac{1}{2}}\}]. \tag{1.36}$$

We adopt the author's name and call this Prentice's constraint. Violating it will result in negative probabilities in the joint probability mass function [8]. Prentice's constraint is why existing methods try to avoid joint specification of random p_i's and a correlation structure: it's likely to result in some improper probability mass functions. We call an expectation set $[p_1, ..., p_n]$ a "valid" set if each pair from $[p_1, ..., p_n]$ satisfies (1.36). In Section 1.6.2.4, we introduce an additional minor constraint.

For an exchangeable correlation matrix with a positive coefficient α, an expectation set $\{p_i\}$ is valid if and only if the lowest upper bound posed by all pairs (p_i, p_j) is greater than $\rho_{ij} = \alpha$, that is:

$$\min_{i,j}\{\min\{(p_i q_j / p_j q_i)^{\frac{1}{2}}, (p_j q_i / p_i q_j)^{\frac{1}{2}}\}\} > \alpha. \tag{1.37}$$

The lower bound in (1.36) plays no part in (1.37), as it is always negative and thus is automatically satisfied. (1.37) can be further simplified, as the lowest upper bound is posed by $(p_{(1)}, p_{(n)})$.

Lemma 1.6.1. $\min_{i,j}\{\min\{(p_i q_j / p_j q_i)^{\frac{1}{2}}, (p_j q_i / p_i q_j)^{\frac{1}{2}}\}\} = (p_{(1)} q_{(n)} / p_{(n)} q_{(1)})^{\frac{1}{2}}$.

Proof. Without loss of generality, we assume the expectation set is sorted: $p_1 \leq ... \leq p_n$.

For any pair (p_i, p_j) such that $0 < p_i < p_j < 1$,

$$\min\{(p_i q_j / p_j q_i)^{\frac{1}{2}}, (p_j q_i / p_i q_j)^{\frac{1}{2}}\} = (p_i q_j / p_j q_i)^{\frac{1}{2}} \tag{1.38}$$

Now consider $0 < p_i < p_j < p_k < 1$. Note that

$$\begin{aligned} &\min\{\min\{(p_i q_j / p_j q_i)^{\frac{1}{2}}, (p_j q_i / p_i q_j)^{\frac{1}{2}}\}, \min\{(p_i q_k / p_k q_i)^{\frac{1}{2}}, (p_k q_i / p_i q_k)^{\frac{1}{2}}\}\} \\ &= \min\{(p_i q_j / p_j q_i)^{\frac{1}{2}}, (p_i q_k / p_k q_i)^{\frac{1}{2}}\} \\ &= (p_i q_k / p_k q_i)^{\frac{1}{2}} \end{aligned} \tag{1.39}$$

Verification of the two identities above is achieved by simply calculating the ratio

of the two items. (1.38) and (1.39) together imply that for a fixed $p_i \in \{p_1, ..., p_{n-1}\}$,

$$\min_{j \in \{i+1,...,n\}} \{\min\{(p_i q_j / p_j q_i)^{\frac{1}{2}}, (p_j q_i / p_i q_j)^{\frac{1}{2}}\}\} = (p_i q_n / p_n q_i)^{\frac{1}{2}}$$

.

In the same way we can show that for a fixed $p_j \in \{p_2, ..., p_n\}$

$$\min_{i \in \{1,j-1\}} \{\min\{(p_i q_j / p_j q_i)^{\frac{1}{2}}, (p_j q_i / p_i q_j)^{\frac{1}{2}}\}\} = (p_1 q_j / p_j q_1)^{\frac{1}{2}}$$

.

Therefore, for any $1 \leq i < j \leq n$, we have:

$$\min\{(p_i q_j / p_j q_i)^{\frac{1}{2}}, (p_j q_i / p_i q_j)^{\frac{1}{2}}\} = (p_i q_j / p_j q_i)^{\frac{1}{2}} \geq (p_i q_n / p_n q_i)^{\frac{1}{2}} \geq (p_1 q_n / p_n q_1)^{\frac{1}{2}}.$$

Equivalently,

$$\begin{aligned} &\min_{1 \leq i \leq j \leq n} \{\min\{(p_i q_j / p_j q_i)^{\frac{1}{2}}, (p_j q_i / p_i q_j)^{\frac{1}{2}}\}\} \\ &= (p_1 q_n / p_n q_1)^{\frac{1}{2}} \\ &= \min\{(p_1 q_n / p_n q_i)^{\frac{1}{2}}, (p_n q_1 / p_1 q_n)^{\frac{1}{2}}\}. \end{aligned}$$

□

Figure 1.8 serves as a visual explanation to Lemma 1.6.1. The curves show how the upper bound $\min\{(p_i q_j / p_j q_i)^{\frac{1}{2}}, (p_j q_i / p_i q_j)^{\frac{1}{2}}\}$ changes with respect to p_i and p_j. Note that for any $p_j > \max p_i$, the upper bound is lower for a smaller p_i.

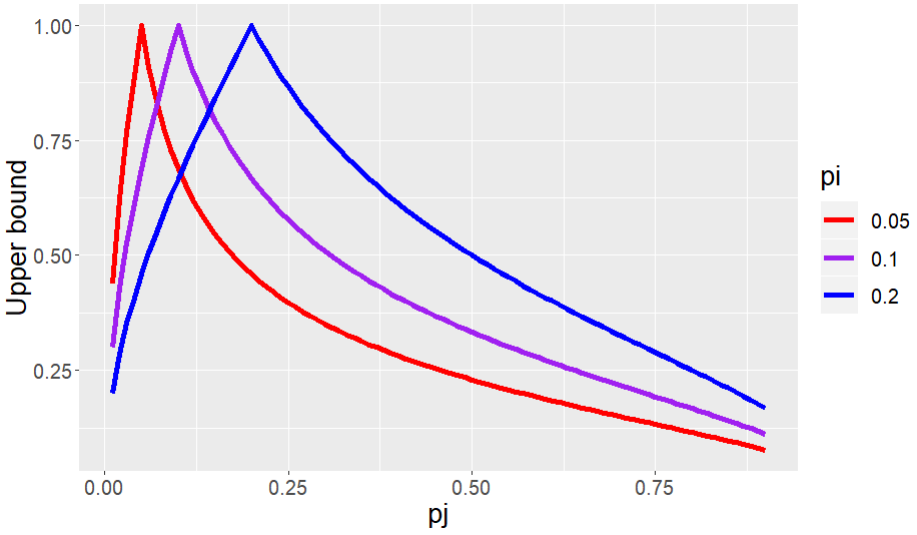

Figure 1.8: Correlation upper bound by marginal probabilities.

The sufficient condition for validity can now be simplified as follows.

Corollary 1.6.1.1. *Given a correlation coefficient α, an expectation set $p_1, ..., p_n$ satisfies Prentice's constraint for all pairs $\{p_i, p_j\}$ if and only if $(p_{(1)}q_{(n)}/p_{(n)}q_{(1)})^{\frac{1}{2}} > \alpha$.*

1.6.2.3 Generating valid expectation sets

There are two obvious approaches for generating a valid expectation set $[p_1, ..., p_n]$:

1. Generate $[p_1, ..., p_n]$ and then validate. Discard and repeat the process if the set is not valid.

2. Generate p_1, calculate range for p_2 that preserves validity, then generate p_2 within the range. Repeat for $p_2, ..., p_n$.

We implemented the first method, as the second one introduces unwanted dependency between cluster members (p_i depends on $p_1, ..., p_{i-1}$). However, unlike in the

second method, the distribution of p_i's needs to be predetermined.

In logistic regression, $p_i = F(\beta_0+\beta_1 X_i)$. The distribution of the p_i can be obtained in two ways:

1. Specify values for the X_i, then calculate the p_i.

2. Specify values for the p_i, then calculate the X_i.

In a regression scenario, it is more natural to consider generating random X_i. One possible procedure would be:

1. Generate $X_1, ..., X_n$ i.i.d. standard normal.

2. Specify a value for $\beta_1 = \log(OR)$. Specify a value for EY and calculate β_0 as the solution to the equation $\int F(\beta_0 + \beta_1 x) f_X(x) dx = EY$.

3. Calculate $p_i = F(\beta_0 + \beta_1 X_i)$ for $i = 1, ..., n$.

It is ideal that the X_i's are i.i.d. across all parameter combinations so the results are comparable. However a test run revealed a significant problem with this procedure: Valid $[p_i]$ sets has a smaller X_i variance than initially assigned (larger and smaller values of X_i had to be discarded). After the validity check, for $EY = 0.5$, the variance of X_i reduces from 1 to 0.8 for $OR = 2$, 0.4 for $OR = 3$ and 0.3 for $OR = 4$. A related problem is computational time: a smaller variance after validation implies that more invalid $[p_i]$ sets were generated, thus more computational time is required. The middle column of Table 1.1 lists the number of invalid generations in our test run. One can see that the rarity of successful generation grows exponentially with larger values of EY and OR. For $EY = 0.1, OR = 4$, it took more than 7 million

$E[Y]$	OR	Prentice Failed	Prentice Passed, PD Failed
0.5	2	9	34
0.1	2	32	91
0.5	3	9633	1098
0.1	3	74496	3636
0.5	4	546458	4173
0.1	4	7096969	14901

Table 1.1: Average failed attempts to generate 40 clusters of size 25, $\alpha = 0.25$.

iterations and about 10 hours to generate 40 valid expectation sets. This was much too slow for generating all the datasets for the GEEIV simulation study.

To reduce computational time and control the variance of X_i, we modified the process to achieve a success rate near one.

A direct result from Equation (1.36) is that a cluster member with expectation p_i poses a range restriction on all other members:

$$p_j \in \left[\frac{\alpha p_i}{q_i + \alpha p_i}, \frac{p_i}{q_i \alpha + p_i} \right].$$

We'll call this the valid range posed by p_i. For an expectation set $p_1, ..., p_n$ to be valid, all members needs to be in the valid range posed by all other members. In other word, all p_i's need to be in the intersection of all valid ranges. Combined with Lemma 1.6.1.1, it follows that the intersection depends only on $p_{(1)}$ and $p_{(n)}$.

Lemma 1.6.2. *The intersection of all valid ranges posed by all pair $(p_1, ..., p_n)$ equals the valid range posed by $p_{(1)}$ and $p_{(n)}$.*

The proof is similar to that of Lemma 1.6.1 and is omitted. Figure 1.9 serves as a visual explanation.

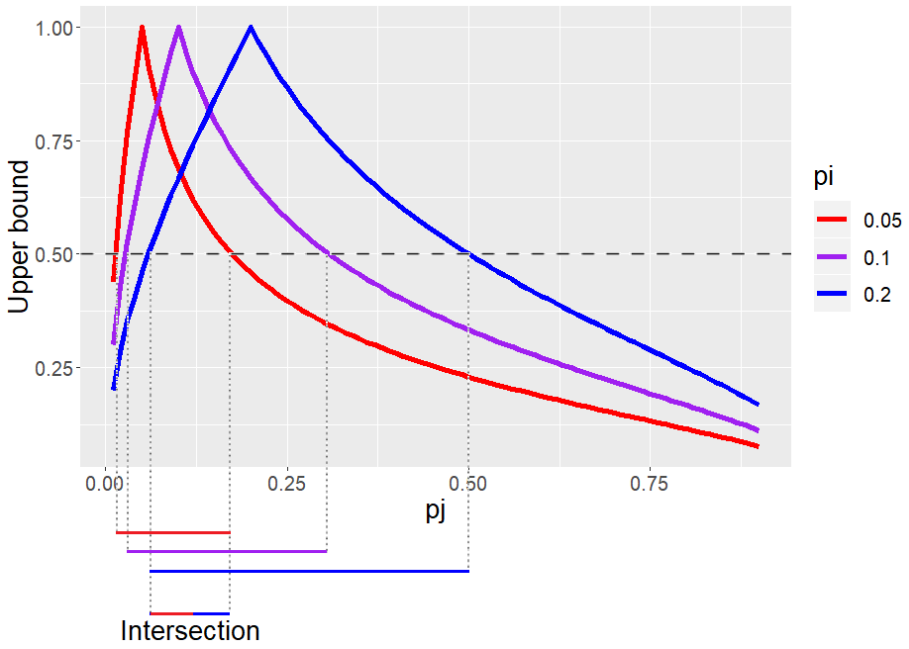

Figure 1.9: For a given α, p_j valid range is determined by maximum and minimum p_i. Intermediate p (Purple) plays no role.

An implication of Lemma 1.6.1 is that the valid range posed by $p_{(1)}$ and $p_{(n)}$ gets narrower as the difference $p_{(n)} - p_{(1)}$ increases, that is, a larger range for the p_i's decreases the probability of generating a valid expectation set.

The disadvantage of generating X_i normally distributed with a predetermined variance is now clear:

1. By the nature of logistic function, a symmetric X_i distribution will result in a skewed p_i distribution, which achieves larger $p_{(n)} - p_{(i)}$ with less p_i variance.

2. The skewness of the p_i distribution increases as EY deviates from 0.5.

3. The standard deviation of p_i is positively related to $\beta_1 \text{SD}(X_i)$, which increases with OR. For a fixed variance of X, a higher OR increases p_i variance, and in turn increases the probability of generation failure. To achieve a success rate

of 1 across all parameter combinations, it is necessary to choose the smallest variance allowed by all parameter combinations.

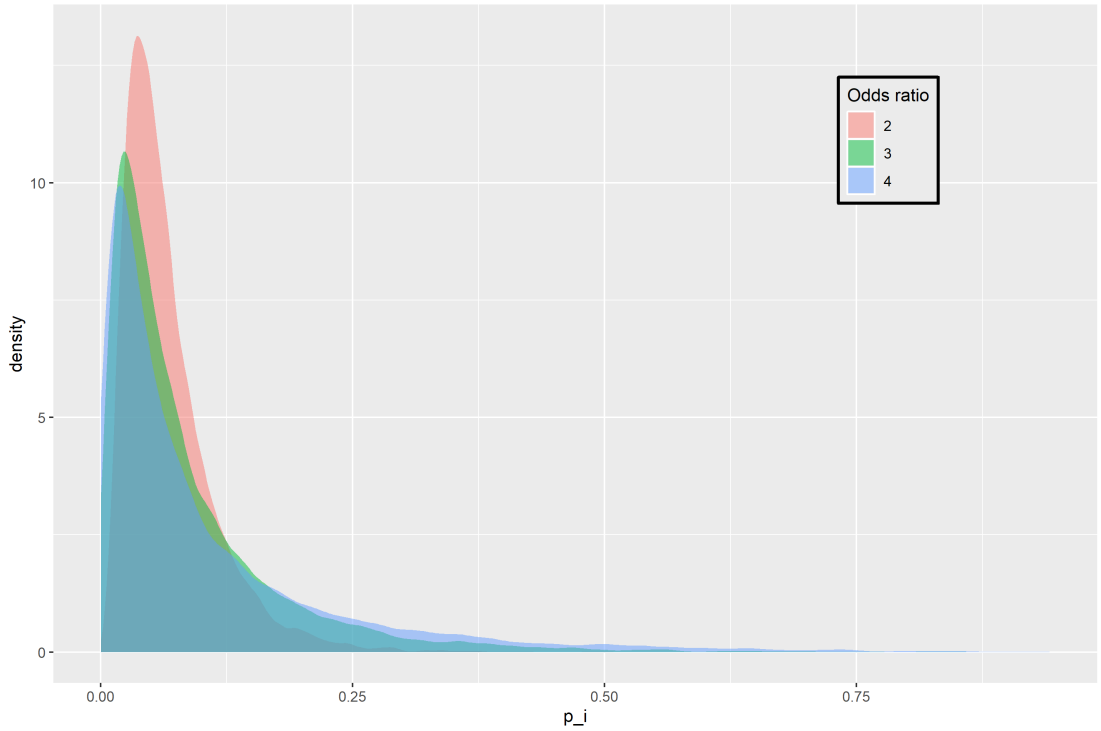

Figure 1.10: Fixing the variance of X causes the p_i to have varied variance and skewness with respect to OR.

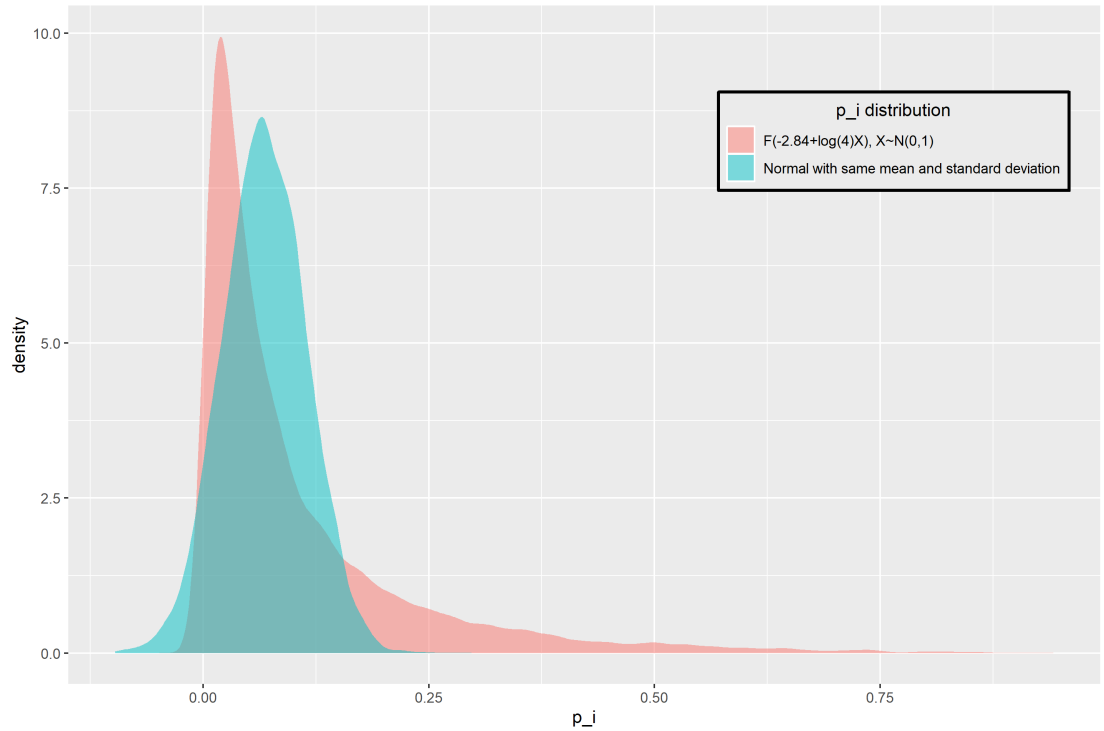

Figure 1.11: p_i generated with standard normal X is likely to have wider range than that from a normal distribution with identical mean and variance.

We reasoned that it is not optimal to begin with i.i.d. X_i's. A second procedure was then designed to begin with random p_i instead since p_i's are directly related to Prentice's constraint. It is a easier way to grasp control over the rate of generating a valid p_i set. By Lemma 1.6.2, the success rate is negatively related to $p_{(n)} - p_{(1)}$. In other word, to achieve a high success rate, we need to restrict the standard deviation of p_i. For a given pair of values for (EY, α), we let the p_i to have a normal distribution with mean EY and standard deviation σ_p such that $6\sigma_p = \frac{p_1}{q_1\alpha+p_1} - p_1$ where $p_1 = EY - 3\sigma_p$. Solving for σ_p yields:

$$\sigma_p = \frac{\frac{1+\sigma_0}{1-\alpha} + \sqrt{\left(\frac{1+\sigma_0}{1-\alpha}\right)^2 - 4(EY - [EY]^2)}}{6} \tag{1.40}$$

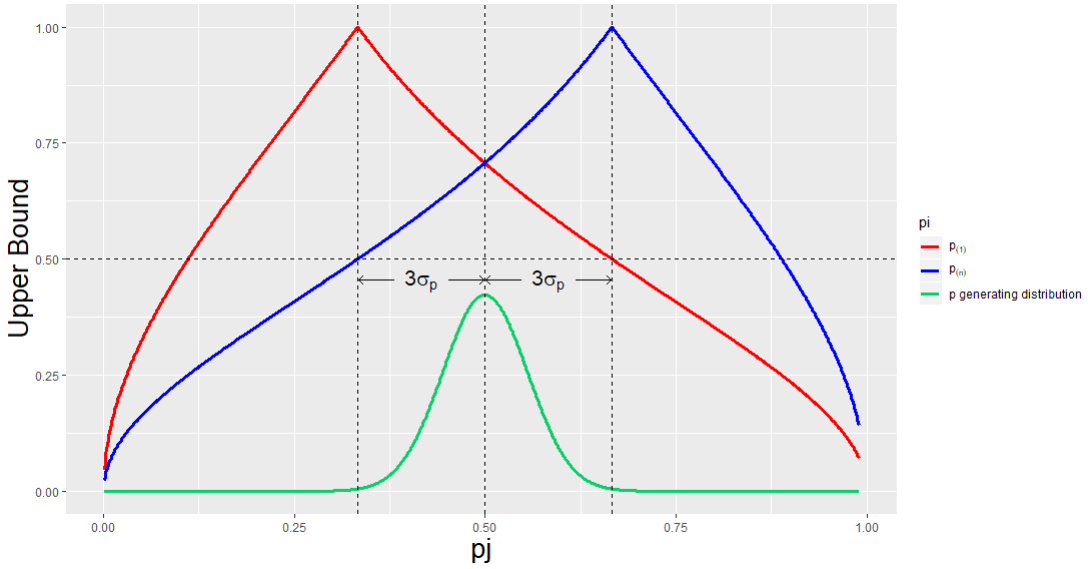

Figure 1.12: For a given (α, EY), σ_p is chosen such that all p's would very likely to fall into the valid region posed by $p_{(1)} = EY - 3\sigma_p$ and $p_{(n)} = EY + 3\sigma_p$.

The procedure is as follows:

1. Generate i.i.d. normal p_i with mean EY and standard deviation σ_p as in (1.40).

2. Let $\beta_1 = \log(OR)$ and β_0 be such that $F(\beta_0) = EY$.

3. Calculate X_i from p_i, β_0 and β_1.

The advantages of the second procedure compared to the first are:

1. About $(0.99)^n * 100\%$ of the attempts will be successful. Recall n is the cluster size. The actual variance of p_i are close to the initially assigned value. Computation time is less than $1/100$ of the prior method.

2. It allows a larger variance for X while satisfying Prentice's constraint.

1.6.2.4 Choice of algorithm

Once a valid set for $p_1, ..., p_n$ are generated, the next step is to generate binary variables $Y_1, \ldots, Y_n$ with expectations $p_1, ..., p_n$ and correlation α. The choice of algorithm for doing so is not nearly as important as the method for generating the expectations $p_1, ..., p_n$. We chose the algorithm proposed by Emrich & Piedmonte [10], which is one of the earliest. Several later algorithms claim to be faster than Emrich & Piedmonte's but for our simulation study, the improvement was not significant. Given a valid expectation set $[p_1, ..., p_n]$, Emrich & Piedmonte's algorithm is as follows:

1. Calculate the latent correlation matrix $\mathbf{\Sigma}_L = [\rho_{ij}]$, where ρ_{ij} is the solution of the following equation:

$$\Phi[\phi(p_i), \phi(p_j), \rho_{ij}] = \delta_{ij}(p_i p_j q_i q_j)^{0.5} + p_i p_j \tag{1.41}$$

   where Φ is the bivariate normal C.D.F and ϕ is the normal quantile function.

2. Generate multivariate normal r.v.'s $\boldsymbol{L} = [L_1, ..., L_n]^T$ with mean $\mathbf{0}$ and latent correlation matrix Σ_L.

3. $Y_i = 1$ if $L_i < \phi(p_i)$. $Y_i = 0$ otherwise.

It is worth noting that an expectation set $[p_i, ..., p_n]$ that satisfies Prentice's constraint does not guarantee a semi positive-definite latent correlation matrix Σ_L, which is required for generating the correlated normal latent variable L's in Step 2. As mentioned in the last section, the criterion for validity must be expanded to two conditions:

1. The expectation set satisfies Prentice's constraint.

2. The expectation set produces a semi positive-definite latent correlation matrix.

For simplicity, we'll call the second condition the "PD constraint". The PD constraint is a necessary condition posed by our choice of algorithm, which denies some expectation sets that might result in successful generation. As in the last section, a cluster failing the PD constraint will be discarded and re-generated, which can affect the variance of X.

We consider the two issues above minor, as the PD condition is rarely unsatisfied for expectation sets that pass Prentice's constraint. A comparison of the frequency of failing either constraint is given in the last column of Table 1.1.

1.6.3 Tables on Newton-Raphson convergence rate

The following tables show the number of times the Newton-Raphson algorithm, used to compute the IV estimates, converged.

EY	OR	IV	No M.Error	Both
0.50	2	1000	1000	1000
0.50	3	1000	1000	1000
0.50	4	1000	1000	1000
0.10	2	1000	1000	1000
0.10	3	999	1000	999
0.10	4	983	985	982

Table 1.2: Cluster Size=25; Number of Cluster=160; Normal Measurement error with attenuation factor 0.8. Reporting number of converged estimation via Newton Rapson method. Total number of cases is 1000 for all rows.

EY	OR	IV	No M.Error	Both
0.50	2	1000	1000	1000
0.50	3	1000	1000	1000
0.50	4	996	996	996
0.10	2	994	994	992
0.10	3	950	950	941
0.10	4	881	880	869

Table 1.3: Cluster Size=25; Number of Cluster=40; Normal Measurement error with attenuation factor 0.8. Reporting number of converged estimation via Newton Rapson method. Total number of cases is 1000 for all rows.

EY	OR	IV	No M.Error	Both
0.50	2	984	985	984
0.50	3	904	922	904
0.50	4	810	839	810
0.10	2	829	840	828
0.10	3	670	702	669
0.10	4	544	597	544

Table 1.4: Cluster Size=6; Number of Cluster=20; Normal Measurement error with attenuation factor 0.8. Reporting number of converged estimation via Newton Rapson method. Total number of cases is 1000 for all rows.

Chapter 2

Estimating the Cardinality of Latent Defective Sets

2.1 Introduction

In an edge search problem, the goal is to identify within a graph a hidden edge set whose members are considered "defective". A graph $G = \{V, E\}$ is a combination of a set of vertices V and edge set E where each edge $e \in E$ links a set of vertices. Our exploration on defective edge search problem was inspired by cascading failure in power system where the defective edges represent transmission line combinations where failure of all its members can result in a chain reaction through the power network and cause major blackout. Knowledge on these combinations are required to assess the risk and potential loss from cascading failure.

The defective combinations are considered latent and can not be observed without an actual cascading failure. With a power grid simulator called DCSIMSEP[29], one can perform a failure test by manually shutting down any vertex subset. If a cascad-

ing failure was observed afterward, one knows that at least one defective edge was contained in the selected vertex subset. Several algorithms have been proposed to utilize such tests to either identify the defective edge set or to estimates its size. However, these algorithms are impractical for graphs with many defective edges and/or defective edges that link many vertices. We introduce an efficient sampling approach to estimate the total number of defective edges for large graphs. The method provides an unbiased estimator of the number of defective edges regardless of their distribution, and a method for assessing the precision to which the number of defective edges has been estimated. The estimator's variance can be well estimated under the assumption that edges are equally likely to be defective.

These defective combinations, called "d-edges", are determined by the network structure and can be identified with a group testing method. Traditional group testing methods aimed at identifying the latent defective set have a time cost growing exponentially with the cardinality of the defective set, rendering them impractical for large networks, including power grids. We leverage statistical sampling ideas and theory to greatly reduce the required number of group tests.

2.2 Background

2.2.1 Power grid and cascading failure

A power grid is an interconnected network of power plants, substations, and transmission lines. The U.S. grid is divided into three major regions - the Western, Eastern and Texas Interconnection [36]. Cascading failure refers to major blackouts started

with the failure of a small number of components in the grid. One may consider electricity as water that flows from high ground (power plants) to low ground (end users). If some pipes (transmission lines) are blocked, the water (load) automatically goes through other passages and potentially overload them. This can cause a chain reaction and take down a large proportion of the grid in a short time. Cascading failures are rare in number, but the losses can be extensive. The largest blackout in the history of North America was due to cascading failure. In the afternoon of August 14, 2003, Maryland, Michigan, Ohio, Ontario, New York, New Jersey, Vermont and Connecticut lost power in an hour after three overheated transmission lines in Ohio sagged into trees and short circuited. The root cause was later identified as a series of events including high ambient temperature, a consequent power consumption surge, untrimmed trees and an unnoticed failure of the Electricity company's alarm system. The blackout lasted for up to 2 days in some areas and caused an estimated total economic loss of $6.4 billion [37].

2.2.2 Group testing and edge search problem

The method of group testing was proposed by Robert Dorfman during the second World War to efficiently screen recruits for syphilitic antigen [3]. Compared to running tests on individual blood samples, the total number of tests could be significantly reduced by testing mixed blood samples from a group of men, as one negative test would indicate that the whole group is not infected. Since then group testing has found application in various area other than blood testing, such as multiple access communication, and coding theory [18]. Recently, the FDA made an announcement on July 16, 2020, allowing testing facilities to pool COVID-19 samples in order to

preserve testing resources [35].

The edge search problem is an extension of group testing, where a positive test indicates the presence of certain combination of elements. Early studies [7] focused on identifying one combination of two elements (an edge). The method was later generalized to finding one combination of any size (a hyperedge) [14]. The problem of identifying all present combinations of any size was addressed by Chen and Hwang [23], which unsurprisingly required a significantly more complicated algorithm. The main challenge was to separate already identified sets so they are not present in later tests.

2.3 Terminology and notation

Here we define key terms and notation used in this chapter.

A few keywords are explained at first:

- Graph/hypergraph: The term "graph" is used as an abbreviation of "hypergraph". A graph is a pair $G = (V, E)$ with vertices V and edges E.
- n: The order of the full graph. $n = ||V||$.
- Vertices: A set of elements.
- Edge/hyperedge: The term "edge" is used as an abbreviation of hyperedge. An edge $e \in E$ is a subset of V of size greater than 1.
- Defective set/d-set: The defective set, or d-set, is defined to be a latent subset of E. The goal of this study is to estimate its cardinality.

- Defective edge/d-edge: The elements of a d-set. A d-edge is minimal, that is, it cannot be a superset of any other d-edge.
- Test: A test can be performed on any vertex subset $V_1 \subset V$. It returns positive if and only if the subgraph induced by V_1, that is, $G_1 = (V_1, E_1)$ such that $E_1 = \{e \in E : e \subset V_1\}$, has at least one common edge with the d-set.
- m-graphs: Random subgraphs being tested.
- d: The cardinality of the d-sets.
- r: The cardinality of a d-edge, must be greater than 1 for meaningful discussion.

2.4 Sampling algorithm for estimating d

The algorithm proposed by Chen and Hwang works well in graphs with only a few d-edges but quickly becomes impractical as d grows. To demonstrate this, we applied Chen and Hwang's algorithm on random graphs with $n = 100$, $r = 3$ and d ranging from 0 to 1,000. In Figure 2.1, each point is the algorithm cost for one random graph, measured in number of tests. The cost grows to the same magnitude as using brute force through all possible trios (100k v.s. $\binom{100}{3} \approx 161$k). A previous study [34] examined the US western grid which has 10,000 vertices. The d-set cardinality for $r = 3$ was estimated to be between $2.0 * 10^5$ and $2.9 * 10^5$. The simulation was performed on DCSIMSEP [29], a load flow simulator that can perform tests by simulating the load redistribution after tripping any combination of transmission

lines (vertices). Each test takes between less than 1 second to a few seconds to finish, depending on whether a cascading failure occurs. Applying Chen and Hwang's algorithm directly is not practically possible because the necessary calculations cannot be completed within a reasonable time frame. In this chapter we study the scenario where the number of d-edges are too numerous to be exhaustively identified, but an estimate of d is still valuable.

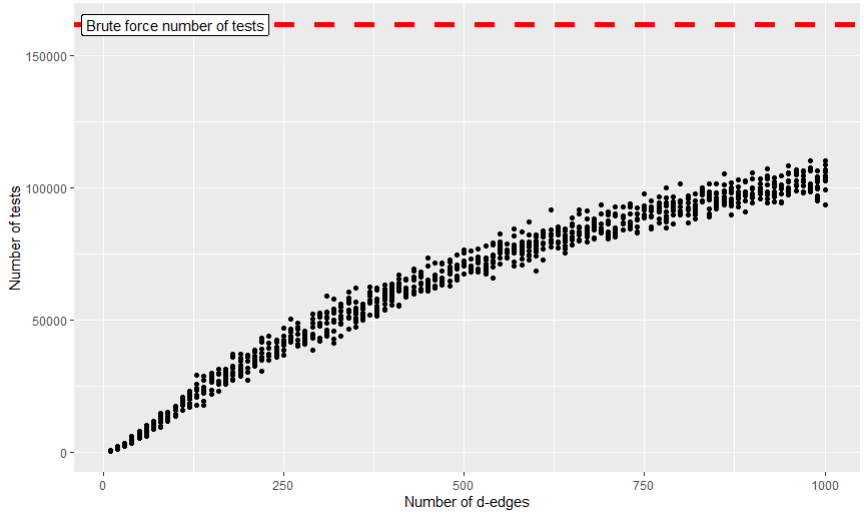

Figure 2.1: A comparison of efficiency of Chen and Hwang's algorithm to edge testing all edges. Uniformly distributed d-edges of size 3 were generated in 750 graphs of order 100. The number of d-edges in each graph is shown on the x-axis.

We develop an algorithm which can be considered a generalized Monte Carlo approach to estimate d for graphs with $d \gg n$. Our algorithm determines an optimum size m such that a uniformly chosen subgraph of size m contains very few d-edges on average. Random subgraphs of size m are then chosen and all d-edges in each subgraph are identified. Using methods from sampling theory, a method of moments estimator for d is defined that is unbiased for d regardless of the distribution of defective edges. We provide a method for constructing confidence intervals and evaluate

their performance through simulation studies.

First, some additional notation and terminology required for the sampling algorithm.

- m: the order of an m-graph. $m \geq r$.
- L: the number of m-graphs sampled.
- x: The d-set cardinality of an m-graph.
- $M : M = \binom{m}{r}$.
- $N : N = \binom{n}{r}$.
- Examine(verb): the action of applying an algorithm to identify the d-set in a m-graph. Examining a m-graph in general takes multiple tests.
- Cost: Number of test taken to examine a m-graph.

In the applications we are considering, it is impractical, if not impossible, to determine d exactly as the computation time is prohibitive. Estimation of d proceeds by sampling subgraphs of a fixed size m, determining all the d-edges in m-graph, and using the resulting data to construct an estimator for d. A straightforward algorithm is as follows.

1. For a fixed value of m, randomly choose L m-graphs for some integer $L \geq 1$.
2. Examine the L m-graphs to obtain $X_1, ..., X_L$.
3. Estimate d with $\hat{d} = g(X_1,...,X_L)$ where g is an estimating function. The function $g(\cdot)$ may be defined implicitly.

The quality of estimation depends on the choice of m, L and the function g. We'll take a closer look at these factors in the next three subsections.

2.4.1 Choice of the estimation function g

Maximum likelihood is an efficient approach to estimating d from $(X_1, X_2, \ldots, X_L)$. However, it requires knowledge of the distribution of the X_i. The distribution of X_i depends on the distribution of the d-edges, which is typically unknown. In the appendix we derive the likelihood function under the assumption that the d-edges follow a uniform distribution, i.e., all edges of size r are equally likely to be defective.

As noted, maximum likelihood is not applicable if the distribution of d-edges is unknown. Therefore, the primary approach we focus on is a method of moments estimator for d. To assess whether there is significant loss in efficiency, we compare the method of moments estimator to the maximum likelihood estimator in our simulation study.

Properties of the estimator for d that we propose rely on basic results from sampling from a finite population. This approach has been used in the context of estimating other graph characteristics, see for example Chapter 5 of [24]. We review the pertinent results here, prior to defining a method of moments estimator for d that assumes m is fixed. An extension of the estimator is then considered for m of variable size.

2.4.1.1 Sampling theory results

Consider a finite population of N^* values denoted $X_1, \ldots, X_{N^*}$. Let μ and σ^2 represent the population mean and variance: $\mu = \bar{X} = \frac{1}{N^*} \sum_{i=1}^{N^*} X_i$ and is $\sigma^2 = \frac{1}{N^*-1} \sum_{i=1}^{N^*} (X_i -$

$\mu)^2$.

Suppose L values are randomly sampled without replacement. The sample mean and variance are defined as $\bar{x} = \frac{1}{L}\sum_{i=1}^{L} X_i$ and $s^2 = \frac{1}{L-1}\sum_{i=1}^{L}(X_i - \bar{x})^2$. Then $E[\bar{x}] = \mu = \bar{X}$ and $E[s^2] = \sigma^2$. Also, $\text{Var}[\bar{x}] = (1 - L/N^*)\sigma^2/L$, and an unbiased estimator of this variance is then $(1 - L/N^*)s^2/L$.

If L values are sampled with replacement, then $E[\bar{x}] = \mu = \bar{X}$, $E[s^2] = (1-\frac{1}{N^*})\sigma^2$, and $\text{Var}[\bar{x}] = (N^* - 1)\sigma^2/N^*L$, and an unbiased estimator of this variance is then s^2/L. Note that the variance of $\bar{x}$ is smaller when sampling without replacement.

When sampling without replacement, a confidence interval for μ is given by

$$\left\{\bar{x} - t\sqrt{(1 - \frac{L}{N^*})s^2/L}, \bar{x} + t\sqrt{(1 - \frac{L}{N^*})s^2/L}\right\}$$

where t is an appropriate critical value. When L and N^* are large enough, a critical value from the normal distribution can be used instead of the $t-$ critical value.

2.4.1.2 Sampling Algorithm for Estimating d for fixed m

The algorithm given above for method of moments estimation is:

1. For a fixed value of m, randomly choose with replacement L m-graphs for some $L \geq 1$.

2. Examine the L m-graphs to obtain $X_1, ..., X_L$.

3. Estimate d with $\hat{d} = (N/M)\bar{x}$ where $\bar{x} = (1/L)\sum_{i=1}^{L} X_i$.

A fixed value of m induces a population of $N^* = \binom{n}{m}$ unique m-graphs, and the above sampling scheme is analogous to sampling L items from a finite population. The

value of X_i can be determined by brute force, or by an efficient algorithm, for example the Chen and Hwang method. Sampling can proceed with or without replacement.

The proposed estimator $\hat{d}$ is a method of moments estimator, which is seen by first noting $\sum_{i=1}^{N^*} X_i = d \cdot \binom{n-r}{m-r}$. Then from the above sampling results, $E[\bar{x}] = d\binom{n-r}{m-r}/N^*$. Noting $N/M = N^*/\binom{n-r}{m-r}$ yields the method of moments estimator $\hat{d} = (N/M)\bar{x}$.

Furthermore, from the standard sampling results reviewed above, it follows that

$$\mathrm{Var}(\hat{d}) = (N/M)^2 \,\mathrm{Var}(\bar{x}) = (N/M)^2 \left(1 - \frac{L}{N^*}\right) \sigma^2/L.$$

An unbiased estimator of $\mathrm{Var}(\hat{d})$ is $(N/M)^2 \left(1 - \frac{L}{N^*}\right) s^2/L$. Note this is an unbiased estimator of $\mathrm{Var}(\hat{d})$ regardless of how the d-edges are distributed in the network.

If the m-graphs are uniformly sampled, then each X_i has the hypergeometric distribution. In that case, the population variance is

$$\sigma^2 = \frac{Md}{N} \frac{(N-d)}{N} \frac{(N-M)}{N-1}.$$

Using this value for σ^2, the variance of $\hat{d}$ when sampling without replacement is then

$$\begin{aligned} \mathrm{Var}(\hat{d}) &= (N/M)^2 \,\mathrm{Var}(\bar{x}) = (N/M)^2 \left(1 - \frac{L}{N^*}\right) \sigma^2/L \\ &= \frac{1}{L} \left(\frac{N}{M}\right)^2 \left(1 - \frac{L}{N^*}\right) \frac{Md}{N} \frac{(N-d)}{N} \frac{(N-M)}{N-1} \\ &= \frac{1}{LM} \left(1 - \frac{L}{N^*}\right) d(N-d)) \frac{(N-M)}{N-1} \\ &\approx \frac{d(N-d)}{LM}. \end{aligned} \tag{2.1}$$

When sampling with replacement, the term $(1 - L/N^*)$ is replaced by $(1 - 1/N^*)$.

The approximation in the last line of (2.1) is useful when $N \gg M$ and $N^* \gg L$. Both these conditions are satisfied in our application, and therefore the approximation is employed below.

The above results are encapsulated in the following theorem.

Theorem 2.4.1. *Let $X_1, \ldots, X_L$ denote the numer of d-sets in L randomly chosen sub-graphs of size m. Let $\hat{d} = (N/M)\bar{x}$ where $\bar{x} = (1/L)\sum_{i=1}^{L} X_i$. Then*

1. $E[\hat{d}] = d$
2. $Var(\hat{d}) = (N/M)^2\, Var(\bar{x}) = (N/M)^2 \left(1 - \frac{L}{N^*}\right) \sigma^2/L$

If the m-graphs are uniformly distributed and $N \gg M$, $N^* \gg L$, then

$$Var(\hat{d}) \approx \frac{d(N-d)}{LM}.$$

2.4.1.3 Sampling Algorithm for Estimating d for variable m

Here a method of moments estimator for d is developed assuming the size of the sampled sets are variable. We'll see in the next section that a varying m is necessary for optimizing the sampling procedure.

The method of moments estimator is defined in Theorem 2.4.3. First, a preliminary lemma.

Lemma 2.4.2. *Let e_d be any d-edge of size r. e_d is contained in exactly $\binom{n-r}{m_i-r}$ order m_i subgraphs.*

Proof. An order m_i subgraph must contain the r vertices that belongs to e_d. The remaining $m_i - r$ vertices for the subgraph can be freely chosen from the $n - r$ total vertices, leaving $\binom{n-r}{m_i-r}$ possible combinations. □

Theorem 2.4.3. *Suppose L m-graphs of order $m_1, ..., m_L$ are randomly selected and examined. Let X_i be the number of d-edges in the ith m-graph and $M_i = \binom{m_i}{r}$. A method of moments estimator for d is*

$$\hat{d} = g(X_1, ..., X_L) = \frac{N}{\sum_{i=1}^{L} M_i} \sum_{i=1}^{L} X_i. \tag{2.2}$$

.

Proof. The method of moment estimator for d is derived by matching sample and population moments:

$$\sum_{i=1}^{L} X_i = \sum_{i=1}^{L} E(X_i|d, m_i). \tag{2.3}$$

Since the m-graphs are randomly chosen, $E(X_i|d, m_i)$ can be calculated regardless of the distribution of the d-edges. To see this, first note that the probability of any subgraph of order m_i being chosen is $1/\binom{n}{m_i}$. We then have

$$E(X_i|d, m_i) = \frac{1}{\binom{n}{m_i}} \sum_{j=1}^{\binom{n}{m_i}} d_j$$

where d_j is the number of d-edges in the jth subgraph, for some permutation of all $\binom{n}{m_i}$ order m_i subgraphs.

Since Lemma 2.4.2 applies to all d d-edges, the d-edge count from all order m_i subgraphs is:

$$\sum_{j=1}^{\binom{n}{m_i}} d_j = d * \binom{n-r}{m_i-r}.$$

The proof is completed by noting that $\frac{\binom{n-r}{m_i-r}}{\binom{n}{m_i}} = \frac{M_i}{N}$. Equation (2.3) can now be

reduced to:

$$\sum_{i=1}^{L} X_i = \sum_{i=1}^{L} E(X_i|d, m_i) = \sum_{i=1}^{L} \frac{M_i}{N} d \tag{2.4}$$

Solving for d yields (2.2). If $m_1 = ... = m_L = m$, the expression can be further simplified to $\hat{d} = \frac{N}{LM} \sum_{i=1}^{L} X_i$, the estimator obtained earlier. □

The following theorem states that the method of moment estimator is an unbiased estimator of d:

Theorem 2.4.4. *Suppose L m-graphs of order $m_1, ..., m_L$ are randomly selected and tested resulting in $X_1, X_2, \ldots, X_L$. The method of moment estimator is unbiased regardless of the d-edge distribution.*

Proof. The proof is essentially the reverse of the derivation of $\hat{d}$:

$$\begin{aligned} E(\hat{d}) &= E\left(\frac{N}{\sum_{i=1}^{L} M_i} \sum_{i=1}^{L} X_i\right) \\ &= \frac{N}{\sum_{i=1}^{L} M_i} \sum_{i=1}^{L} E(X_i) \\ &= \frac{N}{\sum_{i=1}^{L} M_i} \sum_{i=1}^{L} \frac{M_i}{N} d \\ &= d. \end{aligned}$$

□

2.4.2 Determining the optimum subgraph size subject to a computational cost constraint

The precision to which d can be estimated depends on the size of m and the magnitude of L. The following considers how to choose these quantities.

The variance of $\hat{d}$ can be made arbitrary small by increasing the sample size L or the size of the random m-graph (m), see (2.1). However, increasing these values comes at a steep computational cost. Therefore a cost constraint is introduced. In the following, we constrain the total number of edge tests to be T and propose an algorithm for finding optimal values for L and m under this constraint.

An approximation for T is

$$T \approx W_0(n, m, d, r)L$$

where $W_0(n, m, d, r)$ denotes the expected number of tests required to examine a single random m-graph. Obtaining a useful expression for $W_0(n, m, d, r)$ is somewhat complicated. Note that

$$W_0(n, m, d, r) = \sum_{x=0}^{m} p_0(x)u_0(x)$$

where $u_0(x)$ is the expected number of tests required for an m-graph with x d-edges, and $p_0(x)$ is the probability of an m-graph having x d-edges. Under the uniform d-edge assumption, we proved in the appendix that X_i follows a hypergeometric distribution

with probability mass function

$$p_0(x) = \frac{\binom{M}{x}\binom{N-M}{d-x}}{\binom{N}{d}}.$$

The form of $u_0(x)$ is dependent on the method used to find d-edges.

In our simulation, $W_0(n, m, d, r)$ is approximated by

$$W(n, m, d, r) = \sum_{x=0}^{5} p(x)u(x) \tag{2.5}$$

where we use a binomial distribution approximation

$$p_0(x) \approx p(x) = \binom{M}{x}\left(\frac{d}{N}\right)^x\left(1 - \frac{d}{N}\right)^{M-x}.$$

The binomial approximation is for faster computation and concise programming and is totally optional. The efficiency loss is minimal as when x is small, the sampling is essentially independent. An approximation for $u_0(x)$, for the Chen and Hwang algorithm for finding d-edges, is

$$u_0(x) \approx u(x) = 1 + x\log_2 M + \sum_{i=1}^{\min r,x}\left[\binom{x}{i}\sum_{j=0}^{r-i}\binom{i(r-1)}{j}\right].$$

The derivation for $u(x)$ is given in the appendix.

Note that the sum in the definition of $W(n, m, d, r)$ in (2.5) terminates at five. The appendix provides a heuristic explanation for this 'small x assumption'. In short, the probability of having more than 5 d-edges in an optimum m-graph is negligible.

For any fixed T, increasing m would cause an increase in $W(n, m, d, r)$ and as a

consequence, a decrease in L. There is a trade-off between the size of m and L. Using the approximation for T in (2.1) yields:

$$\begin{aligned} Var(\hat{d}) &\approx \frac{d(N-d)}{LM}. \\ &\approx \frac{d(N-d)}{T}\frac{W(n,m,d)}{M} \\ &\propto \frac{W(n,m,d)}{M}. \end{aligned} \tag{2.6}$$

The second line of (2.1) implies that the optimum m minimizing the variance of $\hat{d}$ does not depend on T. This may not be true in a trivial case when T is comparable to N. For example, $Var(\hat{d})$ would be 0 if one allows enough tests to brute force through every edge, or apply Chen and Hwang's algorithm on the whole graph, but those extreme cases are not considered.

An algorithm for finding an (approximate) optimal value for m is defined by the following five steps. In this algorithm, k is used as a loop counter. Let L_0 denote the total number of m-graphs examined at the termination of the algorithm. $\hat{d}_0$ is a rough guess of d generated in the process.

1. Let $k = 1$, $m = r$, $L_{0,k} = 10$.
2. Examine $L_{0,k}$ m-graphs. If $\sum_{i=1}^{L_{0,k}} X_i \geq 1$, go to Step 3, otherwise
   (a) set $k = k + 1$ and $\hat{d}_0 = N(1 - 0.95^{\frac{1}{L_{0,k}M}})$
   (b) obtain m by minimizing $\frac{W(n,m,\hat{d}_0)}{M}$
   (c) set $L_{0,k} = \left\lceil (\log_{1-\hat{d}_0/N} 0.05)/M \right\rceil$
   (d) go to Step 2
3. Let $\hat{d}_0 = \frac{N}{\sum_{i=1}^{L_0} M_i} \sum_{i=1}^{L_0} X_i$.

4. Obtain m by minimizing $\frac{W(n,m,\hat{d}_0)}{M}$.

5. Randomly select and examine an additional m-graph. Repeat steps 3 to 5 until the m's for the last five iterations are identical.

An explanation and justification of the algorithm steps is as follows. The initial guess for m in steps 1 and 2 is conservative in that the smallest possible value for m is used, i.e., $m = r$. This choice for m reduces the risk of investing unnecessary computational cost by unintentionally examining an inefficiently large m-graph. Note that the cost would be exactly 1 when $m = r$ regardless of the outcome. We initially test $L_{0,1} = 10$ small m-graphs to determine whether at least one d-edge can be found. If so, then a $\hat{d}_0$ can be obtained by the method of moments given in Step 3. If not, we then guess that d is small enough such that 95% of the time we would find 0 d-edges. The binomial approximation to the probability of finding 0 d-edges in $L_{0,k}$ m-graphs is

$$p\left(\sum_{i=1}^{L_{0,k}} X_i = 0\right) = \left[\binom{M}{0}\left(\frac{d}{N}\right)^0\left(1-\frac{d}{N}\right)^{M-0}\right]^{L_{0,k}} = \left(1-\frac{d}{N}\right)^{L_{0,k}M}.$$

The $\hat{d}_0$ defined in step 2a above is obtained by setting this probability to 0.95. An optimum m given $\hat{d}_0$ is then calculated by minimizing (2.6). $L_{0,k}$ is then updated such that, given $d = \hat{d}_0$ and m, testing $L_{0,k}$ m-graphs would have an approximate probability of 0.95 to identify at least one d-edge. The identification of the first d-edge ends the loop between steps 1 and 2 and starts the second loop (steps 3-5) to find a stable value for the approximation to the optimal m.

2.4.3 Determining the number of subgraphs subject to a computational cost constraint

Here we determine the total number of m-graphs to examine after finding the optimum m, which we denote as L_1. The L defined above would then equal $L_0 + L_1$. Compared to m, the choice of L_1 is much more flexible and can be customized based on user's need. If one has a cost/time constraint, one can simply examine as many m-graphs as possible. If a specific cost constraint is not imposed, one may choose to stop when a desired precision is reached. One option is through a standard deviation/estimate ratio: Solve for the minimum L_1 such that

$$\sqrt{Var(\hat{d})_0} < k\hat{d}_0$$

for some user defined $k \in (0, 1]$. By (2.2) and (2.6), we have

$$\begin{aligned}(k\hat{d}_0)^2 &> \mathrm{Var}\left(\frac{(N-\hat{d}_0)\sum_{i=1}^{L} X_i}{\sum_{i=1}^{L} M_i}\right) \\ &\approx \left(\frac{N}{\sum_{i=1}^{L_0} M_i + L_1 M}\right)^2 \left[\sum_{i=1}^{L_0} \mathrm{Var}(X_i) + \sum_{i=L_0+1}^{L} \mathrm{Var}(X_i)\right] \qquad (2.7) \\ &\approx \left(\frac{N}{\sum_{i=1}^{L_0} M_i + L_1 M}\right)^2 \left(\sum_{i=1}^{L_0} M_i + L_1 M\right) \frac{\hat{d}_0}{N} \frac{N-\hat{d}_0}{N}.\end{aligned}$$

We approximate the variance of the weighted sum of X_i with the weighted sum of variance. This assumes near independence between X_i's, which is not correct in a strict sense but reasonable in practice given the m-graphs are small. Also $N - \hat{d}_0$ is

replaced by N, which is appropriate under the assumption that $N \gg \hat{d}_0$. The second approximation uses

$$\mathrm{Var}(X_i) \approx M_i \frac{\hat{d}_0}{N} \frac{N - \hat{d}_0}{N}.$$

Solving (2.7) for L_1 yields

$$L_1 = \frac{\frac{N-\hat{d}_0}{k^2 \hat{d}_0} - \sum_{i=1}^{L_0} M_i}{M}. \tag{2.8}$$

2.5 Simulation results

Simulated graphs are used to examine the performance of our estimation algorithm. We examined the affects of the following parameters on the performance of the method: d-edge distribution, d, n, and r. For each parameter configuration, 1000 graphs were generated and analyzed. Details of the simulations and interpretation of results are provided in this section.

2.5.1 Estimation performance for large d

Simulation results for d-edges conforming to a uniform distribution are in Table 2.1, and d-edges following a semi-power law distribution in Table 2.2. Due to technical issues, we were unable to generate d-edge sets following an exact power law distribution. Instead, for fixed values (r,d) and vertices numbered from 1 to n, we generated d-edges by:

1. 80% of the d-edges are forced to include one of the 2% largest (in index) vertices, the remaining $r - 1$ vertices are uniformly generated from the 98% vertices.

2. 20% of the d-edges are uniformly generated from the 98% vertices.

In this way, although the frequency of defective vertices does not strictly follow a trend implied by the power-law distribution, we are able to preserve the characteristic that a small number of vertices occur a large proportion of times, and that there is strong d-edge overlapping.

In Tables 2.1 and 2.2, the graph parameters n, r, and d are shown in the leftmost columns. For each parameter combination, we generated 1000 random graphs seeded with d-edges. For each graph, L is calculated by (2.8) to achieve a 5% standard deviation/estimate ratio, after finding the optimal m. The total number of tests T, varies for each graph as both L and tests spent for finding m are not fixed. The tables display statistics describing the distribution of T and the performance of $\hat{d}$. 95% confidence intervals are calcuated via bootstrapping.

2.5.2 Estimation performance for small d

Although our algorithm is not designed for graphs with few d-edges, we still report simulation result for small d. Result for uniform d-edges is shown in Table 2.3. One can see from the last column, for $d \leq 100$, our algorithm finds most d-edges in the sampling process if not all of them.

2.5.3 Summary of simulation results

When the d-edges follow a uniform distribution, for all (n, d, r) combinations, $\hat{d}$ is unbiased, as expected. The empirical standard deviations are all such that the ratio to $\hat{d}$ are close to the desired value of 5%, implying that our estimation of standard deviation is adequate. The 95% Confidence intervals cover the true d approximately

r	n	d	$\bar{m}$	$\mathbf{E}(\hat{d})$-d	$\mathbf{Std}(\hat{d})$	$\frac{\mathbf{Std}(\hat{d})}{\mathbf{E}(\hat{d})}$	95% C.I. coverage	C.I. width	$\mathbf{E}(T)$	$\mathbf{Std}(T)$
3	100	10^2	16	-0.2	5.06	5.07%	94.15%	20.6	3,778.9	439.24
3	100	10^3	15	1.8	50.95	5.09%	94.57%	208.3	4,971.9	966.27
3	100	10^4	7	-15.3	524.01	5.25%	94.34%	2,095.6	4,072.2	587.13
3	100	10^5	3	798.7	5,942.63	5.90%	94.8%	23,231.0	232.2	97.83
3	1000	10^3	61	3.1	51.42	5.13%	96%	208.5	8,757.1	1,101.01
3	1000	10^4	31	23.0	509.28	5.08%	94.78%	2,101.1	9,549.8	2,594.26
3	1000	10^5	16	526.5	4,934.33	4.91%	96.1%	20,730.7	7,117.4	1,128.57
3	1000	10^6	8	3,795.6	50,242.24	5.01%	95.09%	207,110.6	5,443.1	933.80
4	100	10^2	24	-0.1	5.11	5.12%	93.83%	20.8	7,904.0	1,439.08
4	100	10^3	14	1.5	49.34	4.93%	96.46%	207.9	9,061.2	1,342.00
4	100	10^4	8	16.7	492.76	4.92%	96.08%	2,064.1	9,006.4	1,769.33
4	100	10^5	6	424.3	5,281.81	5.26%	94.23%	20,967.7	5,842.3	939.90
4	100	10^6	4	2,773.7	53,947.72	5.38%	95.05%	215,344.2	1,187.1	298.71
4	1000	10^3	121	5.1	50.68	5.04%	95.2%	208.6	19,564.5	18,807.53
4	1000	10^4	62	8.5	514.53	5.14%	93.83%	2,055.0	15,507.2	2,733.13
4	1000	10^5	44	334.2	5,100.90	5.08%	95.75%	20,891.0	13,305.6	2,414.30
5	100	10^2	28	-0.6	5.04	5.07%	95.13%	20.7	20,015.8	22,189.55
5	100	10^3	19	4.2	52.77	5.25%	94.36%	209.6	18,291.7	3,359.00
5	100	10^4	13	38.5	517.16	5.15%	94.59%	2,031.9	15,516.6	3,371.07
5	100	10^5	8	605.0	5,061.77	5.03%	95.46%	20,746.7	14,596.7	2,617.33
5	100	10^6	6	900.4	51,892.24	5.18%	94.34%	205,872.3	11,162.2	1,673.42
5	1000	10^3	174	4.2	54.01	5.38%	94.99%	209.1	24,006.4	4,837.66
5	1000	10^4	110	11.9	518.93	5.18%	95.1%	2,041.9	24,029.5	5,013.66
5	1000	10^5	62	387.3	5,197.67	5.18%	94.16%	20,969.2	26,709.2	8,980.36

Table 2.1: Simulation results for d-edges following uniform distribution. The statistics on each row are calculated over 1000 random graphs.

95% of the time, and the C.I widths are approximately 2*1.96 times the row's standard deviation. This result suggests bootstrap may be unnecessary and a normal approximation could suffice.

The expected number of tests, $E(T)$, increases as r and/or n increase. The affect of the value for d on $E(T)$ is more complicated, and we are not yet sure how to interpret the relation.

From Table 2.1 and Figure 2.1, an efficiency comparison can be made between

r	n	d	$\mathbf{E}(\hat{d})$	$\mathbf{Std}(\hat{d})$	$\frac{\mathbf{Std}(\hat{d})}{\mathbf{E}(\hat{d})}$	95%C.I. coverage	C.I. width	$\mathbf{E}(T)$	$\mathbf{Std}(T)$
3	100	10^2	100.0	6.56	6.55%	94.69%	25.6	3,648.6	660.48
3	100	10^3	1,016.1	97.82	9.63%	95.29%	328.5	4,156.3	1,008.15
4	100	10^2	100.1	5.79	5.79%	95.5%	23.4	8,274.4	1,625.78
4	100	10^3	1,004.9	64.18	6.39%	93.88%	257.0	9,253.6	5,455.33
4	100	10^4	10,113.0	746.77	7.38%	95.69%	2,681.6	8,049.8	2,073.05
4	100	10^5	100,446.8	8,916.66	8.88%	93.4%	31,879.9	4,550.4	1,132.94
5	100	10^2	99.9	5.68	5.68%	94.66%	22.0	23,126.8	26,569.72
5	100	10^3	1,004.2	57.25	5.70%	93.68%	224.5	17,549.1	3,478.68
5	100	10^4	10,017.1	607.54	6.07%	94.21%	2,302.6	16,401.8	5,130.82
5	100	10^5	100,570.8	6,638.73	6.60%	94.18%	23,727.4	14,001.8	3,499.06
5	100	10^6	1,000,548.2	62,661.91	6.26%	94.68%	248,937.0	9,828.6	1,918.72

Table 2.2: Simulation result for d-edges following power law distribution. For each row 1000 random graphs were estimated. For each graph, L was calculated by the algorithm to achieve a standard deviation 5% of $\hat{d}$, after finding an optimal m.

Chen and Hwang's algorithm and our estimation approach. For moderate to large values for d, our estimation method can be 5 to more than 5000 times more efficient. In general, the more d-edges in a graph, the more efficient our approach can be compared to testing the whole graph. From table 2.3, we can see that when d is small, the improvements are less significant. When there are few d-edges, our estimation method can require more tests than Chen and Hwang's algorithm in order to reach the 5% ratio. However, when d is small, our algorithm, although not guaranteed, finds almost all of the d-edges. Note that a rough estimate for d is obtained when finding an optimum m. If this estimate is small, a decision to switch to Chen and Hwang's algorithm might be warranted.

When the uniform distribution assumption is violated, our estimation method is still unbiased, which was established theoretically in Theorems 2.4.1 and 2.4.4. Furthermore, over the parameter configurations studied, the 95% C.I captures the true d about 95% of the time. $E(T)$ actually decreases a little compared to the

r	**n**	**d**	$\mathbf{E}(\hat{d})$	$\mathbf{Std}(\hat{d})$	$\frac{\mathbf{Std}(\hat{d})}{\mathbf{E}(\hat{d})}$	$\mathbf{E}(T)$	$\mathbf{Std}(T)$	**Unique d-edges identified**
3	100	25	24.7	1.31	5.30%	2,765.1	357.10	25.0
3	100	50	49.7	2.57	5.17%	3,547.7	369.93	50.0
3	100	75	74.8	3.73	4.99%	3,841.7	462.80	74.6
3	100	100	99.8	5.06	5.07%	3,778.9	439.24	97.9
3	100	1,000	1,001.8	50.95	5.09%	4,971.9	966.27	327.9
4	100	25	24.6	1.30	5.27%	6,223.6	1,034.91	25.0
4	100	50	49.5	2.53	5.10%	6,522.2	845.15	50.0
4	100	75	74.6	3.79	5.08%	7,223.2	1,020.50	74.6
4	100	100	99.9	5.11	5.12%	7,904.0	1,439.08	97.9
4	100	1,000	1,001.6	50.12	5.00%	9,151.4	1,364.43	329.1
5	100	25	24.6	1.29	5.23%	13,631.5	2,386.27	25.0
5	100	50	49.5	2.53	5.12%	16,451.7	7,008.61	49.9
5	100	75	74.5	3.82	5.13%	17,805.1	12,988.72	74.3
5	100	100	99.8	5.25	5.26%	18,785.4	20,037.39	97.4
5	100	1,000	1,003.3	55.88	5.57%	17,115.7	2,695.73	322.6

Table 2.3: Simulation results for d-edges following uniform distribution. For each row 1000 random graphs were estimated. For each graph, L was calculated by the algorithm to achieve a standard deviation 5% of $\hat{d}$, after finding an optimal m.

uniform settings. This means that our algorithm is underestimating L, the number of m-graphs sampled. Note that the average C.I widths are still very close to 2*1.96 of the standard deviation.

The C.I width is a good indicator of whether enough m-graphs were sampled and examined to reach a desired precision. If computational time allows, additional graphs can be sampled to decrease the interval width. Consider the second row of Table 2.2, and suppose the algorithm is applied to a graph resulting in $\hat{d} = 1016$ and a C.I width of 328, obtained sampling and testing $L = 500$ m-graphs subsequent to finding the optimum m. The standard deviation can be approximated by 328/2/1.96/1016, which is 8.2% of $\hat{d}$. By Equation 2.1, we can test $500/5\% * 8.2\% - 500 = 320$ more

m-graphs to make the standard deviation closer to 5% of $\hat{d}$.

2.5.4 Assessing the performance of the optimal m and W

In this section, we address three questions about our method for choosing an optimal m:

1. To what degree is the choice of optimal m affected by using $\hat{d}$ in place of d in 2.1?

2. How well is W estimating the average cost of examining an m-graph?

3. Is the optimal m really minimizing $std(\hat{d})$?

For the first question, we compared the optimal m as determined by our algorithm with the optimal m computed by minimizing the variance of $\hat{d}$ when d is known. Results for different values of (n, r, d) are given in Table 2.4. We list in Table 2.4 the optimal m's derived in Table 2.1 (the m column) to the optimal m's calculated using d (the m_d column). One can see that for most settings, m_d and m do not differ significantly.

For the second question, we compare W to the actual average cost T/L of Table 2.1. One can see from Table 2.4 that W is overestimating the average cost for most of the settings.

For the last question, we plot the distribution of m selected by algorithm together with std$(\hat{d})$ for a given value of m and fixed T. The standard deviations of $\hat{d}$ were

computed empirically. In Figure 2.2, one can see that for most (n, d, r) combinations, the algorithm selecting m achieved a highly efficient standard deviation, if not optimal.

r	n	d	m	m_d	W	T/L
3	100	10^2	16	15	6.8	3.0
3	100	10^3	15	8	54.5	38.9
3	100	10^4	7	4	38.3	18.4
3	100	10^5	3	3	1	1
3	1000	10^3	61	58	5.9	4.8
3	1000	10^4	31	31	6.3	5.7
3	1000	10^5	16	15	6.9	5.9
3	1000	10^6	8	8	5.7	4.6
4	100	10^2	24	25	9.4	4.4
4	100	10^3	14	14	7.9	5.8
4	100	10^4	8	8	4.9	3.8
4	100	10^5	6	6	10.2	5.6
4	100	10^6	4	4	1	1
4	1000	10^3	121	116	8.9	7.3
4	1000	10^4	62	73	5.5	5.0
4	1000	10^5	44	41	12.7	11.0
5	100	10^2	28	28	7.2	4.1
5	100	10^3	19	19	8.5	6.6
5	100	10^4	13	13	8.6	6.4
5	100	10^5	8	8	3.6	2.7
5	100	10^6	6	6	3.7	2.5
5	1000	10^3	174	174	10.6	8.6
5	1000	10^4	110	111	9.6	8.7
5	1000	10^5	62	64	4.9	4.5

Table 2.4: Performance of optimal m and W

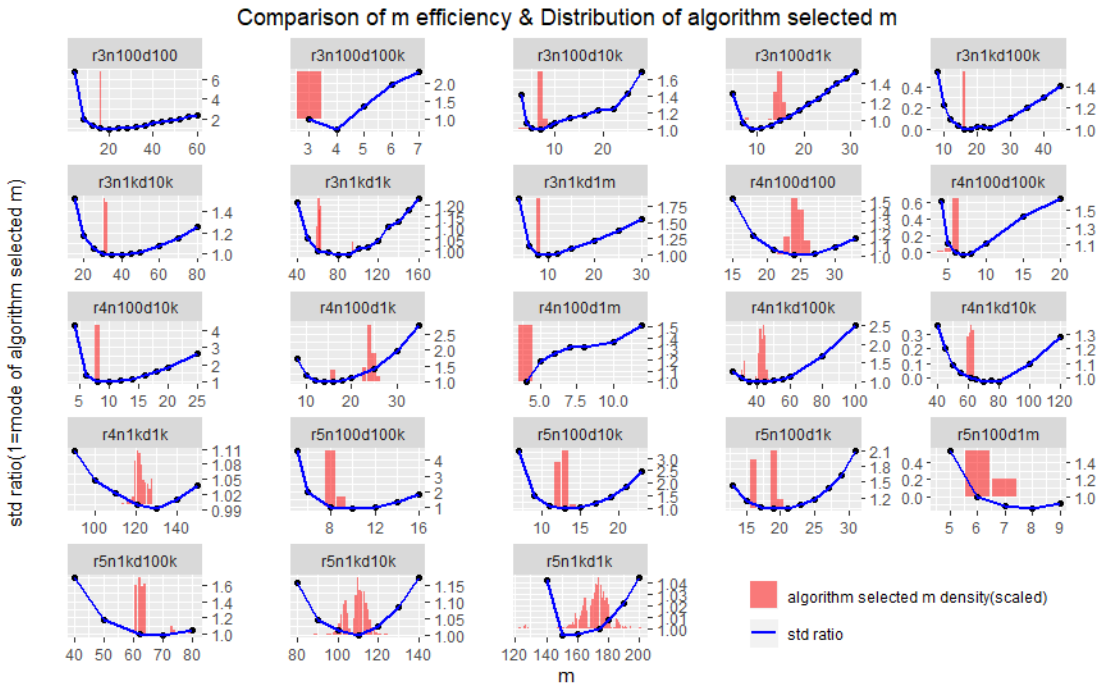

Figure 2.2: A comparison of m efficiency. Standard deviation of adjacent m's are compared to the that of the mode of algorithm selected m.

2.6 Western US power grid test case

We applied our method to Western US grid study with DCSIMSEP [29]. This simulation involves $|V| = 10,000$ vertices and contains d-edges of different sizes. A previous study [34] gave the following estimates: For $r = 2$, $d = 564$ by brute forcing through all pairs. For $r = 3$, d was estimated to be between $2.0 * 10^5$ and $2.9 * 10^5$. The total number of $r = 4$ d-edges were not estimated, as their algorithm did not converge properly.

Having different d-edges sizes poses a problem: The Chen and Hwang's basic method assumes a single r for all d-edges. Their algorithm is not directly applicable when this assumption is violated. For example, assuming $r = 3$ for an m-graph with

$r = 2$ d-edges may result in the algorithm identifying some or all of $r = 3$ edges containing the $r = 2$ d-edges as defective. Chen and Hwang proposed a method for simultaneously identifying d-edges of all sizes with a significantly increased cost and a far more complicated expression for $u(x)$. We streamlined their proposal by using brute force for all edges of size $r = 2, 3, 4$ when Chen and Hwang's algorithm decides to break an m-graph into subgraphs of order $m \leq 4$.

Another major issue, which is also mentioned in the previous study[34], is that DCSIMSEP gives false negatives. This is extremely harmful as it may crash Chen and Hwang's algorithm with a positive on an m-graph but a false negative on its subgraphs. The false negatives become more frequent as m grows. Therefore we manually set m to 6. An m-graph of order less than 6 showed very few false negatives in our pre-run. False negatives would also bias our $\hat{d}$ downward by reducing the total d-edge count. This effect, however, can be partly offset by recounting. Consider an example of examining two m-graphs G_1 and G_2 sharing some d-edges. If a shared d-edge is hidden by G_1 but identified in G_2, then we can add 1 back to X_1. We consider the ability to recount as another advantage of choosing a small m.

The number of m-graphs L, is obtained by calculating $std(\hat{d})/\hat{d}$, as mentioned in Section 2.4. The desired ratio is set to 5%. $\hat{d}$ for (2.1) is set to $d = 2.5 * 10^5$, the previous estimation for $r = 3$ d-set. Plugging the above numbers into (2.1) yields an $L = 8$ million.

All simulations were run on Vermont Advanced Computing Cores(VACC). We take advantage of VACC's multi-core capability and divide the 8 million m-graphs into 1,000 parts, running each part on separate nodes simultaneously. After all nodes finish, we record all identified d-edges, and then do the recount. The recount process

does not involve DCSIMSEP and thus is much faster.

The Western grid estimations are shown in Table 2.5:

Without recounting			
r	2	3	4
$\hat{d}$	517	197,857	76,343,064
C.I.	(478 559)	(156,203 239,511)	(34,701,393 124,925,014)
With recounting			
r	2	3	4
$\hat{d}$	**536**	**236,387**	**138,805,571**
C.I.	**(509 562)**	**(205,147 267,628)**	**(97,163,900 180,447,242)**
Recounting with all 564 $r = 2$ d-edges			
r	2	3	4
$\hat{d}$	546	NA	NA
C.I.	(479 610)	NA	NA

Table 2.5: Estimations for the Western US power grid cascade failure test case. $m = 6$, $L = 8,000,000$.

In table 2.5, we report 5 ME and C.I. calculated with and without recounting. For $d = 2$, we also report estimation from recounting with the 564 d-edges identified by brute forcing from the previous study.

Our algorithm took about 4.5 hours in total: 4 hours was spent on the examination of 8 million m-graphs and the recounting took half an hour. Our estimation for $r = 3$ agrees with the previous study (200,000$\sim$ 290,000 vs $\sim$ 205,147 267,628). We successfully obtained an estimation for $r = 4$ d-edges, which previous study did not. In conclusion, our algorithm fulfilled its purpose: Quick estimation of large number of d-edges.

With recounting, the estimations are adjusted upward by about 5%, 15% for $d = 2, 3$ and almost 50% for $d = 4$. A demonstration of the severity of false negative

and the necessity of recounting. As expected, it wasn't enough to completely remove bias, 95% C.I for $r = 2$ near miss the true value(562 vs 564). The estimation shown in third segment of Table 2.5 was done by recounting with the knowledge of all 564 $r = 2$ d-edges. Without the interference of false negative, the C.I properly contained the true value.

We did some further assessment of the estimators whose results are not shown in the tables. By estimating with first quarter/half of the 8 million m-graphs, we notice that the width of the C.I. for $r = 2, 3, 4$ has a roughly 30%, 30%, 50% increase each time sample size is halved. For an estimator that can be written as an average over sample, it is normal to see a 30% increase in standard deviation with halved sample size($30\% \approx 1 - 1/\sqrt{2}$). It is not the case for $r = 4$, indicating unknown sources of instability.

2.7 Conclusion

This chapter describes our original algorithm for efficiently estimating the total number of latent d-edges. This new algorithm is mainly designed for large graphs and can handle millions of d-edges with ease, while previous methods have trouble reaching stopping criterion within a reasonable amount of time. We look at the problem from a statistical prespective and proposed that efficiency of estimation comes from the efficiency of gaining information, which is measured in number of d-edges identified per test. Our algorithm begins with an initiation process which determines an optimum m-graph size that maximizes the information intake per test. Then it samples random m-graphs and records the d-edge counts. At the conclusion of sampling, a

method of moment estimates is constructed based on the counts. We provide a proof of unbiasedness of this estimator regardless of the d-edge distribution, as well as a variance estimator based on uniform d-edge distribution assumption. Our simulation on synthetic data shows that the algorithm takes even less time to finish as the d-edge count gets larger, a characteristic that is beyond our original expectation.

The core of our algorithm is the sampling theory. Other factors can be flexible. For example, the sample size L, the choice of m, and the algorithm for examining m-graphs can all be customized. One only needs to adjust the variance estimation accordingly. This gives the user a handful of choices to balance their needs. At the same time, the flexibility allows our algorithm to adapt to some critical assumption violations like coexistence of multiple d-edge sizes and false negative test results, as we demonstrate in the US western grid simulation.

In some applications, it is rarely the case to have a complete graph as we assumed in the development of the algorithm. Note that an incomplete graph can be considered as a complete graph with a non-uniform d-edge distribution where some of the edges can never be defective (because they don't exist), and therefore by our results the estimator remains unbiased, with a penalty in terms of underestimated variance. Again the US western grid simulation is a perfect example.

For future work, it is possible to explore methods to recover statistical efficiency in incomplete graphs. The formula for the method of moment estimator already provides some hints: simply replace N and M with the actual number of edges in the graph and m-graphs respectively and we may have a better unbiased estimator. Another direction is to assess the algorithm's efficiency for some other d-edge distributions. For example, non-uniform structures are introduced and programmed in these works

[33][30][32][28].

2.8 Appendix

2.8.1 Hypergeometric distribution and maximum likelihood estimator

The hypergeometric distribution describes the probability of sampling n times without replacement from a population of N and getting exactly k out of K marked items. The probability mass function (pmf) is

$$p_{N,n,K}(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}.$$

An characteristic of this pmf is that the probability does not change if one switches the values of n and K. This can be shown with simple algebra, thus the proof is omitted. This characteristic inspired us to explore the possibility of a tractable solution to finding the maximum likelihood estimator.

Consider a complete graph that has a total number of N edges, of which d are defective. By examining an m-graph, we sample M out of N edges. This matches the case described above, except for the sample not being random. After all, all sampled edges shared a limited vertex set. The probability of finding k d-edges can still be calculated using a hypergeometric model as long as the d-edges are uniformly distributed. As mentioned above, we can interchange n and K. In context, it means to consider the latent d-set as our sample and the edges of the m-graph as "defective".

Knowing the pmf, we can consider the possibility of employing the method of maximum likelihood. Under the uniform d-edge assumption, suppose L m-graphs are examined. Let $\boldsymbol{x} = [x_1, ..., x_L]$. The likelihood is:

$$\begin{aligned}\mathcal{L}(d|\boldsymbol{x}) &= \prod_{i=1}^{L} \frac{\binom{M}{x_i}\binom{N-M}{d-x_i}}{\binom{N}{d}} \\ &= c * \prod_{i=1}^{L} \frac{d!(N-d)!}{(d-x_i)!(N-M+x_i-d)!}\end{aligned} \tag{2.9}$$

where c is a constant with respect to d. The log-likelihood can be maximized directly using numerical methods. The following describes another approach.

Since d takes only integer values, we consider finding the maximum with the following likelihood ratio:

$$\begin{aligned}r(d|\boldsymbol{x}) &= \frac{\mathcal{L}(d+1|\boldsymbol{x})}{\mathcal{L}(d|\boldsymbol{x})} \\ &= \prod_{i=1}^{L} \frac{(d+1)(N-M+x_i-d)}{(d+1-x_i)(N-d)}, \quad d \in \{d \in \mathbf{Z}^+ : \mathcal{L}(d|\boldsymbol{x})\mathcal{L}(d+1|\boldsymbol{x}) \neq 0\}\end{aligned} \tag{2.10}$$

For analysis purposes, we consider an extension of $r(d|\boldsymbol{x})$ to $\mathbb{R}$:

$$r_c(d|\boldsymbol{x}) = \prod_{i=1}^{L} \frac{(d+1)(N-M+x_i-d)}{(d+1-x_i)(N-d)}, \quad \max x_i < d < \min\{N-M+x_i\}$$

Let $\tilde{d}$ be the solution to equation $r_c(\tilde{d}|\boldsymbol{x}) = 1$. We then show that

1. There exists a unique $\tilde{d}$.

2. $\lceil \tilde{d} \rceil$ maximizes $r(d|\boldsymbol{x})$.

Theorem 2.8.1. *If $x_i \ll M \ll N$ for all $i = 1, ..., L$, then $r_c(d|\boldsymbol{x}) = 1$ has a unique root in $(\max x_i, \min\{N - M + x_i\})$.*

Proof. First we explain the support of $r_c(d|\boldsymbol{x})$. Note that with observation $\boldsymbol{x}$, the total number of d-edges can not go below the number of d-edges found in any m-graph ($d \geq \max\{x_i\}$). Similarly, the the total number of non-defective edges can not go below the number of non-defective edges in any m-graph ($N - d \geq \min\{M - x_i\}$). Therefore d must be within $\Big[\max\{x_i\}, \min\{N - M + x_i\}\Big]$ for $\mathcal{L}$ to be non-zero.

Since $r_c(d|\boldsymbol{x})$ is continuous on its support, the existence of $\hat{d}$ can be shown with intermediate value theorem, we only need the two boundary values of $r_c(d|\boldsymbol{x})$ to be on different side of 1.

The two boundary values for $r_c(d|\boldsymbol{x})$ are:

$$\begin{aligned} r(\max x_i|\boldsymbol{x}) &= \prod_{i=1}^{L} \frac{\max x_i + 1}{\max x_i + 1 - x_i} * \frac{N - M + x_i - \max x_i}{N - \max x_i} \\ &\approx \prod_{i=1}^{L} \frac{\max x_i + 1}{\max x_i + 1 - x_i} > 1 \end{aligned}$$

and

$$r(\min\{N - M + x_i\}|\boldsymbol{x}) = 0.$$

The uniqueness of $\tilde{d}$ follows by showing r_c is monotonic. Take the derivative of

$\ln r_c(d|\boldsymbol{x})$:

$$\ln r_c(d|\boldsymbol{x}) = L\ln(d+1) - \sum_{i=1}^{L}\ln(d+1-x_i) - L\ln(N-d) + \sum_{i=1}^{L}\ln(N-M_i+x_i-d)$$

$$\begin{aligned}[\ln r_c(d|\boldsymbol{x})]' &= \frac{L}{d+1} - \sum_{i=1}^{L}\frac{1}{d+1-x_i} + \frac{L}{N-d} - \sum_{i=1}^{L}\frac{1}{N-M+x_i-d} \\ &= \left[\sum_{i=1}^{L}\left(\frac{1}{d+1} - \frac{1}{d+1-x_i}\right)\right] + \left[\sum_{i=1}^{L}\left(\frac{1}{N-d} - \frac{1}{N-M_i+x_i-d}\right)\right] \\ &< 0\end{aligned}$$

The last inequality holds as the previous equality is two summations of non-positive values, and 0 can not be reached at the same time for both components at any x_i. □

Theorem 2.8.2. *If $x_i \ll M \ll N$ for all $i = 1, ..., L$. then $\lceil\tilde{d}\rceil$ among all positive integers maximizes $\mathcal{L}$, where $\lceil\cdot\rceil$ is the ceiling operation.*

Proof. Let $\tilde{d}$ be the root. If $\tilde{d}$ is an integer, then by the definition of $r(d|\boldsymbol{x})$, we have $L(\tilde{d}+1|\boldsymbol{x}) = L(\tilde{d}|\boldsymbol{x})$. For any integer $d^* < \tilde{d}$,

$$\frac{\mathcal{L}(\tilde{d}|\boldsymbol{x})}{\mathcal{L}(d^*|\boldsymbol{x})} = \prod_{d=d^*}^{\tilde{d}-1} r(d|\boldsymbol{x}) > 1$$

and for any integer $d^* > \tilde{d}+1$,

$$\frac{\mathcal{L}(d^*|\boldsymbol{x})}{\mathcal{L}(\tilde{d}+1|\boldsymbol{x})} = \prod_{d=\tilde{d}+1}^{d^*-1} r(d|\boldsymbol{x}) < 1$$

so both $\tilde{d}$ and $\tilde{d}+1$ maximizes $\mathcal{L}(d|\boldsymbol{x})$.

If $\tilde{d}$ is not an integer, for any integer $d^* < \lceil \tilde{d} \rceil$,

$$\frac{\mathcal{L}(\lceil \tilde{d} \rceil | \boldsymbol{x})}{\mathcal{L}(d^* | \boldsymbol{x})} = \prod_{d=d^*}^{\lceil \tilde{d} \rceil - 1} r(d|\boldsymbol{x}) > 1$$

and for any integer $d^* > \lceil \tilde{d} \rceil$,

$$\frac{\mathcal{L}(d^* | \boldsymbol{x})}{\mathcal{L}(\lceil \tilde{d} \rceil | \boldsymbol{x})} = \prod_{d=\lceil \tilde{d} \rceil}^{d^* - 1} r(d|\boldsymbol{x}) < 1$$

Therefore $\lceil \tilde{d} \rceil$ maximizes $\mathcal{L}(d|\boldsymbol{x})$ in either case. □

It is unfortunate that d can only take integer numbers the way we define $\mathcal{L}$. Consider an $\mathcal{L}$ extension to real numbers with gamma function:

$$(d-1)! = \Gamma(d) = \int_0^\infty t^{d-1} e^{-t} dt.$$

The "likelihood" is then

$$\mathcal{L}_c(d|\boldsymbol{x}) = c * \prod_{i=1}^{L} \frac{\Gamma(d+1)\Gamma(N-d+1)}{\Gamma(d-x_i+1)\Gamma(N-M_i+x_i-d+1)}$$

where c is constant with respect to d. Using calculus, one can see that $r_c(d|\boldsymbol{x}) = \mathcal{L}_c(d+1|\boldsymbol{x})/\mathcal{L}_c(d|\boldsymbol{x})$ in its support and therefore Theorem 2.8.1 still holds. Note that Theorem 2.8.1 does not imply a concave $\mathcal{L}_c(d|\boldsymbol{x})$, however it does ensure the MLE is trapped between $\lfloor \tilde{d} \rfloor$ and $\lceil \tilde{d} \rceil$.

Theorem 2.8.3. *If $x_i \ll M \ll N$ for all $i = 1, ..., L$. then $\lceil \tilde{d} \rceil$, then the MLE for $\mathcal{L}_c(d|\boldsymbol{x})$ is in between $(\lfloor \tilde{d} \rfloor, \lceil \tilde{d} \rceil)$.*

Proof. Let d_b be an integer s.t. $\lceil \tilde{d} \rceil < d_b < \min(N - M_i + x_i) - 1$, in Theorem 2.8.1

we showed $r_c(d|\boldsymbol{x}) < 1$ in $(d_b, d_b + 1)$. Let $d_{bm} = \arg\max_d \mathcal{L}_c(d|\boldsymbol{x}), \quad d \in (d_b, d_b + 1)$. We immediately have $\mathcal{L}_c(d_{bm}|\boldsymbol{x}) < \mathcal{L}_c(d_{bm} - 1|\boldsymbol{x}) \leq \max_{d\in(d_b-1,d_b)} \mathcal{L}_c(d|\boldsymbol{x})$. It follows by induction that $\mathcal{L}_c(d_{bm}|\boldsymbol{x}) < \max_{d\in(\lfloor\tilde{d}\rfloor,\lceil\tilde{d}\rceil)} \mathcal{L}_c(d|\boldsymbol{x})$.

Similarly, let d_a be an integer s.t. $\max x_i + 1 < d_a < \lfloor\tilde{d}\rfloor$, since $r_c(d|\boldsymbol{x}) > 1$ on this region, we have $\mathcal{L}_c(d_{am}|\boldsymbol{x}) < \max_{d\in(\lfloor\tilde{d}\rfloor,\lceil\tilde{d}\rceil)} \mathcal{L}_c(d|\boldsymbol{x})$. It suffices to show $\arg\max_d \mathcal{L}_c(d|\boldsymbol{x}) \in (\lfloor\tilde{d}\rfloor, \lceil\tilde{d}\rceil)$. □

In conclusion, while $\tilde{d}$ may not be the MLE, the difference between $\tilde{d}$ and the MLE must be less than 1. For our applications, where d usually ranges from a few hundred to millions, an error of less than 1 is more than acceptable.

2.8.2 A heuristic explanation of why there are very few d-edges in an optimal m-graph

Combining equations 2.2 and 2.1, we have:

$$\begin{aligned} Var(\hat{d}) &\approx \frac{N}{ML} d \\ &= \frac{\hat{d}}{\sum_{i=1}^{L} x_i} d \end{aligned}$$

implying that for a fixed T, the best m maximizes $\sum_{i=1}^{L} x_i$, the total number of identified d-edges.

As will be mentioned in the next section, the cost for identifying one d-edge of size r in an m-graph can be considered a fixed value $r \log_2(m)$. It is appropriate to consider the process as applying the binary search algorithm r times on the m vertices to identify each of the r "defective" vertices, which is exactly what Chen and Hwang's

algorithm does. The loss of efficiency comes from two sources:

1. Examining an m-graph with no d-edges would "waste" 1 test.

2. Examining an m-graph with multiple d-edges that share some vertices would greatly increase the cost.

Simply put, examining a larger m-graph would simultaneously decrease the risk of Type 1 loss and increase the risk for Type 2 loss. The optimum m strikes a balance between the two type of losses, and since the computational cost of shared "defective" vertices is so high, the optimum m almost always lean toward having very few d-edges. In our simulation, optimum m-graphs had an average number of d-edges from 0.4-0.6, depending on the parameters and distribution of d-edges. For the power-law simulation, the average number of d-edges in an m-graph was lower than for the uniform simulation. This is because in the former, d-edges in an m-graph tend to share more vertices.

2.8.3 Brief explanation of Chen and Hwang's algorithm in our setting and the derivation of W

Chen and Hwang's algorithm [23] is designed to identify a latent defect edge set within an incomplete graph. However for easy demonstration, here we present a complete graph as example. The purpose of this brief introduction is to give readers some background information so they can understand the derivation of $W(n, m, d, r)$, our

estimator of the expected cost of Chen and Hwang's algorithm on a random m-graph, as well as why W is typically is an overestimate.

First we explain how a single d-edge is identified. Let G be a complete graph with vertices $\boldsymbol{V} = \{v_1, ..., v_n\}$. Let G_i be the subgraph induced by $(v_1, ...v_i)$. Suppose there is a single d-edge in G, denoted as $(v_{d1}, ..., v_{dr})$ with $1 \leq d1 < ... < dr \leq n$. For $i = 1, ..., r$, we call $v_{d,r-i+1}$ the ith leader of the d-edge.

A d-edge is identified by finding the leaders in order. The first leader can be identified with a binary search:

1. Let $l = 1$, $u = n$.

2. Let $p = \lfloor (l + u)/2 \rfloor$. Test G_p. If positive, $u = p$. If negative, $l = p$.

3. If $l = u - 1$, return $v_{dr} = v_l$. Else, go back to step 2.

The second leader can be identified by the same process, after removing the first leader from the binary search. Note that we still need to attach the first leader to every test, otherwise it would be guaranteed negative. The third leader can be identified by removing the first and second leaders and so on.

We'll adopt the name from Chen and Hwang and call this strategy the "TJ-process" where TJ refers to the method's authors [15][20]. Since it is a binary search performed r times, an upper bound for cost is $r \log_2 n$.

The algorithm becomes much murkier when there are multiple d-edges in the graph. The dichotomous nature of TJ-process has a consequence: Multiple applications will only find the same d-edge, to be more specific, the one with the minimal first leader. If all d-edges share the same first leader, then the one with the minimal second leader will be found, and so on. One may consider a bypass: applying

TJ-process with a different permutation of V. However there is no guarantee a new d-edge will be found, nor can we determine a stopping condition for the algorithm.

Chen and Hwang's algorithm can be roughly described with the following recursion:

1. Apply the TJ-process only on a graph that contains no identified d-edge.
2. Once a d-edge is identified in a graph, find all d-edges in a set of its subgraphs s.t.
   (a) no subgraph contains an identified d-edges.
   (b) any potential d-edge must be contained in at least one of the subgraph.

For readers who wish to skip the technical details below, the takeaway message of the above recursion is that the number of subgraphs required to satisfy (a) and (b) becomes very large when there are multiple d-edges. Even if all these subgraphs do not contain a single d-edges, the 1 test spent on each of these subgraphs would add up to a huge number. The expression for $u(x)$ in (2.4.2), the estimated average cost of an m-graph with x d-edges, is given by (2.11), in which the components excluding $x \log_2 M$ is the number of negative tests.

We use an example to demonstrate how Chen and Hwang's algorithm works. In the example, d-edges will have no shared vertex. This makes the process less complex to explain, and yet the process will have the highest cost. One may follow the same process to see that sharing vertices will slightly reduce the cost.

Let G be a complete graph with vertices $V = \{1, .., 10\}$. E is the set of all edges of size $r = 3$. The d-set is $\{(3, 5, 7), (4, 6, 8)\}$. Let v_0 be the search range, that is,

the only set that contain unidentified d-edge vertices. We use $TJ(v)$ to denotes TJ-process which outputs the d-edge with minimal leader in G_v for $v \in V$. Let t be the set of identified leaders. The pseudocode for solving the main problem of identifying all d-edges in G_V goes as:

1. $v_0 = V$, $t = \emptyset$.
2. Test G_{v0}. If positive, $TJ(v_0)$. Let v_d be the leader of the identified d-edges, $v_0 = v_0 - v_d$, $t = t + v_d$. Repeat Step 2. If negative, go to Step 3.
3. solving sub-problems G_{v_0+k}, where $k \subseteq t$.

Here we step through the process. Step 2 will run twice and the two d-edges $(3, 5, 7)$ and $(4, 6, 8)$ will be identified. The leaders 7 and 8 will be moved out from v_0 to t and the loop will end with a negative test on G_{v0}. This negative test implies that any latent d-edge, if any, must be contained in either $G_{v0+\{7\}}$, $G_{v0+\{8\}}$ or $G_{v0+\{7,8\}}$. We simulate the sub-problem on $G_{v0+\{7\}}$. $G_{v0+\{8\}}$ and $G_{v0+\{7,8\}}$ will be similar.

Note that any potential d-edge in $G_{v0+\{7\}}$ must contain $\{7\}$, otherwise G_{v0} would have had a positive test in the last loop of Step 2. Therefore $\{7\}$ will be attached to every test. Also, since an identified d-edge $\{3, 5, 7\} \subset (v_0 + \{7\})$, $TJ(v_0 + \{7\})$ will be a guaranteed positive and thus is against the rule. The solution is to quarantine some vertices from the d-edge $\{3, 5, 7\}$. Note that any potential d-edge in $G_{v0+\{7\}}$ must be contained in either $G_{v0+\{7\}-\{3\}}$ or $G_{v0+\{7\}-\{5\}}$. Testing these two subgraphs will yield two negatives, implying $G_{v0+\{7\}}$ does not contain any more d-edges. Since no new d-edge is identified in G_{v0}, there is no need to introduce level-3 subproblem.

The total cost can be separated into two parts: the cost on the TJ-process, and the cost of negative tests not in a TJ-process. The TJ-process runs as many times as the

number of d-edges(2, in our example), and always follows a positive test. An upper bound for the first part of the cost is then $d*[1+r*\log(n)] = 2*[1+3*\log(10)] \approx 20$. The cost of the second part, which is the number of negative tests, equals the number of subproblems of all levels, including the main one. This is because a decision to stop or to introduce subproblems is made immediately after a negative test. In our example, we had 1 main problem, followed by 2 subproblems $G_{v_0+\{7\}-\{3\}}$, $G_{v_0+\{7\}-\{5\}}$ from $G_{v_0+\{7\}}$, 2 subproblems $G_{v_0+\{8\}-\{4\}}$, $G_{v_0+\{8\}-\{6\}}$ from $G_{v_0+\{8\}}$ and 4 subproblem $G_{v_0+\{7,8\}-\{3,4\}}$, $G_{v_0+\{7,8\}-\{3,6\}}$, $G_{v_0+\{7,8\}-\{5,4\}}$, $G_{v_0+\{7,8\}-\{5,6\}}$ from $G_{v_0+\{7,8\}}$.

The upper bound for the total cost is then $20 + (1 + 2 + 2 + 4) = 29$. The negative tests already take almost one third of the total cost, and will get larger if there are more than 2 d-edges. Consider 10 d-edges with $r = 3$ not sharing a vertex in a graph with $n = 100$. The first part of the cost is bounded from above by $10 * [1 + 3 * \log(100)] = 310$. $d = 10$ leaders will be found, resulting in $2^d - 1 = 1023$ possible partitions. Each partition will introduce about $(r-1)^k$ subproblems, where k is the number of leaders included. For subgraphs including 1 of the 10 leaders(e.g. $G_{v_0+\{l_1\}}$), $2^1 = 2$ level-2 subproblems will be introduced. For subgraphs including all 10 leaders, $2^{10} = 1024$ level-2 subproblems will be introduced. In this example, the cost of negative tests greatly outnumbers that of the TJ-process.

Chen and Hwang gave an upper bound for the second part cost:

$$(r-1)^{\lfloor \frac{r}{2} \rfloor} d^r + o(d^r)$$

One may compare it to $dr\log(m)$ to see the second part would dominate the cost even for moderate d or r. This upper bound however, is not a good candidate for $u(x)$ as it greatly overestimates the cost and yields an inaccurate optimal m. We follow our

description above and derive a more precise estimate:

$$\begin{aligned} u(x,m,r) &\approx 1 + x(1+\log_2 M) + \sum_{i=1}^{\min r,x} \left[\binom{x}{i} \sum_{j=0}^{r-i} \binom{i(r-1)}{j} \right] - x \\ &= 1 + x\log_2 M + \sum_{i=1}^{\min r,x} \left[\binom{x}{i} \sum_{j=0}^{r-i} \binom{i(r-1)}{j} \right] \end{aligned} \tag{2.11}$$

Conclusion

In this dissertation we proposed and developed two estimation methods for two distinct problems. The first study considered the problem of covariate measurement error in non-linear regression models when data are clustered and estimation in the absence of measurement error is accomplished via generalized estimating equations. In the second study, we employ sampling theories to greatly improve the efficiency of estimating latent edges in large graphs.

In the GEEIV study, we noted that standardized residuals for generalized linear models in canonical form remain unbiased (have expectation zero) in the presence of covariate measurement error. Furthermore, for the logistic model, we showed the residuals retain the same working correlation structure employed in the absence of measurement error. When instrumental variables are available, the standardized residuals can be used to construct unbiased estimating equations. An extensive simulation study showed the GEEIV approach yield essentially unbiased estimators for most data parameter settings. The method is less successful when the outcome probability is low, the odds ratio is large, and the sample size is small. However, even under these more extreme conditions, the estimator removes most of the bias incurred from the measurement error.

The second study developed a statistical method to estimate the total number of latent defective edge in a graph with very large latent edge count. We combined a previous method designed for small graphs with sampling strategies to develop our algorithm. The resulting algorithm is able to determine an optimum subgraph size to sample, and stops when the desired precision is achieved. We tested our algorithm on both synthetic data and a power grid simulator. Unbiased estimators with predetermined precision were obtained on the synthetic data. Estimation on the power grid simulator agreed with results from a previous study using extrapolation for defective edges of size 3. The confidence interval from our method was more precise (narrower). We were also able to obtain an estimate for defective edges of size 4, which was not accomplished by the previous study owing to the very large number of defective edges of this size.

Bibliography

[1] P.G. Wright. *The Tariff on Animal and Vegetable Oils*. Investigations in international commercial policies. Macmillan, 1928.

[2] R. A. Fisher. "The use of multiple measurements in taxonomic problems". In: *Annals of Eugenics* 7.2 (1936), pp. 179–188.

[3] Robert Dorfman. "The Detection of Defective Members of Large Populations". In: 14 (Dec. 1943). Publisher: Institute of Mathematical Statistics.

[4] J. A. Nelder and R. W. M. Wedderburn. "Generalized Linear Models". In: *Journal of the Royal Statistical Society. Series A (General)* 135.3 (1972), pp. 370–384.

[5] Raymond J. Carrol et al. "On errors-in-variables for binary regression models". In: *Biometrika* 71.1 (Apr. 1984), pp. 19–25.

[6] Kung-Yee Liang and Scott L. Zeger. "Longitudinal Data Analysis Using Generalized Linear Models". In: *Biometrika* 73.1 (1986), pp. 13–22.

[7] Martin Aigner. *Combinatorial Search*. 1st edition. Stuttgart : Chichester England ; New York: Wiley, Nov. 1, 1988.

[8] Ross L. Prentice. "Correlated Binary Regression with Covariates Specific to Each Binary Observation". In: *Biometrics* 44.4 (1988), pp. 1033–1048.

[9] Kaoru Ishikawa. *Introduction to quality control / Kaoru Ishikawa.* eng. Tokyo: 3A Corp., 1990.

[10] Lawrence J. Emrich and Marion R. Piedmonte. "A Method for Generating High-Dimensional Multivariate Binary Variates". In: *The American Statistician* 45.4 (1991), pp. 302–304.

[11] Gregory Camilli. "Origin of the Scaling Constant d = 1.7 in Item Response Theory". In: *Journal of Educational and Behavioral Statistics* 19.3 (1994), pp. 293–295.

[12] Garrett M. Fitzmaurice. "A Caveat Concerning Independence Estimating Equations with Multivariate Binary Data". In: *Biometrics* 51.1 (1995), pp. 309–317.

[13] John P. Buonaccorsi. "Measurement Error in the Response in the General Linear Model". In: *Journal of the American Statistical Association* 91.434 (1996), pp. 633–642.

[14] Eberhard Triesch. "A group testing problem for hypergraphs of bounded rank". In: *Discrete Applied Mathematics* 66.2 (Apr. 30, 1996), pp. 185–188.

[15] Eberhard Triesch. "A group testing problem for hypergraphs of bounded rank". In: *Discrete Applied Mathematics* 66.2 (Apr. 30, 1996), pp. 185–188.

[16] Jeffrey S Buzas. "Instrumental variable estimation in nonlinear measurement error models". In: *Communications in Statistics - Theory and Methods* 26.12 (1997), pp. 2861–2877.

[17] C. A. Gotway and W. W. Stroup. "A Generalized Linear Model Approach to Spatial Data Analysis and Prediction". In: *Journal of Agricultural, Biological, and Environmental Statistics* 2.2 (1997), pp. 157–178.

[18] Ding-Zhu Du and Frank K Hwang. *Combinatorial Group Testing and Its Applications*. 2nd ed. Vol. 12. Series on Applied Mathematics. Dec. 1999.

[19] J.W. Hardin and J.M. Hilbe. *Generalized Estimating Equations*. CRC Press, 2002.

[20] Petra Johann. "A group testing problem for graphs with several defective edges". In: *Discrete Applied Mathematics* 117.1 (Mar. 15, 2002), pp. 99–108.

[21] Jeffrey S. Buzas, Leonard A. Stefanski, and Tor D. Tosteson. "Measurement Error". In: *Handbook of Epidemiology*. Ed. by Wolfgang Ahrens and Iris Pigeot. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 729–765.

[22] R.J. Carroll et al. *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2006.

[23] Ting Chen and Frank K. Hwang. "A competitive algorithm in searching for many edges in a hypergraph". In: *Discrete Applied Mathematics* 155 (Feb. 15, 2007).

[24] E.D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer Series in Statistics. Springer New York, 2009.

[25] John Antonakis et al. "On making causal claims: A review and recommendations". In: *The Leadership Quarterly* 21.6 (2010). Leadership Quarterly Yearly Review, pp. 1086–1120.

[26] *Juran's quality handbook : the complete guide to performance excellence / Joseph M. Juran, Joseph A. De Feo.* eng. 6th ed.. New York, 2010.

[27] D.D. Boos and L.A. Stefanski. *Essential Statistical Inference: Theory and Methods*. Springer Texts in Statistics. Springer New York, 2013.

[28] Daniel B. Larremore, Aaron Clauset, and Abigail Z. Jacobs. "Efficiently inferring community structure in bipartite networks". In: *Phys. Rev. E* 90 (1 July 2014), p. 012805.

[29] Pooya Rezaei, Paul D. H. Hines, and Margaret J. Eppstein. "Estimating Cascading Failure Risk with Random Chemistry". In: *IEEE Transactions on Power Systems* 30.5 (Sept. 2015), pp. 2726–2735.

[30] Sinan G. Aksoy, Tamara G. Kolda, and Ali Pinar. "Measuring and modeling bipartite graphs with community structure". In: *Journal of Complex Networks* 5.4 (Mar. 2017), pp. 581–603.

[31] C.A. Moser and G. Kalton. *Survey Methods in Social Investigation*. Taylor & Francis, 2017.

[32] Megan Dewar et al. *Subhypergraphs in non-uniform random hypergraphs*. 2018.

[33] Federico Battiston et al. "Networks beyond pairwise interactions: Structure and dynamics". In: *Physics Reports* 874 (2020). Networks beyond pairwise interactions: Structure and dynamics, pp. 1–92.

[34] Laurence A. Clarfeld et al. "Risk of Cascading Blackouts Given Correlated Component Outages". In: *IEEE Transactions on Network Science and Engineering* 7.3 (July 2020). Conference Name: IEEE Transactions on Network Science and Engineering, pp. 1133–1144.

[35] Office of the Commissioner. *Coronavirus (COVID-19) Update: Facilitating Diagnostic Test Availability for Asymptomatic Testing and Sample Pooling*. FDA. June 18, 2020. URL: https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-facilitating-diagnostic-test-availability-asymptomatic-testing-and.

[36] *U.S. Grid Regions*. Jan. 24, 2022. URL: https://www.epa.gov/green-power-markets/us-grid-regions.

[37] *Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations*. URL: https://www.energy.gov/sites/default/files/oeprod/DocumentsandMedia/BlackoutFinal-Web.pdf.