R-squared inference under non-normal error

Lei Xu

A dissertationsubmitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2014

Reading Committee:

Ross L Prentice, Chair

Ross L Prentice

Thomas Richardson

Jon A Wellner

Program Authorized to Offer Degree:
Department of Statistics

University of Washington

Abstract

R-squared inference under non-normal error

Lei Xu

Chair of the Supervisory Committee:
Professor Ross L Prentice
Department of Biostatistics

Assessment of the relationship between diet and health status, especially association between diet and chronic disease risk, has attracted lot of research interest in statistical and epidemiologic studies. However, due to measurement errors in commonly utilized self-reported assessment approaches, an expected strong relationship was not identified in most studies. Developments in biomarker measures provide objective consumption assessment for specific dietary components which are utilized to develop calibrated dietary consumption function to remove bias embedded in those self-reported dietary measures. Researchers are interested in the explanatory strength of calibration equations and comparison of the strengths among various self-report measures. Thus, as a common metric used in these studies, reliable estimation of R^2 and of its confidence interval are important.

Inference for R^2, including confidence intervals for R^2 has not attracted much attention in the statistical literature. In this dissertation we proposed two methods to estimate confidence intervals for R^2 under errors from normal and non-normal distributions: the first method is based on asymptotic theories and entails the development of the asymptotic distribution of R^2, and its relevant functions, when sample size becomes large; the second approach is based on a general F-test applied to linear regression but adjusts degree of freedom parameters in the F-test statistics using the empirical skewness and kurtosis of regression errors. In addition, when there are measurement errors in the independent variables, R^2 directly estimated from the regression can be biased and may, for example,

underestimate the relationship between dependent and independent variables even with normally distributed errors. This dissertation also proposes a correction methodology to reduce the bias in R^2 estimation in the presence of classical additive measurement errors. The proposed methodologies have been evaluated in simulation and applied to nutritional biomarker studies in the Women's Health Initiative.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to express my sincere and heartfelt thanks to my advisor, Ross Prentice, for his vast reserve of patience and knowledge and for his friendship. I would also like to thank my supervisory committee members, Drs. Doug Martin, Thomas Richardson and Jon Wellner, for their guidance and advice on the dissertation. Lastly I would like to express my gratitude to my parents in China and to my wife, Hua(Judy), for their patience and support throughout my studies.

DEDICATION

to my dear wife, Judy

Chapter 1

INTRODUCTION

In the statistical and public health communities, the relationship between diet and health status has attracted a lot of attention, especially the assessment of association between diet and chronic disease risk. However, due to measurement errors in dietary measures, these studies mostly did not show the expected strong relationships (Hunter et al, 1996, Fuchs et al, 1999). The development of biomarker measure allows for the objective assessment of the consumption of specific dietary components. These biomarker measures are utilized to calibrate the dietary consumption function and to remove bias present in self-reported dietary measures. Researchers are interested in the explanatory strength of calibration equations and comparison in the strength of various self-report measures. Thus, as a common metric used in these studies, reliable estimation of R^2 and the development of associated confidence intervals are important.

Inference for R^2, including confidence intervals has not attracted much attention. There are few publications on this topic. In this dissertation we propose two ways to estimate confidence intervals for R^2 under errors from normal and non-normal distributions: the first method is based on asymptotic theory and develops the asymptotic distribution of R^2 and its relevant functions when sample size is large; the second approach is based on the general F-test applied to linear regression but adjusts degree of freedom parameters in the F-test statistics using the empirical skewness and kurtosis of regression errors. In addition, when there are measurement errors in the independent variables, R^2 directly estimated from a regression can be biased and may underestimate the relationship between dependent and independent variables even with normally distributed errors. This dissertation also proposes a correction methodology to mitigate the bias in R^2 estimation. The proposed methodologies have been tested in simulation and applied to nutritional biomarker studies in the Women's Health Initiative.

In this chapter, we first briefly introduce the study background, and explains different dietary measurement methodologies and their pros and cons. Next we describe the calibration method and define different measurement error model structures for various dietary measures. Then we provide the definition of R^2 and a brief summary of proposed methodologies in this research.

1.1 Calibrated consumption estimation and biomarker development in nutritional epidemiology

A healthy diet is of great interest to the general population. Many people seek dietary recommendations to maintain or improve their general health status. There are numerous articles published in magazines, newspapers and television that claim to provide guidelines for a healthful diet. It is generally accepted that there are strong relationships between food and nutrient consumption and a favorable health status. There are already many statistical and epidemiologic studies published on this topic. However, counter-intuitively, these studies mostly did not show the expected strong relationships. For example, a study published in 1996 (Hunter et al, 1996) combined seven large cohort studies in four countries and did not find an association between the intake of total dietary fat and the risk of breast cancer. A similar observation was reported in 1999 (Fuchs et al, 1999) which did not find statistical significant relationship between the dietary fiber intake and the risk of colorectal cancer. These results have been questioned because of deficiencies in the underlying dietary data assessment methodology. One common limitation embedded in nearly all these studies may help explain this phenomenon: measurement error in dietary assessment (Fraser 2003, Kipnis et al, 2001,2009). A brief description of commonly used dietary assessment methods will be given here to help anticipate their measurement characteristics.

- Food-frequency questionnaires (FFQs), due to their efficiency and low cost, have been ubiquitous in nutritional epidemiology research for the past 25 years (Prentice et al, 2011). FFQs contain a list of beverage and food for subjects to report consumption frequency and portion size category of each item during a specified time period (e.g. past six months). Usually FFQs provides standard frequency options for subjects to

select. Ease of implementation and low cost are the major advantages of the FFQ method. Its self-administered and machine-readable features make it practical for application to large study cohorts. However, systematic bias in energy and protein consumption, for which objective biomarkers are available, have been reported (Subar et al, 2003, Neuhouser et al, 2008, Prentice et al 2011), and repeat application by the same study subject reveals a substantial "noise" component in the FFQ assessment.

- 24-hour dietary recalls (24HR) are another commonly used assessment technique. This method collects food and drink information retrospectively over the preceding 24 hours, often in an unannounced fashion via telephone. U.S. national nutritional surveys traditionally have used 24HR to collect information on food intake as the primary assessment instrument (Dwyer et al., 2003). The main purpose of such surveys is to estimate the distribution of usual (that is, average long-term) intake of nutrients and foods in the population, and to monitor such intakes over time. Another important purpose is to relate individual usual intakes to health-related outcomes such as blood pressure. There has been concern over its use for assessing intake of foods that are not typically consumed every day. Consumers of such episodically consumed foods naturally report zero intake on the 24HR if the report happens to be on a nonconsumption day. Consequently, with typically only one or two administrations of 24HRs in surveys, usual intake of such foods is difficult to estimate (Kipnis, 2009). Similar to the FFQ method, the method also entails systematic bias (e.g. Subar 2003, Prentice et al. 2011).

- Dietary records are another self-reported dietary assessment method. This approach requests participants to record the types and amounts of food and beverage at each consumption over a specified time period, often one or multiple consecutive days (usually no more than 7 days). Computerized programs have been developed to facilitate the food record utilization and entry of consumption amounts. Ideally, if the respondent records the food and beverage intakes at the time of eating, this approach has potential advantages over other self-reported measures since it may report more ac-

curate portion size and more fully list food some of which may be missed by other methods that rely on memory such as FFQs and 24 hour recalls. A successful dietary record requires cooperation and commitments from the respondents. Thus the approach is subject to potential sample selection bias and measurement error as well. Several studies (Subar et al, 2003, Prentice et al, 2011) already showed underreporting of energy and protein intakes on the diet records by comparing reported intakes against objective measures obtained from biomarker methods: doubly-labeled water for total energy consumption and urinary nitrogen for protein intake. For each of these self-reported measures, studies have found that underreporting of energy and protein was associated with participant characteristics such as body mass index (BMI), gender and age (Subar et al, 2003, Prentice et al,2011).

- Biomarker measures of specific dietary components provides an objective consumption assessment, typically over a time period of one or a few days. A biomarker is an indicator of biological status, and it is often considered as providing an objective assessment of its targeted dietary factor. There are a small number of nutrients for which a well established biomarker of short-term consumption has been developed, including a doubly-labeled water (DLW) assessment of energy consumption over a 14 day period (Schoeller 1999), and a urinary nitrogen (UN) assessment of protein consumption from a 24 hour urine collection (Bingham 2003). These urinary recovery biomarkers (Kaaks et al. 2002) provide objective estimates of short-term consumption among persons in energy balance, with measurement error that is plausibly independent of study subject characteristics such as body mass index (BMI), age, and gender, and importantly is plausibly independent also of dietary self-report measurement error. However, it is not practical to obtain these biomarkers for the tens of thousands of persons in a typical epidemiology cohort study, and application in advance of disease diagnosis is essential to avoid influences due to the presence of disease, or its sequelae. Other putative biomarkers include 24-hour urinary recovery for sodium and potassium consumption.

Due to assessment method-specific properties of the measurement error, different mathematical models are required to develop a measurement error model for various data collection methods. Let Z denote the actual target dietary assessment measure such as average nutrient consumption over a defined time period, Q denote the corresponding biomarker measure, and W represent the dietary assessment measure obtained from self-reported assessment methods such as FFQs, 24 hour recalls or diet records. Also let V denote individual characteristics that may relate to measurement error properties of the self-report, or may usefully augment the self-report. Assuming that the biomarker measure is an objective assessment that adheres to a classical measurement error model, Q can be represented as

$$Q = Z + \epsilon \tag{1.1}$$

But for the self-reported measure W, the measurement error model needs to be more flexible. Misreporting has been found in many studies. For example, one observational study (Subar et al, 2003) conducted from September 1999 to March 2000 showed that males and females underreported substantially both intake of energy and protein. That is, beyond random error, systematic bias has been identified. Maurer et al (2006) reviewed studies on energy misreporting from self-reported assessment methodologies. By investigating characteristics of energy misreporters and summarizing evidence from published papers, the review showed that several psychosocial factors had significant influence on the reported dietary intake. For example, females were more likely to underreport energy intake than males; and history of dieting and overweight might be associated with underreporting. Other studies also found similar associations between measurement error with participant characteristics (e.g., Heitmann and Lissner 1995, Horner et al, 2002, Neuhouser et al, 2008 and Prentice et al 2011). Due to systematic bias emebbed in the measurement error, a classical measurement error model is not appropriate. Instead, participant characteristics typically need to be included in the model and W can be modeled as (Prentice et al. 2002):

$$W = \beta_0 + \beta_Z Z + \beta_V V + \mu \tag{1.2}$$

In 1.1 and 1.2, ϵ and μ are random errors that are independent from other right side variables given V. In addition, it is reasonable to assume that measurement error associated with biomarker measure ϵ is independent from self-reported dietary assessment measurement error μ and independent of Z given V.

In order to improve quality of analytic results, it is necessary to take advantage of the strengths of each data collection method, while compensating for their shortcomings at the same time. To achieve this objective, some research papers propose methods to develop calibrated dietary consumption using the biomarker measure in a subsample of a study cohort having data (W, Q, V) available. Then the method uses the calibration equation to estimate dietary consumption from the self-reported measure on the larger study cohort for use in disease association analyses. In a practical study design, biomarkers will be measured on a subsample of a study cohort, then the calibration equation will be developed on the subsample, which will be applied to concurrent self-report data of the whole cohort in order to produce measurement error corrected consumption estimates. In the calibration method, the property of the calibration equation will affect the performance and properties of the calibrated consumption estimates in the overall study cohort. Thus, it is critical to have a good measure to assess the statistical properties of calibration equations. A useful approach to calibration equation assessment includes examining R^2, the fraction of biomarker (usually log-transformed) explained by the calibration equation. Additional considerations, including whether available data include all needed components in V, may also be important.

1.2 Choices for the definition of R-squared

In statistics, the coefficient of determination R^2 is used in the context of statistical models whose main purpose is the prediction of future outcomes on the basis of other related information. It is the proportion of variability in a data set that is accounted for by the statistical model. It provides a measure of how well future outcomes are likely to be pre-

dicted by the model. The assessment of a potential calibration equation typically involves a data set having values y_i, with $i = 1, \ldots, n$ and each of which has an associated predicted value $\widehat{y_i}$. The average value of the y values in the dataset is denoted as $\overline{y}$. Then variability of the data set can be measured through various sums of squares:

- Total sum of squares, denoted as SST, is defined as $\sum_{i=1}^{n}(y_i - \overline{y})^2$
- Sum of squares of residuals, denoted as SSE, is defined as $\sum_{i=1}^{n}(y_i - \widehat{y_i})^2$
- Explained sum of squares, denoted as SSM, is defined as $\sum_{i=1}^{n}(\widehat{y_i} - \overline{y})^2$

In general, R^2 is defined as $1 - \frac{SSE}{SST}$. In some situations, $SST = SSM + SSE$. In such cases, $R^2 = 1 - \frac{SSE}{SST} = \frac{SSM}{SST}$, and it is between 0 and 1. In the regression model context, this happens when the dependent variable Y is a linear function of predictor variables X with random error ϵ, i.e., $Y = \beta X + \epsilon$ in mathematical form. β is then estimated by minimizing the sum of squared error, $\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n}(y_i - \beta X_i)^2$. The solution to this optimization problem is well known, $\widehat{\beta} = (X'X)^{-1}X'Y$, and the predicted $\widehat{Y}$ is $\widehat{Y} = X(X'X)^{-1}X'Y$. Without loss of generality and for convenience, the average value $\bar{Y}$ can be set to 0. In addition, the SST, SSE and SSM can be written in matrix form as:

$$SST = Y'Y, SSE = Y'(I - X(X'X)^{-1}X')Y, SSM = Y'X(X'X)^{-1}X'Y \tag{1.3}$$

Correspondingly, the R^2 of a linear regression model in matrix form is $R^2 = 1 - \frac{SSE}{SST} = \frac{Y'X(X'X)^{-1}X'Y}{Y'Y}$.

R^2 in this form is easy to understand and has an intuitive appeal. Thus this measure has been used extensively, especially in linear regression. However, due to its definition, R^2 is typically used in relation to continuous response variables. For binary or categorical responses, R^2 may not be very meaningful. Nagelkerke (1991) provided a generalized definition of R^2 that is based on likelihood ratios. This R^2 is defined as one minus the ratio of the likelihood under the null model to the likelihood of the fitted model. This generalized R^2 is consistent with classical R^2 when the data is from a linear model with normal distribution,

but it has a much broader application scope. This dissertation will focus on estimation of the linear model form of R^2.

1.3 Confidence intervals for R-squared

Development of confidence interval for R^2 has not attracted much attention. There are few publications on this topic. Since R^2 increases when adding variables to the regression model, some may think that confidence interval for R^2 may not be meaningful. However, when R^2 is used to assess model calibration equation performance, such a confidence interval may be quite helpful, especially when there are measurement errors in the data in which case R^2 from naive regression is biased even with normally distributed error. Correction of R^2 and estimation of the scorresponding interval is especially important in the presence of measurement error.

Supppose that the independent variable X is k dimensional, i.e. there are k variables in the regression model. Two methodologies can be considered to estimate the confidence interval for R^2:

- First, when the underlying data are from a normal distribution, it is well known that $\frac{R^2}{(1-R^2)}\frac{(n-k)}{k}$ follows an F distribution with degree of freedom k and $(n-k)$ where n is the number of observations. The confidence interval of R^2 can then be derived from the corresponding F-distribution.

- Another way is to calculate the interval from an asymptotic distributional approximation. Based on a normal distribution assumption for y, it is possible to estimate asymptotic distributions of R^2, or a suitable function of R^2, which then is used as a base to calculate corresponding confidence intervals.

However, limitation of these methods is that they are based on normal distributional assumption. When the response data are from other non-normal distributions, especially those distributions that are high skewed and/or heavy tailed, these methods may not work well. The estimation of R^2 under a non-normal error distribution, with or without measurement error in predictor variables, is the focus of this dissertation.

1.4 Dissertation outline

The dissertation is focusing on statistical inference on R-squared, especially on the confidence interval development under linear models with error distributions that may be non-normal. Theory development is discussed in Chapter 2 with asymptotic distributions and F distribution approximation after degrees of freedom adjustment. Chapter 3 shows simulation results under various distributional assumptions to assess the theoretical approximation in Chapter 2 in moderate sized samples. The R-squared correction and inference under measurement error are discussed in Chapter 4. Chapter 5 presents simulation results under measurement error. The application of the proposed methods to the Women's Health Initiative nutritional biomarker studies is discussed in Chapter 6. Limitations of the research and future research topics are discussed in Chapter 7.

Chapter 2

CONFIDENCE INTERVAL ESTIMATION FOR R-SQUARED UNDER NON-NORMAL ERROR

As discussed in Chapter 1, the coefficient of determination R^2 in a linear model context is the proportion of variability in a data set that is accounted for by the statistical model. Considering a data set having values y_i, with $i = 1, , n$, each of which with an associated predicted value $\widehat{y_i}$, its total sum of squares defined as $SST = \sum_{i=1}^{n}(y_i - \overline{y})^2$ can be decomposed to two components, SSE and SSM, defined as $SSE = \sum_{i=1}^{n}(y_i - \widehat{y_i})^2$ and $SSM = \sum_{i=1}^{n}(\widehat{y_i} - \overline{y})^2$ respectively. In these formulas, $\overline{y}$ denotes the average of y_i. R^2 is usually defined as $1 - \frac{SSE}{SST}$.

In the linear regression model context, it assumes that the dependent variable y is a function of predictor variables X, represented as $Y = \beta X + \epsilon$. β is estimated by minimizing the sum of squared error, $\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n}(y_i - \beta X_i)^2$. To solve the optimization problem, we can get the well known result that: $\widehat{\beta} = (X'X)^{-1}X'Y$, and the predicted $\widehat{Y}$ is $\widehat{Y} = X(X'X)^{-1}X'Y$. Without loss of generality and for convenience, average value $\bar{Y}$ is set to 0. The SST, SSE and SSM can be formulated in matrix form as presented below

$$SST = Y'Y, SSE = Y'(I - X(X'X)^{-1}X')Y, SSM = Y'X(X'X)^{-1}X'Y \tag{2.1}$$

Correspondingly, the R^2 of a regresion model in a matrix form is, $R^2 = 1 - \frac{SSE}{SST} = \frac{Y'X(X'X)^{-1}X'Y}{Y'Y}$. Atiqullah(1962,1964) developed theorems about the mean and variance for matrix forms, $Y'AY$. Our study follows Atiqullah's work to develop degree of freedom adjustments for approximating F distributions under nonnormal errors, and to approximates the variance of R^2 using a matrix form. This work proceeds by calculating the variance of the denominator and numerator, and the covariance between the denominator and nu-

merator in matrix form; from which approximations to the mean and variance of R^2, or its suitable functions are developed. Confidence intervals are derived two ways: asymptotic approximation assuming large sample size, and F-distribution with adjusted degree of freedom for possible moderate sample size improvement to confidence interval approximation. Section 2.1 discusses variance and covariance of matrix calculation, and confidence interval development in asymptotic form and F-distribution adjustment is discussed in section 2.2 and section 2.3 respectively.

2.1 Variance and Covariance under Matrix Calculation

Atiqullah(1962,1964) calculated the mean and variance of $Y'AY$, where Y is a vector of n independent random variables, $y_1, ..., y_n$, and A denotes a n by n matrix. In this thesis, skewness for a random variable X with mean μ and variance σ^2 is defined as $E[\frac{(X-\mu)^3}{\sigma^3}]$, and kurtosis for X is defined as the extra kurtosis from normal distribution, $E[\frac{(X-\mu)^4}{\sigma^4}] - 3$.The result is presented below.

Theorem 1. *Let y be a column vector of n independent observations from an identical distribution, and let A and B be n by n matrices. Then*

$$E(y'Ay) = \sigma^2 tr(A) + \mu' A\mu \tag{2.2}$$

$$Var(y'Ay) = 2\sigma^4\{tr(A^2) + \frac{1}{2}\gamma_2 a'a + 2\mu' A^2\mu/\sigma^2 + 2\gamma_1\mu' Aa/\sigma\} \tag{2.3}$$

$$Cov(y'Ay, y'By) = 2\sigma^4\{tr(AB) + \frac{1}{2}\gamma_2 a'b + 2\mu' AB\mu/\sigma^2 + \gamma_1(\mu' Ba + \mu' Ab)/\sigma\} \tag{2.4}$$

where μ is the expectation of the vection y, i.e. $E(y) = \mu$, and $Var(y) = \sigma^2$, and $tr(A)$ is the trace of the matrix A, that is the sum of the diagonal values of A. In equation (2.3) and (2.4), γ_2 is the kurtosis of y, γ_1 is the skewness of y, and a and b are the diagonal vectors for A and B respectively.

In Atiqullah's development, each y_i is from an identical distribution with specified moments. We relax this feature and extend Atiqullah's results by allowing y_is to be from different distributions with various moments, including variance, skewness and kurtosis.

The following theorem shows the mean, variance and covariance under these more general conditions.

Theorem 2. *Let y be a column vector of n independent observations, and A and B are n by n matrices. Let $U(y) = \{\mu_1, ..., \mu_n\}$ where $\mu_i = E(y_i)$; $\Gamma_1 = \{\gamma_1^1, ..., \gamma_1^n\}$, where $\gamma_1^i = skewness(y_i)$; $\Gamma_2 = \{\gamma_2^1, ..., \gamma_2^n\}$, where $\gamma_2^i = kurtosis(y_i)$; $\Sigma = \{\sigma_1^2, ..., \sigma_n^2\}$, where $\sigma_i^2 = var(y_i)$. And a is diagonal vector of the matrix A, and b is diagonal vector of the matrix B; Also denote $\Sigma_3 = \{\sigma_1^3, ..., \sigma_n^3\}$. Then*

$$E(y'Ay) = \Sigma' a + U'AU, \tag{2.5}$$

$$\begin{aligned} Var(y'Ay) &= \Sigma'(diag(a) \times diag(a) \times diag(\Gamma_2))\Sigma + 2\Sigma'(A \otimes A)\Sigma \\ &+ 4\Sigma_3'(diag(\Gamma_1) \times diag(a) \times A)U + 4U'(A \times diag(\Sigma) \times A)U, and \end{aligned} \tag{2.6}$$

$$\begin{aligned} Cov(y'Ay, y'By) &= \Sigma'(diag(a) \times diag(b) \times diag(\Gamma_2))\Sigma + 2\Sigma'(A \otimes B)\Sigma \\ &+ 2\Sigma_3'(diag(\Gamma_1) \times diag(a) \times B + diag(\Gamma_1) \times diag(b) \times A)U + 4U'(A \times diag(\Sigma) \times B)U \end{aligned} \tag{2.7}$$

where $diag(a)$ is a diagonal matrix with a as the diagonal vector. $diag(\Gamma_1)$, $diag(\Sigma)$ and $diag(\Gamma_2)$ are diagonal matrices with defined diagonal vectors, and $A \otimes B = (a_{ij}b_{ij})_{n\times n}$ and $A \otimes A = (a_{ij}^2)_{n\times n}$.

Proof. $y'Ay = \sum_{i=1}^n a_{ii}y_i^2 + \sum_{i=1}^n \sum_{\substack{j=1 \\ j\neq i}}^n a_{ij}y_iy_j$, so that

$E(y'Ay) = \sum_{i=1}^n a_{ii}\sigma_i^2 + \sum_{i=1}^n \sum_{j=1}^n a_{ij}\mu_i\mu_j = \Sigma' a + U'AU$, and

$var(y'Ay) = cov(\sum_{i=1}^n a_{ii}y_i^2 + \sum_{i=1}^n \sum_{\substack{j=1 \\ j\neq i}}^n a_{ij}y_iy_j, \sum_{i=1}^n a_{ii}y_i^2 + \sum_{i=1}^n \sum_{\substack{j=1 \\ j\neq i}}^n a_{ij}y_iy_j)$

$= \sum_{i=1}^n a_{ii}^2(\gamma_2^i\sigma_i^4 + 2\sigma_i^4 + 4\sigma_i^2\mu_i^2 + 4\gamma_1^i\sigma_i^3\mu_i)$

$+ 4\sum_{i=1}^n \sum_{\substack{j=1 \\ j\neq i}}^n (\gamma_1^i\sigma_i^3\mu_m + 2\sigma_i^2\mu_i\mu_m)a_{ii}a_{im}$

$+ 2\sum_{i=1}^n \sum_{\substack{j=1 \\ j\neq i}}^n (\sum_{\substack{m=1 \\ m\neq i \\ m\neq j}} a_{ij}a_{im}\sigma_i^2\mu_j\mu_m + \sum_{\substack{m=1 \\ m\neq i \\ m\neq j}} a_{ij}a_{jm}\sigma_j^2\mu_i\mu_m + a_{ij}^2\mu_i^2\sigma_j^2 + a_{ij}^2\mu_j^2\sigma_i^2 + a_{ij}^2\sigma_j^2\sigma_i^2)$

$= \sum_{i=1}^n a_{ii}^2\gamma_2^i\sigma_i^4 + 2\sum_{i=1}^n \sum_{j=1}^n a_{ij}^2\sigma_i^2\sigma_j^2 + 4\sum_{i=1}^n \sum_{m=1}^n \gamma_1^i\sigma_i^3\mu_m a_{ii}a_{im}$

$+ 4\sum_{i=1}^n \sum_{j=1}^n \sum_{m=1}^n a_{ij}a_{im}\sigma_i^2\mu_j\mu_m$

$$= \Sigma'(diag(a) \times diag(a) \times diag(\Gamma_2))\Sigma + 2\Sigma'(A \otimes A)\Sigma$$
$$+ 4\Sigma_3'(diag(\Gamma_1) \times diag(a) \times A)U + 4U'(A \times diag(\Sigma) \times A)U$$
$$cov(y'Ay, y'By) = cov(\textstyle\sum_{i=1}^n a_{ii}y_i^2 + \sum_{i=1}^n \sum_{\substack{j=1 \\ j\neq i}}^n a_{ij}y_iy_j, \sum_{i=1}^n b_{ii}y_i^2 + \sum_{i=1}^n \sum_{\substack{j=1 \\ j\neq i}}^n b_{ij}y_iy_j)$$
$$= \textstyle\sum_{i=1}^n a_{ii}b_{ii}(\gamma_2^i\sigma_i^4 + 2\sigma_i^4 + 4\sigma_i^2\mu_i^2 + 4\gamma_1^i\sigma_i^3\mu_i)$$
$$+ 2\textstyle\sum_{i=1}^n \sum_{\substack{j=1 \\ j\neq i}}^n (\gamma_1^i\sigma_i^3\mu_m + 2\sigma_i^2\mu_i\mu_m)a_{ii}b_{im}$$
$$+ 2\textstyle\sum_{i=1}^n \sum_{\substack{j=1 \\ j\neq i}}^n (\gamma_1^i\sigma_i^3\mu_m + 2\sigma_i^2\mu_i\mu_m)b_{ii}a_{im}$$
$$+2\textstyle\sum_{i=1}^n \sum_{\substack{j=1 \\ j\neq i}}^n (\sum_{\substack{m=1 \\ m\neq i \\ m\neq j}} a_{ij}b_{im}\sigma_i^2\mu_j\mu_m + \sum_{\substack{m=1 \\ m\neq i \\ m\neq j}} a_{ij}b_{jm}\sigma_j^2\mu_i\mu_m + a_{ij}b_{ij}\mu_i^2\sigma_j^2 + a_{ij}b_{ij}\mu_j^2\sigma_i^2 + a_{ij}b_{ij}\sigma_j^2\sigma_i^2)$$
$$= \textstyle\sum_{i=1}^n a_{ii}b_{ii}\gamma_2^i\sigma_i^4 + 2\sum_{i=1}^n \sum_{j=1}^n a_{ij}b_{ij}\sigma_i^2\sigma_j^2 + 2\sum_{i=1}^n \sum_{m=1}^n \gamma_1^i\sigma_i^3\mu_m a_{ii}b_{im}$$
$$+ 2\textstyle\sum_{i=1}^n \sum_{m=1}^n \gamma_1^i\sigma_i^3\mu_m b_{ii}a_{im} + 4\sum_{i=1}^n \sum_{j=1}^n \sum_{m=1}^n a_{ij}b_{im}\sigma_i^2\mu_j\mu_m$$
$$= \Sigma'(diag(a) \times diag(b) \times diag(\Gamma_2))\Sigma + 2\Sigma'(A \otimes B)\Sigma$$
$$+ 2\Sigma_3'(diag(\Gamma_1) \times diag(a) \times B + diag(\Gamma_1) \times diag(b) \times A)U + 4U'(A \times diag(\Sigma) \times B)U$$

□

2.2 Moderate sample size improvements using an F-distribution approximation

For linear regression model, $Y = \beta X + \epsilon$ with normal error and k predictors, $\frac{R^2/k}{1-R^2/(n-k)} = \frac{Y' \times X(X'X)^{-1}X' \times Y/k}{Y' \times (I-X(X'X)^{-1}X') \times Y/(n-k)}$ follows F distribution with degree of freedoms as k and $(n-k)$. Noncentrality component of the F distribution is $\beta'X'X\beta$; k is number of paramters used in the model; and n is the overall sample size. However, when the error term does not follow normal distribution, simulation results show that the agreement between F-distribution and the distribution of this function of R^2 may be poor. One intuitive way to adapt R^2 distribution to non-normal error is to adjust the degrees of freedom and the non-central component of the F-distribution to approximate the actual distribution. Here we use a moment-matching method to develop degrees of freedom approximations. This entails approximation of first order and second order moments of the actual distribution, and adjustment of parameters in the F-distribution to match these first and second-order moments. The first order and the second order moments are approximated through Taylor expansion and are calculated based on the matrix form results presented in Theorem 1 and Theorem 2. Considering the statistic $\frac{y'Ay/n_A}{y'By/n_B}$, with n_A and n_B the sum of diagonal values of

A and B respectively. For convenience, let X_A define $y'Ay/n_A$ and X_B define $y'By/n_B$, and the corresponding mean and variance are defined as (μ_A, σ_A^2), and (μ_B, σ_B^2). σ_{AB} denotes the covariance between the two statistics $(y'Ay/n_A, y'By/n_B)$. Then by Taylor expansion techinique to second-order derivative around $\frac{\mu_A}{\mu_B}$, $\frac{y'Ay/n_A}{y'By/n_B}$ can be approximated by,

$$\frac{y'Ay/n_A}{y'By/n_B} = \frac{X_A}{X_B} \approx \frac{\mu_A}{\mu_B} - \frac{\mu_A}{\mu_B^2}(X_B - \mu_B) + \frac{1}{\mu_B}(X_A - \mu_A) + \frac{\mu_A}{\mu_B^3}(X_B - \mu_B)^2 - \frac{1}{\mu_B^2}(X_A - \mu_A)(X_B - \mu_B) \tag{2.8}$$

Thus, the approximation of $E(\frac{y'Ay/n_A}{y'By/n_B})$ is

$$E(\frac{y'Ay/n_A}{y'By/n_B}) \approx \frac{\mu_A}{\mu_B} + \frac{\mu_A}{\mu_B^3}\sigma_B^2 - \frac{1}{\mu_B^2}\sigma_{AB} \tag{2.9}$$

If $A \times B = 0$ and replace the mean and variance with the formulas presented in Theorem 1 and 2, then

$$\begin{aligned} E(\frac{y'Ay/n_A}{y'By/n_B}) \approx \frac{1}{\sigma^6}\{\sigma^6[1 + \frac{2}{n_B^2}(n_B + \frac{1}{2}\gamma_2 b'b) - \frac{2}{n_A n_B}(\frac{1}{2}\gamma_2 a'b + \gamma_1 u'Ab/\sigma)] \\ + \sigma^4[\frac{\mu'A\mu}{n_A} + \frac{\mu'A\mu}{n_A} \times \frac{2}{n_B^2}(n_B + \frac{1}{2}\gamma_2 b'b)]\} \end{aligned} \tag{2.10}$$

Similar process is applied to $E((\frac{y'Ay/n_A}{y'By/n_B})^2)$ to get the corresponding approximation, as

$$\begin{aligned} (\frac{y'Ay/n_A}{y'By/n_B})^2 = \frac{X_A^2}{X_B^2} \approx \frac{\mu_A^2}{\mu_B^2} - 2\frac{\mu_A^2}{\mu_B^3}(X_B - \mu_B) + 2\frac{\mu_A}{\mu_B^2}(X_A - \mu_A) + \frac{1}{\mu_B^2}(X_A - \mu_A)^2 \\ + 3\frac{\mu_A^2}{\mu_B^4}(X_B - \mu_B)^2 - 4\frac{\mu_A}{\mu_B^3}(X_A - \mu_A)(X_B - \mu_B) \end{aligned} \tag{2.11}$$

Thus, the approximation of $E[(\frac{y'Ay/n_A}{y'By/n_B})^2]$ is

$$E[(\frac{y'Ay/n_A}{y'By/n_B})^2] \approx \frac{\mu_A^2}{\mu_B^2} + \frac{1}{\mu_B^2}\sigma_A^2 + 3\frac{\mu_A^2}{\mu_B^4}\sigma_B^2 - 4\frac{\mu_A}{\mu_B^3}\sigma_{AB} \tag{2.12}$$

When $A \times B = 0$ and one puts the variance and mean formulas to the above equation,

the approximation reduces to

$$E((\frac{y'Ay/n_A}{y'By/n_B})^2) = \frac{1}{\sigma^8}\{\sigma^8[1+\frac{2}{n_A^2}(n_A+\frac{1}{2}\gamma_2 a'a+2\gamma_1\mu'Aa/\sigma)-\frac{8}{n_An_B}(\frac{1}{2}\gamma_2 a'b+\gamma_1 u'Ab/\sigma)]$$
$$+\sigma^6[2*\frac{\mu'A\mu}{n_A}+4*\frac{\mu'A\mu}{n_A^2}-\frac{\mu'A\mu}{n_A}\times\frac{8}{n_An_B}(\frac{1}{2}\gamma_2 a'b+\gamma_1\mu'Ab/\sigma)+\frac{12*\mu'A\mu}{n_An_B^2}(n_B+\frac{1}{2}\gamma_2 b'b)]$$
$$+\sigma^4[\frac{\mu'A\mu}{n_A}+\frac{\mu'A\mu}{n_A}\times\frac{2}{n_B^2}(n_B+\frac{1}{2}\gamma_2 b'b)]+\frac{6\sigma^8}{n_B^2}(n_B+\frac{1}{2}\gamma_2 b'b)]\} \tag{2.13}$$

Then one can adjust the degree freedom according to:

$$\frac{2}{n_B^{adj}} = \frac{\sigma^2[1+\frac{2}{n_B^2}(n_B+\frac{1}{2}\gamma_2 b'b)-\frac{2}{n_An_B}(\frac{1}{2}\gamma_2 a'b+\gamma_1\mu'Ab/\sigma)+\lambda+\frac{2\lambda}{n_B^2}(n_B+\frac{1}{2}\gamma_2 b'b)]}{\sigma^2+\lambda}-1 \tag{2.14}$$

$$n_A^{adj} = \frac{2\sigma^8-\frac{6\sigma^8}{n_B}(\frac{1}{2}\gamma_2 a'b+\gamma_1\mu'Ab/\sigma)-\frac{6\sigma^6\lambda}{n_B}(\frac{1}{2}\gamma_2 a'b+\gamma_1\mu'Ab/\sigma)+4\sigma^6\lambda}{2\frac{\sigma^8}{n_A^2}(n_A+\frac{1}{2}\gamma_2 a'a+2\gamma_1\mu'Ab/\sigma)-\frac{(8\sigma^8+8\sigma^6\lambda)}{n_An_B}(\frac{1}{2}\gamma_2 a'b+\gamma_1\mu'Ab/\sigma)+\frac{4\sigma^6\lambda}{n_A}} \tag{2.15}$$

and the adjusted non-centrality parameter ncp_{adj} is

$$ncp_{adj} = \mu'A\mu\frac{n_A^{adj}}{n_A} \tag{2.16}$$

2.3 Confidence interval estimation using asymptotic distributional approximation

Degree of freedom adjustments are considered for moderate sample size adjustment to the distribution of R^2 under nonnormal error. For large sample size datasets, asymptotic distributional approximation may have better performance. Thus in this section, the confidence interval estimation for R^2 based on its asymptotic distribution is developed. For linear regression model for $Y = \beta X + \epsilon$, then R-squared can be calculated as $R^2 = \frac{Y'\times X(X'X)^{-1}X'\times Y}{Y'\times I\times Y}$, where the I denotes the identity matrix of dimension $n \times n$; and in correspondence $\frac{R^2}{(1-R^2)} = \frac{Y'X(X'X)^{-1}X'Y}{Y'(I-X(X'X)^{-1}X')Y)}$.

Theorem 3. *When $Y = \beta X + \epsilon$ and assuming X is from a k dimensional distibution with mean 0, finite covariance matrix and fourth moment, and residual error ϵ is from a distribution with mean 0, finite variance denoted as σ^2 and finite fourth moment, then $\frac{R^2}{(1-R^2)}$ converges to a normal distribution when sample size n goes to ∞,*

$$\sqrt{n}(\frac{R^2}{1-R^2} - \frac{\beta'\Sigma_{XX}\beta}{\sigma^2}) \sim N\{0, \frac{Var_1}{\sigma^4} + \frac{Var_d \times (\beta'\Sigma_{XX}\beta)^2}{\sigma^8}\}, where \tag{2.17}$$

$$Var_1 = 4\sigma^2\beta'\Sigma_{XX}\beta + \sum_{h=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p}\sum_{m=1}^{p}\beta_h\beta_k\beta_l\beta_m(E(X_hX_kX_lX_m) - E(X_hX_k)E(X_lX_m)), and$$

$$Var_d = (\gamma_\epsilon^2 + 2)\sigma^4$$

Proof. $\frac{R^2}{(1-R^2)} = \frac{Y'X(X'X)^{-1}X'Y}{Y'(I-X(X'X)^{-1}X')Y)} = \frac{Y'X(X'X)^{-1}X'Y/n}{Y'(I-X(X'X)^{-1}X')Y)/n}$

X is $n \times p$ matrix denoted as $\begin{pmatrix} x_1 \\ ... \\ x_n \end{pmatrix}$, where x_i is a $1 \times p$ vector, and β denote a $p \times 1$ coefficient vector.

$Y'X(X'X)^{-1}X'Y = (X\beta)'X\beta + 2(X\beta)'\epsilon + \epsilon'X(X'X)^{-1}X'\epsilon$

$= \epsilon'X(X'X)^{-1}X'\epsilon + \sum_{i=1}^{n}\beta'x_i'(2\epsilon_i + x_i\beta)$

For convenience, let V_1 denote $\sum_{i=1}^{n}\beta'x_i'(2\epsilon_i + x_i\beta)$ and V_2 be $\epsilon'X(X'X)^{-1}X'\epsilon$. The denominator, $Y'(I - X(X'X)^{-1}X')Y = \epsilon'\epsilon = \sum_{i=1}^{n}\epsilon_i^2$. Note that

ϵ_i^2 follows distribution with mean σ^2 and variance as $(\gamma_\epsilon^2 + 2)\sigma^4$ denoted as Var_d.

Also, $\beta'x_i'(2\epsilon_i + x_i\beta)$ has mean $\beta'\Sigma_{XX}\beta$ and variance

$4\sigma^2\beta'\Sigma_{XX}\beta + \sum_{h=1}^{p}\sum_{k=1}^{p}\sum_{l=1}^{p}\sum_{m=1}^{p}\beta_h\beta_k\beta_l\beta_m(E(X_hX_kX_lX_m) - E(X_hX_k)E(X_lX_m))$ denoted as Var_1. $Cov(\beta'x_i'(2\epsilon_i + x_i\beta), \epsilon_i^2) = 0$

Then, when n goes to ∞, by central limit theorem(Casella et al,2001),

$$\sqrt{n}\begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}\beta'x_i'(2\epsilon_i + x_i\beta) - \beta'\Sigma_{XX}\beta \\ \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 - \sigma^2 \end{pmatrix} \sim N\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} Var_1 & 0 \\ 0 & Var_d \end{pmatrix}\}$$

Then by delta method, we can get,

$$\sqrt{n}(\frac{\sum_{i=1}^{n}\beta' x_i'(2\epsilon_i + x_i\beta)}{\sum_{i=1}^{n}\epsilon_i^2} - \frac{\beta'\Sigma_{XX}\beta}{\sigma^2}) \sim N\{0, \frac{Var_1}{\sigma^4} + \frac{Var_d * (\beta'\Sigma_{XX}\beta)^2}{\sigma^8}\}$$

For the remaining term V_2 as $\epsilon' X(X'X)^{-1}X'\epsilon$, since trace of $X(X'X)^{-1}X'$ is number of parameters in the regression model, p, and $\frac{p}{n} \to 0$ as $n \to \infty$. In addition $X(X'X)^{-1}X'$ is a symmetric positive definite matrix, similar to diagonal matrix with only p 1s. Thus $\frac{\epsilon' X(X'X)^{-1}X'\epsilon}{\sqrt{n}} \to 0$ as $n \to \infty$.
Thus, by adding back the remaining item, we can get

$$\sqrt{n}(\frac{R^2}{1-R^2} - \frac{\beta'\Sigma_{XX}\beta}{\sigma^2}) \sim N\{0, \frac{Var_1}{\sigma^4} + \frac{Var_d \times (\beta'\Sigma_{XX}\beta)^2}{\sigma^8}\}$$

□

Therefore, when sample size n is sufficient large, it will be reasonable to use normal distribution to approximate function of $\frac{R^2}{(1-R^2)}$, such as $log(\frac{R^2}{(1-R^2)})$, by the delta method. In real application, it often assumed that Xs are given and fixed. In this case, theorems defined in this chapter can be used to approximate the asymptotic distribution of $\frac{R^2}{1-R^2}$. We can also apply delta method to the asymptotic distributions to get the asymptotic distribution of $\frac{R^2}{(1-R^2)}$ and its differentiable L^1 type functions. The asymptotic approximation in the following lemma will be evaluated in simulation studies in the next chapter.

Lemma 1. *When $Y = \beta X + \epsilon$ and assuming X given and let Σ_{XX} denote the covariance matrix of X, $Y'X(X'X)^{-1}X'Y$ and $Y'(I - X(X'X)^{-1}X')Y$ converge to normal distribution when sample size n goes to ∞. Specifically,*

$$\begin{aligned}
&\sqrt{n}(Y'X(X'X)^{-1}X'Y/n - \beta'\Sigma_{XX}\beta) \to \mathcal{N}(0, 4\sigma^2\beta'\Sigma_{XX}\beta) \\
&\sqrt{n}(Y'(I - X(X'X)^{-1}X')Y/n - \sigma^2) \to \mathcal{N}(0, (\gamma_\epsilon^2 + 2)\sigma^4) \\
&\sqrt{n}(\frac{R^2}{1-R^2} - \frac{\beta'\Sigma_{XX}\beta}{\sigma^2}) \to \mathcal{N}(0, \frac{4\sigma^2\beta'\Sigma_{XX}\beta}{\sigma^4} + \frac{(\gamma_\epsilon^2 + 2)(\beta'\Sigma_{XX}\beta)^2}{\sigma^4})
\end{aligned} \tag{2.18}$$

$$\sqrt{n}(log(\frac{R^2}{1-R^2}) - log(\frac{\beta'\Sigma_{XX}\beta}{\sigma^2})) \to \mathcal{N}(0, \frac{4\sigma^2}{\beta'\Sigma_{XX}\beta} + (\gamma_\epsilon^2 + 2)) \tag{2.19}$$

.

Chapter 3

SIMULATION EVALUATION OF R-SQUARED ESTIMATES AND CONFIDENCE INTERVALS

As described in the previous chapter, the matrix form for R^2 is $R^2 = \frac{Y'X(X'X)^{-1}X'Y}{Y'Y}$, and it has been shown in Chapter 2 that when sample size goes to infinity, the distributions of R^2 converges to a normal distribution. In Chapter 2, we also showed that after degree of freedom adjustments, first and second-order moments of an adjusted F-distribution can match those from the distribution of R-squared. Confidence interval of the R^2 hence will be estimated in two ways: based on the asymptotic normal distribution with kurtosis and skewness estimated from empirical data and regression residuals, and based on the adjusted F distribution. A simulation study was conducted to test the moderate sample size properties of the approximations discussed in the previous chapter.

In real applications, the confidence interval derived from empirical estimation using asymptotic normal distribution or adjusted F distribution, can violate the constraints of R^2 (over 1 or negative). Thus, in order to mitigate this potential risk, instead of directly calculating confidence interval for R^2, we also developed the interval for monotonic functions of R^2 that have fewer constraints such as $R^2/(1-R^2)$ and $log(R^2/(1-R^2))$ which extends the value from the interval $\lceil 0, 1 \rceil$ to the positive line and the whole range respectively. Using the delta method, it was noted (Chapter 2) that distributions of those functions also converge to normal distributions as $n \to \infty$. $R^2/(1-R^2)$ can also be presented in matrix form as $\frac{(YX(X'X)^{-1}X'Y)}{(Y'(I-X(X'X)^{-1}X')Y)}$.

In the simulation, $Y = \beta X + \epsilon$ is assumed. Two sets of simulations were performed: the first simulation test considered X's as given and not random but ϵ's were sampled from various distributions, normal and non-normal, skewed and heavy-tail; in the second simulation test, X's were also treated as random variables and were sampled from different distributions including normal and non-normal distributions. These simulations assess the

asymptotic distribution approximation presented in Theorem 3 discussed in Chapter 2. For each simulation scenario, we ran 10000 simulations for each sample size. There are two steps in the simulation test for asymptotic properties: first, quantiles of numerators, denominators and target functions including $Y'X(X'X)^{-1}X'Y, Y'(I - X(X'X)^{-1}X')Y, Y'Y, R^2/(1-R^2)$ and $log(R^2/(1-R^2))$ are compared to those for a normal distribution based on QQ plots, as well as by examination of sample skewness and kurtosis. Secondly, the variance derived from Theorem 1 and Lemma 1 in Chapter 2 are used to calculate standard error estimates for comparison to those derived from simulation results, and confidence interval coverage ratio is benchmarked to the specified significance level. Sample sizes used in the simulation studies ranged from 100, to 1000, and to 10000.

3.1 Choices of error distributions

Since a major purpose of this work is to develop confidence intervals for general error distributions in the linear model, simulation studies were not limited to the normal distribution. To generate random sample from non-normal distributions, the method described in the paper of Vale (1987) was used. This work generates non-normal distribution with given skewness and kurtosis. The sampling methodology is based on Fleishmans(1978) framework, $Y = a + bX + cX^2 + dX^3$, where X is a standard normal variable. By specifying the four parameters, (a, b, c, d), it is possible to generate a Y sample with the desired first four moments, including skewness and kurtosis.

For this study, we have tested several combinations of skewness and kurtosis, including (0,0), i.e. normal distribution, (1,3) and (2,7). We regard the distribution with (2,7) property as being highly non-normal. Besides this, we also tested one commonly used distribution, the Weibull distribution due to its ability to include a variety of error distribution types. In the Weibull distribution simulation, scale and shape parameters were set to 1 and 0.8 respectively which lead to skewness 2.8 and kurtosis 12.8.

3.2 Simulation results when X's are given

This section presents simulation results when X's are considered as given and fixed during the simulation. X values, without of loss generality, were sampled from distributions having

mean 0 and variance 1. To test the impact of the distribution of X values on the estimation results, X's from normal distribution and X's from lognormal distribution are both considered. Table 3.1 and 3.2 lists summary statistics for X values used in the simulation test. The X's were sampled from normal and lognormal distributions respectively at a set of sample sizes: 100, 1000 and 10000.

Table 3.1: Summary of X values derived from a normal distribution used in the simulation

	Sample size=100	Sample size=1000	Sample size=10000
Mean	0.21	-0.03	-0.02
Median	0.18	-0.01	-0.02
Variance	0.95	0.97	1.01
Skewness	-0.35	-0.14	0.02
Excess Kurtosis	-0.06	0.00	0.03
Min Value	-2.70	-2.96	-4.27
Max Value	2.10	2.62	3.93

Table 3.2: Summary of X values from lognormal distribution used in the simulation

Statistics	Sample size=100	Sample size=1000	Sample size=10000
Mean	0.02	-0.03	-0.01
Median	-0.32	-0.30	-0.31
Variance	0.64	0.88	0.93
Skewness	1.57	5.07	4.61
Excess Kurtosis	1.94	41.80	34.17
Min Value	-0.68	-0.75	-0.75
Max Value	3.02	11.82	14.89

3.2.1 Normal distribution approximation when given X from normal distribution

According to the theory described in chapter 2, when sample size becomes large, the distributions of numerator and denominator of R^2 converge jointly to a normal distribution. In this section, we checked the distributions of simulated $Y'X(X'X)^{-1}X'Y, Y'(I - X(X'X)^{-1}X')Y, R^2/(1-R^2)$ and $log(R^2/(1-R^2))$ via QQ plot against normal distribution

and calculated its skewness and kurtosis. The simulation has been done at sample size 100, 1000 and 10000 to check the impact of sample size on the quality of asymptotic approximations. In the simulation, X was sampled from a standard normal distribution. As usual, skewness is defined as $\frac{E(X-EX)^3}{\sigma^3}$, and kurtosis is defined as $\frac{E(X-EX)^4}{\sigma^4} - 3$. Under normal distribution, the skewness and kurtosis are 0.

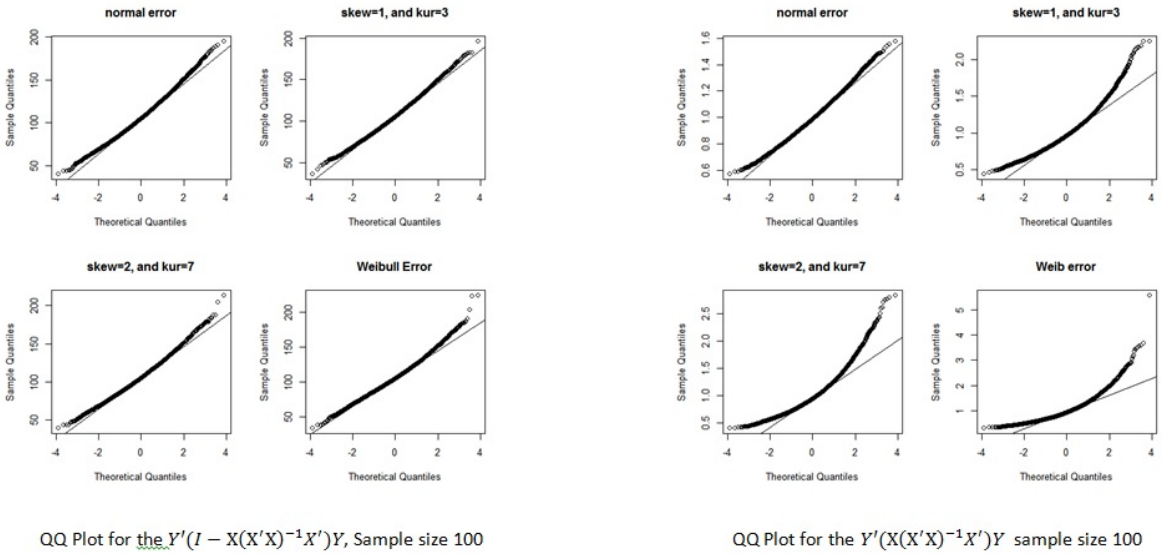

Figure 3.1: QQ Plot for $Y'(I - X(X'X)^{-1}X')Y$ and $Y'(X(X'X)^{-1}X')Y$ at sample size 100

Table 3.3: Sample skewness and kurtosis at sample size 100

	Normal Error		Skew=1,Kurt=3		Skew=2,Kurt=7		Weibull Error	
	skewness	kurtosis	skewness	kurtosis	skewness	kurtosis	skewness	kurtosis
$Y'(I - X(X'X)^{-1}X')Y$	0.28	0.12	0.92	1.69	1.17	2.36	1.56	5.42
$Y'(X(X'X)^{-1}X')Y$	0.34	0.16	0.27	0.09	0.33	0.25	0.37	0.55
$\frac{R^2}{(1-R^2)}$	0.62	0.84	0.64	0.69	0.63	0.56	0.81	1.15
$log(\frac{R^2}{(1-R^2)})$	-0.16	0.22	-0.26	0.23	-0.40	0.45	-0.44	0.46

Figure 3.1 to 3.6 above were QQ plots for denominator, $Y'(I - X(X'X)^{-1}X')Y$, numerator, $Y'(X(X'X)^{-1}X')Y$, $\frac{R^2}{(1-R^2)}$ and $log(\frac{R^2}{(1-R^2)})$ at different sample sizes (100, 1000 and 10000) under errors from various distributions with normal distribution as reference,

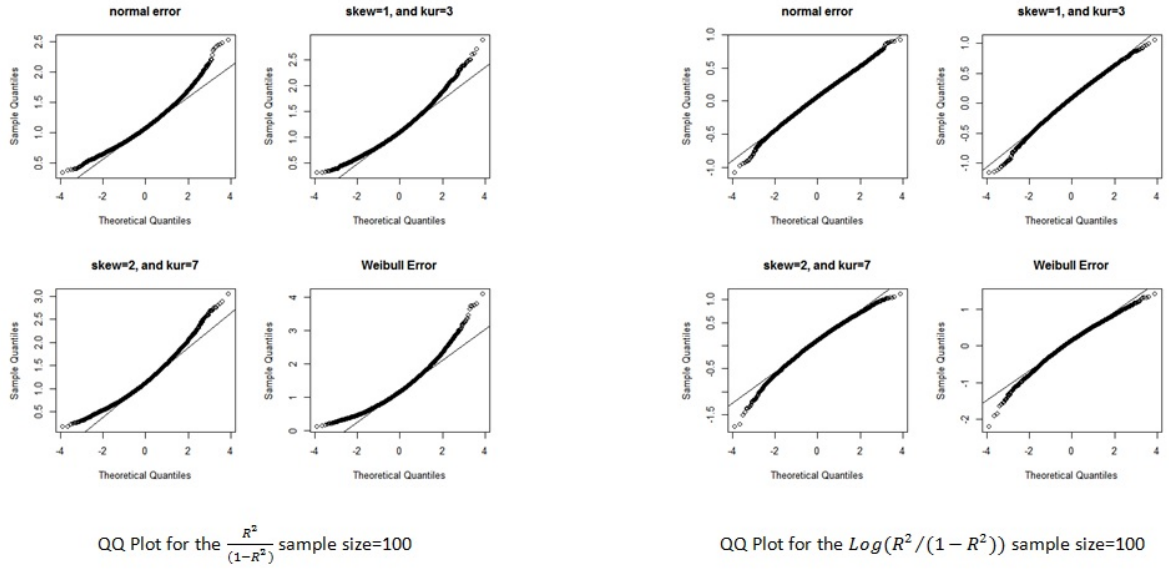

Figure 3.2: QQ Plot for $\frac{R^2}{(1-R^2)}$ and $log(\frac{R^2}{(1-R^2)})$ at sample size 100

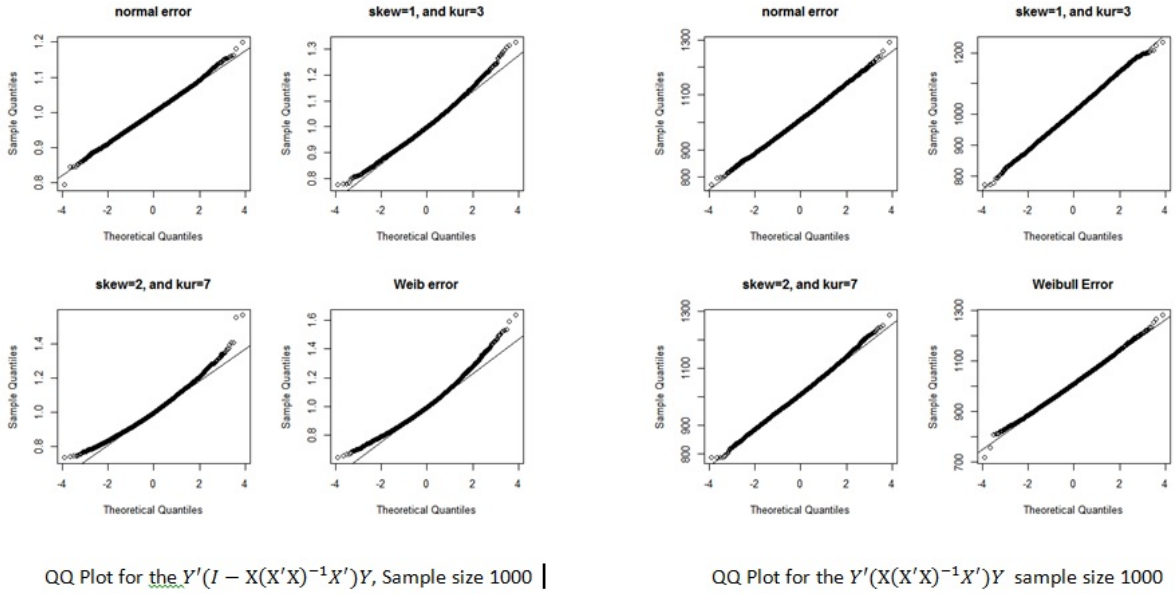

Figure 3.3: QQ Plot for $Y'(I - X(X'X)^{-1}X')Y$ and $Y'(X(X'X)^{-1}X')Y$ at sample size 1000

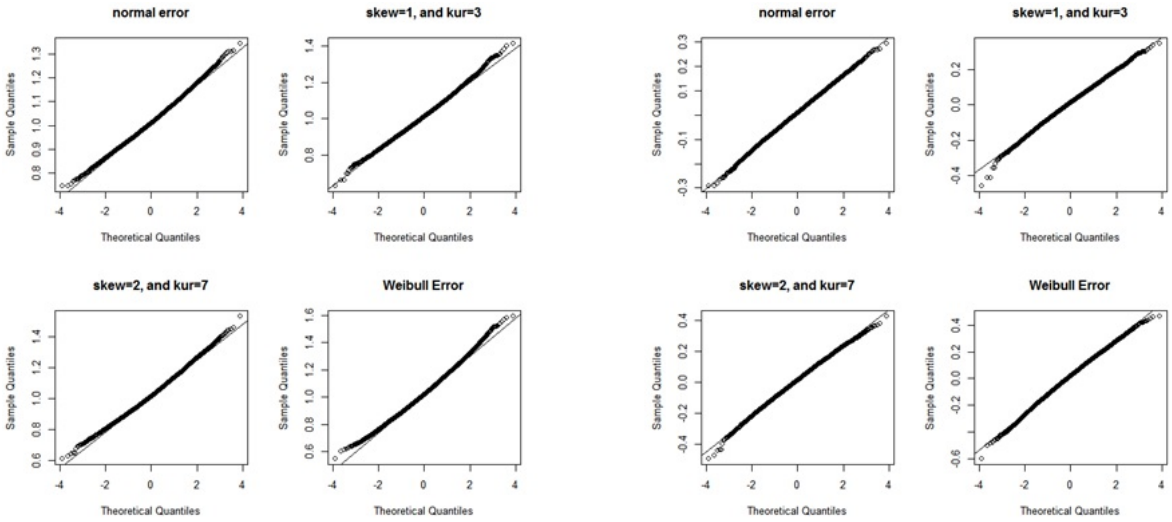

Figure 3.4: QQ Plot for $\frac{R^2}{(1-R^2)}$ and $log(\frac{R^2}{(1-R^2)})$ at sample size 1000

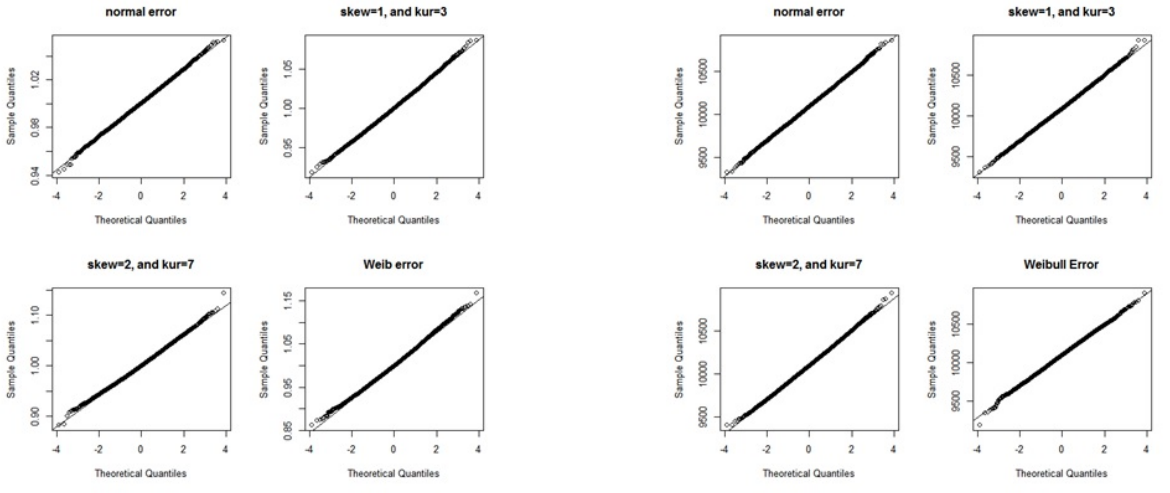

Figure 3.5: QQ Plot for $Y'(I - X(X'X)^{-1}X')Y$ and $Y'(X(X'X)^{-1}X')Y$ at sample size 10000

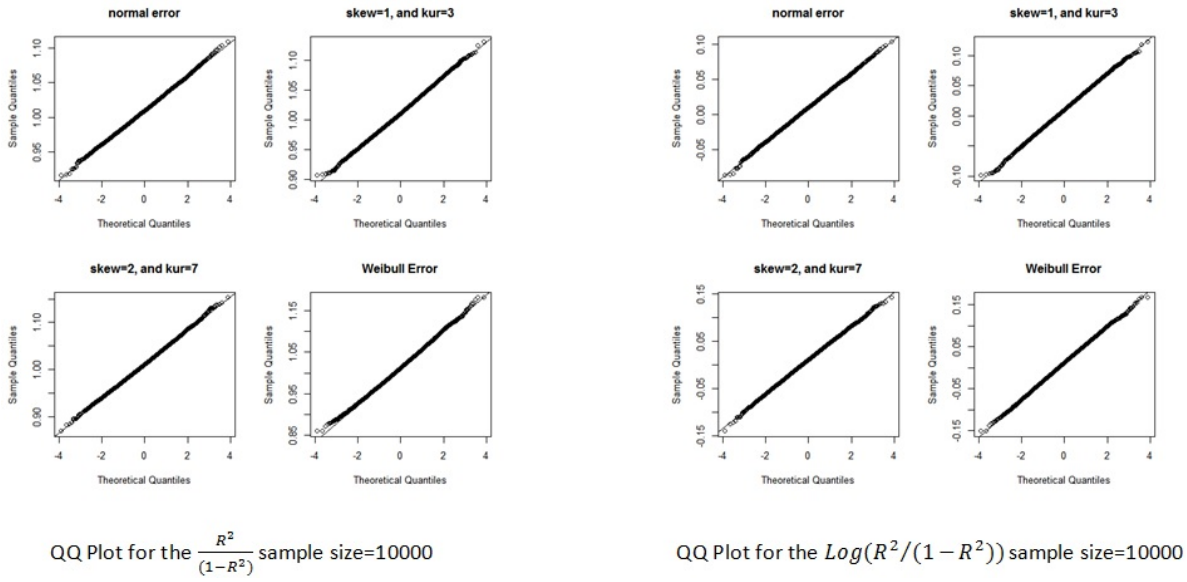

Figure 3.6: QQ Plot for $\frac{R^2}{(1-R^2)}$ and $log(\frac{R^2}{(1-R^2)})$ at sample size 10000

Table 3.4: Sample skewness and kurtosis at sample size 1000

	Normal Error		Skew=1,Kurt=3		Skew=2, Kurt=7		Weibull Error	
	skewness	kurtosis	skewness	kurtosis	skewness	kurtosis	skewness	kurtosis
$Y'(I-X(X'X)^{-1}X')Y$	0.07	0.16	0.31	0.27	0.41	0.34	0.52	0.48
$Y'(X(X'X)^{-1}X')Y$	0.09	0.00	0.03	-0.01	0.12	0.17	0.11	0.05
$\frac{R^2}{(1-R^2)}$	0.19	0.10	0.17	0.16	0.22	0.02	0.24	0.10
$log(\frac{R^2}{(1-R^2)})$	-0.05	0.06	-0.14	0.18	-0.12	0.03	-0.17	0.07

Table 3.5: Sample skewness and kurtosis at sample size 10000

	Normal Error		Skew=1,Kurt=3		Skew=2,Kurt=7		Weibull Error	
	skewness	kurtosis	skewness	kurtosis	skewness	kurtosis	skewness	kurtosis
$Y'(I-X(X'X)^{-1}X')Y$	0.02	0.06	0.08	0.02	0.13	-0.01	0.15	0.02
$Y'(X(X'X)^{-1}X')Y$	0.02	0.02	0.05	0.05	0.06	0.03	-0.03	-0.05
$\frac{R^2}{(1-R^2)}$	0.05	0.02	0.07	0.00	0.06	-0.01	0.08	-0.05
$log(\frac{R^2}{(1-R^2)})$	-0.02	0.01	-0.02	0.00	-0.05	0.00	-0.04	-0.06

while Table 3.3 to 3.5 presents the corresponding skewness and excess kurtosis of simulation results of these variables. For samples from a normal distribution, it should have zero skewness and zero excess kurtosis, and its QQ plot should be a straight line. The simulation showed that distribution of $Y'(I - X(X'X)^{-1}X')Y$ was close to normal distribution under various errors even under sample size 100, given that the QQ plots were close to a straight line and its skewness and kurtosis were around zero. This can be explained by the fact that $Y'(I - X(X'X)^{-1}X')Y$ can be presented as $\sum_{i=1}^{n} \epsilon_i^2$ that fits the central limit theorem framework. As described in Chapter 2, $Y'X(X'X)^{-1}X'Y = (X\beta)'X\beta + 2(X\beta)'\epsilon + \epsilon'X(X'X)^{-1}X'\epsilon$, comprises of two random components, and only when $n \to \infty$, the second component $\epsilon'X(X'X)^{-1}X'\epsilon$ converges to 0. Thus under small to moderate sample size, we observed deviations in simulated $Y'(X(X'X)^{-1}X')Y$ from a normal distribution, as illustrated by the deviation from straight lines in the QQ plots and non-zero skewness and kurtosis in the tables. Similar deviations from normal distribution were observed for $\frac{R^2}{1-R^2}$ and $log(\frac{R^2}{1-R^2})$. As expected, when errors were sampled from non-normal distribution, the deviation was more obvious. Another observation from simulation results was that the performance of the normal approximation to its log transformation was more stable when compared to that of $\frac{R^2}{(1-R^2)}$, especially when the sample size is small, at 100 or 1000. When sample size increased to 10000, the performance of normal approximation to these variables were reasonable, with almost straight lines in QQ plots and with skewness and kurtosis close to zero. This agrees with our expectations, based on Chapter 2.

3.2.2 Normal distribution approximation when given X from a lognormal distribution

The previous section presented the simulation results when the given X's are sampled from a normal distribution. To assess the impact of the distribution of X's on the estimation results, we conducted another simulation test with X's are sampled from a lognormal distribution. The summary statistics of the X's used in the simulation are summarized in Table 3.2. The simulation results, QQ plots are presented in Figure 3.7 to 3.12 and skewness and kurtosis of related variables are listed in Tables 3.6 to 3.8. Similar to the previous

section in which X's are sampled from a normal distribution, the distributions of denominator, numerator. $\frac{R^2}{(1-R^2)}$ and $log(\frac{R^2}{(1-R^2)})$ converge to normal distribution when sample size increases from 100 to 10000. Also, with moderate sample size (100 or 1000), distribution of $log(\frac{R^2}{(1-R^2)})$ was closer to normal distribution compared to the statistics, $\frac{R^2}{(1-R^2)}$. There were no significant differences in the convergence speed to normal distribution when X's are sampled from a normal distribution or from a lognormal distribution. This is consistent with our expectation. Since in the simulation, the X's are considered as given and fixed, error distributions should have a more substantial impact on asymptotic approximation.

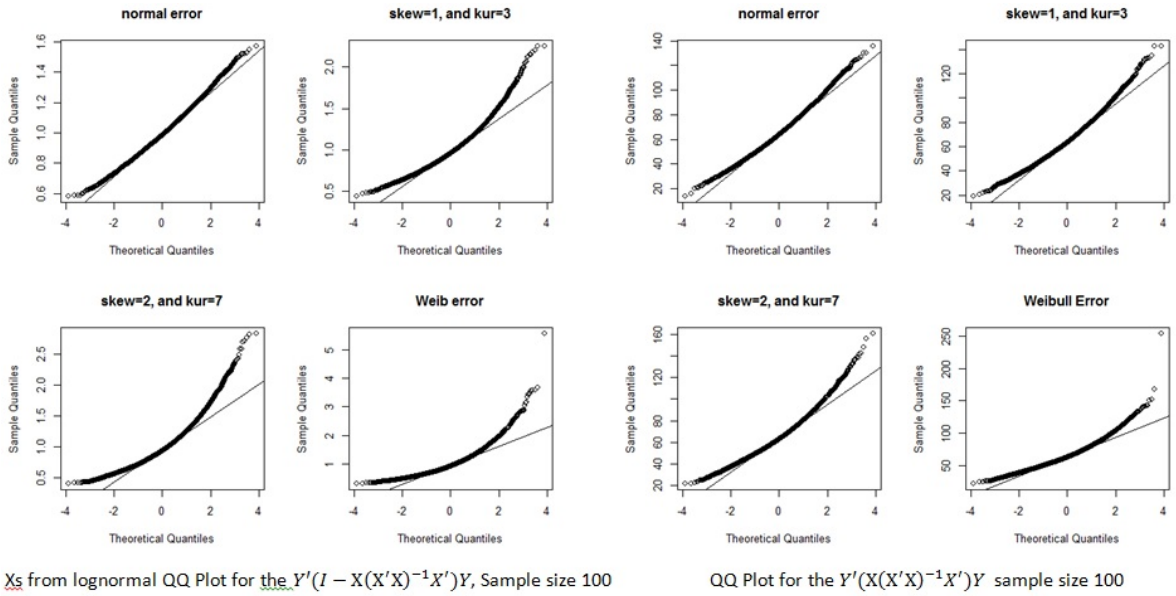

Figure 3.7: QQ Plot for $Y'(I - X(X'X)^{-1}X')Y$, $Y'(X(X'X)^{-1}X')Y$ at sample size 100 and lognormal distributed X.

3.2.3 Variance estimation when X sampled from normal distribution

Standard errors from simulation data were compared against those derived from the formulas listed in Chapter 2. In the calculation, first we estimated skewness and kurtosis using the residual error from the fitted regression model, and then the estimated skewness and kurtosis are put into the asymptotic distributional formulas to calculate the corresponding mean

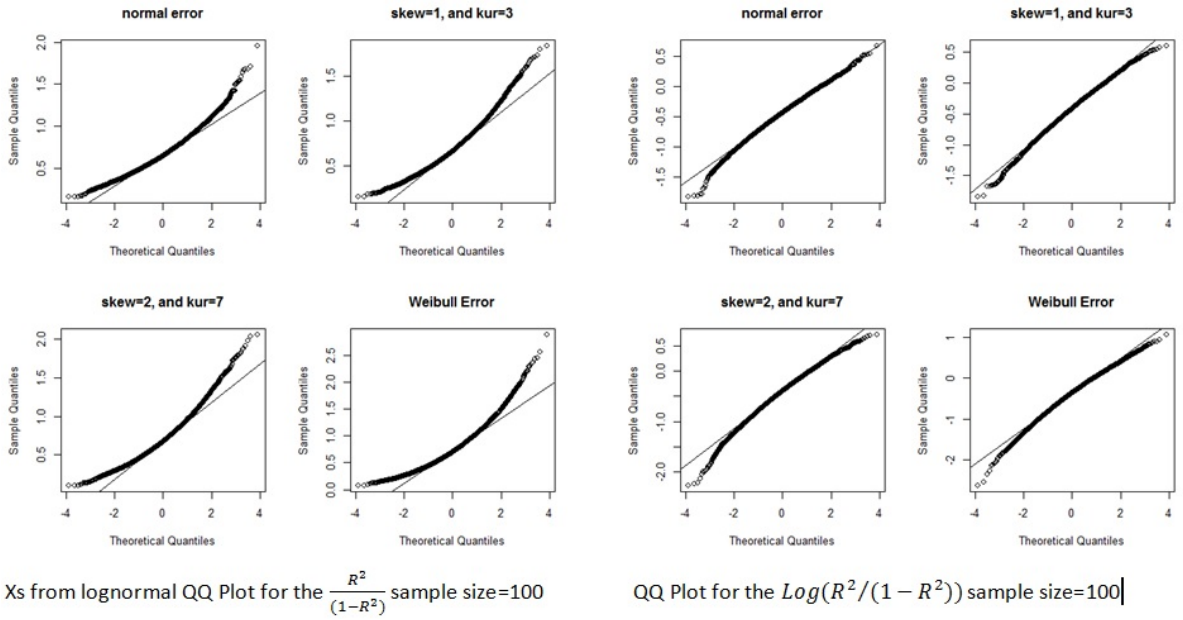

Figure 3.8: QQ Plot for $\frac{R^2}{(1-R^2)}$ under sample size 100, X Lognormal

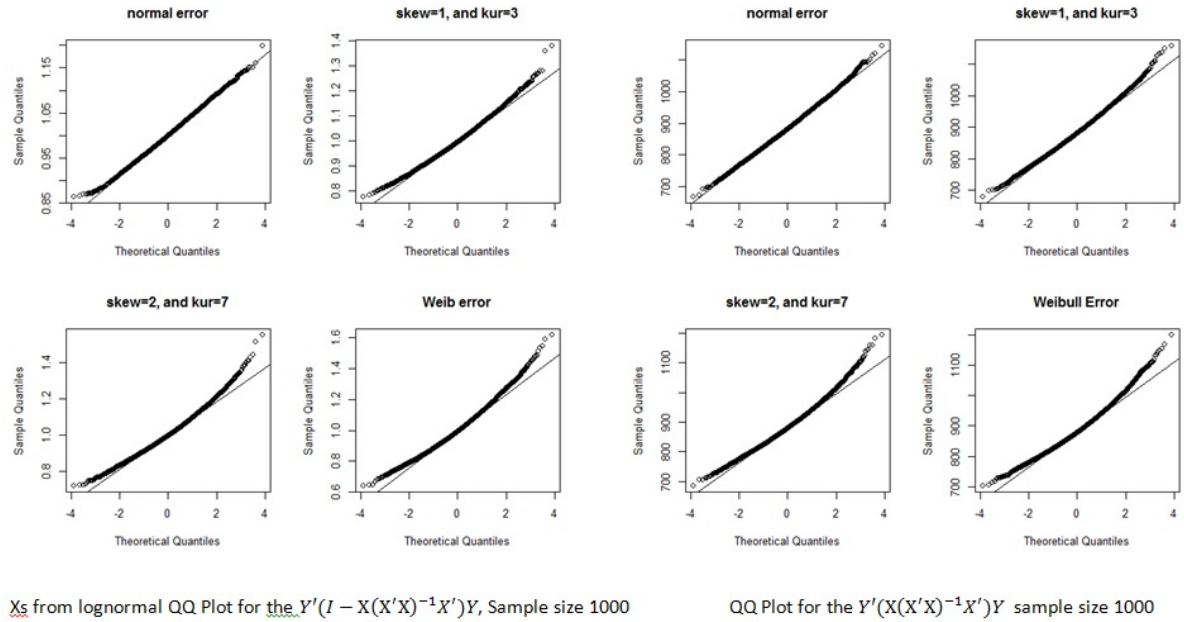

Figure 3.9: QQ Plot for $Y'(I - X(X'X)^{-1}X')Y$, $Y'(X(X'X)^{-1}X')Y$ sample size 1000, X Lognormal

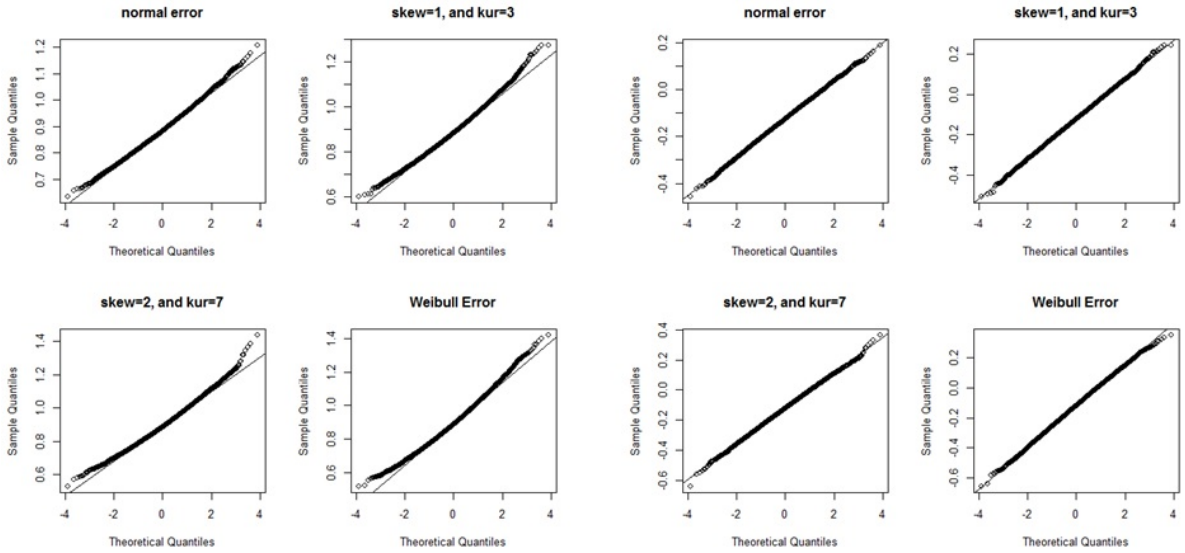

Figure 3.10: QQ Plot for $\frac{R^2}{(1-R^2)}$ under sample size 1000, X Lognormal

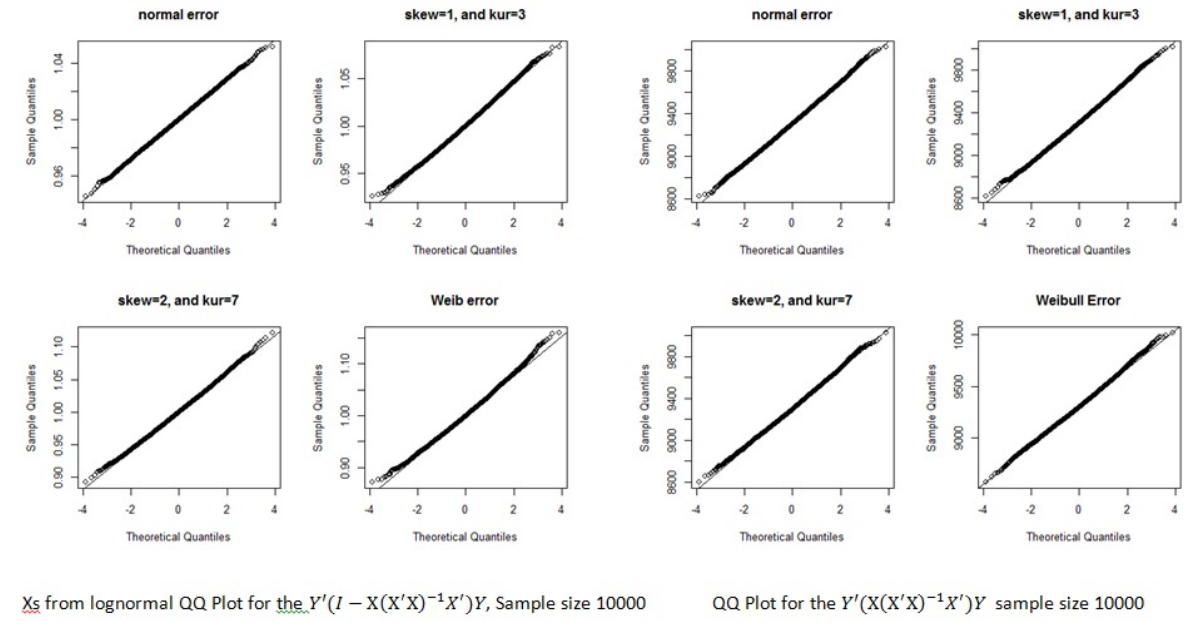

Figure 3.11: QQ Plot for $Y'(I - X(X'X)^{-1}X')Y$, $Y'(X(X'X)^{-1}X')Y$ sample size 10000, X Lognormal

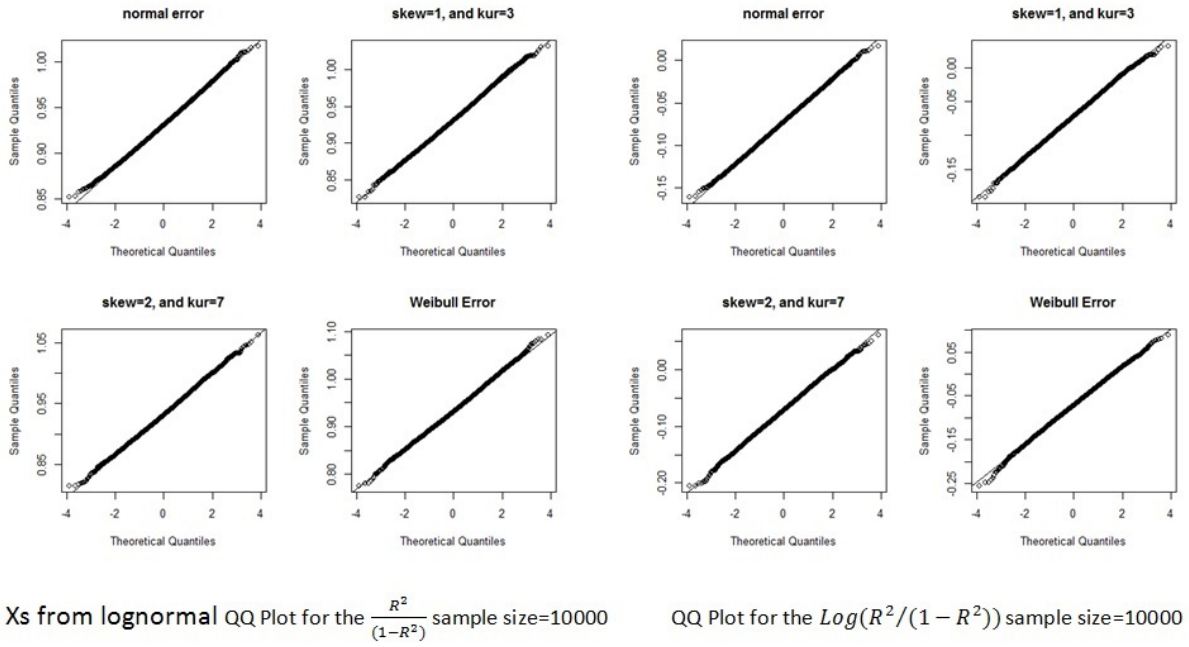

Figure 3.12: QQ Plot for $\frac{R^2}{(1-R^2)}$ under sample size 10000, X Lognormal

Table 3.6: Skewness and kurtosis at sample size 100 when given X from lognormal

	Normal Error		Skew=1,Kurt=3		Skew=2,Kurt=7		Weibull Error	
	skew	kurtosis	skew	kurtosis	skew	kurtosis	skew	kurtosis
$Y'(I-X(X'X)^{-1}X')Y$	0.29	0.13	0.92	1.69	1.17	2.38	1.56	5.41
$Y'(X(X'X)^{-1}X')Y$	0.39	0.21	0.54	0.51	0.70	0.99	0.98	3.04
$\frac{R^2}{(1-R^2)}$	0.68	1.04	0.74	0.86	0.76	0.87	0.93	1.54
$log(\frac{R^2}{(1-R^2)})$	-0.28	0.36	-0.26	0.18	-0.38	0.40	-0.40	0.33

Table 3.7: Skewness and kurtosis at sample size 1000 when given X from lognormal

	Normal Error		Skew=1,Kurt=3		Skew=2,Kurt=7		Weibull Error	
	skew	kurtosis	skew	kurtosis	skew	kurtosis	skew	kurtosis
$Y'(I-X(X'X)^{-1}X')Y$	0.09	-0.07	0.30	0.18	0.44	0.51	0.45	0.33
$Y'(X(X'X)^{-1}X')Y$	0.12	0.07	0.25	0.20	0.42	0.49	0.51	0.56
$\frac{R^2}{(1-R^2)}$	0.19	0.09	0.29	0.19	0.29	0.15	0.33	0.10
$log(\frac{R^2}{(1-R^2)})$	-0.06	0.04	-0.02	0.04	-0.06	-0.00	-0.07	-0.05

Table 3.8: Skewness and kurtosis at sample size 10000 when given X from lognormal

	Normal Error		Skew=1,Kurt=3		Skew=2,Kurt=7		Weibull Error	
	skew	kurtosis	skew	kurtosis	skew	kurtosis	skew	kurtosis
$Y'(I - X(X'X)^{-1}X')Y$	0.00	-0.02	0.12	-0.04	0.08	0.02	0.18	0.08
$Y'(X(X'X)^{-1}X')Y$	0.05	-0.01	0.09	-0.01	0.10	0.01	0.12	0.08
$\frac{R^2}{(1-R^2)}$	0.07	-0.02	0.08	-0.03	0.08	-0.05	0.08	0.05
$log(\frac{R^2}{(1-R^2)})$	0.00	-0.04	-0.01	-0.03	-0.03	-0.04	-0.05	0.07

and variance. Standard deviation estimates from asymptotic approximations and sample standard error from simulated data are compared in Tables 3.9 to 3.12. The comparison shows good agreement between average of standard deviation estimates and sample standard deviations from simulated data. The comparison was based on sample size 100 and 1000.

Since $log(\frac{R^2}{(1-R^2)})$ has more stable performance, its function was used to derive a confidence interval for R^2. The calculation estimated variance of $log(\frac{R^2}{(1-R^2)})$, and generated the confidence intervals assuming $log(\frac{R^2}{(1-R^2)})$ to follow a normal distribution. The upper and lower bound of the derived confidence interval can be transformed back to the corresponding confidence interval for R^2. The coverage ratio of the process was compared against the corresponding ratio when assuming errors from normal distribution with zero skewness and zero kurtosis. The comparison results were presented in Tables 3.13 and 3.14. In the comparison, the specified significance level was 95%. The comparison results showed deviations from the specified significance level when sample size was 100. But still the coverage ratio from the proposed methodology was significantly better than the ratio of confidence interval assuming errors from a normal distribution. When sample size was 1000, the coverage ratio using sample skewness and kurtosis was consistent with the nominal significance level.

3.2.4 Variance estimation when X sampled from lognormal distribution

We also performed a simulation test for the given X's from a nonnormal distribution (lognormal) to test the asymptotic approximation for the estimation of variance statistics. Simulation results are presented and discussed in this section. Similar to the results described in the previous section, we observed reasonable agreement between variance estimation based

Table 3.9: Simulation vs formula estimation in variance when given X from normal with sample size=100

	Normal Error		Non-normal Error(skew=1,kurt=3)	
	Simulation	Formula	Simulation	Formula
Mean of $Y'(I - X(X'X)^{-1}X')Y$	98.938	97.965	98.718	97.751
Std Err of $Y'(I - X(X'X)^{-1}X')Y$	14.053	13.402	21.854	19.165
Mean of $Y'(X(X'X)^{-1}X')Y$	99.247	100.237	99.204	100.191
Std Err of $Y'(X(X'X)^{-1}X')Y$	19.894	19.721	19.868	19.677
Mean of $\frac{R^2}{(1-R^2)}$	1.024	1.044	1.048	1.069
Std Err of $\frac{R^2}{(1-R^2)}$	0.254	0.249	0.291	0.275
Mean of $Log(\frac{R^2}{(1-R^2)})$	-0.007	0.013	0.008	0.028
Std Err of $Log(\frac{R^2}{(1-R^2)})$	0.249	0.244	0.282	0.265

Table 3.10: Simulation vs formula estimation in variance when given X from normal with sample size=100

	Non-normal Error (skew=2, kurt=7)		Weibull Error	
	Simulation	Formula	Simulation	Formula
Mean of $Y'(I - X(X'X)^{-1}X')Y$	98.588	97.626	98.853	97.857
Std Err of $Y'(I - X(X'X)^{-1}X')Y$	29.167	24.303	37.430	29.969
Mean of $Y'(X(X'X)^{-1}X')Y$	99.191	100.177	100.322	101.311
Std Err of $Y'(X(X'X)^{-1}X')Y$	19.824	19.625	19.988	19.438
Mean of $\frac{R^2}{(1-R^2)}$	1.079	1.102	1.152	1.179
Std Err of $\frac{R^2}{(1-R^2)}$	0.336	0.304	0.466	0.397
Mean of $Log(\frac{R^2}{(1-R^2)})$	0.026	0.047	0.058	0.080
Std Err of $Log(\frac{R^2}{(1-R^2)})$	0.322	0.288	0.420	0.359

Table 3.11: Simulation vs formula estimation in variance when given X from normal with sample size=1000

	Normal Error		Non-normal Error(skew=1,kurt=3)	
	Simulation	Formula	Simulation	Formula
Mean of $Y'(I - X(X'X)^{-1}X')Y$	999.4	998.4	999.2	998.2
Std Err of $Y'(I - X(X'X)^{-1}X')Y$	44.7	44.5	70.5	68.8
Mean of $Y'(X(X'X)^{-1}X')Y$	1005.7	1006.7	1005.7	1006.7
Std Err of $Y'(X(X'X)^{-1}X')Y$	64.6	63.4	64.5	63.3
Mean of $\frac{R^2}{(1-R^2)}$	1.01	1.01	1.01	1.01
Std Err of $\frac{R^2}{(1-R^2)}$	0.079	0.078	0.097	0.094
Mean of $Log(\frac{R^2}{(1-R^2)})$	0.01	0.01	0.01	0.01
Std Err of $Log(\frac{R^2}{(1-R^2)})$	0.079	0.077	0.096	0.093

Table 3.12: Simulation vs formula estimation in variance when given X from normal with sample size=1000

	Non-normal Error (skew=2, kurt=7)		Weibull Error	
	Simulation	Formula	Simulation	Formula
Mean of $Y'(I - X(X'X)^{-1}X')Y$	999.0	998.0	998.2	997.2
Std Err of $Y'(I - X(X'X)^{-1}X')Y$	94.4	91.1	121.5	115.7
Mean of $Y'(X(X'X)^{-1}X')Y$	1005.8	1006.8	1004.6	1005.6
Std Err of $Y'(X(X'X)^{-1}X')Y$	64.3	63.3	63.0	63.1
Mean of $\frac{R^2}{(1-R^2)}$	1.02	1.02	1.02	1.02
Std Err of $\frac{R^2}{(1-R^2)}$	0.116	0.112	0.138	0.132
Mean of $Log(\frac{R^2}{(1-R^2)})$	0.01	0.01	0.012	0.014
Std Err of $Log(\frac{R^2}{(1-R^2)})$	0.115	0.111	0.136	0.131

Table 3.13: Comparison of the 95% confidence interval coverage ratio for different error distributions when X from normal and sample size=100

	Normal	Skew=1,Kurt=3	Skew=2,Kurt=7	Weibull
Assuming normal error (set skew=0 and kurt=0)	0.93	0.89	0.83	0.73
Using skew and kurtosis estimated from empirical data	0.93	0.91	0.88	0.87

Table 3.14: Comparison of the 95% confidence interval coverage ratio for different error distributions when X from normal and sample size=1000

	Normal	Skew=1,Kurt=3	Skew=2,Kurt=7	Weibull
Assuming normal error (set skew=0 and kurt=0)	0.94	0.89	0.81	0.73
Using skew and kurtosis estimated from empirical data	0.94	0.94	0.93	0.93

on formula and variance estimated from simulation results. Also when the sample size was small, at 100, we observed deviations from specified significance level, but obvious improvements from the traditional method. When sample size increased to 1000, the coverage ratio was consistent with the nominal significance level at 95%.

Table 3.15: Simulation vs formula estimation in variance when given X from lognormal with sample size=100

	Normal Error		Non-normal Error(skew=1,kurt=3)	
	Simulation	Formula	Simulation	Formula
Mean of $Y'(I - X(X'X)^{-1}X')Y$	98.951	97.934	98.723	97.712
Std Deviation of $Y'(I - X(X'X)^{-1}X')Y$	14.067	13.398	21.878	19.167
Mean of $Y'(X(X'X)^{-1}X')Y$	98.637	99.626	98.664	99.651
Std Deviation of $Y'(X(X'X)^{-1}X')Y$	19.698	19.656	19.940	19.799
Mean of $\frac{R^2}{(1-R^2)}$	1.018	1.039	1.047	1.068
Std Deviation of $\frac{R^2}{(1-R^2)}$	0.254	0.248	0.307	0.288
Mean of $Log(\frac{R^2}{(1-R^2)})$	-0.013	0.007	0.003	0.024
Std Deviation of $Log(\frac{R^2}{(1-R^2)})$	0.250	0.244	0.297	0.279

Table 3.16: Simulation vs formula estimation in variance when given X from lognormal with sample size=100

	Non-normal Error (skew=2, kurt=7)		Weibull Error	
	Simulation	Formula	Simulation	Formula
Mean of $Y'(I - X(X'X)^{-1}X')Y$	98.586	97.580	98.892	97.891
Std Deviation of $Y'(I - X(X'X)^{-1}X')Y$	29.194	24.306	37.499	30.133
Mean of $Y'(X(X'X)^{-1}X')Y$	98.691	99.677	98.811	99.800
Std Deviation of $Y'(X(X'X)^{-1}X')Y$	20.167	19.928	20.634	20.453
Mean of $\frac{R^2}{(1-R^2)}$	1.082	1.106	1.130	1.157
Std Deviation of $\frac{R^2}{(1-R^2)}$	0.365	0.330	0.453	0.390
Mean of $Log(\frac{R^2}{(1-R^2)})$	0.021	0.043	0.044	0.065
Std Deviation of $Log(\frac{R^2}{(1-R^2)})$	0.346	0.314	0.403	0.360

3.2.5 Mass at zero studies and proposal for the estimation

In nutritional epidemiology studies, there is a common problem with response data for some nutrients or foods having a mass at zero. We assume that the underlying model is $Y = Z(\beta X + \epsilon)$ where Z is a binary variable with probability p of being zero. We assumed Z to be independent of X and ϵ. In standard regression, it is common to assume that

Table 3.17: Simulation vs formula estimation in variance when given X from lognormal with sample size=1000

	Normal Error		Non-normal Error(skew=1,kurt=3)	
	Simulation	Formula	Simulation	Formula
Mean of $Y'(I - X(X'X)^{-1}X')Y$	999.463	998.468	999.206	998.215
Std Deviation of $Y'(I - X(X'X)^{-1}X')Y$	44.746	44.488	70.514	68.820
Mean of $Y'(X(X'X)^{-1}X')Y$	1032.343	1035.342	1034.328	1035.327
Std Deviation of $Y'(X(X'X)^{-1}X')Y$	64.887	64.273	65.012	64.360
Mean of $\frac{R^2}{(1-R^2)}$	1.037	1.039	1.040	1.042
Std Deviation of $\frac{R^2}{(1-R^2)}$	0.080	0.079	0.097	0.096
Mean of $Log(\frac{R^2}{(1-R^2)})$	0.033	0.035	0.035	0.037
Std Deviation of $Log(\frac{R^2}{(1-R^2)})$	0.077	0.077	0.094	0.093

Table 3.18: Simulation vs formula estimation in variance when given X from lognormal with sample size=1000

	Non-normal Error (skew=2, kurt=7)		Weibull Error	
	Simulation	Formula	Simulation	Formula
Mean of $Y'(I - X(X'X)^{-1}X')Y$	999.001	998.011	998.233	997.214
Std Deviation of $Y'(I - X(X'X)^{-1}X')Y$	94.356	91.134	121.495	115.665
Mean of $Y'(X(X'X)^{-1}X')Y$	1034.377	1035.376	906.820	907.819
Std Deviation of $Y'(X(X'X)^{-1}X')Y$	65.055	64.442	59.970	60.370
Mean of $\frac{R^2}{(1-R^2)}$	1.044	1.046	0.922	0.924
Std Deviation of $\frac{R^2}{(1-R^2)}$	0.117	0.112	0.125	0.121
Mean of $Log(\frac{R^2}{(1-R^2)})$	0.037	0.039	-0.091	-0.089
Std Deviation of $Log(\frac{R^2}{(1-R^2)})$	0.112	0.110	0.136	0.133

Table 3.19: Comparison of the 95% confidence interval coverage ratio for different error distributions when X from lognormal and sample size=100

	Normal	Skew=1,Kurt=3	Skew=2,Kurt=7	Weibull
Assuming normal error (set skew=0 and kurt=0)	0.891	0.831	0.856	0.687
Using skew and kurtosis estimated from empirical data	0.892	0.895	0.929	0.928

Table 3.20: Comparison of the 95% confidence interval coverage ratio for different error distributions when X from lognormal and sample size=1000

	Normal	Skew=1,Kurt=3	Skew=2,Kurt=7	Weibull
Assuming normal error (set skew=0 and kurt=0)	0.950	0.889	0.818	0.749
Using skew and kurtosis estimated from empirical data	0.950	0.945	0.938	0.936

given X, observations, Y , all have the same moments in variance, kurtosis and skewness as distribution of ϵ. However when there are mass at zero from the underlying model, even though for each observation, the error ϵ is from the same distribution, the corresponding variance and higher order moments are not the same across Y values and they depend on the mean value, βX. In the calculation, vector of variance, skewness and kurtosis, with individual value for each observation, will be used in the calculation as presented in Theorem 1. Details of the calculation procedure are listed as below:

- The probability of Y being zero was estimated as the ratio of number of observed zeros to total observations, $\bar{p}$.

- Build the regression model using the subset of observations without zeros to estimate the coefficient β and standard error of the residuals, ϵ.

- Based on the results from step 1 and step 2, we calculated the corresponding mean, variance, and higher moments including skewness and kurtosis after incorporating the mass at zero probability.

- The vector of moments for each observation that fed into the formula in Theorem 1.

We performed simulation studies to check the variance estimation from our formula and compared it against the empirical variance from our simulation results. The comparison was presented below. The simulation results (Tables 3.21-3.22) showed reasonable agreement between empirical data and the moment's value derived from formula. In the simulation, sample size is set to 1000, and the probability of Y being zero is 20%.

3.2.6 F-distribution approximation

As presented in Figure 3.2 and simulation results (for example, Table 3.10) in previous sections, when errors are from non-normal distributions and sample size is small at 100, QQ plots show deviation in distributions of simulated $\frac{R^2}{(1-R^2)}$ and $log(\frac{R^2}{(1-R^2)})$ from a normal distribution and their standard errors are underestimated using a normal approximation. Thus, we tried another approximation option: through adjusted F distribution by matching

Table 3.21: Simulation vs formula estimation in variance with mass at zero, the given X from normal distribution, sample size=1000

	Normal Error		Non-normal Error(skew=1,kurt=3)	
	Simulation	Formula	Simulation	Formula
Mean of $Y'(I - X(X'X)^{-1}X')Y$	644.6	645.3	643.7	644.8
Std Deviation of $Y'(I - X(X'X)^{-1}X')Y$	57.4	57.0	56.8	56.9
Mean of $Y'(X(X'X)^{-1}X')Y$	958.6	958.1	957.7	956.9
Std Deviation of $Y'(X(X'X)^{-1}X')Y$	42.4	42.9	108.8	103.5
Mean of $\frac{R^2}{(1-R^2)}$	0.674	0.674	0.681	0.682
Std Deviation of $\frac{R^2}{(1-R^2)}$	0.0632	0.0631	0.095	0.091

Table 3.22: Simulation vs formula estimation in variance with mass at zero, the given X from normal distribution, sample size=1000

	Non-normal Error (skew=2, kurt=7)		Weibull Error	
	Simulation	Formula	Simulation	Formula
Mean of $Y'(I - X(X'X)^{-1}X')Y$	644.7	645.2	644.7	645.2
Std Deviation of $Y'(I - X(X'X)^{-1}X')Y$	57.3	56.9	57.3	56.9
Mean of $Y'(X(X'X)^{-1}X')Y$	958.1	957.6	957.8	957.2
Std Deviation of $Y'(X(X'X)^{-1}X')Y$	63.6	63.3	84.1	82.5
Mean of $\frac{R^2}{(1-R^2)}$	0.675	0.676	0.678	0.678
Std Deviation of $\frac{R^2}{(1-R^2)}$	0.0714	0.0711	0.0811	0.0802

first-order and second-order moments, which helps improve the fit of the approximation. As discussed in the Chapter 2, when the errors are from a normal distribution, $\frac{R^2/k}{(1-R^2)/(n-k)}$ follows an F distribution with degrees of freedom being k and $(n-k)$. However when the random errors are not from normal distribution, the statistic, $\frac{R^2/k}{(1-R^2)/(n-k)}$, does not follow an F distribution. By matching the first and second moment of the F distribution to the corresponding moments derived from the matrix calculation, we estimated the adjusted degrees of freedom and non-centrality parameter. We compared the distributions before and after adjustments. Figure 3.13 to 3.15 present plots of the percentiles from simulated R^2, unadjusted F-distribution with degree of freedom k and $(n-k)$, and the adjusted F-distribution. The plot showed that when errors were from non-normal distribution, the percentile line of unadjusted F-distribution deviated from the line of simulated R^2, while it was improved after adjusting degrees of freedom and non-central component. The simulation results presented in the plots were based on sample size 100.

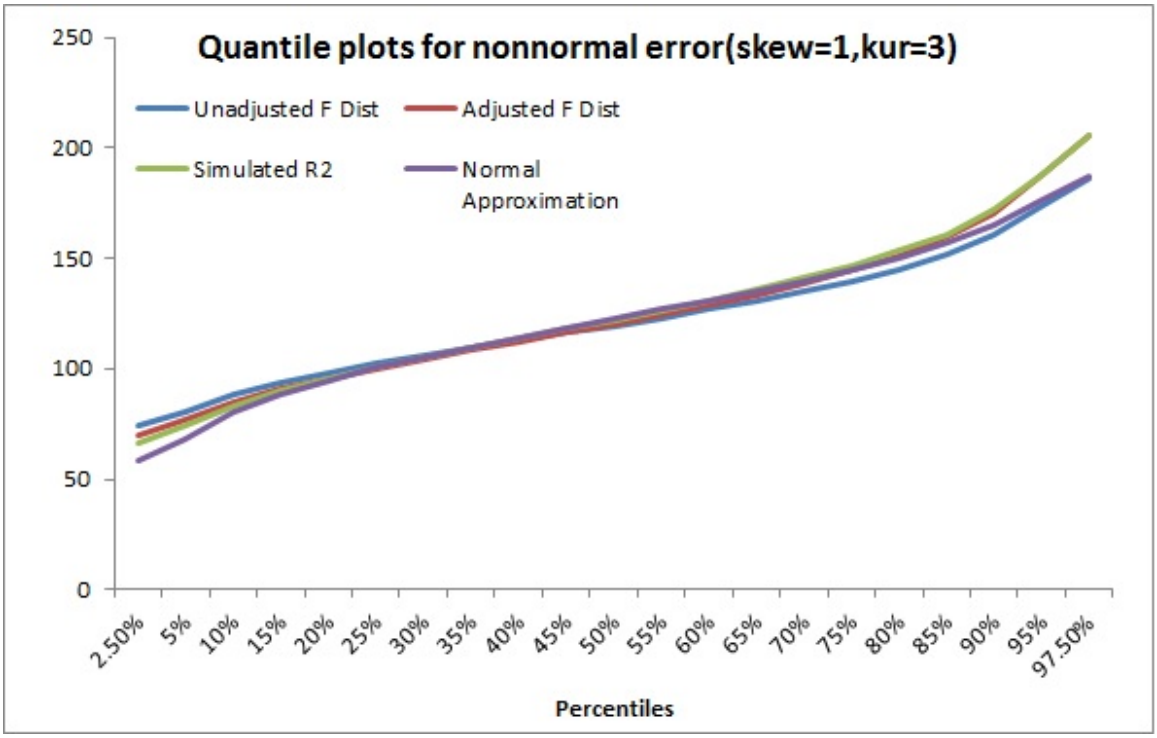

Figure 3.13: QQPlot of F distribution approximation with adjusted degrees of freedom vs simulated data, Error from nonnormal distribution(skew=1,Kurt=3), sample size at 100

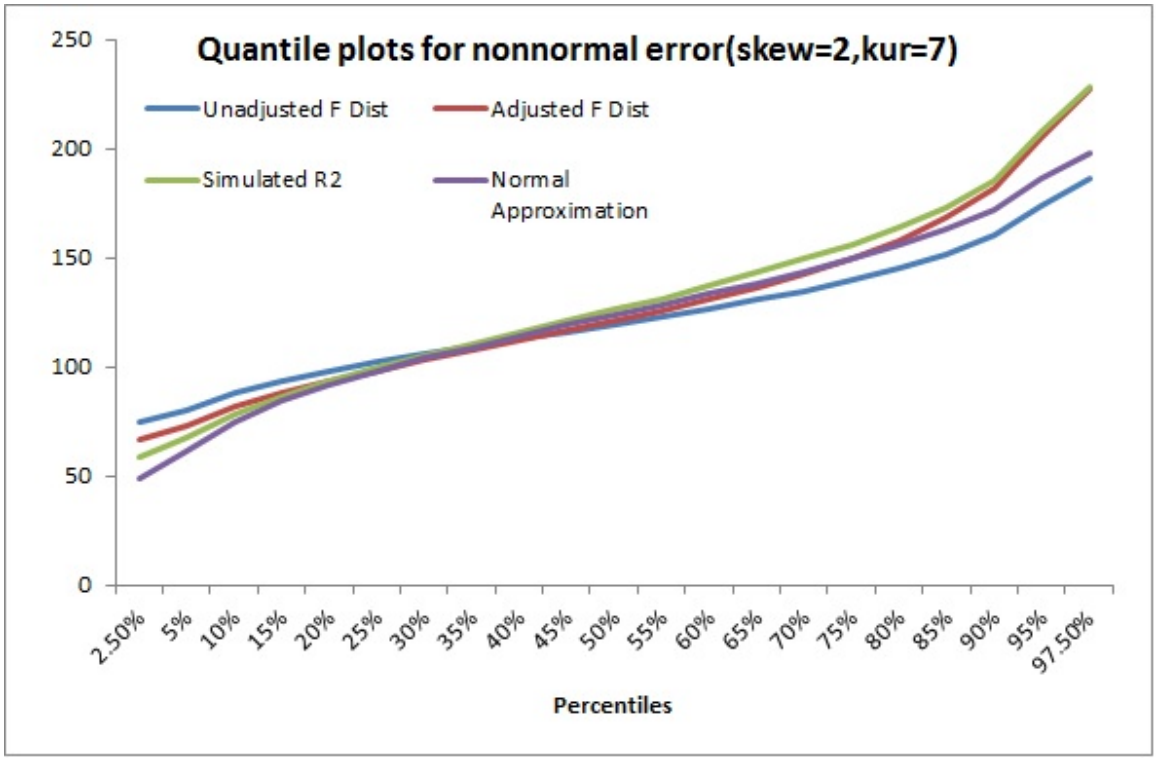

Figure 3.14: QQPlot of F distribution approximation with adjusted degrees of freedom vs simulated data, Error from nonnormal distribution(skew=2,Kurt=7), sample size at 100

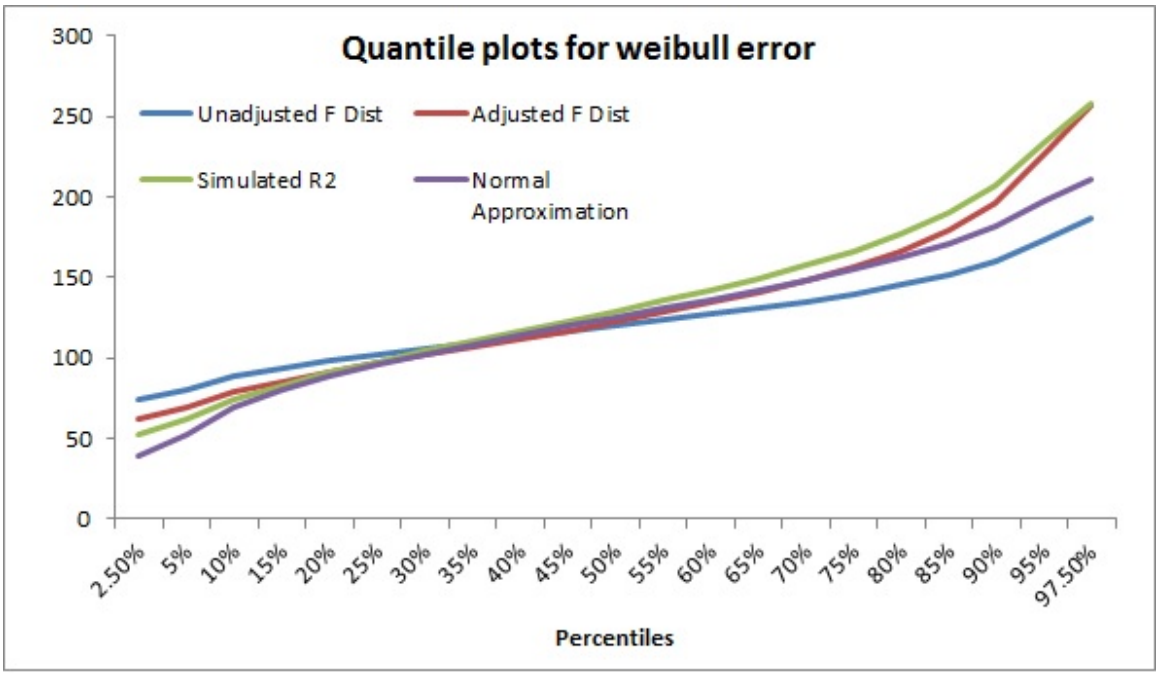

Figure 3.15: QQPlot of F distribution approximation with adjusted degrees of freedom vs simulated data, Error from Weibull distribution, sample size at 100

3.3 Simulation study under unconditional distribution of X

In previous sections, the X's were assumed given and fixed. In Chapter 2, it has been proved that when sample size $n \to \infty$, $\frac{R^2}{(1-R^2)}$ and its function also converge to a normal distribution even when X is random. We also performed simulation tests to check the theory's performance when X's were random and sampled from distributions, and the simulation results were summarized in this section. In this simulation, we assumed that $Y = 0.5X_1 + X_2 + 1.5X_3 + \epsilon$. The X's were sampled from the four pre-defined distributions with various skewness and kurtosis values. The error term ϵ is sampled from the corresponding distribution having mean 0 and standard error, $\sqrt{3}$. In the simulation, error term and predictors are independent, but different correlation levels between X's are tested in the simulation. The correlated X's are sampled based on a normal copula and its correlation structure. Simulation results at sample size 1000 are summarized in Tables 3.23 to 3.24, and Figures 3.16 to 3.19. Table 3.23 and 3.24 compare variance estimation from Theorem 3 in Chapter 2 and variance estimates using the simulated R^2 values which showed good agreement between simulated variance and variance from formula. Figure 3.16 to 3.19 present QQPlots of $\frac{R^2}{(1-R^2)}$ and $Log(\frac{R^2}{(1-R^2)})$ with normal distribution as reference. These plots show that the normal distribution has good approximation to the simulated results when sample size is 1000, especially for $Log(\frac{R^2}{(1-R^2)})$.

Table 3.23: Variance comparison between simulated results and theory when correlation between Xs is zero, sample size=1000

			nonnormal Error	
	Normal	Weibull	Skew=1,Kurt=3	Skew=2,Kurt=7
Var of Numerator from simulation	68.115	143.723	86.334	111.038
Var of Numerator from theory	66.844	144.561	85.119	109.874
Var of Denominator from simulation	17.619	129.116	44.366	79.963
Var of Denominator from theory	17.854	129.990	44.566	80.171
Var of $\frac{R^2}{(1-R^2)}$ from simulation	0.010	0.037	0.016	0.025
Var of $\frac{R^2}{(1-R^2)}$ from theory	0.010	0.036	0.016	0.024
Var of $log(\frac{R^2}{(1-R^2)})$ from simulation	0.007	0.026	0.012	0.018
Var of $log(\frac{R^2}{(1-R^2)})$ from theory	0.007	0.025	0.012	0.017

Table 3.24: Variance comparison between simulated results and theory when correlation between Xs is 0.5, sample size=1000

			nonnormal Error	
	Normal	Weibull	Skew=1,Kurt=3	Skew=2,Kurt=7
Var of Numerator from simulation	156.754	431.618	227.766	318.164
Var of Numerator from theory	153.674	430.558	224.805	314.432
Var of Denominator from simulation	17.619	129.109	44.363	79.957
Var of Denominator from theory	17.854	129.985	44.567	80.173
Var of $\frac{R^2}{(1-R^2)}$ from simulation	0.026	0.106	0.047	0.073
Var of $\frac{R^2}{(1-R^2)}$ from theory	0.026	0.103	0.046	0.071
Var of $log(\frac{R^2}{(1-R^2)})$ from simulation	0.006	0.027	0.011	0.017
Var of $log(\frac{R^2}{(1-R^2)})$ from theory	0.006	0.026	0.011	0.017

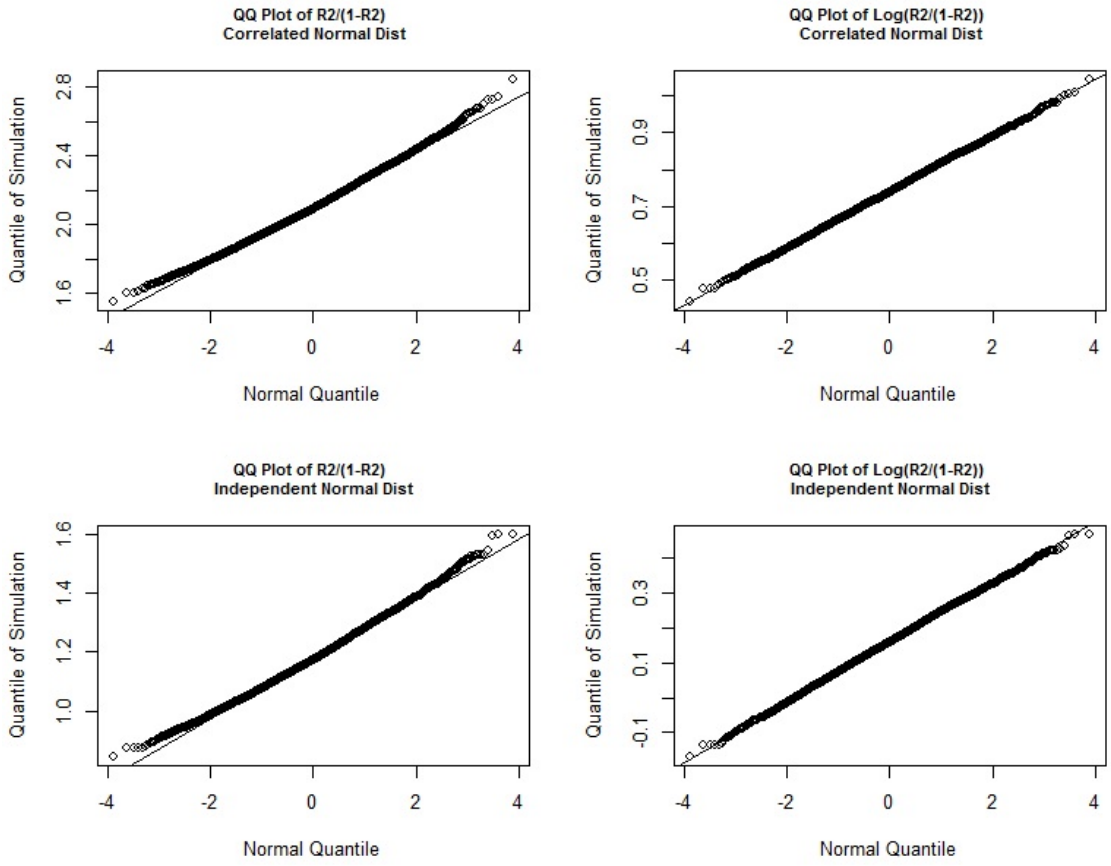

Figure 3.16: QQ Plots for Normal Distribution, Sample Size 1000

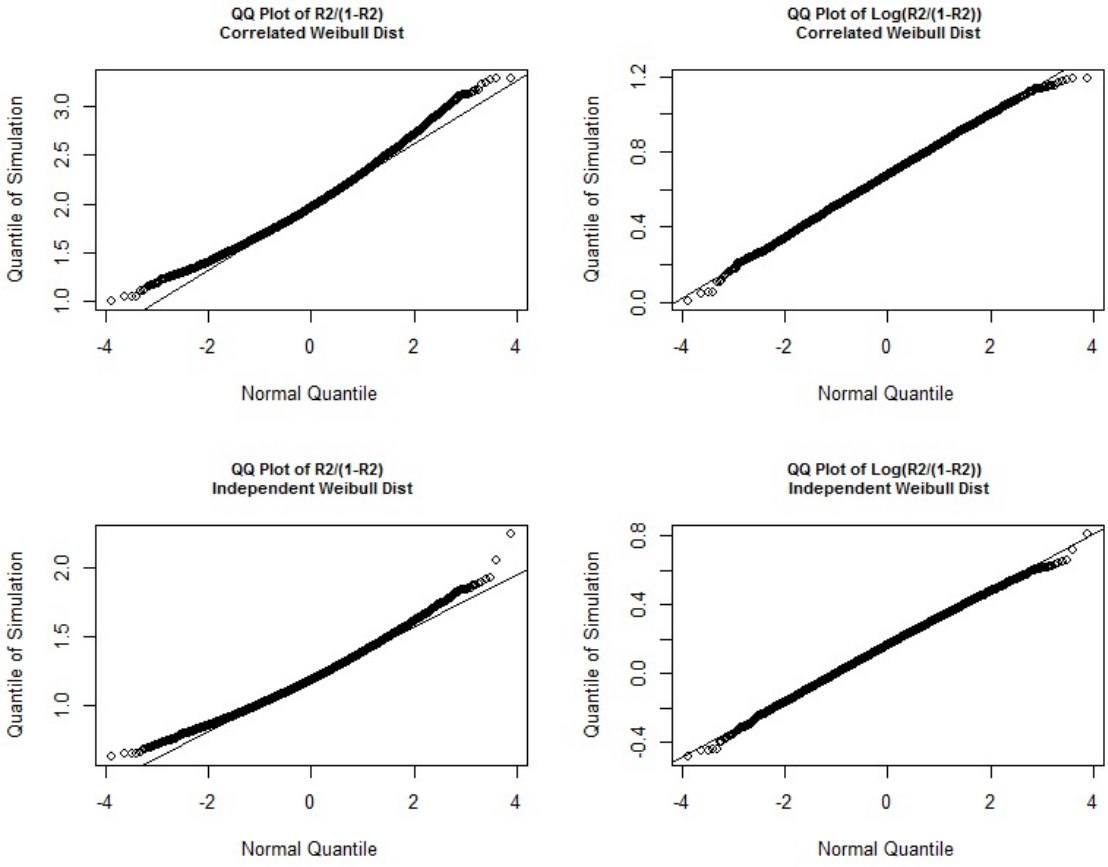

Figure 3.17: QQ Plots for Weibull Distribution, Sample Size 1000

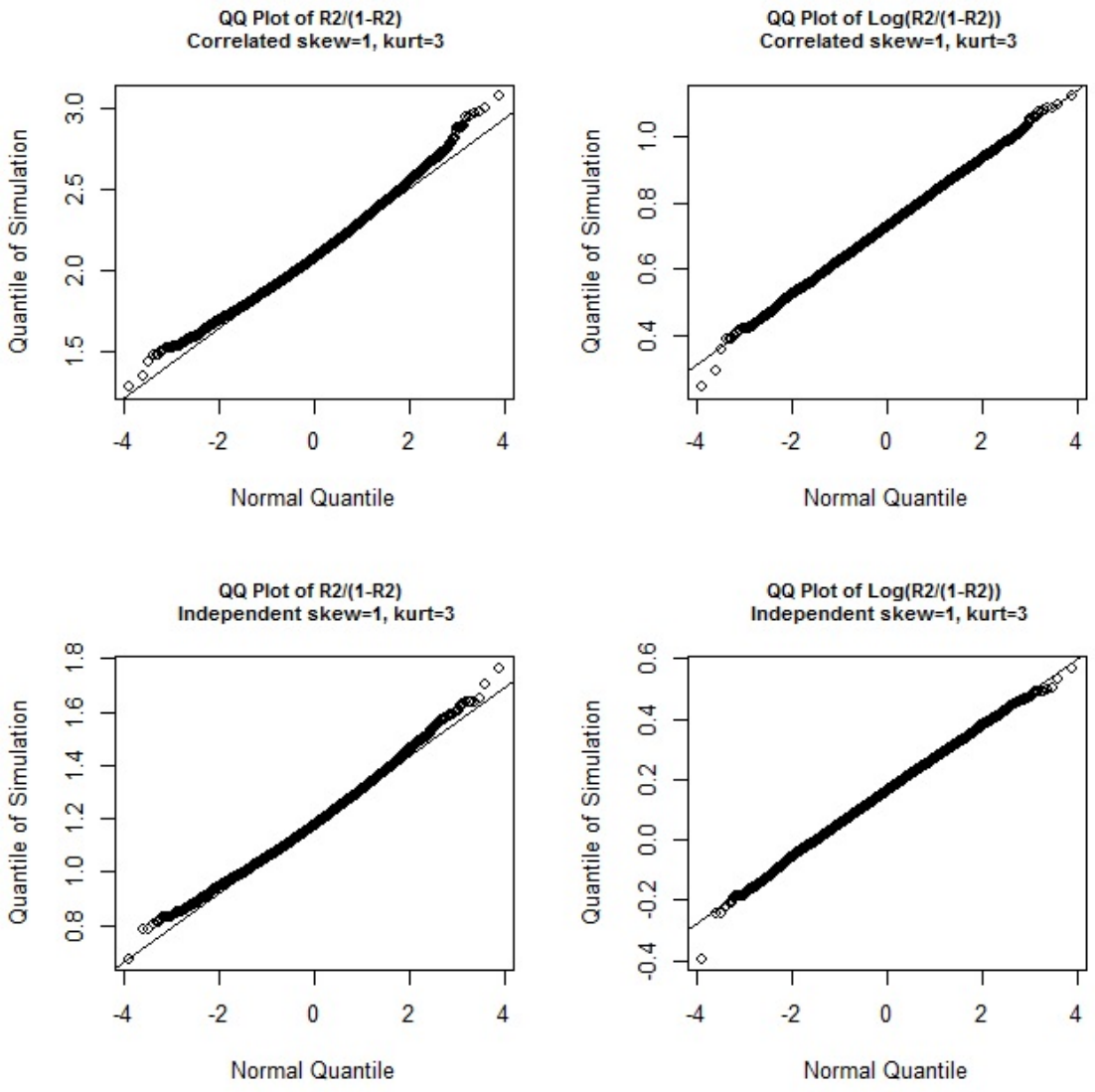

Figure 3.18: QQ Plots for nonnormal Distribution with skew=1 and kurt=3, Sample Size 1000

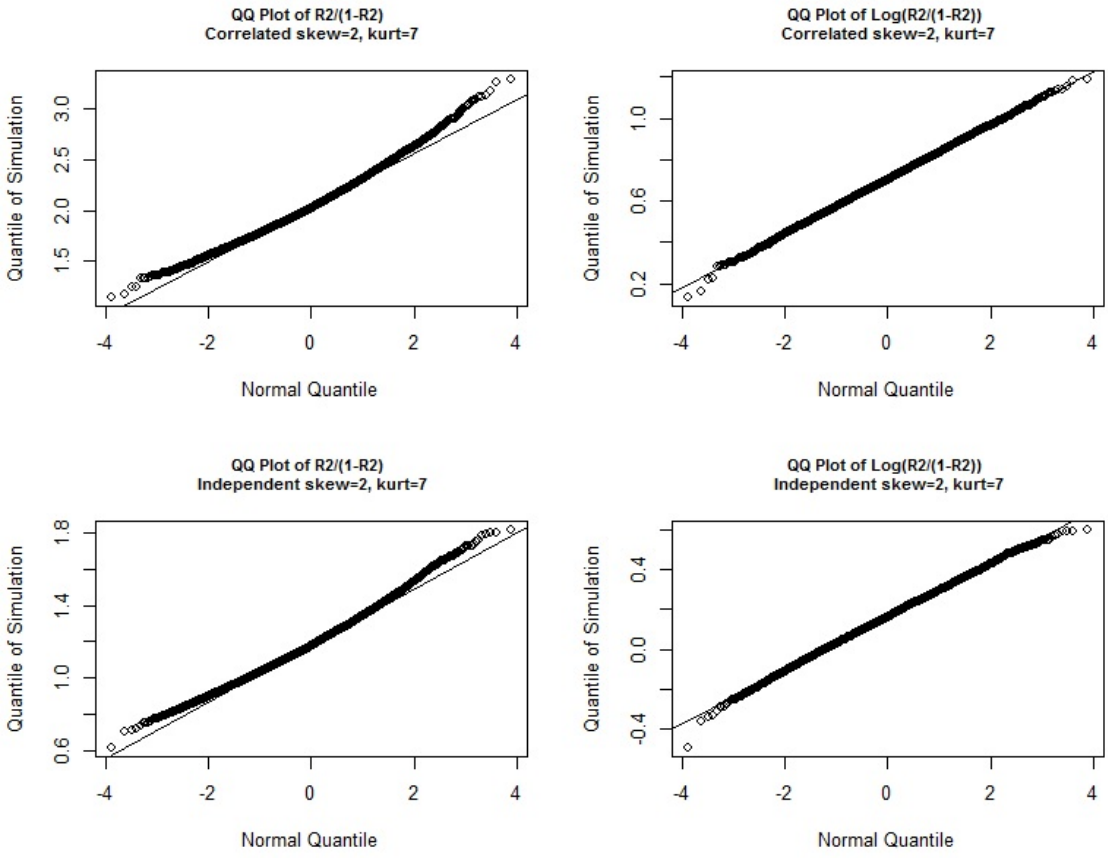

Figure 3.19: QQ Plots for nonnormal Distribution with skew=2 and kurt=7, Sample Size 1000

3.4 Summary of simulation studies

In this Chapter, simulation results were analyzed and presented which were used to test the performance of theories described in Chapter 2. In the simulation, two scenarios were considered: X was given and fixed, and X was random sampled from pre-specified distributions. Besides normal distributions, non-normal distributions were also tested to assess the impact on performance from non-normal error distributions. Simulation results showed that when sample size was reasonable large (1000 or more), there was good agreement between variance derived from formulas described in Chapter 2 and variance estimated from simulation results in both scenarios. Also, simulated $\frac{R^2}{(1-R^2)}$ and $log(\frac{R^2}{(1-R^2)})$ converged to normal distribution at large sample size, while the rate of convergence to normal distribution of $log(\frac{R^2}{(1-R^2)})$ was faster. Based on the asymptotic distribution of $log(\frac{R^2}{(1-R^2)})$, a confidence interval for R^2 was calculated, and the coverage ratio of a nominal 95% confidence interval was tested, which was consistent with the pre-determined level when sample size increased to 1000. A common problem in nutritional studies, with mass zero in the response variable, was also tested in the simulation. The distributional approximations of Chapter 2 also showed reasonable performance. When sample size was small or moderate, for example when sample size was 100, distributions of simulated $\frac{R^2}{(1-R^2)}$ and $log(\frac{R^2}{(1-R^2)})$ deviated from normal distribution. We also tested the performance of the adjusted F distribution by comparing percentiles of simulated results against quantiles from our adjusted F-distribution, showing improved performance when compared to that for an unadjusted F distribution.

Chapter 4

ESTIMATION OF R-SQUARED WITH PREDICTOR VARIABLE MEASUREMENT ERROR

In many regression applications, researchers are confronted with measurement error in key study variables. It is well known that measurement error in covariates can cause bias in the estimated regression coefficients. In-depth analysis of impact of measurement error on linear regression has been provided by Fuller (1987). Carroll et al (2006) expanded the analysis to cover non-linear models. Measurement errors in the covariates typically also lead to bias in R-squared estimation, which has not attracted much attention in previous studies. In this chapter, we first review the existing literature on measurement error methods and then discuss methodology for R-squared estimation in the presence of measurement error.

4.1 Literature review

Measurement error is a common issue in many studies. Specification of measurement error in mathematical form by defining a relationship between actual measurement with error and the latent variable; that is, the true value, typically is the first step in existing methodology. Different measurement error structures do not have the same impact on parameter estimation, and require model-specific analysis to estimate the impact, and to correct biases caused by the error. In addition, the complexity of analysis and data requirement on the correction methodology vary under different types of measurement error. In the literature, measurement error can be classified into two main classes : non-differential and differential error. The distinction between the two types of errors is discussed below in section 4.1.1. Furthermore, measurement error models that define the relationship between true value and observed measure can be categorized into two main classes: classical error and Berkson error. The definition of each error type and the distinction between them are described in section 4.1.2. Correction methodologies that have been developed in the existing literature can be distinguished by how they treat the latent variable (true value) and they can

be grouped into two major classes: functional methods and structural methods. Popular methodologies in each group are discussed in section 4.1.3. Finally, there is a brief introduction to the studies of measurement error and correction methodologies in nutritional studies. In the following discussion, W denotes the observed actual measurement possibly with error on the true value X, and Y defines the response variable. In addition, Z includes other individual characteristics that are assumed to be without measurement errors.

4.1.1 Differential and Nondifferential Measurement Error

If the error in W contains extra information about the response variable Y that is not explained by true value X and Z, then the conditional distribution of Y given W, X and Z will be different from the distribution of Y given X and Z only. With this kind of error, without observing true values for some study subjects, it is typically not possible to estimate the relationship between Y and X from Y and W. Thus, in measurement error analysis, it is important to distinguish two types of measurement error, differential and nondifferential. The error in W is nondifferential if the distribution of Y conditional on (Z, X) does not change after including the observed measurement,W, i.e., $f_{Y|ZXW} = f_{Y|ZX}$ where f denotes the corresponding distribution function for Y. If $f_{Y|ZXW}$ and $f_{Y|ZX}$ are not equal, the error is differential. Nondifferential error indicates that the related measurement of X, W, does not contain more information about Y after accounting for the associations between Y and (Z, X). W with nondifferential error is sometimes called a surrogate for X.

To estimate the relationship between the response Y and the true predictor value, X, the analysis methodology will depend on the type of measurement error. For nondifferential error, W contains the same information when predicting Y when observing X, thus with minimal assumptions, it is possible to derive corresponding regression parameter estimates for X from W without observing the true measurements X. However, with differential error, it is necessary to record X values at least in a validation subsample to generate the corresponding parameter estimates. As discussed in Buzas (2003), differential error problems are typically addressed using missing data techniques, while measurement error research mainly focuses on nondifferential errors. In this thesis, we focus on nondifferential

measurement error.

4.1.2 Measurement Error Model

There are various ways to model and present the relationship between a surrogate variable W and X. Various error structures can have different effects on model estimation and statistical methods to analyze measurement error also may need to be adjusted for different model structures. Some types of measurement error structure are summarized below:

- Classical error model: Classical measurement error is an additive model, assuming a simple structure $W = X + U$. The error model has been studied by many researchers, for example, the OPEN study report by Subar, et al.(2001). Usually the measurement error U is assumed to be independent from X and Y so that the error is nondifferential and W is a surrogate. It is also often assumed that the error term has mean 0 indicating that W is an unbiased measurement of X. Due to its simple structure, the model is convenient for statistical analysis. However in real applications, it may happen that W is biased, and is associated with other characteristics denoted by Z. Especially in nutritional studies, systematically biased measurement errors were identified in intake measurements from FFQ, food record and 24-hour recall procedures. Thus the classical error model was extended to be applicable to a broader set of applications. In the extended model structure, W is presented as $W = \gamma_0 + \gamma_X X + \gamma_Z Z + U$, being a function of both X and Z. In dietary studies, the Z's often denote personal characteristics such as body mass index (BMI) and age. Through error calibration, the extended linear error model can be mapped to a classical error model. For example, let $W^* = \frac{W-\gamma_0-\gamma_Z Z}{\gamma_X}$ and $U^* = \frac{U}{\gamma_X}$, then $W^* = X + U^*$ while W^* represents a measure with classical error.

- Berkson error model: In the classical measurement error model, the actual observed W values are more variable than the underlying true values. Berkson (1950) defined another type of error model, in which X is varying around W with larger variability. The model can be presented in mathematical formula as $X = W + U$, where U is

independent from W with mean zero. Under this model, $E(X|W) = W$. Examples of Berkson model arise from experimental studies or exposure assessment in environment studies. For example, W may represent average exposure to a group of objects, such as air pollution level measured in one area while X is the actual exposure for each individual. There are more examples explained in the book by Carroll et al (2006). The Berkson error model can also be extended to a more complicated form, as $X = \theta_0 + \theta_W W + \theta_Z Z + V$ where V is independent from W and Z with mean 0. One popular correction methodology, regression calibration, transforms a surrogate variable W under a classical measurement model into a Berkson error model.

- Customized measurement error model: the two measurement error strcutures defined above are both standard and additive. For a specific study, neither error structure may be appropriate. Thus, there are some ad hoc error structures that have been developed in the literature and they generally are nonlinear and more complicated. For example, Pierce (1992) proposed a multiplicative error structure, W defined as $W = XU$. A log-transformation of this model may yield a classical measurement model. Transformation models are frequently used in measurement error modeling.

4.1.3 Measurement Error's Impact on Linear Regression

The impact of measurement error on linear regression has been extensively studied, and in-depth discussion is provided in Caroll et al (2006) and Fuller (1987). Here we will briefly summarize the measurement error effects within simple linear regression and multiple linear regression models. In a simple linear regression, $Y = \beta_x X + \epsilon$, then $E(Y|X) = \beta_x X$. Then for the surrogate variable W, $E(Y|W) = E\{E(Y|X,W)|W\} = E\{E(Y|X)|W\} = \beta_x E(X|W)$. Under classicial error model and joint normal distribution assumption, $W = X + U$, we can derive that $E(X|W) = \lambda_X W$, where $\lambda_X = \frac{\sigma_X^2}{\sigma_W^2} = \frac{\sigma_X^2}{(\sigma_X^2+\sigma_U^2)}$. We can observe the attenuation effects of β with the ratio λ_X called the reliability ratio. When the error follows Berkson structure, $X = W + U$, then $E(X|W) = W$. Then there will no attenuation effect on β_X, $\beta_W = \beta_X$.

For a multiple linear regression model, the effect of classical measurement error on

parameter estimation is not so simple. Assuming a linear relationship between Y and (X, Z) presented as $Y = \beta_X X + \beta_Z Z + \epsilon$, after replacing X with the surrogate W, the conditional expectation of Y on given W, Z is also a linear function as $E(Y|W, Z) = \beta_W W + \beta_Z^* Z$. Let $\sum_{\mu\mu}$ denote the covariance matrix of measurement error, U, then the coefficients under measurement error, β_W and β_Z^*, in matrix form are

$$\begin{pmatrix} \beta_W \\ \beta_Z^* \end{pmatrix} = \begin{pmatrix} \sum_{WW} & \sum_{XZ} \\ \sum_{XZ} & \sum_{ZZ} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{XX} & \sum_{XZ} \\ \sum_{XZ} & \sum_{ZZ} \end{pmatrix} \begin{pmatrix} \beta_X \\ \beta_Z \end{pmatrix} \tag{4.1}$$

$$\begin{aligned} \begin{pmatrix} \beta_X \\ \beta_Z \end{pmatrix} &= \begin{pmatrix} \sum_{XX} & \sum_{XZ} \\ \sum_{XZ} & \sum_{ZZ} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{XX} + \sum_{\mu\mu} & \sum_{XZ} \\ \sum_{XZ} & \sum_{ZZ} \end{pmatrix} \begin{pmatrix} \beta_W \\ \beta_Z^* \end{pmatrix} \\ &= \begin{pmatrix} \beta_W \\ \beta_Z^* \end{pmatrix} + \begin{pmatrix} \sum_{XX} & \sum_{XZ} \\ \sum_{XZ} & \sum_{ZZ} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{WW} - \sum_{XX} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_W \\ \beta_Z^* \end{pmatrix} \end{aligned} \tag{4.2}$$

Thus, coefficients can be inflated or attenuated towards zero. The direction of impact depends on the correlation structure between X and Z. If (X, Z, W) is from a multivariate normal distribution, expected value of X conditional on (Z, W) is a linear function $E(X|Z, W) = \gamma_Z Z + \gamma_W W$, and Expression (4.3) was also similar to simple linear regression:

$$E(Y|Z, W) = E\{E(Y|X, W, Z)|Z, W\} = (\beta_Z + \beta_X \times \gamma_Z)Z + \beta_X \times \gamma_W W \tag{4.3}$$

So, we can obtain the relationship between (β_W^*, β_Z^*) and true coefficients (β_X, β_Z) as $\beta_W^* = \beta_X \gamma_W$ and $\beta_Z^* = (\beta_Z + \beta_X \times \gamma_Z)$. Caroll et al (2006) and Fuller (1987) provided more details about the variance of the corrected estimates.

4.1.4 Measurement Error Correction Methodology

Methods to correct biases caused by measurement error have been an interesting topic for a long time and there has been considerable related statistical research. The developments in the literature can be classified using two distinct types of models: functional methods and structural methods. These two classes of models make different assumptions on the latent variable X. Functional methods treats the unobserved true values X as model parameters

to be estimated, while structural methods consider X as a random variable drawing from a specific distribution. Usually structural model correction procedures are based on likelihood methods. One advantage of the functional approach is that it makes fewer assumptions on X. This section describes some standard structural model correction methods developed in the literature.

- Regression calibration: Regression calibration is a structural method, and has been successfully applied to a broad range of applications. This method was introduced by Prentice (1982) in the context of a proportional hazard regression model. The methodology has been generalized to different applications such as logistic regression and generalized linear models. The method replaces X with its conditional expectation given W and Z, denoted by $E(X|W,Z)$. If the dependent variable Y has expectation that is a function of X and Z, defined as $E(Y|X,Z) = f(X,Z)$, then this method approximates Y as $E(Y|X,Z) \simeq f(E(X|W,Z),Z)$. To implement this method, there are two steps:

  - step 1, estimate the conditional expectation function of unobserved X given observed measurement W and other covariates Z. The estimation generally is based on repeated measures or validation data.
  - step 2, use the estimation results from step 1, $E(X|W,Z)$, to approximate X, and proceed with the analysis to estimate the relationship between Y and X,Z.

  In step 2, the standard errors of estimated parameters should be adjusted since the replacement $E(X|W,Z)$ is estimated from data in step 1. The Sandwich method or bootstrap method can be used to estimate the variance. In some nutrition studies, there are minor changes on the implementation. Actual nutrition intake X is not observable, however, in a biomarker subsample, there is a measurement of X, denoted as Q, that follows a classical error model being unbiased estimator for X, such as energy or protein biomarkers. The measurement W from FFQ or 24-hour recall is assumed to follow an extended classical error model as described above. Thus, instead of calibrating to X directly in step 1, the estimation of $E(X|W,Z)$ is approximated

by regressing Q on W, Z. When the measurement error in Q is independent from both W and Z, it is straightforward to show that $E(Q|W,Z) = E(X|W,Z)$. The regression calibration model is consistent for linear models, but can't remove all bias for non-linear models, and is only approximately consistent in this context.

- SIMEX method: SIMEX refers to simulation extrapolation. It was first developed by Cook and Stefanski (1994, 1995). Caroll et al (2006) provided a comprehensive summary of this method. It is also a structural method, which does not make any assumption on the distribution of the underlying unobserved X. This method assumes that bias in the estimated parameters induced by measurement error is a function of the variance of measurement errors. To get the function and to correct bias, there are also two steps in the implementation:

  - Simulation step: in this step, measurement error is simulated and added to observed measurement W. Then the corresponding parameters are estimated based on the simulated measurement. The simulation process will repeat several times with incremental variance in the measurement error.
  - Extrapolation step: This step uses estimation results from the simulation step to develop the relationship between variance of measurement error and the corresponding bias in parameter estimation. Then the function will be extrapolated to the case with zero measurement error variance to calculate parameters without measurement error.

We use the simple additive measurement error model as an example to illustrate the calculation methodology. Under the classical error model, $W = X + U$, where U is from a distribution with mean 0 and variance σ_u^2. Then in each simulation, additional variance will be added to the measurement error, so that $W_\alpha = W + U_\alpha$, where U_α have zero mean and variance $\alpha\sigma_u^2$. Then the estimated coefficients θ is considered as a function of α, denoted as $\theta(\alpha)$. Simulation results can be used to develop the function, and extrapolate to $\alpha = -1$ under which there are no biases. The major

limitation for this methodology is that it assumes that either there is prior knowledge of measurement error variance or the variance can be estimated from another dataset.

- Score function method: Score function methodology is based on estimation equations that are used to estimate model parameters. Estimation equation often is derived from a score function, as the derivative of a corresponding likelihood function. Expectation of the score function is zero denoted as $E[\psi(Y, X, Z, \Theta)] = 0$ where ψ indicates the score function and Θ is the model parameters. When there are no measurement errors and true X values are available, we can solve Θ through the estimation equation: $\sum_{i=1}^{n} \psi(Y_i, X_i, Z_i, \Theta) = 0$. However, when replacing X with W with measurement error, $E[\psi(Y, W, Z, \Theta)] \neq 0$. Thus when using the same estimation equations to solve Θ, it will result in bias in parameter estimation. To correct the bias, there are two popular ways:

  - Corrected score method: this method was first developed and discussed by Stefanski (1989) and Nakamura (1990). The underlying idea is to find a corrected score function that is based on the observed measurement W instead of the true X. The corrected function has the property that its expectation is the same as the uncorrected score function using X, i.e. $E[\psi^*(Y, W_i, Z, \Theta)] = \psi(Y, X_i, Z, \Theta)$. For the simple linear regression, and classical error $W = X + U$, it has been shown that $\psi^*(Y, W_i, \theta) = \psi(Y, W_i, \Theta) + \sigma_u^2 * \theta$, where σ_u is the measurement error variance. In some nonlinear models, the corrected score function is not easy to identify. Novick and Stefanski (2002) developed a general approach that is based on Montecarlo simulation and is applicable to a much broader range of applications. It is called Monte Carlo Corrected Score (MCCS). This method is similar to Simex. Instead of adding additional random error to W, MCCS introduces a complex number with real part equal to the observation W_i but the image part is from the added random variance. The whole calculation process follows:

    * Assuming the simluation repeats k times, and for one sample, a random

measurement error $U_{j,i}$ is generated from the distribution of measurement error, and form a complex variable as $\tilde{W_{j,i}} = W_i + img(U_{j,i})$.

* The corrected estimation score function is the average of the real part of the score function $\psi_{mccs}(Y_i, W_i, Z_i, \Theta) = k^{-1} \sum_{j=1}^{k} \psi(Y_i, \tilde{W_{j,i}}, Z_i, \Theta)$.

* Solve the equation $\sum_{i=1}^{n} \psi_{mccs}(Y_i, W_i, Z_i, \Theta) = 0$ to estimate Θ.

It has been proved that corrected score method estimation is consistent (Caroll et al 2006).

– Conditional score method: the conditional score function method is based on the theory of sufficient statistics. The underlying approach for this functional method is to find a sufficient statistic that is a function of W, and then the score function conditional on W will remove the nuisance variable X. The conditional score function can be used to estimate model parameters. This methodology was proposed by Stafanski and Caroll (1987). It has good properties for the linear regression and logistic regression (Stefanski and Caroll, 1990). Conditional score function for other models besides linear and logistic regression model is complicated and hard to compute. Research in this literature focuses on measurement error from normal distribution, and there is evidence of lack of robustness to departures from normality(e.g. Shaw and Prentice, 2012).

- Likelihood methodology: Those methods discussed above are mainly structural methods. The desirable properties of maximum likelihood analysis on estimation efficiency and confidence interval estimation have also attracted interest of many researchers. In the literature, some studies make distributional assumptions on X and then develop the full parameterized likelihood function to estimate parameters as $f(Y|Z, W) = f(Y|Z, X, \beta_X)f(X|W, Z)$. Instead of assuming the full distribution of underlying X, some studies just make assumptions on the first and second moments and estimate parameters via the quasi likelihood method. One major limitation of this methodology is that it requires that X is observable in a subset of a study cohort to fit the distribution or to estimate the first or second moment. This limits its application

scope.

4.2 Measurement Error in nutritional studies

Measurement error is a common issue in nutritional studies. Nutritional research includes interest in the actual intake of a given nutrient, some type of food, or energy. As discussed in Chapter 1, FFQ, self-reported food frequency questionnaires, and 24-hour recalls are popular methodologies considering their easy implementation and low cost. However, systematic bias and measurement error were identified in those dietary measurements. It is challenging to correct bias since it is very difficult or impossible to estimate the true intake which may, for example, be defined as average daily consumption over the preceding months or years that may be relevant to health outcomes. Recent development in related biomarkers provides a way to perform the regression calibration. Those biomarkers of dietary intake were considered as an unbiased measure of the true consumption (Kaaks et al, 1997) that adhered to a classical measurement model. These biomarkers have been used as the reference in regression calibration to calibrate the function $E(X|W)$ and then use this function to estimate the true relationship between Y and X (Prentice et al, 2009, Prentice et al, 2011). However, most of these studies use a joint normal distribution assumption so that $E(X|W)$ can be presented in a linear form. In the next section, we expand the application scope and show that under some conditions, regression calibration still leads to asymptotically unbiased parameter estimation. Some research also extends the usage of biomarkers, using it as a benchmark to compare various intake measurement methodologies such as food record, FFQ and recalls based on the metric, R-Squared (Prentice et al, 2011). R-squared of a linear regression model will be skewed in the presence of measurement error either in response variable or independent variables. This chapter also includes a final section to discuss R-squared correction methodology.

4.3 Regression calibration in linear regression under non-normal measurement error

In a linear regression model, Y is a linear function of (X, Z), denoted as $Y = \beta_X X + \beta_Z Z + \epsilon$, where ϵ has mean 0 and variance σ_ϵ^2. W, the actual measurement of X, generally contains

measurement error. In the context of a nutritional study, W follows an extended classical error model, presented as $W = \gamma_0 + \gamma_X X + \gamma_Z Z + U$. When measurement error is from a normal distribution, and (W, Z, X) are from a joint normal distribution, $E(X|W, Z)$ is a linear function of (W, Z) denoted as $E(X|W, Z) = \alpha + \lambda_W W + \lambda_Z Z$. This is the usual first step in regression calibration methodology. However, when the error is from a non-normal distribution, $E(X|W, Z)$ generally will not be linear. Thus when following the regression calibration method, the estimated linear function from step 1 does not represent $E(X|W, Z)$. Hence, whether the corrected parameter estimation is still asymptotically consistent is a research issue. The following theorem shows that under a non-normal error distribution, the methodology still generates consistent parameter estimation.

Theorem 4. *Assuming that* $Y = \beta_X X + \beta_Z Z + \epsilon$, *and* $\tilde{X} = \lambda_W W + \lambda_Z Z$ *as a linear estimator of* X *that minimizes squared error defined as* $E(\tilde{X} - X)^2$. *Then* $(\hat{\beta_X}, \hat{\beta_Z}) = argmin[E(Y - \beta_1\tilde{X} - \beta_2 Z)^2]$, *is consistent for* (β_X, β_Z).

Proof. Since λ_W and λ_Z are minimizing the objective function $E(X - \lambda_1 W - \lambda_2 Z)^2$, then λ_W and λ_Z solve two equations $E\{(X-\lambda_1 W-\lambda_2 Z)W\} = 0$ and $E\{(X-\lambda_1 W-\lambda_2 Z)Z\} = 0$.

$$
\begin{aligned}
& E\{(Y - \beta_1\tilde{X} - \beta_2 Z)^2\} \\
&= E\{(\beta_X X + \beta_Z Z - \beta_1\tilde{X} - \beta_2 Z + \epsilon)^2\} \\
&= E\{(\beta_X(X - \tilde{X}) + (\beta_X - \beta_1)\tilde{X} + (\beta_Z - \beta_2)Z + \epsilon)^2\} \\
&= E\{\beta_X^2(X - \tilde{X})^2 + (\beta_X - \beta_1)^2\tilde{X}^2 + (\beta_Z - \beta_2)^2 Z^2 + \epsilon^2\} \\
&+ 2 * E\{\beta_X(\beta_X - \beta_1)(X - \tilde{X})\tilde{X} + \beta_X(\beta_Z - \beta_2)(X - \tilde{X})Z + (\beta_Z - \beta_2)(\beta_X - \beta_1)\tilde{X}Z\} \\
&= E\{\beta_X^2(X - \tilde{X})^2 + ((\beta_X - \beta_1)\tilde{X} + (\beta_Z - \beta_2)Z)^2\}
\end{aligned}
$$

Thus, when $\beta_1 = \beta_X$ and $\beta_2 = \beta_Z$, the objective function is minimized. □

Since β_X and β_Z minimize the objective function $E\{(Y - \beta_1\tilde{X} - \beta_2 Z)^2\}$ where $\tilde{X}$ is a linear combination of W and Z that minimizes the function $E(\tilde{X} - X)^2$, taking the first derivative of the objective function, we can get that $E\{(Y - \beta_X\tilde{X} - \beta_Z Z)\tilde{X}\} = 0$ and $E\{(Y - \beta_X\tilde{X} - \beta_Z Z)Z\} = 0$. Without losing generality, we assume that X,W, and Z have

zero mean. Now β_X and β_Z can be solved as:

$$\begin{pmatrix} \beta_X \\ \beta_Z \end{pmatrix} = \begin{pmatrix} \lambda_W^2\sigma_{WW}^2 + \lambda_Z^2\sigma_{ZZ}^2 + 2 * \lambda_W\lambda_Z\sigma_{WZ} & \lambda_W\sigma_{WZ} + \lambda_Z\sigma_{ZZ}^2 \\ \lambda_W\sigma_{WZ} + \lambda_Z\sigma_{ZZ}^2 & \sigma_{ZZ}^2 \end{pmatrix}^{-1} \begin{pmatrix} \lambda_W\sigma_{YW} + \lambda_Z\sigma_{YZ} \\ \sigma_{YZ} \end{pmatrix} \tag{4.4}$$

where (λ_W, λ_Z) minimizes $E(\tilde{X} - X)^2$, and by solving the two equations $E\{(X - \lambda_1 W - \lambda_2 Z)W\} = 0$ and $E\{(X - \lambda_1 W - \lambda_2 Z)Z\} = 0$ we can get that:

$$\begin{pmatrix} \lambda_W \\ \lambda_Z \end{pmatrix} = \begin{pmatrix} \sigma_{WW}^2 & \sigma_{WZ} \\ \sigma_{WZ} & \sigma_{ZZ}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sigma_{XW} \\ \sigma_{XZ} \end{pmatrix} \tag{4.5}$$

In a real application, (λ_W, λ_Z) are estimated based on a biomarker data set, solving the following two equations, and the estimators are denoted as $(\widehat{\lambda_W}, \widehat{\lambda_Z})$,

$$\begin{aligned} &\sum_{i=1}^{n}(x_i - \lambda_W w_i - \lambda_Z z_i)w_i = 0 \\ &\sum_{i=1}^{n}(x_i - \lambda_W w_i - \lambda_Z z_i)z_i = 0 \end{aligned} \tag{4.6}$$

By solving equation (4.6), we obtain the solution

$$\begin{pmatrix} \widehat{\lambda_W} \\ \widehat{\lambda_Z} \end{pmatrix} = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}(w_i w_i) & \frac{1}{n}\sum_{i=1}^{n}(w_i z_i) \\ \frac{1}{n}\sum_{i=1}^{n}(w_i z_i) & \frac{1}{n}\sum_{i=1}^{n}(z_i z_i) \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n}(x_i w_i) \\ \frac{1}{n}\sum_{i=1}^{n}(x_i z_i) \end{pmatrix} \tag{4.7}$$

As $n \to \infty$, it is straightforward to show that $(\hat{\lambda_W}, \hat{\lambda_Z})$ are asymptotically consistent for (λ_W, λ_Z) from equation (4.5) and (4.7). (β_X, β_Z) are estimated by solving the following equations on study population for given estimated (λ_W, λ_Z)

$$\begin{aligned} &\sum_{i=1}^{N}(x_i - \beta_X(\widehat{\lambda_W}w_i + \widehat{\lambda_Z}z_i) - \beta_Z z_i)(\widehat{\lambda_W}w_i + \widehat{\lambda_Z}z_i) = 0 \\ &\sum_{i=1}^{N}(x_i - \beta_X(\widehat{\lambda_W}w_i + \widehat{\lambda_Z}z_i) - \beta_Z z_i)z_i = 0 \end{aligned} \tag{4.8}$$

As shown in (4.4) and the equation (4.8), (β_X, β_Z) is a function of (λ_W, λ_Z). Thus by the continuous mapping theorem, $(\hat{\beta_X}, \hat{\beta_Z})$ is consistent for (β_X, β_Z) when $N \to \infty$.

4.4 R-squared estimation under a flexible model for measurement error in predictor variables

As described above, R-squared is a common metric to measure the explanatory power of predictors in a linear regression model. However, in real application, the true value of predictor variables X may not be observable because of measurement error embedded in the actual measurement W. This may cause bias in R-squared estimates resulting in improper inference in the relationship between response variable Y and predictor variables X. In the following presentation, we assume that in the underlying true model Y is a linear function of X and Z defined as $Y = \beta_X X + \beta_Z^t Z + \epsilon$. To accomodate non-normal measurement error, we consider the version of generalized R-squared discussed above, given by $\tilde{R}^2_{Y|XZ} = 1 - \frac{\sigma^2_{Y|X,Z}}{\sigma^2_Y}$, where $\sigma^2_{Y|X,Z}$ is the variance of conditional distribution of Y on given (X, Z).

Theorem 5. *Assuming that $Y = \beta_X X + \beta_Z^t Z + \epsilon$, and W is a surrogate measurement for X, then $\tilde{R}^2_{Y|XZ} = \tilde{R}^2_{Y|WZ} + \frac{\beta_X^2 \sigma^2_{X|W,Z}}{\sigma^2_Y}$*

Proof.

$$
\begin{aligned}
&E(Y|W,Z) = \beta_X E(X|W,Z) + \beta_Z^t Z \\
&\sigma^2_{Y|W,Z} = \beta_X^2 \sigma^2_{X|W,Z} + \sigma^2_\epsilon \\
&\tilde{R}^2_{Y|X,Z} = 1 - \frac{\sigma^2_\epsilon}{\sigma^2_Y} \\
&\tilde{R}^2_{Y|W,Z} = 1 - \frac{\sigma^2_{Y|W,Z}}{\sigma^2_Y} = \tilde{R}^2_{Y|X,Z} - \frac{\beta_X^2 \sigma^2_{X|W,Z}}{\sigma^2_Y}
\end{aligned}
$$

□

When (X, W, Z) are from a joint normal distribution, the generalized R-squared is equivalent to traditional R-squared and the theorem can be applied to linear regression model for R-squared correction.

4.5 R-squared correction estimation

The theorem presented in the previous section can be used to correct R-squared estimators in the presence of measurement error. However this correction requires knowledge of actual coefficient β_X and the variance term, $\sigma^2_{X|WZ}$ which are generally estimated from biomarker subsample or are based on prior knowledge. Actual parameter values and variance are typically not available, thus this section focuses on the estimation using a biomarker dataset. In many nutritional studies, instead of observing the true value in the biomarker subsample, unbiased objective measurements using biomakers are available, and typically there are repeated measures on the same subjects. In the following discussion, biomarker measures are denoted as Q and we assume that there are two repeat measures defined by Q_1 and Q_2. The biomarker measure is considered as an unbiased measure of X and is defined in the mathematical formula as $Q = X + \upsilon$. In a typical study design, we assume that biomarker subsample dataset have n subjects and include (Y, W, Z) and biomarker measure Q, while there are N subjects in the study population with (Y, W, Z) only. Then the biomarker subsample can be used to estimate (β_X, β_Z) and $\sigma^2_{X|WZ}$ to calculate the R-squared correction item. The estimation procedure of the two terms, β_X and $\sigma^2_{X|WZ}$ is explained below:

- Estimation of $\sigma^2_{X|WZ}$: since $Q = X+\upsilon$, $\sigma^2_{Q|W,Z} = \sigma^2_{X|W,Z}+\sigma^2_\upsilon$, so $\sigma^2_{X|W,Z} = \sigma^2_{Q|W,Z}-\sigma^2_\upsilon$. $\sigma^2_{Q|W,Z}$ is estimated from the regression of Q_n on W_n and Z_n, and the σ^2_υ is estimated by $\widetilde{\sigma^2_\upsilon} = \sum_{i=1}^{n}(Q_{1i} - Q_{2i})^2/2n$. In multidimensional study, first regress each individual variable, $Q_1, ..., Q_p$ on W and Z to get vectors of residuals, $\upsilon_1, ..., \upsilon_p$. And define $\overline{\sum_{\upsilon\upsilon}} = Cov(\upsilon_1, ..., \upsilon_p)$. Then $Cov(X|W, Z) = Cov(Q|W, Z) - \overline{\sum_{\upsilon\upsilon}}$ where $\sum_{\upsilon\upsilon}$ is estimated as $Var(Q_1 - Q_2)/2n$.

- Estimation of β_X: the estimation can be implemented via two different ways:

  - the first methodology is based on a calibration equation:

    * step 1: Develop a calibration equation by regressing Q_n on W_n and Z_n based on the biomarker subsample.

  - step 2: Put W_N and Z_N in the calibration equation to get an estimator of X, denoted $\overline{X_N}$, then develop regression model by regressing Y_N on $\overline{X_N}$ and Z_N.
  - In multidimensional case, the estimation follows the same procedure. Regress $Q_1, ..., Q_p$ on W and Z to get calibration equations, and use the calibration equations to estimate the corresponding X values. Estimated X values plus observed Z are put in the regression with Y to get the estimation results.

- The second methodology is based on the biomarker measurements, Q via method of moment.
  - Regress Y_n on Q_n and Z_n to get an estimator of $\overline{\beta_Q}$.
  - Correct $\overline{\beta_Q}$ to estimate β_X. $\beta_X = \frac{\overline{\beta_Q}}{\lambda}$ where $\lambda = \frac{Var(X|Z)}{Var(Q|Z)} = \frac{(Var(Q|Z) - \sigma_v^2)}{Var(Q|Z)}$.
  - In multidimensional study, the calculation is similar but in matrix format. Assuming $Y = Q\beta_Q + Z\beta_Z^* + e$, then

$$\begin{pmatrix} \beta_Q \\ \beta_Z^* \end{pmatrix} = \begin{pmatrix} \sum_{XX} + \sum_{vv} & \sum_{XZ} \\ \sum_{XZ} & \sum_{ZZ} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{XX} & \sum_{XZ} \\ \sum_{XZ} & \sum_{ZZ} \end{pmatrix} \begin{pmatrix} \beta_X \\ \beta_Z \end{pmatrix}, \tag{4.9}$$

$$\begin{aligned} \begin{pmatrix} \beta_X \\ \beta_Z \end{pmatrix} &= \begin{pmatrix} \sum_{XX} & \sum_{XZ} \\ \sum_{XZ} & \sum_{ZZ} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{XX} + \sum_{vv} & \sum_{XZ} \\ \sum_{XZ} & \sum_{ZZ} \end{pmatrix} \begin{pmatrix} \beta_Q \\ \beta_Z^* \end{pmatrix} \\ &= \begin{pmatrix} \beta_Q \\ \beta_Z^* \end{pmatrix} + \begin{pmatrix} \sum_{XX} & \sum_{XZ} \\ \sum_{XZ} & \sum_{ZZ} \end{pmatrix}^{-1} \begin{pmatrix} \sum_{vv} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \beta_Q \\ \beta_Z^* \end{pmatrix}. \end{aligned} \tag{4.10}$$

In the formula, β_Q and β_Z^* are estimated by regressing on (Q, Z) using the data with biomarker measure, and $Cov(X, Z)$ is estimated through standard convariance estimation method. Estimation of $\sum_{vv}$ is explained above.

Lemma 2. *Assume measurement error to be from a normal distribution,* $Y = \beta_X X + \beta_Z Z$

and (X, Z, W) *are from a joint normal distribution, as validation dataset sample size* $n \to \infty$ *and* $N \to \infty$*, the corrected R-squared* $\hat{R_{Y|XZ}}$ *is asymtotically unbiased for* $R_{Y|XZ}$

Proof. As $n \to \infty$, by standard statistical theory, $\widetilde{\sigma_v^2} = \sum_{i=1}^{n}(Q_{1i} - Q_{2i})^2/2n$ converges to σ_v^2 in probability, and $\widetilde{\sigma^2_{Q|W,Z}}$ estimated from the regression model of Q on (W, Z) converges in probability to $\sigma^2_{Q|W,Z}$, thus the estimator based on validation dataset $\widetilde{\sigma^2_{X|W,Z}} = \widetilde{\sigma^2_{Q|W,Z}} - \widetilde{\sigma_v^2}$ is asymptotically consistent for $\sigma^2_{X|W,Z}$. Theorem 4 described in the previous section showed that as $(n, N) \to \infty$, the estimated $\widetilde{\beta_X}$ is asymptotically consistent for the true β_X. Since the method of moment is used to estimate β_X, when the size of the estimation dataset goes to infinity, the estimate will converge to the true β_X as well based on standard statistical theory. Per Theorem 5, $\tilde{R}^2_{Y|XZ} = \tilde{R}^2_{Y|WZ} + \frac{\beta_X^2 \sigma^2_{X|W,Z}}{\sigma_Y^2}$. The above proof showed that estimation of each component on the right side of the equation is consistent. Thus, under asymptotically consistent estimation of β_X and $\sigma^2_{Q|W,Z}$ the corrected R-squared is asymptotically unbiased for the true $R^2_{Y|XZ}$. □

Chapter 5

SIMULATION EVALUATION OF R-SQUARED ESTIMATION WITH MEASUREMENT ERROR IN PREDICTOR VARIABLES

5.1 Introduction

Simulation studies have been performed to test the theories described in the previous chapter and to evaluate the performance of proposed R-squared correction methodologies. In the simulation, we selected the same distributions employed in previous simulation studies as explained in Chapter 3 to assess performance under normal and nonnormal distributions, including normal distribution, non-normal distributions using Vale's method having various skewness and kurtosis specifications, and Weibull distribution. Section 5.2 presented coefficient correction results based on the approaches proposed in Chapter 4, regression calibration method and method of moments. Section 5.3 shows simulation results for R-squared correction. Single dimension and multiple dimension regression models are both tested in the simulation, also the impact on performance from correlation between measurement errors in the multiple dimension case is also tested. For covenience, in the following discussion, Y defines the dependent variable, Q denotes the unbiased measure following a classic measurement error model while W is the measure with measurement error from a more complicated error structure. X is the true value, Z defines other characteristics that are assumed to be accurately measured and are included in the regression model. Single dimension and multi-dimension model structures are defined as follows. We ran 10000 simulations for every simulation setup defined by sample size, dimension and error distribution.

- In the single dimension simulation, $Q = X + \epsilon$, $Y = 1.5X + 0.75Z_1 + 0.5Z_2 + 1.5 * e$, and $W = X + Z_1 + Z_2 + \mu$. Z_1 and Z_2 are independent and simulated from standard normal distributions, and X is generated as $X = \sqrt{1/6}Z_1 + \sqrt{1/3}Z_2 + \sqrt{1/2}\nu$. All the error terms, ϵ, e, μ and ν, are simulated from various distributions with mean 0 and standard deviation 1.

- In a multidimensional simulation study, there are multiple variables with measurement error, while the measurement errors for different variables can be independent or correlated. The underlying mathematical model in the simulation study is defined as:

$$\begin{aligned} Y &= X_1 + 1.25X_2 + 1.5X_3 + 1.4Z_1 + 2Z_2 + 1.8Z_3 + e \\ X_1 &= Z_1 + 0.5Z_2 + 0.2Z_3 + \delta_1 \\ X_2 &= 0.2Z_1 + Z_2 + 0.5Z_3 + \delta_2 \\ X_3 &= 0.5Z_1 + 0.2Z_2 + Z_3 + \delta_3 \\ W_1 &= 2X_1 + X_2 + 0.5X_3 + 0.4Z_1 + 0.6Z_2 + 0.8Z_3 + \mu_1 \\ W_2 &= 0.5X_1 + 2X_2 + X_3 + 0.6Z_1 + 0.8Z_2 + 0.4Z_3 + \mu_2 \\ W_3 &= X_1 + 0.5X_2 + 2X_3 + 0.8Z_1 + 0.4Z_2 + 0.6Z_3 + \mu_3 \end{aligned} \tag{5.1}$$

  In the simulation, $Q = X + \epsilon$, and variance of ϵ is set to 5. The δ terms, $\delta_1, \delta_2, \delta_3$, have mean equal to zero and variance of 4; for μ errors, their means are zero and variances are 25; the error term e of Y on X and Z has variance 49. The simulation assumes that e and δ, μ, are independent, but relevant measurement errors can be correlated, i.e., $\delta_1, \delta_2, \delta_3$ can be correlated. The two scenarios are both tested in the study: error terms are independent and error terms are correlated.

5.2 Coefficients correction test

Chapter 4 describes two approaches to correct model coefficients under measurement errors: based on regression calibration method, first develop calibration equation using biomarker data and then apply the calibrated equation to the general population to estimate coefficients; the second method is based on method of moment methodology, based on the biomarker data, Q, unbiased measure of real values. The method first regresses Y on Q, Z to estimate β_Q, then follows formula (4.9) and (4.10) in Chapter 4 to estimate corresponding β_X. This section presents simulation results of coefficient correction via both methodologies. The tests were performed in a single dimension model as well as in a multi-dimension model. Different sets of calibration data size and general population size were tested: calibration

data size increases from 100, to 250 and 500, and general population had two different sizes: 500 and 1000.

5.2.1 Coefficient Correction under Single Dimension

In this subsection, we present simulation results for a single dimensional prediction. Measurement errors are simulated from four distributions: Normal, non-normal distributions with various skewness and kurtosis, and Weibull distribution. Table 5.1 to 5.3 present uncorrected regression coefficients of Y on W, Z, and corrected model coefficients through regression calibration and method of moment (MoM) under various sample sizes for calibration data and general population size. As defined in section 5.1, in the single dimension simulation, the true β_X is 1.5. The simulation results show that the method of moment approach has reasonably good correction performance under various calibration dataset sizes. Also the performance was persistent under different measurement error distributions. Given that the underlying model is linear and Q is from a classical measurement error model, unbiased correction is expected from a method of moment approach. For the regression calibration method, the performance was not good when sample size was small at 100, however the bias in the corrected coefficient became smaller as the sample size increased. This was consistent with the proof presented in Chapter 4, as sample size $n \to \infty$, regression calibration correction will be asymptotically consistent even under non-normal distributions.

Table 5.1: Regression coefficient correction under single dimension calibration size=100

	Sample size =500 for general population			Sample size=1000 for general population		
	Before Correction	Correction based on Calibration	Correction based on MoM	Before Correction	Correction based on Calibration	Correction based on MoM
Normal	0.499 (0.063)	1.544 (0.362)	1.512 (0.268)	0.500 (0.045)	1.545 (0.324)	1.512 (0.278)
Non-normal (Skew=1, Kurtosis=3)	0.500 (0.073)	1.576 (0.489)	1.512 (0.274)	0.500 (0.052)	1.579 (0.462)	1.512 (0.284)
Non-normal (Skew=2, Kurtosis=7)	0.502 (0.084)	1.619 (0.631)	1.512 (0.281)	0.500 (0.060)	1.625 (0.639)	1.513 (0.291)
Weibull	0.504 (0.098)	1.705 (0.862)	1.514 (0.300)	0.501 (0.068)	1.685 (0.814)	1.511 (0.299)

Table 5.2: Regression coefficient correction under single dimension calibration size=250

	Sample size =500 for general population			Sample size=1000 for general population		
	Before Correction	Correction based on Calibration	Correction based on MoM	Before Correction	Correction based on Calibration	Correction based on MoM
Normal	0.500 (0.064)	1.515 (0.256)	1.503 (0.167)	0.501 (0.045)	1.521 (0.215)	1.505 (0.168)
Non-normal (Skew=1, Kurtosis=3)	0.500 (0.073)	1.532 (0.328)	1.504 (0.169)	0.501 (0.052)	1.534 (0.282)	1.505 (0.169)
Non-normal (Skew=2, Kurtosis=7)	0.502 (0.084)	1.553 (0.409)	1.505 (0.171)	0.502 (0.060)	1.552 (0.357)	1.505 (0.171)
Weibull	0.503 (0.096)	1.579 (0.495)	1.506 (0.172)	0.501 (0.069)	1.570 (0.463)	1.505 (0.171)

Table 5.3: Regression coefficient correction under single dimension with large calibration size with general population size=1000

	Calibration size =500			Calibration size=1000		
	Before Correction	Correction based on Calibration	Correction based on MoM	Before Correction	Correction based on Calibration	Correction based on MoM
Normal	0.500 (0.045)	1.507 (0.177)	1.503 (0.118)	0.499 (0.045)	1.502 (0.156)	1.500 (0.083)
Non-normal (Skew=1, Kurtosis=3)	0.501 (0.052)	1.513 (0.226)	1.503 (0.119)	0.500 (0.052)	1.507 (0.191)	1.500 (0.083)
Non-normal (Skew=2, Kurtosis=7)	0.501 (0.060)	1.522 (0.279)	1.502 (0.119)	0.500 (0.060)	1.512 (0.231)	1.500 (0.083)
Weibull	0.501 (0.070)	1.537 (0.341)	1.502 (0.120)	0.502 (0.069)	1.520 (0.281)	1.502 (0.083)

5.2.2 Coefficient Correction under Multi-Dimension

Figures 5.1 to 5.8 present coefficient correction results when the Xs are multidimensional with corresponding measurement errors. In the simulation, measurement errors were sampled from four distributions as defined in section 5.1, from Normal to Weibull. The simulation tested two correlation assumptions among measurement errors: independent measurement errors and correlated measurement errors with correlation set to 0.5. Each table summarized simulation results for one selected distribution and one assumed correlation structure under various sample sizes, calibration sample size from 100 to 250 and 500, while general population sample size was either 500 or 1000. For each tested simulation setting defined by error distribution, sample sizes and correlation strcutrue, the table presents three results: the first is the uncorrected coefficients, Y directly regressed on W and Z; the second one is corrected coefficients through method of moments, based on regression results of Y on unbiased measure Q and Z; the last is corrected results through calibration method, using the calibration sample to develop a calibration equation (based on Q and W) and regressing Y on calibrated W. Both correction methods showed reasonable improvement in coefficient estimation. The regression calibration method showed better convergence to true coefficient values, while similar to single dimension simulation results, corrected estimation through calibration method in general had a larger variance compared to the corrected estimation through the method of moment approach. There was no obvious impact on performance of the two correction methodologies after introducing correlation between measurement errors, besides a slight increase in the estimation variance.

5.3 R-squared correction simulation

5.3.1 Single Dimension

Following the same simulation setting, we tested the performance of the R-squared correction methodologies described in Chapter 4. Tables 5.4 to 5.8 present simulation results when X is a single dimensional random variable under various simulation sample sizes. Three R-squared values are presented in each table: the original R-squared without correction, R-quared correction when the coefficient β_X was corrected through the calibration

		Calibration size 100 Total size 500			Calibration size 100 Total size 1000			Calibration size 250 Total size 500			Calibration size 250 Total size 1000			Calibration size 500 Total size 500			Calibration size 500 Total size 1000		
Vars	Actual Coef	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z
X1	1.00	0.189 (0.053)	1.084 (0.855)	0.996 (0.626)	0.189 (0.038)	1.076 (1.29)	1.008 (0.99)	0.19 (0.054)	1.024 (0.42)	1.003 (0.436)	0.189 (0.038)	1.029 (0.42)	0.994 (0.363)	0.19 (0.053)	1.006 (0.281)	1.006 (0.369)	0.189 (0.038)	1.014 (0.279)	0.999 (0.299)
X2	1.25	0.252 (0.054)	1.368 (1.109)	1.253 (0.622)	0.252 (0.037)	1.377 (1.122)	1.258 (0.609)	0.252 (0.054)	1.292 (0.42)	1.25 (0.432)	0.252 (0.038)	1.301 (0.415)	1.253 (0.361)	0.252 (0.053)	1.279 (0.282)	1.247 (0.371)	0.252 (0.038)	1.281 (0.28)	1.248 (0.297)
X3	1.50	0.267 (0.054)	1.62 (1.69)	1.495 (0.645)	0.268 (0.038)	1.608 (1.09)	1.493 (0.687)	0.269 (0.054)	1.526 (0.426)	1.509 (0.435)	0.268 (0.038)	1.527 (0.425)	1.5 (0.363)	0.268 (0.053)	1.508 (0.282)	1.5 (0.378)	0.268 (0.038)	1.504 (0.284)	1.505 (0.296)
Z1	1.40	1.59 (0.381)	1.271 (1.371)	1.419 (0.838)	1.575 (0.271)	1.28 (1.962)	1.383 (1.012)	1.574 (0.38)	1.396 (0.746)	1.391 (0.601)	1.576 (0.268)	1.378 (0.749)	1.402 (0.495)	1.577 (0.384)	1.421 (0.502)	1.392 (0.52)	1.582 (0.27)	1.413 (0.508)	1.403 (0.421)
Z2	2.00	2.24 (0.384)	1.734 (1.451)	2.007 (0.834)	2.238 (0.268)	1.713 (1.397)	1.996 (0.785)	2.24 (0.382)	1.857 (0.738)	2.001 (0.594)	2.232 (0.271)	1.852 (0.736)	1.993 (0.504)	2.239 (0.39)	1.885 (0.505)	2.004 (0.525)	2.236 (0.266)	1.876 (0.502)	1.997 (0.407)
Z3	1.80	2.267 (0.381)	1.669 (2.678)	1.814 (0.845)	2.261 (0.264)	1.674 (1.492)	1.797 (0.799)	2.265 (0.377)	1.807 (0.721)	1.8 (0.595)	2.262 (0.268)	1.791 (0.724)	1.799 (0.504)	2.26 (0.378)	1.831 (0.497)	1.798 (0.524)	2.264 (0.268)	1.829 (0.495)	1.801 (0.414)

Figure 5.1: Model Coefficients Correction when errors are independent and from a normal distribution

Vars	Actual Coef	Calibration size 100 Total size 500: Coef of Y regressed on W and Z	Calibration size 100 Total size 500: Corrected Coef using Y on Q and Z	Calibration size 100 Total size 500: Corrected Coefs using Calibrated W and Z	Calibration size 100 Total size 1000: Coef of Y regressed on W and Z	Calibration size 100 Total size 1000: Corrected Coef using Y on Q and Z	Calibration size 100 Total size 1000: Corrected Coefs using Calibrated W and Z	Calibration size 250 Total size 500: Coef of Y regressed on W and Z	Calibration size 250 Total size 500: Corrected Coef using Y on Q and Z	Calibration size 250 Total size 500: Corrected Coefs using Calibrated W and Z	Calibration size 250 Total size 1000: Coef of Y regressed on W and Z	Calibration size 250 Total size 1000: Corrected Coef using Y on Q and Z	Calibration size 250 Total size 1000: Corrected Coefs using Calibrated W and Z	Calibration size 500 Total size 500: Coef of Y regressed on W and Z	Calibration size 500 Total size 500: Corrected Coef using Y on Q and Z	Calibration size 500 Total size 500: Corrected Coefs using Calibrated W and Z	Calibration size 500 Total size 1000: Coef of Y regressed on W and Z	Calibration size 500 Total size 1000: Corrected Coef using Y on Q and Z	Calibration size 500 Total size 1000: Corrected Coefs using Calibrated W and Z
		0.189	1.094	1.004	0.19	1.107	1.017	0.19	1.034	1.002	0.19	1.031	0.999	0.189	1.015	1.002	0.189	1.014	1.001
X1	1.00	(0.054)	(0.861)	(0.694)	(0.039)	(1.054)	(0.8)	(0.055)	(0.425)	(0.458)	(0.038)	(0.424)	(0.384)	(0.055)	(0.292)	(0.393)	(0.038)	(0.288)	(0.312)
		0.252	1.367	1.26	0.252	1.381	1.254	0.252	1.295	1.254	0.252	1.301	1.263	0.252	1.278	1.252	0.253	1.282	1.254
X2	1.25	(0.056)	(0.901)	(0.702)	(0.039)	(0.927)	(0.676)	(0.056)	(0.428)	(0.473)	(0.039)	(0.424)	(0.396)	(0.056)	(0.283)	(0.394)	(0.039)	(0.286)	(0.317)
		0.268	1.604	1.519	0.267	1.621	1.508	0.269	1.527	1.514	0.268	1.523	1.511	0.269	1.507	1.503	0.268	1.512	1.503
X3	1.50	(0.056)	(0.9)	(0.75)	(0.039)	(0.982)	(0.727)	(0.056)	(0.439)	(0.476)	(0.04)	(0.43)	(0.397)	(0.056)	(0.288)	(0.395)	(0.039)	(0.292)	(0.321)
		1.575	1.284	1.38	1.579	1.241	1.39	1.576	1.378	1.389	1.578	1.379	1.393	1.583	1.404	1.401	1.577	1.407	1.391
Z1	1.40	(0.383)	(1.401)	(0.892)	(0.268)	(1.524)	(0.831)	(0.383)	(0.744)	(0.62)	(0.272)	(0.749)	(0.517)	(0.387)	(0.514)	(0.538)	(0.272)	(0.514)	(0.427)
		2.242	1.732	1.989	2.236	1.7	1.977	2.238	1.851	1.997	2.237	1.84	1.992	2.234	1.888	1.993	2.235	1.88	1.995
Z2	2.00	(0.391)	(1.401)	(0.894)	(0.276)	(1.535)	(0.819)	(0.387)	(0.736)	(0.62)	(0.277)	(0.737)	(0.528)	(0.39)	(0.498)	(0.537)	(0.272)	(0.499)	(0.427)
		2.258	1.686	1.762	2.266	1.659	1.8	2.25	1.786	1.782	2.258	1.799	1.781	2.262	1.821	1.795	2.261	1.821	1.8
Z3	1.80	(0.385)	(1.408)	(0.935)	(0.272)	(1.456)	(0.871)	(0.378)	(0.742)	(0.616)	(0.266)	(0.732)	(0.515)	(0.386)	(0.508)	(0.537)	(0.271)	(0.502)	(0.425)

Figure 5.2: Model Coefficients Correction when errors are independent and from non-Normal distribution with skew=1 and kurtosis=3

Vars	Actual Coef	Calibration size 100 Total size 500			Calibration size 100 Total size 1000			Calibration size 250 Total size 500			Calibration size 250 Total size 1000			Calibration size 500 Total size 500			Calibration size 500 Total size 1000		
		Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z
X1	1.00	0.189 (0.056)	1.112 (1.115)	1.009 (0.788)	0.19 (0.04)	1.122 (1.173)	1.016 (0.823)	0.19 (0.056)	1.035 (0.436)	1.004 (0.489)	0.19 (0.039)	1.034 (0.437)	1.002 (0.411)	0.189 (0.056)	1.015 (0.295)	1.005 (0.411)	0.19 (0.039)	1.016 (0.292)	1.003 (0.329)
X2	1.25	0.252 (0.058)	1.056 (32.507)	1.266 (0.95)	0.252 (0.04)	1.396 (1.051)	1.263 (0.786)	0.252 (0.058)	1.299 (0.442)	1.258 (0.517)	0.252 (0.041)	1.306 (0.438)	1.27 (0.437)	0.252 (0.058)	1.28 (0.288)	1.252 (0.422)	0.253 (0.04)	1.283 (0.291)	1.258 (0.344)
X3	1.50	0.268 (0.059)	1.47 (15.252)	1.534 (0.909)	0.267 (0.041)	1.651 (1.57)	1.526 (0.925)	0.269 (0.059)	1.53 (0.451)	1.524 (0.523)	0.268 (0.042)	1.528 (0.447)	1.519 (0.438)	0.269 (0.058)	1.508 (0.294)	1.51 (0.427)	0.268 (0.042)	1.513 (0.298)	1.508 (0.349)
Z1	1.40	1.574 (0.384)	1.367 (10.904)	1.367 (0.973)	1.578 (0.27)	1.21 (1.709)	1.38 (0.93)	1.577 (0.387)	1.373 (0.754)	1.383 (0.642)	1.577 (0.274)	1.374 (0.758)	1.385 (0.536)	1.583 (0.391)	1.404 (0.517)	1.395 (0.553)	1.575 (0.274)	1.405 (0.518)	1.385 (0.439)
Z2	2.00	2.241 (0.395)	2.163 (45.8)	1.972 (1.107)	2.236 (0.278)	1.661 (1.871)	1.962 (0.928)	2.238 (0.39)	1.846 (0.75)	1.991 (0.644)	2.236 (0.279)	1.833 (0.748)	1.98 (0.554)	2.234 (0.394)	1.887 (0.501)	1.99 (0.55)	2.234 (0.275)	1.877 (0.504)	1.99 (0.442)
Z3	1.80	2.257 (0.388)	1.822 (17.012)	1.74 (1.079)	2.266 (0.275)	1.617 (2.117)	1.774 (1.013)	2.25 (0.384)	1.782 (0.75)	1.77 (0.65)	2.258 (0.268)	1.795 (0.741)	1.771 (0.542)	2.261 (0.389)	1.82 (0.512)	1.787 (0.555)	2.259 (0.273)	1.819 (0.506)	1.792 (0.443)

Figure 5.3: Model Coefficients Correction when errors are independent and from non-Normal distribution with skew=2 and kurtosis=7

Vars	Actual Coef	Calibration size 100 Total size 500			Calibration size 100 Total size 1000			Calibration size 250 Total size 500			Calibration size 250 Total size 1000			Calibration size 500 Total size 500			Calibration size 500 Total size 1000		
		Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z
X1	1.00	0.19	1.041	1.012	0.19	1.222	1.026	0.19	1.033	1.002	0.19	1.04	1.01	0.189	1.013	1.001	0.19	1.017	1.009
		(0.058)	(6.216)	(0.987)	(0.041)	(5.489)	(0.912)	(0.058)	(0.454)	(0.536)	(0.041)	(0.454)	(0.465)	(0.057)	(0.297)	(0.432)	(0.041)	(0.294)	(0.355)
X2	1.25	0.253	1.39	1.289	0.252	1.438	1.273	0.253	1.303	1.273	0.252	1.31	1.275	0.253	1.281	1.266	0.252	1.284	1.26
		(0.061)	(11.375)	(1.049)	(0.043)	(2.43)	(0.974)	(0.06)	(0.453)	(0.577)	(0.043)	(0.466)	(0.506)	(0.061)	(0.297)	(0.465)	(0.043)	(0.302)	(0.38)
X3	1.50	0.268	1.681	1.554	0.268	1.62	1.547	0.267	1.535	1.524	0.268	1.543	1.525	0.267	1.508	1.512	0.268	1.509	1.512
		(0.062)	(6.485)	(1.22)	(0.044)	(2.623)	(0.968)	(0.061)	(0.465)	(0.578)	(0.044)	(0.474)	(0.516)	(0.063)	(0.306)	(0.483)	(0.044)	(0.308)	(0.391)
Z1	1.40	1.574	1.255	1.351	1.577	1.076	1.354	1.583	1.379	1.383	1.579	1.358	1.376	1.578	1.411	1.391	1.577	1.412	1.38
		(0.396)	(10.902)	(1.073)	(0.276)	(5.696)	(1.016)	(0.389)	(0.763)	(0.664)	(0.275)	(0.779)	(0.576)	(0.39)	(0.522)	(0.559)	(0.277)	(0.514)	(0.454)
Z2	2.00	2.236	1.714	1.947	2.233	1.629	1.962	2.232	1.835	1.971	2.231	1.833	1.96	2.236	1.877	1.981	2.236	1.875	1.983
		(0.396)	(12.966)	(1.232)	(0.278)	(2.754)	(1.076)	(0.391)	(0.76)	(0.683)	(0.274)	(0.775)	(0.589)	(0.393)	(0.509)	(0.572)	(0.282)	(0.516)	(0.468)
Z3	1.80	2.259	1.651	1.72	2.254	1.59	1.72	2.26	1.785	1.762	2.261	1.779	1.755	2.27	1.823	1.79	2.254	1.818	1.777
		(0.393)	(5.091)	(1.204)	(0.278)	(2.934)	(1.105)	(0.388)	(0.762)	(0.697)	(0.277)	(0.771)	(0.613)	(0.391)	(0.513)	(0.582)	(0.274)	(0.515)	(0.471)

Figure 5.4: Model Coefficients Correction when errors are independent and from Weibull distribution

Vars	Actual Coef	Calibration size 100 Total size 500			Calibration size 100 Total size 1000			Calibration size 250 Total size 500			Calibration size 250 Total size 1000			Calibration size 500 Total size 500			Calibration size 500 Total size 1000		
		Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z
X1	1.00	0.19 (0.074)	1.067 (1.051)	0.986 (1.089)	0.19 (0.053)	1.067 (1.111)	1.015 (1.19)	0.189 (0.074)	1.028 (0.547)	0.991 (0.7)	0.19 (0.053)	1.03 (0.551)	1.003 (0.608)	0.19 (0.075)	1.011 (0.372)	0.994 (0.571)	0.19 (0.052)	1.016 (0.372)	1.005 (0.477)
X2	1.25	0.252 (0.075)	1.374 (1.059)	1.256 (1.251)	0.252 (0.052)	1.371 (1.085)	1.247 (1.025)	0.254 (0.075)	1.311 (0.543)	1.263 (0.704)	0.251 (0.052)	1.313 (0.547)	1.25 (0.605)	0.252 (0.074)	1.283 (0.37)	1.254 (0.571)	0.252 (0.053)	1.285 (0.374)	1.246 (0.475)
X3	1.50	0.267 (0.074)	1.584 (1.072)	1.501 (1.149)	0.269 (0.052)	1.602 (1.341)	1.487 (1.203)	0.267 (0.073)	1.507 (0.554)	1.507 (0.694)	0.268 (0.053)	1.512 (0.565)	1.5 (0.609)	0.268 (0.074)	1.5 (0.376)	1.499 (0.573)	0.268 (0.053)	1.495 (0.374)	1.5 (0.476)
Z1	1.40	1.574 (0.381)	1.362 (1.429)	1.405 (1.166)	1.578 (0.269)	1.339 (1.442)	1.392 (1.124)	1.575 (0.386)	1.4 (0.775)	1.395 (0.757)	1.576 (0.271)	1.4 (0.778)	1.4 (0.669)	1.586 (0.388)	1.418 (0.519)	1.41 (0.634)	1.577 (0.27)	1.411 (0.523)	1.397 (0.522)
Z2	2.00	2.239 (0.389)	1.785 (1.42)	2.011 (1.349)	2.237 (0.276)	1.79 (1.406)	2.002 (1.152)	2.237 (0.39)	1.858 (0.754)	2.002 (0.76)	2.237 (0.275)	1.852 (0.752)	2.001 (0.662)	2.234 (0.383)	1.896 (0.515)	1.998 (0.629)	2.233 (0.273)	1.891 (0.519)	1.991 (0.523)
Z3	1.80	2.259 (0.384)	1.761 (1.42)	1.793 (1.183)	2.265 (0.271)	1.727 (1.757)	1.821 (1.291)	2.252 (0.38)	1.816 (0.769)	1.777 (0.764)	2.256 (0.267)	1.82 (0.764)	1.801 (0.663)	2.261 (0.383)	1.837 (0.524)	1.803 (0.631)	2.261 (0.268)	1.842 (0.521)	1.801 (0.524)

Figure 5.5: Model Coefficients Correction when errors are correlated and from a normal distribution

Vars	Actual Coef	Calibration size 100 Total size 500			Calibration size 100 Total size 1000			Calibration size 250 Total size 500			Calibration size 250 Total size 1000			Calibration size 500 Total size 500			Calibration size 500 Total size 1000		
		Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z
		0.189	1.107	1.006	0.19	1.094	1.015	0.189	1.037	1	0.19	1.034	0.988	0.188	1.018	0.997	0.189	1.014	1
X1	1.00	(0.079)	(1.194)	(1.211)	(0.056)	(1.481)	(1.588)	(0.079)	(0.571)	(0.779)	(0.056)	(0.559)	(0.68)	(0.079)	(0.387)	(0.626)	(0.056)	(0.385)	(0.528)
		0.252	1.377	1.255	0.252	1.392	1.263	0.252	1.306	1.254	0.252	1.315	1.264	0.252	1.289	1.253	0.253	1.289	1.253
X2	1.25	(0.078)	(1.181)	(1.304)	(0.054)	(1.198)	(1.491)	(0.078)	(0.566)	(0.754)	(0.055)	(0.561)	(0.657)	(0.078)	(0.377)	(0.602)	(0.054)	(0.377)	(0.505)
		0.267	1.578	1.521	0.267	1.593	1.497	0.269	1.514	1.514	0.268	1.507	1.517	0.269	1.495	1.507	0.267	1.502	1.506
X3	1.50	(0.077)	(1.194)	(1.348)	(0.054)	(1.548)	(1.139)	(0.077)	(0.573)	(0.753)	(0.055)	(0.563)	(0.655)	(0.076)	(0.379)	(0.597)	(0.054)	(0.381)	(0.505)
		1.576	1.32	1.376	1.58	1.304	1.391	1.577	1.39	1.39	1.579	1.393	1.401	1.585	1.406	1.403	1.577	1.413	1.39
Z1	1.40	(0.384)	(1.61)	(1.234)	(0.269)	(1.72)	(1.308)	(0.386)	(0.777)	(0.813)	(0.273)	(0.771)	(0.709)	(0.389)	(0.537)	(0.664)	(0.273)	(0.539)	(0.557)
		2.243	1.775	1.991	2.236	1.75	1.969	2.24	1.861	1.996	2.237	1.848	1.992	2.235	1.891	1.992	2.235	1.886	1.995
Z2	2.00	(0.399)	(1.506)	(1.313)	(0.28)	(1.88)	(1.361)	(0.392)	(0.768)	(0.799)	(0.281)	(0.768)	(0.704)	(0.395)	(0.52)	(0.653)	(0.276)	(0.521)	(0.546)
		2.258	1.744	1.756	2.266	1.733	1.809	2.25	1.808	1.785	2.257	1.825	1.773	2.263	1.836	1.79	2.26	1.835	1.797
Z3	1.80	(0.385)	(1.501)	(1.256)	(0.273)	(1.962)	(1.171)	(0.378)	(0.776)	(0.784)	(0.265)	(0.763)	(0.688)	(0.386)	(0.53)	(0.647)	(0.271)	(0.526)	(0.539)

Figure 5.6: Model Coefficients Correction when errors are correlated and from non-Normal distribution with skew=1 and kurtosis=3

		Calibration size 100 Total size 500			Calibration size 100 Total size 1000			Calibration size 250 Total size 500			Calibration size 250 Total size 1000			Calibration size 500 Total size 500			Calibration size 500 Total size 1000		
Vars	Actual Coef	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z
		0.189	1.143	1.017	0.19	1.117	1.035	0.188	1.038	0.999	0.189	1.035	0.991	0.188	1.019	1	0.189	1.015	1.001
X1	1.00	(0.085)	(1.583)	(1.413)	(0.06)	(1.465)	(1.44)	(0.085)	(0.594)	(0.87)	(0.06)	(0.584)	(0.764)	(0.085)	(0.398)	(0.687)	(0.06)	(0.396)	(0.585)
		0.253	1.386	1.263	0.252	1.402	1.25	0.252	1.311	1.258	0.252	1.32	1.272	0.252	1.29	1.251	0.253	1.291	1.257
X2	1.25	(0.082)	(1.526)	(1.496)	(0.057)	(1.37)	(1.353)	(0.083)	(0.585)	(0.814)	(0.058)	(0.578)	(0.711)	(0.082)	(0.384)	(0.644)	(0.057)	(0.385)	(0.543)
		0.268	1.587	1.528	0.267	1.591	1.516	0.27	1.516	1.526	0.268	1.51	1.524	0.27	1.495	1.516	0.268	1.503	1.511
X3	1.50	(0.079)	(1.38)	(1.419)	(0.056)	(2.424)	(1.412)	(0.08)	(0.584)	(0.811)	(0.056)	(0.581)	(0.707)	(0.079)	(0.384)	(0.636)	(0.056)	(0.387)	(0.539)
		1.576	1.273	1.362	1.578	1.273	1.366	1.579	1.385	1.386	1.578	1.39	1.394	1.586	1.405	1.398	1.576	1.411	1.385
Z1	1.40	(0.387)	(1.788)	(1.378)	(0.272)	(1.739)	(1.335)	(0.39)	(0.792)	(0.867)	(0.276)	(0.785)	(0.757)	(0.394)	(0.544)	(0.703)	(0.276)	(0.547)	(0.591)
		2.243	1.752	1.971	2.236	1.737	1.966	2.24	1.856	1.991	2.236	1.843	1.98	2.236	1.89	1.991	2.235	1.883	1.99
Z2	2.00	(0.404)	(1.829)	(1.476)	(0.283)	(2.099)	(1.31)	(0.396)	(0.784)	(0.833)	(0.284)	(0.779)	(0.741)	(0.401)	(0.525)	(0.677)	(0.279)	(0.527)	(0.569)
		2.257	1.71	1.742	2.266	1.712	1.786	2.251	1.806	1.772	2.257	1.822	1.763	2.262	1.835	1.781	2.258	1.833	1.788
Z3	1.80	(0.388)	(1.659)	(1.391)	(0.275)	(2.667)	(1.32)	(0.383)	(0.785)	(0.825)	(0.267)	(0.771)	(0.723)	(0.388)	(0.536)	(0.667)	(0.273)	(0.532)	(0.56)

Figure 5.7: Model Coefficients Correction when errors are correlated and from non-Normal distribution with skew=2 and kurtosis=7

Vars	Actual Coef	Calibration size 100 Total size 500			Calibration size 100 Total size 1000		
		Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z
		0.19	1.076	1.036	0.189	1.301	1.001
X1	1.00	(0.084)	(4.553)	(1.871)	(0.059)	(11.688)	(1.812)
		0.252	1.433	1.271	0.253	2.042	1.295
X2	1.25	(0.086)	(2.196)	(1.819)	(0.06)	(63.214)	(1.809)
		0.269	1.668	1.598	0.268	1.578	1.597
X3	1.50	(0.088)	(3.695)	(1.724)	(0.061)	(11.501)	(1.996)
		1.579	1.28	1.306	1.579	0.849	1.35
Z1	1.40	(0.395)	(3.181)	(1.696)	(0.282)	(28.634)	(1.593)
		2.236	1.659	1.942	2.235	1.094	1.938
Z2	2.00	(0.408)	(4.464)	(1.632)	(0.284)	(54.682)	(1.658)
		2.255	1.619	1.686	2.26	1.836	1.683
Z3	1.80	(0.398)	(4.31)	(1.722)	(0.282)	(27.104)	(1.708)

Vars	Actual Coef	Calibration size 250 Total size 500			Calibration size 250 Total size 1000		
		Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z
		0.19	1.034	1.018	0.189	1.034	1.008
X1	1.00	(0.083)	(0.594)	(0.877)	(0.059)	(0.588)	(0.791)
		0.253	1.337	1.289	0.253	1.326	1.273
X2	1.25	(0.086)	(0.589)	(0.926)	(0.061)	(0.603)	(0.826)
		0.267	1.531	1.528	0.268	1.534	1.543
X3	1.50	(0.087)	(0.609)	(0.935)	(0.063)	(0.607)	(0.841)
		1.577	1.361	1.358	1.575	1.378	1.352
Z1	1.40	(0.393)	(0.82)	(0.886)	(0.282)	(0.812)	(0.784)
		2.235	1.82	1.941	2.229	1.827	1.955
Z2	2.00	(0.401)	(0.81)	(0.919)	(0.283)	(0.825)	(0.807)
		2.261	1.792	1.749	2.257	1.798	1.747
Z3	1.80	(0.395)	(0.815)	(0.918)	(0.281)	(0.811)	(0.81)

Vars	Actual Coef	Calibration size 500 Total size 500			Calibration size 500 Total size 1000		
		Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z	Coef of Y regressed on W and Z	Corrected Coef using Y on Q and Z	Corrected Coefs using Calibrated W and Z
		0.19	1.019	1.018	0.19	1.016	1.009
X1	1.00	(0.085)	(0.387)	(0.706)	(0.059)	(0.386)	(0.591)
		0.252	1.298	1.257	0.252	1.295	1.256
X2	1.25	(0.087)	(0.386)	(0.718)	(0.061)	(0.39)	(0.602)
		0.269	1.497	1.517	0.269	1.509	1.534
X3	1.50	(0.088)	(0.394)	(0.731)	(0.062)	(0.392)	(0.624)
		1.574	1.409	1.372	1.569	1.407	1.368
Z1	1.40	(0.401)	(0.545)	(0.715)	(0.281)	(0.543)	(0.594)
		2.234	1.871	1.983	2.233	1.87	1.985
Z2	2.00	(0.404)	(0.538)	(0.713)	(0.283)	(0.541)	(0.602)
		2.253	1.831	1.768	2.263	1.822	1.765
Z3	1.80	(0.403)	(0.544)	(0.738)	(0.279)	(0.539)	(0.616)

Figure 5.8: Model Coefficients Correction when errors are correlated and from Weibull distribution

method, and R-quared correction when the coefficient β_X was corrected through method of moment. In all simulations, the true R-squared was 0.683. As illustrated in these simulation results, original R-squared values without correction underestimated R-squared due to impact of measurement error. Both correction methodologies showed over-correction under small sample size, but as calibration sample size increased, corrected R-squared moved close to the true value. As expected, correction methods worked best when measurement errors were from normal distributions, however, they also showed reasonable performance when errors were from non-normal distributions. In the single dimension simulation, method of moment correction methodology showed better performance.

Table 5.4: R-Squared correction and standard error under single dimension, calibration size 100 and population size 500

Mean of Estimated R-squared (Standard Error of Estimation)	Regress on W and Z	Correction based on β through calibration	Correction based on β based on MoM
Normal	0.579 (0.029)	0.695 (0.079)	0.686 (0.045)
NonNormal skewness=1 kurtosis=3	0.580 (0.032)	0.702 (0.107)	0.685 (0.050)
NonNormal skewness=2 kurtosis=7	0.581 (0.036)	0.711 (0.136)	0.684 (0.055)
Weibull	0.594 (0.073)	0.768 (0.300)	0.698 (0.095)

Table 5.5: R-Squared correction and standard error under single dimension, calibration size 100 and population size 1000

Mean of Estimated R-squared (Standard Error of Estimation)	Regress on W and Z	Correction based on β through calibration	Correction based on β based on MoM
Normal	0.578 (0.020)	0.693 (0.062)	0.685 (0.043)
NonNormal skewness=1 kurtosis=3	0.579 (0.022)	0.700 (0.093)	0.685 (0.046)
NonNormal skewness=2 kurtosis=7	0.579 (0.025)	0.711 (0.157)	0.684 (0.050)
Weibull	0.593 (0.072)	0.764 (0.506)	0.697 (0.094)

5.3.2 Simulation results under multidimensional measurement errors

The simulation also tested R-squared correction under multi-dimensional measurement errors. As in other simulation tests, independent and correlated measurement errors were

Table 5.6: R-Squared correction and standard error under single dimension, calibration size 250 and population size 500

Mean of Estimated R-squared (Standard Error of Estimation)	Regress on W and Z	Correction based on β through calibration	Correction based on β based on MoM
Normal	0.579 (0.029)	0.689 (0.050)	0.685 (0.035)
NonNormal skewness=1 kurtosis=3	0.580 (0.032)	0.692 (0.062)	0.685 (0.039)
NonNormal skewness=2 kurtosis=7	0.580 (0.036)	0.697 (0.078)	0.685 (0.045)
Weibull	0.585 (0.052)	0.711 (0.120)	0.691 (0.064)

Table 5.7: R-Squared correction and standard error under single dimension, calibration size 250 and population size 1000

Mean of Estimated R-squared (Standard Error of Estimation)	Regress on W and Z	Correction based on β through calibration	Correction based on β based on MoM
Normal	0.578 (0.020)	0.686 (0.035)	0.684 (0.025)
NonNormal skewness=1 kurtosis=3	0.579 (0.023)	0.691 (0.051)	0.684 (0.033)
NonNormal skewness=2 kurtosis=7	0.579 (0.025)	0.694 (0.064)	0.684 (0.037)
Weibull	0.585 (0.049)	0.712 (0.121)	0.690 (0.063)

Table 5.8: R-Squared correction and standard error under single dimension, calibration size 500 and population size 1000

Mean of Estimated R-squared (Standard Error of Estimation)	Regress on W and Z	Correction based on β through calibration	Correction based on β based on MoM
Normal	0.578 (0.020)	0.686 (0.035)	0.684 (0.025)
NonNormal skewness=1 kurtosis=3	0.579 (0.023)	0.688 (0.042)	0.685 (0.028)
NonNormal skewness=2 kurtosis=7	0.579 (0.026)	0.690 (0.051)	0.685 (0.032)
Weibull	0.581 (0.037)	0.696 (0.076)	0.686 (0.046)

included. Tables 5.9 to 5.14 present R-squared correction results with independent measurement errors under various sample sizes, while Tables 5.15 to 5.20 list simulation results for correlated measurement errors. Under the independence assumption, the real R-squared is around 0.568 while the R-squared under correlated data is 0.629. The tables also included the R-squared of regression of Y on X and Z for reference purpose. As expected, without correction, the regression model of Y on W and Z generated biased R-squared estimation, both correction methodologies improved R-squared estimation. Correction using the method of moment approach showed better performance, smaller variance and was stable over different distributions. The correction method through calibration showed overcorrection under smaller sample sizes especially for non-normal distributions, however when the calibration sample size increased, the corrected R-squared valued moved closer to the true values.

Table 5.9: R-Squared correction and standard error under multi-dimension case with independent errors, calibration size 100 and population size 500

Mean of Estimated R-squared (Standard Error of Estimation)	Regress on X and Z	Regress on W and Z	Correction based on MoM	Correction through calibration
Normal	0.572 (0.029)	0.512 (0.031)	0.572 (0.059)	0.582 (0.074)
Weibull	0.584 (0.086)	0.523 (0.080)	0.582 (0.114)	0.617 (0.128)
NonNormal skewness=1 kurtosis=3	0.572 (0.035)	0.512 (0.036)	0.572 (0.064)	0.591 (0.090)
NonNormal skewness=2 kurtosis=7	0.573 (0.041)	0.513 (0.041)	0.572 (0.074)	0.597 (0.090)

Table 5.10: R-Squared correction and standard error under multi-dimension case with independent errors, calibration size 100 and population size 1000

Mean of Estimated R-squared (Standard Error of Estimation)	Regress on X and Z	Regress on W and Z	Correction based on MoM	Correction through calibration
Normal	0.569 (0.021)	0.509 (0.022)	0.569 (0.055)	0.576 (0.064)
Weibull	0.583 (0.086)	0.521 (0.078)	0.580 (0.114)	0.613 (0.127)
NonNormal skewness=1 kurtosis=3	0.570 (0.025)	0.509 (0.025)	0.568 (0.060)	0.582 (0.069)
NonNormal skewness=2 kurtosis=7	0.570 (0.029)	0.509 (0.029)	0.568 (0.067)	0.588 (0.079)

Table 5.11: R-Squared correction and standard error under multi-dimension case with independent errors, calibration size 250 and population size 500

Mean of Estimated R-squared (Standard Error of Estimation)	Regress on X and Z	Regress on W and Z	Correction based on MoM	Correction through calibration
Normal	0.571 (0.029)	0.512 (0.031)	0.574 (0.037)	0.582 (0.045)
Weibull	0.576 (0.062)	0.517 (0.058)	0.579 (0.068)	0.596 (0.082)
NonNormal skewness=1 kurtosis=3	0.572 (0.035)	0.513 (0.035)	0.574 (0.042)	0.585 (0.052)
NonNormal skewness=2 kurtosis=7	0.573 (0.041)	0.514 (0.040)	0.575 (0.048)	0.589 (0.061)

Table 5.12: R-Squared correction and standard error under multi-dimension case with independent errors, calibration size 250 and population size 1000

Mean of Estimated R-squared (Standard Error of Estimation)	Regress on X and Z	Regress on W and Z	Correction based on MoM	Correction through calibration
Normal	0.569 (0.020)	0.509 (0.022)	0.570 (0.031)	0.574 (0.036)
Weibull	0.574 (0.059)	0.514 (0.055)	0.576 (0.065)	0.588 (0.076)
NonNormal skewness=1 kurtosis=3	0.570 (0.025)	0.510 (0.025)	0.572 (0.034)	0.577 (0.041)
NonNormal skewness=2 kurtosis=7	0.570 (0.029)	0.511 (0.029)	0.572 (0.039)	0.580 (0.047)

Table 5.13: R-Squared correction and standard error under multi-dimension case with independent errors, calibration size 500 and population size 500

Mean of Estimated R-squared (Standard Error of Estimation)	Regress on X and Z	Regress on W and Z	Correction based on MoM	Correction through calibration
Normal	0.571 (0.029)	0.512 (0.031)	0.573 (0.033)	0.580 (0.040)
Weibull	0.574 (0.048)	0.515 (0.047)	0.576 (0.052)	0.589 (0.061)
NonNormal skewness=1 kurtosis=3	0.572 (0.034)	0.513 (0.035)	0.574 (0.038)	0.582 (0.045)
NonNormal skewness=2 kurtosis=7	0.572 (0.040)	0.514 (0.040)	0.575 (0.044)	0.585 (0.052)

Table 5.14: R-Squared correction and standard error under multi-dimension case with independent errors, calibration size 500 and population size 1000

Mean of Estimated R-squared (Standard Error of Estimation)	Regress on X and Z	Regress on W and Z	Correction based on MoM	Correction through calibration
Normal	0.570 (0.020)	0.510 (0.022)	0.571 (0.026)	0.574 (0.030)
Weibull	0.572 (0.045)	0.513 (0.042)	0.574 (0.048)	0.582 (0.054)
NonNormal skewness=1 kurtosis=3	0.569 (0.024)	0.510 (0.025)	0.571 (0.030)	0.575 (0.034)
NonNormal skewness=2 kurtosis=7	0.570 (0.029)	0.511 (0.028)	0.571 (0.034)	0.577 (0.039)

Table 5.15: R-Squared correction and standard error under multi-dimension case with correlated errors, calibration size 100 and population size 500

Mean of Estimated R-squared (Standard Error of Estimation)	Regress on X and Z	Regress on W and Z	Correction based on MoM	Correction through calibration
Normal	0.632 (0.026)	0.532 (0.031)	0.618 (0.069)	0.637 (0.103)
Weibull	0.635 (0.084)	0.540 (0.075)	0.619 (0.122)	0.677 (0.144)
NonNormal skewness=1 kurtosis=3	0.632 (0.032)	0.533 (0.035)	0.617 (0.076)	0.653 (0.098)
NonNormal skewness=2 kurtosis=7	0.633 (0.038)	0.533 (0.039)	0.617 (0.087)	0.659 (0.108)

Table 5.16: R-Squared correction and standard error under multi-dimension case with correlated errors, calibration size 100 and population size 1000

Mean of Estimated R-squared (Standard Error of Estimation)	Regress on X and Z	Regress on W and Z	Correction based on MoM	Correction through calibration
Normal	0.630 (0.019)	0.530 (0.022)	0.616 (0.068)	0.633 (0.096)
Weibull	0.635 (0.083)	0.539 (0.073)	0.616 (0.127)	0.674 (0.142)
NonNormal skewness=1 kurtosis=3	0.630 (0.023)	0.530 (0.025)	0.614 (0.075)	0.645 (0.091)
NonNormal skewness=2 kurtosis=7	0.630 (0.027)	0.530 (0.028)	0.613 (0.081)	0.651 (0.100)

Table 5.17: R-Squared correction and standard error under multi-dimension case with correlated errors, calibration size 250 and population size 500

Mean of Estimated R-squared (Standard Error of Estimation)	Regress on X and Z	Regress on W and Z	Correction based on MoM	Correction through calibration
Normal	0.632 (0.026)	0.534 (0.030)	0.628 (0.041)	0.642 (0.055)
Weibull	0.628 (0.059)	0.536 (0.055)	0.629 (0.070)	0.657 (0.095)
NonNormal skewness=1 kurtosis=3	0.632 (0.032)	0.535 (0.034)	0.630 (0.046)	0.646 (0.062)
NonNormal skewness=2 kurtosis=7	0.633 (0.038)	0.535 (0.039)	0.630 (0.051)	0.651 (0.071)

Table 5.18: R-Squared correction and standard error under multi-dimension case with correlated errors, calibration size 250 and population size 1000

Mean of Estimated R-squared (Standard Error of Estimation)	Regress on X and Z	Regress on W and Z	Correction based on MoM	Correction through calibration
Normal	0.630 (0.018)	0.532 (0.021)	0.627 (0.036)	0.635 (0.047)
Weibull	0.626 (0.057)	0.533 (0.052)	0.625 (0.068)	0.647 (0.090)
NonNormal skewness=1 kurtosis=3	0.630 (0.023)	0.532 (0.025)	0.627 (0.040)	0.639 (0.052)
NonNormal skewness=2 kurtosis=7	0.631 (0.027)	0.533 (0.028)	0.627 (0.044)	0.643 (0.059)

Table 5.19: R-Squared correction and standard error under multi-dimension case with correlated errors, calibration size 500 and population size 500

Mean of Estimated R-squared (Standard Error of Estimation)	Regress on X and Z	Regress on W and Z	Correction based on MoM	Correction through calibration
Normal	0.631 (0.026)	0.534 (0.030)	0.631 (0.035)	0.640 (0.045)
Weibull	0.626 (0.047)	0.535 (0.046)	0.628 (0.053)	0.647 (0.071)
NonNormal skewness=1 kurtosis=3	0.632 (0.032)	0.535 (0.034)	0.632 (0.039)	0.643 (0.051)
NonNormal skewness=2 kurtosis=7	0.632 (0.038)	0.536 (0.039)	0.632 (0.045)	0.646 (0.058)

Table 5.20: R-Squared correction and standard error under multi-dimension case with correlated errors, calibration size 500 and population size 1000

Mean of Estimated R-squared (Standard Error of Estimation)	Regress on X and Z	Regress on W and Z	Correction based on MoM	Correction through calibration
Normal	0.630 (0.018)	0.532 (0.021)	0.629 (0.028)	0.635 (0.036)
Weibull	0.625 (0.043)	0.532 (0.040)	0.625 (0.049)	0.638 (0.062)
NonNormal skewness=1 kurtosis=3	0.630 (0.022)	0.532 (0.024)	0.629 (0.032)	0.636 (0.046)
NonNormal skewness=2 kurtosis=7	0.630 (0.027)	0.533 (0.027)	0.629 (0.036)	0.639 (0.046)

5.4 Summary of simulation studies

In this Chapter, we performed simulation tests to check the proposed correction methods' performance on R-squared correction and model coefficients correction. We tested two correction approaches: method of moment (MoM) and regression calibration. Simulation results showed that without any correction, R-squared was underestimated and estimation of model coefficients was also biased. Both correction methodologies improved estimation of model coefficients and R-squared. Over-correction was observed when sample size was small, however, as calibration sample size increasing, both correction methods worked better. As discussed in Chapter 4, when measurement errors were from a normal distribution, the correction methodology generated asymptotic unbiased estimation as evidenced by the simulation results. The correction methodologies also showed reasonable correction performance even when errors were from non-normal distributions. The simulation tests also showed that at small sample sizes, correction using method of moment approaches had better performance, smaller variance and were more stable over different distributions compared

to the performance of the correction approach using calibration.

Chapter 6

APPLICATION TO CALIBRATION EQUATION EVALUATION IN NUTRITIONAL BIOMARKER STUDIES IN THE WOMEN'S HEALTH INITIATIVE

6.1 Introduction

As discussed in Chapter 1, there are several popular data collection methodologies to assess dietary consumption, including food frequency questionnaires, food records and dietary recalls. Systematic bias from these assessment methodologies have been reported in previous studies. With the advent of good quality biomarkers, such as the doubly-labeled water (DLW) technique for energy consumption and urinary nitrogen for protein consumption, it has been proposed to improve measurement quality by calibrating self-reported consumption measures to biomarker assessments. Previous research (Prentice et. al, 2011) raised questions on the use of calibration equations:

- Based on the calibration equation, are there significant differences in the explanatory strength of the three assessment procedures, FFQ, food records and recalls?

- For the three self-reported measures, what are their strengths to recover actual unobserved values, such as actual energy consumption?

Prentice (2011) performed a further biomarker study, the Nutrition and Physical Activity Assessment Study (NPAAS) to address these questions. The study collected FFQ, food records and 24HR recalls as well as related biomarkers, doubly-labeled water and urinary nitrogen assessments of energy and protein consumption. R-squared of calibration equations was used to answer these questions. This section extends the analysis by developing confidence interval for the corresponding R-squared based on the theory proposed in Chapter 2.

6.2 Data collection

The NPAAS study collected relevant nutritional consumptions data from 450 postmenopausal women enrolled in the Women's Health Initiative (WHI) observational study. The U.S. National Institutes of Health (NIH) launched the WHI study in 1991. The purpose of the WHI study was to address health issues that are associated with major causes of mobility, mortality and quality of life in postmenopausal women – including cardiovascular disease, cancer, and osteoporosis. 40 clinics were contracted to help collect relevant data and the Fred Hutchinson Cancer Research Center (FHCRC) was assigned the role of Clinical Coordinating Center (CCC) to coordinate the research program. The WHI study included three clinical trials and an observational study. The three clinical trials are:

- The Hormone therapy (HT) study: the study was to assess the health risks and benefits of using menopausal hormone therapy (MHT).

- The Dietary modification (DM) study: The purpose of the study is to analyze the health-related effects of the diet pattern based on low-fat, high fruit , grain and vegetable.

- The Calcium and Vitamin D (CAD) Supplementation study: This study is to test whether increasing intake of calcium and vitamin D supplementation helps reduce risk of hip and other fractures as well as colorectal cancer.

The WHI observational study (OS) aimed to develop reliable estimates of effects of known risk factors on postmenopausal women's health, and to identify new risk factors and biological indicators that are associated with diseases. The OS study recruited 93,676 postmenopausal women between the ages of 50 to 79, and tracked them for an average of eight years. OS women were drawn from extensively the same population as for the clinical trials. In order to develop calibration equations using biomarkers and to evaluate the measurement properties of self-reported assessment procedures, Neuhouser (2008) performed a study called the Nutrient Biomarker Study (NBS) on a subsample from the DM trial study, containing 544 postmenopausal women. This study developed calibration equations for the

consumption of energy, protein and protein density, and these calibration equations were applied throughout WHI cohorts in order to evaluate the relationship between diseases and diet patterns. NPAAS is one additional biomarker study conducted in 2007-2009 to address these questions related to calibration equations as listed above.

The NPAAS study enlisted 450 postmenopausal women that were enrolled in the WHI OS. This study collected diet data from these women via various assessment approaches: self-reported measures including a food frequency questionnaire, a 4-day food record, and three 24-hour dietary recalls; and related biomarkers including doubly labeled water and urinary nitrogen assessments of energy and protein consumptions. Prentice (2011) explained details about the data collection process. A brief description of the procedure is explained in this section. In the first clinic visit, participants received training for recording a 4-day food record, received doubly labeled water dosing, completed a food frequency questionnaire and an other relevant questionnaire, and provided blood specimen and spot urine samples. Before the second visit, participants completed the 4-day food records and provided 24-hour urine sample one day before. The second clinic visit collected a 4-day food record, urine samples and additional blood specimens and spot urine from participants. Three 24-hour recalls were collected after the second clinic visit, this first happened after 1-3 weeks, and the other two visits took place monthly thereafter. The study also repeated the entire protocol on 88 participants from the 450 population after about 6 months. This study oversampled Black and Hispanic women for comparison between different ethnic groups; younger postmenopausal women and overweight women were also oversampled in the population.

In the statistical analysis, consumption estimates for each of energy, protein and protein density were log-transformed to approximate normal distributions. The calibration equations were developed by regressing log-transformed biomarker measures on log-transformed self-report assessment and other related personal characteristics, body mass index (BMI), age and ethnicity. R-squared of the calibration equations was reported and used to evaluate the signal strength of these self-report assessments.

6.3 Data analysis

Table 6.1 summarizes the distribution of characteristics in the NPAAS study population, as well as summary information for the subsample having repeated biomarker measures. The table showed that most women involved in the NPAAS study were white women about 70 years of age with BMI higher than normal range (i.e. ≥ 25). The participants with repeated biomarker measures had similar distributions in these characteristics compared to those for NPAAS as a whole. Table 6.2 and 6.3 list the average of diet measures, before and after log-transformation respectively, on the population and the subsample including energy, protein and protein density from biomarkers and self-report assessments. For energy consumption, the biomarker measure is based on doubly labeled water methodology, while the protein biomarker measure is through urinary nitrogen. The tables showed that, for energy and protein consumption, biomarker measures were higher than assessments collected from self-reported procedures, as reported in previous studies, however the biomarker-based measure of protein density was smaller than estimates from other methodologies. Also the tables showed average dietary estimates in the subsample were similar to the corresponding measures from the whole NPAAS population. Table 6.4 compares skewness and kurtosis of relevant dietary measures before and after log-transformation to present the transformation effects. In general, all original measures have slightly positive skewness, however the skewness of almost all measures are changed to negative after log-transformation, except protein density measure using the 24 hour recall approach. Most of the dietary measures do not show high kurtosis, which do not change much even after log-transformation. Energy consumption measured by the FFQ method, and protein density using FFQ and 24 hour recall approaches have excess kurtosis more than 3, and log-transformation leads to a smaller level of kurtosis.

Table 6.5 to 6.8 present the R-squared of calibration equations and their 95% confidence intervals estimated through two approaches: using normal asymptotic distribution and using adjusted F-distribution. Calibration equations included self-reported measures and related participants' characteristics, age, BMI and ethnicity. In general, these tables showed that compared to other self-reported measures, the four-day food record had stronger

Table 6.1: Distribution of characteristics in the population and subsample

		NPAAS Sample (Sample size 450)		Reliability Subsample (Sample size 88)	
		No	%	No	%
Age					
	< 70	244	54.2%	49	55.7%
	70 - 80	169	37.6%	31	35.2%
	80+	37	8.2%	8	9.1%
	Mean	70.4		70.3	
	Median	69.4		69.2	
BMI					
	< 25	156	34.7%	34	38.6%
	25 - 35	224	49.8%	42	47.7%
	35+	70	15.6%	12	13.6%
	Mean	28.5		28.0	
	Median	27.5		26.6	
Ethnity					
	White	288	64.0%	60	68.2%
	Black	84	18.7%	14	15.9%
	Hispanic	64	14.2%	13	14.8%
	Other	14	3.1%	1	1.1%

Table 6.2: Summary of dietary assessments before log-transformation

	Assessment Method	NPAAS Sample (Sample size 450)		Reliability Subsample (Sample size 88)	
		Sample Size	Avg of measure	Sample Size	Avg of measure
Energy					
	Biomarker-Doubly Labeled Water	415	2055.35	82	2047.83
	FFQ	450	1584.09	88	1605.81
	4-Day Food Record	450	1661.06	88	1648.33
	24HR Recall	447	1607.28	88	1612.40
Protein					
	Biomarker-Urinary nitrogen	443	72.56	88	75.07
	FFQ	450	69.93	88	70.53
	4-Day Food Record	450	69.14	88	68.27
	24HR Recall	446	64.33	88	65.37
Protein Density					
	Biomarker	408	14.37	82	14.89
	FFQ	450	17.60	88	17.68
	4-Day Food Record	449	16.93	88	16.96
	24HR Recall	447	16.38	88	16.58

Table 6.3: Summary of dietary assessments after log-transformation

	Assessment Method	NPAAS Sample (Sample size 450)		Reliability Subsample (Sample size 88)	
		Sample Size	Avg of measure	Sample Size	Avg of measure
Energy					
	Biomarker-Doubly Labeled Water	415	7.61	82	7.61
	FFQ	450	7.28	88	7.31
	4-Day Food Record	450	7.39	88	7.38
	24HR Recall	447	7.35	88	7.35
Protein					
	Biomarker-Urinary nitrogen	443	4.24	88	4.28
	FFQ	450	4.14	88	4.18
	4-Day Food Record	450	4.20	88	4.20
	24HR Recall	446	4.13	88	4.14
Protein Density					
	Biomarker	408	2.63	82	2.67
	FFQ	450	2.85	88	2.86
	4-Day Food Record	449	2.81	88	2.81
	24HR Recall	447	2.77	88	2.79

Table 6.4: Comparison in skewness and kurtosis before and after transformation

	Assessment Method	Before Log-transformation		After Log-transformation	
		Skewness	Kurtosis	Skewness	Kurtosis
Energy					
	Biomarker-Doubly Labeled Water	0.82	3.64	-0.17	1.40
	FFQ	0.54	0.06	-0.52	0.13
	4-Day Food Record	0.35	0.16	-0.32	-0.00
	24HR Recall	0.29	-0.26	-0.48	0.71
Protein					
	Biomarker-Urinary nitrogen	0.65	0.64	-0.25	0.08
	FFQ	0.80	1.10	-0.53	0.18
	4-Day Food Record	0.46	0.27	-0.40	0.47
	24HR Recall	0.61	0.78	-0.34	0.66
Protein Density					
	Biomarker	0.73	0.70	-0.16	0.26
	FFQ	0.89	3.21	-0.13	1.00
	4-Day Food Record	0.54	0.77	-0.19	0.57
	24HR Recall	1.17	3.82	0.26	0.63

association with biomarker measures, indicating more explanatory power. FFQ had poorer performance compared to other measures. Confidence intervals using asymptotic approximation have both slightly smaller lower bound and upper bound when compared to the interval derived from the adjusted F-distribution. Table 6.5 listed the R-squared and confidence intervals based on the original dietary measure without log-transformation and their corresponding 95% confidence intervals using the asymptotic normal distribution method estimating empirical skewness and kurtosis as well as zero skewness and kurtosis, i.e. normal distribution assumption on residuals. Table 6.6 lists the corresponding results for dietary measures after log-transformation. The two tables show that there is no significant impact on R-squared after log-transformation. As to confidence interval estimation, as expected, the interval based on empirical skewness and kurtosis is wider than the interval under a normal distribution assumption, especially when kurtosis and skewness of calibration equation residuals are further from zero. We also estimated confidence intervals based on the adjusted F-distribution as illustrated in Table 6.7 and 6.8, for dietary measures before and after log-transformation respectively. Table 6.7 and 6.8 also compared adjusted degree of freedom and non-central parameters to those without adjustment. For protein and protein density, confidence intervals after adjustment are very close to the intervals based on the normal assumption, especially when dietary measures are transformed. There are some differences in confidence intervals of Energy consumption after adjustment.

Table 6.5: R-squared and its confidence interval of calibration equations using asymptotic normal distribution before log-tranformation

		R-squared	empirical skewness of residuals	empirical kurtosis of residuals	95% confidence interval using empirical skewness and kurtosis	95% confidence interval using zero skewness and kurtosis
Energy						
	4Day Food Record	0.449	0.87	5.10	(0.370,0.531)	(0.388,0.511)
	FFQ	0.423	0.85	4.65	(0.345,0.504)	(0.373,0.499)
	24-Hour Recall	0.435	0.23	0.80	(0.370,0.502)	(0.374,0.499)
Protein						
	4Day Food Record	0.329	0.38	0.52	(0.267,0.398)	(0.269,0.396)
	FFQ	0.213	0.48	0.34	(0.157,0.281)	(0.158,0.280)
	24-Hour Recall	0.300	0.34	0.23	(0.239,0.369)	(0.240,0.368)
Protein Density						
	4Day Food Record	0.151	0.54	0.48	(0.102,0.219)	(0.103,0.218)
	FFQ	0.095	0.63	0.46	(0.058,0.154)	(0.058,0.153)
	24-Hour Recall	0.118	0.56	0.40	(0.075,0.181)	(0.076,0.181)

Table 6.6: R-squared and its confidence interval of calibration equations using asymptotic normal distribution after log-tranformation

		R-squared	empirical skewness of residuals	empirical kurtosis of residuals	95% confidence interval using empirical skewness and kurtosis	95% confidence interval using zero skewness and kurtosis
Energy						
	4Day Food Record	0.449	-0.31	2.26	(0.379,0.521)	(0.388,0.511)
	FFQ	0.420	-0.30	2.23	(0.349,0.494)	(0.359,0.489)
	24-Hour Recall	0.423	-0.47	2.00	(0.353,0.495)	(0.361,0.487)
Protein						
	4Day Food Record	0.327	-0.34	0.61	(0.265,0.396)	(0.266,0.394)
	FFQ	0.206	-0.31	0.19	(0.151,0.273)	(0.152,0.273)
	24-Hour Recall	0.285	-0.34	0.07	(0.225,0.353)	(0.225,0.353)
Protein Density						
	4Day Food Record	0.145	-0.24	0.24	(0.097,0.211)	(0.097,0.210)
	FFQ	0.090	-0.19	0.15	(0.054,0.147)	(0.054,0.147)
	24-Hour Recall	0.106	-0.24	0.14	(0.066,0.166)	(0.066,0.166)

Table 6.7: R-squared and its confidence interval of calibration equations based on adjusted F-distribution before log-transformation

		R-squared	n_A^{adj} (n_A)	n_B^{adj} (n_B)	ncp_{adj} (ncp)	95% confidence interval using empirical skewness and kurtosis	95% confidence interval using zero skewness and kurtosis
Energy							
	4Day Food Record	0.449	8.20 (8.0)	149.93 (406.0)	345.40 (337.10)	(0.385,0.537)	(0.396,0.520)
	FFQ	0.423	8.18 (8.0)	166.83 (406.0)	309.61 (302.92)	(0.359,0.509)	(0.369,0.496)
	24-Hour Recall	0.435	8.01 (8.0)	153.83 (403.0)	317.42 (316.86)	(0.371,0.524)	(0.381,0.508)
Protein							
	4Day Food Record	0.329	8.00 (8.0)	343.65 (434.0)	217.22 (217.19)	(0.276,0.406)	(0.276,0.406)
	FFQ	0.213	7.98 (8.0)	362.30 (434.0)	119.23 (119.47)	(0.163,0.293)	(0.164,0.292)
	24-Hour Recall	0.300	7.98 (8.0)	332.30 (430.0)	187.16 (187.68)	(0.245,0.381)	(0.247,0.379)
Protein Density							
	4Day Food Record	0.151	7.94 (8.0)	312.61 (398.0)	71.84 (72.35)	(0.103,0.226)	(0.108,0.233)
	FFQ	0.095	7.90 (8.0)	324.11 (399.0)	42.34 (42.89)	(0.062,0.172)	(0.063,0.171)
	24-Hour Recall	0.118	7.92 (8.0)	345.42 (396.0)	53.80 (54.32)	(0.080,0.198)	(0.081,0.197)

Table 6.8: R-squared and its confidence interval of calibration equations based on adjusted F-distribution after log-transformation

		R-squared	n_A^{adj} (n_A)	n_B^{adj} (n_B)	ncp_{adj} (ncp)	95% confidence interval using empirical skewness and kurtosis	95% confidence interval using zero skewness and kurtosis
Energy							
	4Day Food Record	0.449	8.06 (8.0)	194.87 (406.0)	339.47 (337.11)	(0.388,0.531)	(0.396,0.520)
	FFQ	0.420	8.01 (8.0)	196.41 (406.0)	286.95 (286.62)	(0.347,0.494)	(0.355,0.484)
	24-Hour Recall	0.423	8.07 (8.0)	205.76 (403.0)	303.42 (300.97)	(0.362,0.506)	(0.368,0.497)
Protein							
	4Day Food Record	0.327	8.01 (8.0)	336.42 (434.0)	215.06 (214.72)	(0.272,0.406)	(0.273,0.404)
	FFQ	0.206	8.00 (8.0)	399.86 (434.0)	114.42 (114.38)	(0.157,0.285)	(0.157,0.285)
	24-Hour Recall	0.285	8.01 (8.0)	416.94 (430.0)	174.94 (174.64)	(0.232,0.364)	(0.232,0.364)
Protein Density							
	4Day Food Record	0.145	8.01 (8.0)	360.28 (398.0)	69.04 (68.93)	(0.103,0.226)	(0.103,0.226)
	FFQ	0.090	8.01 (8.0)	376.29 (399.0)	40.37 (40.32)	(0.059,0.165)	(0.059,0.165)
	24-Hour Recall	0.106	8.02 (8.0)	372.82 (396.0)	48.03 (47.93)	(0.071,0.184)	(0.071,0.183)

Chapter 7

SUMMARY AND DISCUSSION

The relationship between dietary habits and health status is always a hot research topic. However it is difficult to obtain reliable measures to assess dietary consumption. Self-reported measures, including food-frequency questionnaires (FFQ), 24-hour dietary recalls and dietary records, are commonly used assessment methods. However, systematic biases have been found in these self-reported measures (e.g. Subar 2003, Prentice et al, 2011). Development of biomarker measures, such as a urinary nitrogen (UN) assessment of protein consumption, provide unbiased and objective measures of a specific nutrient consumption. But the high cost of performing biomarker assessment limits its application to large cohorts. Researchers instead have proposed a calibration approach that takes biomarker measures on a subsample and calibrate self-reported measures to biomarker assessments to mitigate systematic bias embedded in the self-reported measures (Prentice, et al, 2011). Questions are then raised concerning calibration equations, including the assessment of explanatory strength of self-reported dietary measures, and comparison in the strength between different measures. R^2 is a common metric used to assess the explanatory strength. Non-normal distributions will distort the distribution of R^2 under a normal assumption, and measurement error causes bias in R^2 estimation.

This dissertation addresses the two issues mentioned above: development of confidence intervals for R-squared under various non-normal distribution assumptions and correction methods for R-squared when there are measurement errors in independent variables. Two methodologies were proposed in the dissertation to estimate confidence intervals: based on an asymptotic distributional approximation for moderate and large sample size; or based on an F-distribution using adjusted degree of freedom and non-central parameter as a possible means of improving asymptotic approximations when sample size is moderate. Simulation studies showed reasonable approximation performance. Method of moment and regression

calibration were both tested in R-squared correction under different measurement error assumptions, and both approaches reduced bias in R-squared estimation.

There are still some improvement opportunities that can be explored in future research. For small sample size, current research is based on F-distribution and adjusts distributional parameters including the non-central parameter and degrees of freedom in order to match first and second-order moments. Although the approximation showed reasonable performance, there may be some other distributions that may have better fit and be more appropriate for confidence interval estimation. Also the current research assumes a homogeneous distribution in regression residuals, which can extended to more complex error structure such as correlated data and heterogeneous errors. As for R-squared correction under measurement error, the dissertation considers measurement error in independent variables. However in real applications, measurement errors can exist in both dependent and independent variables, thus correction methodology to improve R-squared in this situation should be studied. In addition, further research should develop theories to estimate the confidence interval of corrected R-squared with measurement error in both independent and dependent variables.

BIBLIOGRAPHY

[1] Adams, K.F., Schatzkin, A., Harris, T.B., Kipnis, V., Mouw, T., Ballard-Barbash, R., Hollenbeck,A., Leitzmann, M.F., Overweight, Obesity, and Mortality in a Large Propective Cohort of Persons 50-71 Years Old. The New England Journal of Medicine, Aug 2006, Vol. 355, No.8: 763-778.

[2] Atiqullah, M., The Estimation of Residual Variance in Quadratically Balanced Least-Squares Problems and the Robustness of the F-test. Biometrika, 1962, 49, 1 and 2: 83-91.

[3] Atiqullah, M., The Robustness of the Covariance Analysis of a One-Way Classification. Biometrika, 1964, 51, 3 and 4: 365-372.

[4] Beresford, S.A., Johnson, K.C., et al. Low-Fat Dietary Pattern and Risk of Colorectal Cancer. Journal of the American Medical Association, Feb 2006, Vol.295, No.6: 643-654.

[5] Berkson, J., Are there two regressions? Journal of the American Statistical Association, 1950, 45: 164-180.

[6] Bingham, S.A., Urine nitrogen as a biomarker for the validation of dietary protein intake. Journal of Nutrition. 2003, 133: 921S-924S.

[7] Binukumar,B., Mathew, A., Dietary fat and risk of breast cancer. World Journal of Surgical Oncology, Jul 2005, 3:45

[8] Buzar, J.S., Tosteson, T.D., Stefanski, L.A., Measurement error, Institute of Statistics Mimeo Series No. 2544., 2003.

[9] Byers, T. Food Frequency Dietary Assessment: How Bad is Good Enough. American Journal of Epidemiology. Dec 2001, Vol.164, No.12: 1087-1088.

[10] Carroll, R.J., Ruppert, D., Stefanski, L.A., Crainiceanu, C.M., Measurement Error in Nonlinear Models: A Modern Perspective. 2nd Edition, Chapman Hall, 2006.

[11] Casella, G., Berger, R.L., Statistical Inference, Cengage Learning, 2001.

[12] Cook, J., Stefanski, L.A., A simulation extrapolation method for parametric measurement error models. Journal of the American Statistical Association, 1995, 89: 1314-1328.

[13] Fleishman,A.I., A method for simulating non-normal distributions. Psychometrika, 43(4):521-532.

[14] Fraser, G., A search for truth in dietary epidemiology. The American Journal of Clinical Nutrition. 2003, Vol.78, No.3.: 521S-525S.

[15] Friedenreich, C.M., Orenstein, M.R., Physical Activity and Cancer Prevention: Etiology Evidence and Biological Mechanisms. The Journal of Nutrition, 2002: 3456-3464.

[16] Fuchs, C.S., Giovannucci, E.L., et al. Dietary Fiber And The Risk of Colorectal Cancer and Adenoma In Women. The New England Journal of Medicine, 1999, Vol.340, No.3: 169-176.

[17] Fuller, W.A., Measurement Error Models, Wiley, New York, 1987.

[18] World Cancer Research Fund/American Institute for Cancer Research, Food, Nutrition, Physical Activity, and the Prevention of Cancer: a Global Perspective. 2007.

[19] Heitmann, B.L., Lissner, L., Dietary Underreporting by Obese Individuals -Is it specific or non-specific. British Medical Journal, Oct 1, 5, Vol.311: 986-989.

[20] Horner, N.K, Patterson, R.E., Neuhouser, M.L., Lampe, J.W., Beresford, S.A., Prentice, R.L., Participant characteristics associated with errors in self-reported energy intake from the Women's Health Initiative food-frequency questionnaire. The American Journal of Clinical Nutrition, 2002, 76: 766-773.

[21] Hunter, D., Spiegelman, D., et al. Cohort Studies of Fat Intake and The Risk of Breast Cancer- A Pooled Analysis. The New England Journal of Medicine, Feb 1996, Vol.334, No.4: 356-361.

[22] Kaaks, R.J., Riboli, E., Sinha, R., Biochemical markers of dietary intake. IARC Scientific Publications, 1997, 142: 103-126.

[23] Kaaks, R., Ferrari, P., Ciampi,A., Plummer, M., Riboli, E., Uses and limitations of random error correlations in the validation of dietary questionnaire assessments. Public Health Nutrition. 2002, 5: 969-976.

[24] Keogh, R.H., White, I.R., Rodwell, S.A., Using surrogate biomarkers to improve measurement error models in nutritional epidemiology. Statistics in Medicine, 2013, 32: 3838-3861.

[25] Kipnis, V., Midthune, D., Freedman, L.S., Bingham, S., Schatzkin, A., Subar, A., Carroll, R.J., Empirical Evidence of Correlated Biases in Dietary Assessment Instruments and Its Implications. American Journal of Epidemiology, 2001, Vol.153, No.4, 394-403.

[26] Kipnis, V., Subar, A.F., Midthune, D., et al. Structure of Dietary Measurement Error: Results of the OPEN Biomarker Study. American Journal of Epidemiology, 2003 Vol.158, No.1: 14-21.

[27] Kipnis, V., Midtune, D., et al. Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. Biometrics, 2009 Dec, Vol.65, No.4.: 1003-10.

[28] Kushi, H.L., Byers, T., et al. American Cancer Society Guidelines on Nutrition and Physical Activity for Cancer Prevention: Reducing the Risk of Cancer With Healthy Food Choices and Physical Activity. CA A Cancer Journal for Clinicians, May 2006, 56: 254-281.

[29] Maurer, J., Taren, D.L., Teixeira, P.J., Thomson, C.A., Lohman, T.G., Going, S.B., Houtkooper, L.B., The Psychosocial Behavioral Characteristics Related to Energy Misreporting. Nutrition Reviews, 2006, Vol.64, No.2: 53-66.

[30] Nagelkerke, N.J.D., A Note on a General Definition of the Coefficient of Determination. Biometrika, 1991,78,3: 691-692.

[31] Nakamura, T., Corrected score functions for errors-in-variables models: methodology and application to generalized linear models. Biometrika, 1990, 77: 127-137.

[32] Neuhouser, M.L., Tinker, L., Shaw, S.A., et al. Use of Recovery Biomarkers to Calibrate Nutrient Consumption Self-Reports in the Women's Health Initiative. American Journal of Epidemiology, 2008.

[33] Novick, S.J., Stefanski, L.A., Corrected score estimation via complex variable simulation extrapolation. Journal of the American Statistical Association, 2002, 458: 472-481.

[34] Pesaran, M.H., Smith, R.J., A Generalized R^2 Criterion For Regression Models Estimated By the Instrumental Variables Method. Econometrica, Vol.62, No.3: 705-710.

[35] Pierce, D.A., Stram, D.O., Vaeth, M., Schafer, D., Some insights into the errors in variables problem provided by consideration of radiation dose-response analyses for the A-bomb survivors. Journal of the American Statistical Association, 1992, 87: 351-359.

[36] Prentice, R.L., Degrees of Freedom Modifications for F tests Based on Nonnormal Errors, Biometrika, 1974, 61,3: 559-563.

[37] Prentice, R.L., Covariate measurement errors and parameter estimation in a failure time regression model, Biometrika, 1982, 69: 331-342.

[38] Prentice, R.L., Sugar, E., Wang, C.Y., Neuhouser, M., Patterson, R., Research strategies and the use of nutrient biomarkers in studies of diet and chronic disease. Public Health Nutrition, 2002,5: 977-984.

[39] Prentice, R.L., Shaw, P.A, et al., Biomarker-calibrated energy and protein consumption and increased cancer risk among postmenopausal women, American Journal of Epidemiology, 2009, 169, 977-989.

[40] Prentice, R.L., Huang, Y., et al., Statistical Aspects of the Use of Biomarkers in Nutritional Epidemiology Research. Statistics in BioSciences, 2009, 1, 112-123.

[41] Prentice, R.L., Huang, Y., Kuller, L.H., Tinker, L.F., Vam Horn, L., Stefanick, M.L., Sarto, G., Ockene, J. Biomarker-Calibrated energy and protein consumption and cardiovascular disease risk among postmenopausal women. Epidemiology, 2011, 22,170-179.

[42] Prentice, R.L., Huang, Y., Measurement error modeling and nutritional epidemiology association analyses. The Canadian Journal of Statistics, 2011, 39: 498-509.

[43] Prentice, R.L., Yasmin, M.R., Huang, Y., et al. Evaluation and Comparison of Food Records, Recalls, and Frequencies for Energy and Protein Assessment by Using Recovery Biomarkers. American Journal of Epidemiology, 2011, Vol. 174, No.5, 591-603..

[44] Schoeller, D.A., Recent advances from application of doubly-labeled water to measurement of human energy expenditure. Journal of Nutrition, 1999, 129: 1765-1768.

[45] Shaw, P.A., Prentice, R.L., Hazard ratio estimation for biomarker-calibrated dietary exposures. Biometrics, 2012, 68: 397-407.

[46] Stefanski, L.A., Caroll, R.J., Conditional scores and optimal scores in generalized linear measurement error models. Biometrika, 1987, 74: 703-716.

[47] Stefanski, L.A., Correcting data for measurement error in generalized linear models. Communications in Statistics-Theory and Methods, 1989, 18:1715-1733.

[48] Stefanski, L.A., Caroll, R.J., Structural logistic regression measurement error models, in Proceedings of the Conference on Measurement Error Models, P.J.Brown, W.A. Fuller, eds, Wiley, New York, 1990.

[49] Stefanski, L.A., Cook, J., Simulation extrapolation: the measurement error jackknife. Journal of the American Statistical Association, 1995, 90: 1247-1256.

[50] Subar, A.F., Kipnis, V. , Troiano, P.R., A.H., et al. Using Intake Biomakers to Evaluate the Extent of Dietary Misreporting in a Large Sample of Adults: The OPEN Study. American Journal of Epidemiology, 2003 Vol.158, No.1: 1-13.

[51] Thompson, F.E, M., Subar, A.F., Dietary Assessment Methodology. Chapter 1, Nutrition in the Prevention and Treatment of Disease, 2nd ed, Academic Press, 2008.

VITA

Lei Xu was born in China. In 2002, he earned a Master of Engineering in Automation from Tsinghua University of China in Beijing. In 2007, he received a Master of Science in Statistics from University of Washington in Seattle, Washington. In 2014, he graduated with a Doctor of Philosophy in Statistics also from the University of Washington.