Generalized Method of Moments

Theoretical, Econometric and Simulation Studies

by

Yitian Liang

B.Sc. (Honor), Jinan University, 2004
M.Sc. (Honor), City University of Hong Kong, 2009

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

The Faculty of Graduate Studies

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

Abstract

The GMM estimator is widely used in the econometrics literature. This thesis mainly focus on three aspects of the GMM technique. First, I derive the prooves to study the asymptotic properties of the GMM estimator under certain conditions. To my best knowledge, the original complete prooves proposed by Hansen (1982) is not easily available. In this thesis, I provide complete prooves of consistency and asymptotic normality of the GMM estimator under some stronger assumptions than those in Hansen (1982). Second, I illustrate the application of GMM estimator in linear models. Specifically, I emphasize the economic reasons underneath the linear statistical models where GMM estimator (also referred to the Instrumental Variable estimator) is widely used. Third, I perform several simulation studies to investigate the performance of GMM estimator under different situations.

Table of Contents

List of Tables

List of Figures

Acknowledgments

I would like to express my sincere gratitude to my supervisor Professor Jiahua Chen and my thesis second reader Professor Paul Gustafson whose support and suggestions have been an immeasurable help throughout my master study.

I would also like to thank Professor John Petkau for his support throughout my master program and Eugenia Yu for her support of my TA works.

I am also grateful to my fellow graduate students Corrine and Eric who brought me into real Canadian culture and help to reshape my concept of value.

I would also like to express my gratitude to Mike Danilov, who helped me a lot during the first year and whom I always enjoy talking to. I would also like to thank all graduate students. Last but not the least, thanks to all secretary staffs who make my life much easier.

To my parents, aunts and Ling Ip.

Chapter 1

Introduction of GMM

In statistics, we often wish to learn about some aspects of the world from a data set. For example, marketing researchers are often interested in analyzing what factors affect a consumer's decision on buying one brand versus another within a product category. Typical consumer scanner data from stores provide a basis for such studies. In many cases, we assume an underlying model which potentially generates the observed data. The model could come from our prior experiences, or some theoretical arguments. Following the previous example, marketing researchers usually model consumers' decisions based on the theory of utility maximization. Specifically, they assume consumers make their decision by maximizing some underlying utility function which depends on unknown parameters (usually called preference parameters or structural parameters in the econometrics literature). Hence, the observed decisions are ultimately functions of observed covariates and a set of unknown parameters of interest. One main goal of statistical inference is to estimate the unknown parameters by effectively using the observed data. This domain in statistical inference is usually called estimation. In addition, we often wish to utilize the data to test some of our beliefs, or some implications drawn from theoretical models. In other words, we hope to see whether the observed data provides inconsistent evidence against some statements. This domain in statistical inference is usually called hypothesis testing. Both domains admit the inherent randomness embedded in the observed data. Out of many statistical techniques, the Method of Moment (MM) and Maximum Likelihood Estimation (MLE) are two popular candidates used in statistical inferences. Both methods represent our knowledge or assumptions about the mechanism that generates the observed data. For instance, the MM method reflects our knowledge or assumptions about the moment conditions belonging to the mechanism that generates the data. On the other hand, the use of MLE method requires more knowledge about the mechanism, i.e. the joint distribution of the observed data. The Generalized Method of Moment, from some point of view, lies between the MM and MLE methods. It generally requires more information about the data than MM does, yet leaving the complete joint distribution of the data unspecified.

Suggested by its name, the cornerstone of GMM is a set of population moment conditions. These conditions could come from the assumptions made by researchers, or implications drawn from some theoretical models. From this point of view, it is obvious that there is a strong connection between GMM and Method of Moment (MM). Section 1.1 briefly reviews the idea of Method of Moment. It also discusses the directions in which GMM extends the idea of MM. It is also interesting to compare GMM with another popular estimation method MLE. Section 1.2 briefly reviews the idea of MLE and provides some motivations for the fact that GMM is preferred to MLE in some cases. After Hansen's 1982 paper, GMM has become a widely used estimation method in econometrics, especially macroeconomics and finance. Section 1.3 provides an example to illustrate how moment conditions can be used in econometrics problems.

1.1 Method of Moment

In statistical analysis, the population moments of a random variable are often functions of the unknown parameters of interest. The idea of MM is simply to equate the population moment conditions to the analogous sample moments and define the estimates of the unknown parameters to be the solutions of the resulting equations. To illustrate the idea, consider a simple example. Suppose we have a random sample of annual incomes in 2010 of a city (e.g. Vancouver). Denote X_i as the annual income of the i^{th} individual in the sample. Further assume all individual annual incomes in the sample come from a common population distribution and they are independent of each other. Notationally, we have

$$X_i \overset{i.i.d}{\sim} N\left(\mu, \sigma^2\right), i = 1, 2, ..., n,$$

where μ and σ^2 are the population mean and variance, respectively. In this simple example, our interested parameters are μ and σ^2. Due to the i.i.d structure, by the classical Law of Large Number, when the sample size n is large, we have

$$\begin{aligned} n^{-1}\Sigma_{i=1}^{n} x_i &\approx E\left(X_1\right) \\ n^{-1}\Sigma_{i=1}^{n} x_i^2 &\approx E\left(X_1^2\right), \end{aligned}$$

where x_i denotes the observed values of X_i. The Method of Moment involves estimating $\left(\mu, \sigma^2\right)$ by the values $\left(\hat{\mu}, \hat{\sigma}^2\right)$ defined as the solution to the analogous sample moment conditions

$$\begin{aligned} n^{-1}\Sigma_{i=1}^{n} x_i - \hat{\mu} &= 0 \\ n^{-1}\Sigma_{i=1}^{n} x_i^2 - \hat{\mu}^2 - \hat{\sigma}^2 &= 0. \end{aligned}$$

It follows that

$$\begin{aligned} \hat{\mu} &= n^{-1}\Sigma_{i=1}^{n} x_i \\ \hat{\sigma}^2 &= n^{-1}\Sigma_{i=1}^{n} (x_i - \hat{\mu})^2 . \end{aligned} \tag{1.1}$$

This approach is very intuitive but not without its weaknesses. For example, all the higher moments of the normal distribution are also functions of $\left(\mu, \sigma^2\right)$. Therefore, this technique could have been applied as effectively to the third and the fourth moments of the distribution. Yet the resulting estimators of $\left(\mu, \sigma^2\right)$ would be different from those given by (1.1). There would be an issue on which estimators should be used within the MM framework. On the other hand, there is another potential weakness inherent in the MM framework. Suppose a researcher would like to base the estimation of $\left(\mu, \sigma^2\right)$ on the first three moments of X_1, that is the first two moment equations introduced previously plus

$$E\left(X_1^3\right) - 3E\left(X_1^2\right)\mu + 3E\left(X_1\right)\mu^2 - \mu^3 = 0.$$

In this case, the three moment conditions form a system of three equations with only two parameters. Such a system typically has no solution. Therefore, the Method of Moment is infeasible in this case. This is one motivation for the origination of GMM, which is able to utilize information presented in a large number of moment equations, often larger than the number of parameters. In addition, historically speaking, we usually refer the moment conditions in MM to the expectations of the polynomial powers of a random variable that is directly observed. This serves as another motivation for GMM in which the moment conditions are usually referred to the expectation of some general functions of the (observed) random variables.

1.2 Maximum Likelihood Estimator

Although GMM is widely used in many econometric studies, MLE remains the main statistical toolkit in some areas. I will first present an example of demand analysis which is a classical topic in marketing. Some discussion about the potential weaknesses of MLE is presented next.

1.2.1 Demand Analysis

In marketing, the market share of a brand within a product category is of great interest to the managers and the researchers. They are usually interested in the following questions: how does my brand's market share change if we raise our price by one percent? How does my brand's market share change if one of our competitors decreases its price by one percent? How does my brand's sales change if there is a promotion? Many of this type of questions could be summarized in such a way: how does our marketing strategy affect the profit. To answer these questions, it is crucial to understand how consumers' purchase decisions are made. In other words, we hope to understand what variables and how they affect consumers' demand. The random utility model is usually employed for such analysis.

To illustrate the model, denote J as the number of brands in a product category. Suppose a marketing researcher collects a random sample of N consumers at a specific time. Denote $\mathbf{W}_i$ as the information collected from the i^{th} consumer, $\mathbf{Y}_i$ as a J-dimensional vector with the j^{th} component being one if the i^{th} consumer purchased the j^{th} brand and zero otherwise, $\mathbf{p}$ as a J-dimensional vector containing the prices for all brands, and $\mathbf{X}_i$ as a Q dimensional vector containing the demographic information of the i^{th} consumer. Notationally, we have $\{\mathbf{W}_i = (\mathbf{Y}_i, \mathbf{p}, \mathbf{X}_i), \; i = 1, 2, ..., N\}$, where $\{\mathbf{W}_i\}$ are *i.i.d* random vectors. Further suppose a consumer only purchases one brand at a time. The random utility model postulates that for the i^{th} consumer, his/her utility of consuming the j^{th} brand (u_{ij}) can be expressed as follows[1]

$$u_{ij} = \alpha_j + \beta \cdot p_j + \delta_j' \mathbf{X}_i + \varepsilon_{ij}, \; j = 1, 2, ..., J,$$

where α_j is the brand-specific parameter, β is usually called price sensitivity, δ_j is a vector containing the parameters of the demographic variables for the j^{th} brand, ε_{ij} contains all other factors that determine the utility. In the econometrics literature, ε_{ij} is often referred to "demand shock" which is assumed to be known by the consumer but unobserved by the researcher. Hence, from the perspective of the researcher, ε_{ij} is assumed to be random, which is the reason for the name "random utility model". It is assumed that a consumer would choose to purchase brand j if it gives him/her the highest utility. Hence, the probability (from the perspective of the researcher) of observing the i^{th} consumer purchased the j^{th} brand is

$$P_{ij} = P\left\{u_{ij} \geq u_{ik}, \text{for any}\, k \neq j\right\}.$$

If $\{\varepsilon_{i1}, \varepsilon_{i2}, ..., \varepsilon_{iJ}\}$ are *i.i.d* random variables across brands, which follows the Type I Extreme Value distribution, the above probability has a closed form expression

$$P_{ij} = \frac{\exp\left(\alpha_j + \beta \cdot p_j + \delta_j' \mathbf{X}_i\right)}{\Sigma_k \exp\left(\alpha_k + \beta \cdot p_k + \delta_k' \mathbf{X}_i\right)}. \tag{1.2}$$

[1] The detailed specification varies from study to study. The specification presented here is just for illustration.

Further assume the demand shocks are independent across consumers. Hence, the probability of observing the data set is

$$L = \Pi_i \left\{ \Pi_j \left(P_{ij}^{y_{ij}} \right) \right\}.$$

When L is viewed as a function of the unknown parameters, we call it the likelihood function. The MLE is a statistical method in which we estimate

$$\{\alpha_j, \beta, \delta_j, j = 1, 2, ..., J\}$$

by

$$\left\{\hat{\alpha}_j, \hat{\beta}, \hat{\delta}_j, j = 1, 2, ..., J\right\}$$

at which the likelihood function is maximized.

There are two points worth notice in this example. First, the randomness in the above model framework is from the perspective of the researcher not the consumer. Since the demand shock (ε_{ij}) is assumed to be known by the consumer, there is no uncertainty when the consumer is making a decision. Second, based on the observed data, we can analyze the problem from a usual GLM multinomial logit model and obtain the same likelihood function. Specifically, the multinomial response in this case is $\mathbf{Y}_i$ and the covariates are $\mathbf{p}$ and $\mathbf{X}_i$. If we assume the following link function (choosing the first brand as the baseline)

$$\log \frac{P_{ij}}{P_{i1}} = \alpha_j + \beta \cdot p_j + \delta_j' \mathbf{X}_i,$$

the choice probability P_{ij} can also be expressed as (1.2). So from the perspective of the final likelihood function, the random utility model introduced above is equivalent as the multinomial logit model introduced in many GLM textbooks. However, the construction of random utility model reflects the theoretical arguments from the economics literature and will make a difference when the underlying economic problem becomes more complicated.

1.2.2 Potential Weakness of MLE

In general, the data set is assumed to come from a distribution family which is indexed by a vector of unknown parameters of interest. The idea of MLE is to pick up the point in the parameter space that maximizes the likelihood function as the estimate. Under some regularity conditions, the MLE is optimal in the sense that its asymptotic variance attains the Cramer-Rao lower bound. Despite its optimality, there are some potential weaknesses of MLE, compared to GMM.

1. The optimality of MLE stems from its basis on the joint probability distribution of the data. Under certain conditions, we know that MLE is the most efficient (asymptotically) estimator, provided the population distribution is correctly specified. However, in some circumstances, this dependence on the probability distribution can become a weakness. The desirable statistical properties of MLE might not be realized if the distribution is not correctly specified. However, an economic theory rarely provides a complete knowledge of the probability distribution of the data. On the other hand, the GMM method does not require full specification of the distribution of the data. The cornerstone of GMM is a set of moment conditions, which might be deduced from the economics theories or the assumed econometric model. Hence, the GMM method is often regarded to be more robust to the MLE method, in terms of misspecification of the joint distribution.

2. In many econometric applications, the computation of MLE could be very cumbersome. Two types of problems tend to occur. First, the econometric model implied by the economic theory reasonably coincides with the specified joint distribution. But the likelihood function is extremely difficult to evaluate numerically with the current computer technology. In other cases, the economic theory only implies some aspects of the joint distribution. But the complete specification of the joint distribution involves some additional parameters which must also be estimated. Often in these cases, the likelihood function must be maximized subject to a set of nonlinear constraints implied by the economics model, resulting in more computational burden. In contrast, in many econometric cases, the GMM framework provides a computationally convenient method of conducting statistical inferences without completely specifying the likelihood function.

Besides these two potential reasons, there are others that motivate the usage of GMM instead of MLE in many econometric applications. I will provide an examples in the next section.

1.3 An Example of Moment Conditions

The method of Instrumental Variables (IV) is probably the most popular empirical tool in econometrics. In fact, the method of IV is a special case of GMM. In this section, I will present a simple example to illustrate the reasons accounting for the popularity of IV in econometrics. In addition, I will briefly discuss the difficulty of MLE encountered in this example.

The estimation of the relationship between demand and price is a classical problem in econometrics. In many cases, the problem can be simply described by one sentence: does higher price *cause* lower demand? The keyword of the problem is "cause". The definition of causation varies across different subject areas, even within a same academic area. Due to its ambiguity, I do not aim to give a precise and rigorous definition of causation in this essay. Instead, I will present one of its definition loosely which is accepted by many scholars in the field. In the demand and price example, the causal effect of price in the economics literature is usually (loosely) defined as: the units demand changes in response to a unit change of price holding all other relevant factors fixed. In the economics literature, such causal effect is also referred to "the *ceteris paribus*[2] effect of price".

With this in mind, the example proceeds as follows. Suppose a researcher collects a random sample of quantity and price pairs $\{q_i, p_i\}$ $(i = 1,2,...,N)$ across N geographical markets for a specific commodity (e.g. a specific brand of coffee). Further assume demand and price follow a linear relationship as follows

$$q_i = \alpha + \beta \cdot p_i + \varepsilon_i, \tag{1.3}$$

where α is the intercept, β is the price sensitivity, ε_i contains all other demand relevant factors in the i^{th} market which are not observed by the researcher. Without loss of generality, further assume $E(\varepsilon_i) = 0$. The causal parameter of interest is β, which captures the effect of price on demand while holding other factors constant. Intuitively, we expect the sign of the coefficient β should be negative. However, the OLS estimate of β based on many real data is often positive. The underlying reason is that prices are not randomly assigned to different geographical areas. Instead, firms in different areas set their prices according to the demand relevant factors. In other words, firms set their prices according to ε_i which is not observed by the researcher. For example, the wealth level of a city certainly affects the demand while it is also a factor determines the firms pricing strategy. In fact, we expect the wealthier a city is, the higher the price and the demand would be. In such case, the positive OLS estimate of β is largely driven by the positive correlations between price and ε_i, as well as between demand and ε_i. In statistics, we often refer such problem to the existence of unobserved confounders. In econometrics, researchers often refer such reason to a terminology "endogeneity". To be more specific, the endogeneity issue explained above can be mathematically expressed as[3]

$$E(p_i \varepsilon_i) \neq 0. \tag{1.4}$$

[2] *Ceteris paribus* is a Latin phrase, literally translated as "with other things the same" or "all other things being equal or held constant".

[3] Note that this condition is equivalent to $corr(p_i, \varepsilon_i) \neq 0$ under the assumption $E(\varepsilon_i) = 0$.

One way to identify the interested parameter β is the use of instrumental variables. Specifically, if we could find an observed variable z_i, which is correlated with price but uncorrelated with ε_i, then it implies the following moment condition according to (1.3)

$$Cov\left(z_i, q_i\right) - \beta \cdot Cov\left(z_i, p_i\right) = 0.$$

If we can consistently estimate $Cov\left(z_i, q_i\right)$ and $Cov\left(z_i, p_i\right)$, we can also consistently estimate β. Suppose we can find K ($K > 2$) such observed variables, denoted by a vector $\mathbf{z}_i = \left(z_{i1}, z_{i2}, ..., z_{iK}\right)$. Since $E\left(\mathbf{z}_i \varepsilon_i\right) = \mathbf{0}$, from (1.3) we have

$$E\left(\mathbf{z}_i q_i\right) - \alpha E\left(\mathbf{z}_i\right) - \beta E\left(\mathbf{z}_i p_i\right) = 0. \tag{1.5}$$

(1.5) implies a system of K ($K > 2$) equations with only two parameters α and β. The GMM technique introduced in latter chapters can be applied to consistently estimate the parameters.

The economic meaning of instrumental variable is as follows. Since z_i is correlated with price but uncorrelated with ε_i, z_i affects demand *only* through price. For example, suppose the commodity is coffee, z_i could be the price of the raw material used in producing coffee. In such a case, the price of coffee is correlated with z_i, but most economists believe z_i is uncorrelated with any other factors that affect demand[4]. In the econometrics terminology, z_i is said to be satisfy the "exclusion restriction".

The intuition for why the instrumental variable z can help to identify the causal effect of p on q is as follows. Suppose we observe when z increases one unit, q and p increases four and two units respectively. Then it is *as if* we hypothetically "observe" when p increases one unit, q increases two units. Since z is uncorrelated with all other factors that affect q, all other relevant factors can be treated as fixed in the above hypothetical observation. In summary, with the help of instrumental variable, we could hypothetically observe when p increases one unit with all other variables constant, q increases two units. The causal effect of p on q is then identified.

To illustrate the difficulty of applying MLE in the above problem, notice that (1.4) implies $E\left(\varepsilon_i | p_i\right) \neq 0$. In general, due to the endogeneity issue, $E\left(\varepsilon_i | p_i\right)$ is a highly non-linear function of p_i. However, in order to apply MLE, we need to specify the functional form of the conditional expectation which is a controversial task. Therefore, MLE is generally not applied in the above problem.

[4] Whether this assumption is realistic or not is beyond the scope of this essay.

1.4 Definition of GMM Estimator

In this section, I will formally define GMM estimator. In the next chapter, I will continue to study its asymptotic properties. Assume the researcher has collected data on realized values of $\{X_n\}, n = 1, 2, ..., N$, where $\{X_n\}$ is a set of $p \times 1$ *i.i.d* random vectors. Denote S as the parameter space which is a subset of R^q. Consider a function $f : R^p \times R^q \to R^r$, for some r and we are most interested in $r \geq q$. The population model is defined through the following moment condition

$$E\{f(X_1, \beta)\} = 0, \text{ for some } \beta \in R^q.$$

Denote a_N as a non-singular $r \times r$ matrix (possibly depends on the data), which satisfies

$$a_N \overset{a.s}{\to} a_0,$$

where a_0 is a constant $r \times r$ non-singular matrix. Denote

$$d(b) = \left\{ \left(a_N N^{-1} \Sigma_{i=1}^N f(X_i, b)\right)' \left(a_N N^{-1} \Sigma_{i=1}^N f(X_i, b)\right) \right\}. \tag{1.6}$$

DEFINITION 1.1: The function $d(b)$ in (1.6) is defined as the *GMM objective function.*

DEFINITION 1.2: The GMM estimator is defined as

$$\hat{\beta}_N^{GMM} = \text{argmin}_{b \in S} d(b).$$

Note that in the above definition, the population moment condition implies a system of r equations, containing only q $(r \geq q)$ parameters. If we directly form the analogous sample moment equations, generally there does not exist a solution. To illustrate the rationale of GMM estimator, let's consider a simple example. Suppose the population moment conditions are

$$\begin{aligned} E(X_1) - \beta_1 &= 0 \\ E\left(X_1^2\right) - \beta_2 &= 0 \\ E\left(X_1^3\right) - 3\beta_1\beta_2 - 2\beta_1^3 &= 0. \end{aligned}$$

Hence we have

$$[E(X_1) - \beta_1]^2 + \left[E\left(X_1^2\right) - \beta_2\right]^2 + \left[E\left(X_1^3\right) - 3\beta_1\beta_2 - 2\beta_1^3\right]^2 = 0.$$

Since in general there does not exist a pair $\left(\hat{\beta}_1, \hat{\beta}_2\right)$ such that

$$\begin{aligned} & d\left(\hat{\beta}_1, \hat{\beta}_2\right) \\ = & \left[N^{-1}\left(\Sigma_{i=1}^{N} X_i\right) - \hat{\beta}_1\right]^2 + \left[N^{-1}\left(\Sigma_{i=1}^{N} X_i^2\right) - \hat{\beta}_2\right]^2 \\ & + \left[N^{-1}\left(\Sigma_{i=1}^{N} X_i^3\right) - 3\hat{\beta}_1\hat{\beta}_2 - 2\hat{\beta}_1^3\right]^2 \\ = & \ 0, \end{aligned}$$

instead the idea of GMM estimator is to minimize the above quadratic distance:

$$\left(\hat{\beta}_{1,GMM}, \hat{\beta}_{2,GMM}\right) = \text{argmin}_{(\beta_1, \beta_2)} d\left(\beta_1, \beta_2\right),$$

in which the weighting matrix a_N is a 3×3 identity matrix.

Chapter 2

Theoretical Development of GMM

The Generalized Method of Moment was first proposed by Professor Hansen in *Econometrica* 1982. In the original paper, Hansen studies the large sample properties of GMM estimators under the setup where the observed data is assumed to be a realization of some stochastic process. There are four main results in the paper. First, the author shows the GMM estimator is strongly consistent under some conditions. Secondly, the GMM estimator is asymptotically normally distributed under certain conditions. Thirdly, there is a lower bound for the asymptotic variance of GMM estimators. Fourth, the author proposed a statistical test for the validity of some economic modeling specifications based on the GMM framework[5]. However, the author did not provide detailed proofs in the paper. For example, the author only presented the main assumptions and outlined the idea of the proof. The promised details are not easily available based on my best effort. For instance, it is not in the technical report as we may have expected. In this section, I provide a proof of my own. For simplicity, I will work with the situation where the observed data are realizations of some i.i.d random variables. In fact, his rough proof is very similar to Wald, Wilks and Cramer's proofs of consistency and asymptotic normality of MLE. The rest of this section is organized as follows. Section 2.1 provides the proof of consistency of the GMM estimator under i.i.d situation. Section 2.2 derives the proof of asymptotic normality of the GMM estimator under i.i.d situation. Section 2.3 introduces the notion of efficient GMM.

2.1 Consistency

In this section, I discuss the proof of consistency of the GMM estimator under the i.i.d situation. Before going into technical details, it is worthwhile to sketch the basic idea of the proof. Similarly to Wald's proof of consistency of the MLE, the idea could be summarized in one sentence: the GMM estimator always picks up

[5] The material on this issue is beyond the scope of this essay.

the “winner”, which coincides with the true parameter that we wish to estimate in the limit. The main assumptions used in the proof are listed below.

ASSUMPTION 2.1.1: The model introduced in Section 1.4 is identifiable. That is, the moment condition

$$E\{f(X_1,\beta)\}=0$$

is satisfied *if and only if* $\beta=\beta_0$, where $\beta_0\in R^q$ and is often referred to the “true” parameter.

Under assumption 2.1.1, together with the assumption that a_0 is non-singular, $a_0E\{f(X_1,\beta)\}=0$ holds *if and only if* $E\{f(X_1,\beta)\}=0$. Hence it implies $\beta=\beta_0$. Therefore, as a function of β, the following function

$$D(\beta)=(a_0E\{f(X_1,\beta)\})'(a_0E\{f(X_1,\beta)\}),$$

attains its lower bound zero, if and only if $\beta=\beta_0$. Since the GMM estimator always picks up the “winner”, which is the minimum of the sample analogue of $D(\beta)$, the winner will eventually fall into an infinitesimal neighborhood of the true parameter in the limit. Denote the sample analogue of $D(\beta)$ as

$$d(\beta)=\left(a_NN^{-1}\Sigma_{i=1}^{N}f(X_i,b)\right)'\left(a_NN^{-1}\Sigma_{i=1}^{N}f(X_i,b)\right).$$

In fact, Hansen (1982) only proves $d(\beta)$ converges almost surely to $D(\beta)$ under the situation where the data is a realization of some stochastic process. Now we continue to list more assumptions.

ASSUMPTION 2.1.2: The parameter space S is compact.

ASSUMPTION 2.1.3: $f(\cdot,\beta)$ is Borel measurable for each β in S.

ASSUMPTION 2.1.4: $f(x,\cdot)$ is continuous on S for each x in R^p.

ASSUMPTION 2.1.5: $lim_{\delta\to 0}E\{\varepsilon(X_1,\beta,\delta)\}=0,\ for\ every\ \beta\in S$, where

$$\varepsilon(X_1,\beta,\delta)=sup\{|f(X_1,\beta)-f(X_1,\alpha)|:\alpha\in S,\|\beta-\alpha\|<\delta\}.$$

Assumption 2.1.2 implies if S is covered by a collection of open sets, then it is also covered by a finite sub-collection of them.. Assumption 2.1.3 is a technical requirement. Assumption 2.1.4 means the transformation of the observed data, when viewed as a function of the parameters, is smooth. Assumption 2.1.5 means: in the limit, if α and β is sufficiently close, the value of $E\{f(X_1,\alpha)\}$ can be arbitrarily close to $E\{f(X_1,\beta)\}$. Let B be a measurable subset of S. Define

$$g(x,B) = inf_{\beta\in B}\left(a_0 f(x,\beta)\right)'\left(a_0 f(x,\beta)\right),$$

and

$$d(B) = inf_{b\in B}\left\{\left(a_N N^{-1}\Sigma_{n=1}^{N} f(X_n,b)\right)'\left(a_N N^{-1}\Sigma_{n=1}^{N} f(X_n,b)\right)\right\}.$$

For any $b \in S$, by SLLN, we have

$$N^{-1}\Sigma_{n=1}^{N} f(X_n,b) \overset{a.s}{\to} E\{f(X_1,b)\}.$$

According to the GMM definition in section 1.4, we have

$$a_N \overset{a.s}{\to} a_0.$$

Therefore, for every $b \in S$, we have

$$\begin{aligned} &\left(a_N N^{-1}\Sigma_{n=1}^{N} f(X_n,b)\right)'\left(a_N N^{-1}\Sigma_{n=1}^{N} f(X_n,b)\right) \\ \overset{a.s}{\to}\ &\left(a_0 E\{f(X_1,b)\}\right)'\left(a_0 E\{f(X_1,b)\}\right). \end{aligned}$$

Since it's true for all $b \in S$ and by the continuity of f in β (assumption 2.1.4), we have

$$d(B) \overset{a.s}{\to} E\{g(X_1,B)\}.$$

The following lemma plays an important role in the proof of consistency later.

Lemma 2.1.1: Suppose: (i) The parameter space $S = \cup_{j=0}^{k} B_j$; (ii) The true parameter is in B_0 and for all $j > 0$, $E\left\{g\left(X_1,B_j\right)\right\} > E\left\{g\left(X_1,\beta_0\right)\right\} = 0$. Then as $N \to \infty$,

$$Pr\left(\hat{\beta}_N^{GMM} \in B_0\right) \to 1.$$

Proof:
Denote $\left\{\hat{\beta}_N^{GMM} \in B_j, i.o.\right\}$ as $\cap_{N=1}^{\infty}\cup_{n=N}^{\infty}\left\{\hat{\beta}_n^{GMM} \in B_j\right\}$. We need only show

$$Pr\left(\hat{\beta}_N^{GMM} \in B_j, i.o.\right) = 0,$$

for each $j = 1, 2, ..., k$. Without loss of generality, assume $k = 1$ and hence $j = 1$.

$$Pr\left(\hat{\beta}_N^{GMM} \in B_1, i.o.\right) \leq Pr\left(d\left(B_1\right) \leq d\left(B_0\right), i.o.\right).$$

Since

$$\begin{aligned} &d\left(B_1\right) \overset{a.s}{\to} E\left\{g\left(X_1, B_1\right)\right\}, \\ &d\left(B_0\right) \overset{a.s}{\to} E\left\{g\left(X_1, B_0\right)\right\}, \\ &E\left\{g\left(X_1, B_1\right)\right\} > E\left\{g\left(X_1, B_0\right)\right\}, \end{aligned}$$

we have

$$Pr\left(\hat{\beta}_N^{GMM} \in B_1, i.o.\right) \leq 0.$$

It is equivalent as $Pr\left(\hat{\beta}_N^{GMM} \in B_0, i.o.\right) = 1$. Therefore $Pr\left(\hat{\beta}_N^{GMM} \in B_0\right) \to 1$. *Q.E.D.*

THEOREM 2.1: Suppose Assumptions 2.1.1-2.1.5 are satisfied, the GMM estimator defined in Section 1.4 converges almost surely to β_0.

Proof:
For a sufficiently large r, denote $B_1 = \{\beta : \|\beta - \beta_0\| > r\}$. we have

$$E\left\{g\left(X_1, B_1\right)\right\} > E\left\{g\left(X_1, \beta_0\right)\right\} = 0.$$

Hence, by Lemma 2.1.1, $Pr\left(\hat{\beta}_N^{GMM} \in B_1, i.o.\right) = 0$. For an arbitrary $\varepsilon > 0$, let $B_2 = \{\beta : \varepsilon \leq \|\beta - \beta_0\| \leq r\}$. For every $\beta \in B_2$, by Assumption 2.1.1 (identification), Assumption 2.1.4 (continuity) and Assumption 2.1.5, we can find small enough δ_β, such that

$$E\left\{\left(a_0 f\left(x, \beta'\right)\right)'\left(a_0 f\left(x, \beta'\right)\right)\right\} > E\left\{\left(a_0 f\left(x, \beta_0\right)\right)'\left(a_0 f\left(x, \beta_0\right)\right)\right\} = 0,$$

for all β' where $\|\beta' - \beta\| \leq \delta_\beta$. Denote $A_\beta = \left\{\beta' : \|\beta' - \beta\| \leq \delta_\beta\right\}$. By Assumption 2.1.2 (compactness), B_2 will be covered by finitely many such A_β, say A_j, $j = 1, 2, ..., m$. Then by Lemma 2.1.1,

$$Pr\left(\hat{\beta}_N^{GMM} \in A_j, i.o.\right) = 0,$$

for $j = 1, 2, ..., m$. So, we have shown, for every $\varepsilon > 0$,

$$Pr\left(\left\|\hat{\beta}_N^{GMM} - \beta_0\right\| > \varepsilon, i.o.\right) = 0.$$

This implies $\hat{\beta}_N^{GMM} \to \beta_0$ almost surely. *Q.E.D.*

2.2 Asymptotic Normality

In this section, I derive the asymptotic distribution of the GMM estimator after properly scaled under the i.i.d situation. Similarly to the proof of asymptotic normality of MLE, the original proof in the paper re-defines the GMM estimator to be a sequence of solutions to the first order condition of the GMM objective function. In my proof here, however, I simply assume the GMM estimator defined in Section 1.4 satisfies the first order condition of the GMM objective function, which means

$$\left(N^{-1}\Sigma_{n=1}^{N}\left.\frac{\partial f(X_n,b)}{\partial b}\right|_{b=\hat{\beta}_N^{GMM}}\right)' a_N' a_N N^{-1}\Sigma_{n=1}^{N} f\left(X_n,\hat{\beta}_N^{GMM}\right) = 0. \quad (2.1)$$

Recall that the GMM estimator is defined as the point that minimizes the GMM objective function. Equation (2.1) indicates the GMM estimator being studied in this section satisfies the necessary condition of being a minimum point. In other words, the GMM estimator is further assumed to be the solution that equates the first derivative of the GMM objective function to zero. Before going to the proof, the important assumptions are listed below.

ASSUMPTION 2.2.1: S is an open subset of R^q that contains β_0, which is an interior point of S.

ASSUMPTION 2.2.2: $\frac{\partial f(X_n,b)}{\partial b'}$ $(n = 1, ,2, ..., N)$ is Borel measurable for each $b \in S$.

ASSUMPTION 2.2.3: $\left(N^{-1}\Sigma_{n=1}^{N}\frac{\partial f(X_n,\beta)}{\partial b'}\right) \xrightarrow{p} E\left\{\frac{\partial f(X_1,\beta_0)}{\partial b'}\right\}$ uniformly in a small neighborhood of β_0. $E\left\{\frac{\partial f(X_1,\beta_0)}{\partial b'}\right\}$ is finite, and has full rank.

ASSUMPTION 2.2.4: $E\left\{f(X_1,\beta_0)\,f(X_1,\beta_0)'\right\}$ exists, is finite, and has full rank.

ASSUMPTION 2.2.5: The GMM estimator defined in Section 1.4, which is also assumed to satisfy the first order condition of the GMM objective function as stated in (2.1), is strongly consistent.

Assumption 2.2.1 rules out the possibility of boundary solution which simplifies the proof. Assumption 2.2.2 is a technical requirement. In the original paper, assumption 2.2.3 is obtained through a lemma under some conditions. Here, I simply state it as an assumption. Since the proof of consistency of GMM estimator in the previous section is based on the definition in section 1.4, assumption 2.2.5 indicates the consistency result follows in the case where (2.1) is also satisfied.

THEOREM 2.2: Suppose Assumptions 2.2.1-2.2.5 are satisfied. Then the GMM estimator defined in Section 1.4, which is also assumed to satisfy the first order condition of the GMM objective function as stated in (2.1), converges in distribution to a normally distributed random vector after properly scaled. Specifically, denote

$$\begin{aligned} V &= \left(Q'a_0'a_0Q\right)^{-1} Q'a_0'a_0\Omega a_0'a_0Q\left(Q'a_0'a_0Q\right)^{-1}, \\ Q &= E\left\{\frac{\partial f\left(X_1, \beta_0\right)}{\partial b'}\right\}, \\ \Omega &= E\left\{f\left(X_1, \beta_0\right) f\left(X_1, \beta_0\right)'\right\}, \end{aligned}$$

then

$$\sqrt{N}\left(\hat{\beta}_N^{GMM} - \beta_0\right) \xrightarrow{d} N(0, V).$$

Proof:

By assumption 2.2.5, for sufficiently large N, we can expand $f\left(X_n, \hat{\beta}_N^{GMM}\right)$ around $f\left(X_n, \beta_0\right)$. We obtain

$$f\left(X_n, \hat{\beta}_N^{GMM}\right) = f\left(X_n, \beta_0\right) + \tfrac{\partial f\left(X_n, \beta^*\right)}{\partial b'}\left(\hat{\beta}_N^{GMM} - \beta_0\right),$$

where β^* is between $\hat{\beta}_N^{GMM}$ and β_0. Substitution of the above expansion into the first order condition (2.1) yields

$$\begin{aligned} 0 = & \left(N^{-1}\Sigma_{n=1}^N \frac{\partial f\left(X_n, \hat{\beta}_N^{GMM}\right)}{\partial b'}\right)' a_N' a_N N^{-1}\Sigma_{n=1}^N f\left(X_n, \beta_0\right) \\ & + \left(N^{-1}\Sigma_{n=1}^N \frac{\partial f\left(X_n, \hat{\beta}_N^{GMM}\right)}{\partial b'}\right)' a_N' a_N \\ & \times \left(N^{-1}\Sigma_{n=1}^N \frac{\partial f\left(X_n, \beta^*\right)}{\partial b'}\right)\left(\hat{\beta}_N^{GMM} - \beta_0\right), \end{aligned}$$

from which we obtain

$$\begin{aligned}
&\sqrt{N}\left(\hat{\beta}_N^{GMM}-\beta_0\right)\\
=\quad &-\left(\left(N^{-1}\Sigma_{n=1}^{N}\frac{\partial f\left(X_n,\,\hat{\beta}_N^{GMM}\right)}{\partial b'}\right)' a_N' a_N\left(N^{-1}\Sigma_{n=1}^{N}\frac{\partial f\left(X_n,\,\beta^*\right)}{\partial b'}\right)\right)^{-1}\\
&\times\left(N^{-1}\Sigma_{n=1}^{N}\frac{\partial f\left(X_n,\,\hat{\beta}_N^{GMM}\right)}{\partial b'}\right)' a_N' a_N\frac{1}{\sqrt{N}}\Sigma_{n=1}^{N}f\left(X_n,\,\beta_0\right).
\end{aligned}$$

Due to $E\left\{f\left(X_n,\,\beta_0\right)\right\}=0$ and the i.i.d structure of $\{X_n\}$, together with Assumption 2.2.4, applying CLT gives us

$$\frac{1}{\sqrt{N}}\Sigma_{n=1}^{N}f\left(X_n,\,\beta_0\right)\xrightarrow{d}N\left(0,\,\Omega\right).$$

Under Assumption 2.2.3, when N is large enough, due to consistency of $\hat{\beta}_N^{GMM}$, we have

$$\left(N^{-1}\Sigma_{n=1}^{N}\frac{\partial f\left(X_n,\,\hat{\beta}_N^{GMM}\right)}{\partial b'}\right)\xrightarrow{p}Q.$$

Since $\hat{\beta}_N^{GMM}$ is consistent, as a result, $\beta^*\xrightarrow{p}\beta_0$ as well. Hence, we have

$$\left(N^{-1}\Sigma_{n=1}^{N}\frac{\partial f\left(X_n,\,\beta^*\right)}{\partial b'}\right)\xrightarrow{p}Q.$$

Since $a_N\xrightarrow{p}a_0$ by definition and the fact that $(Q'a_0'a_0Q)$ is invertible due to Q and a_0 are of full rank, we have

$$\left(\left(N^{-1}\Sigma_{n=1}^{N}\frac{\partial f\left(X_n,\hat{\beta}_N^{GMM}\right)}{\partial b'}\right)' a_N' a_N\left(N^{-1}\Sigma_{n=1}^{N}\frac{\partial f\left(X_n,\beta^*\right)}{\partial b'}\right)\right)^{-1}\xrightarrow{p}(Q'a_0'a_0Q)^{-1},$$

$$\left(N^{-1}\Sigma_{n=1}^{N}\frac{\partial f\left(X_n,\hat{\beta}_N^{GMM}\right)}{\partial b'}\right)' a_N' a_N\xrightarrow{p}Q'a_0'a_0.$$

Then, by Slutsky's Theorem, we have

$$\sqrt{N}\left(\hat{\beta}_N^{GMM}-\beta_0\right)\xrightarrow{d}N\left(0,V\right).\qquad Q.E.D.$$

2.3 Efficiency

From the previous sections, we could see that the asymptotic variance of the GMM estimator after properly scaled depends on the weight matrix a_N. The following theorem gives us a lower bound and identify a situation where it is attained. I will assume Ω (defined in theorem 2.2) is positive definite.

THEOREM 2.3: The lower bound for the asymptotic variance of the class of GMM estimator indexed by a_N is given by $\left(Q'\Omega^{-1}Q\right)^{-1}$. The lower bound is achieved if $a_N'a_N \xrightarrow{p} \Omega^{-1}$.

Proof:
The first part amounts to say that

$$\left(Q'\Omega^{-1}Q\right)^{-1} - \left(Q'a_0'a_0Q\right)^{-1} Q'a_0'a_0\Omega a_0'a_0Q \left(Q'a_0'a_0Q\right)^{-1}$$

is negative semi-definite for any a_0 that has full rank. This clam is equivalent to

$$Q'\Omega^{-1}Q - Q'a_0'a_0Q\left(Q'a_0'a_0\Omega a_0'a_0Q\right)^{-1} Q'a_0'a_0Q \quad \geq \quad 0. \tag{2.2}$$

Since Ω is positive definite, we can write

$$\Omega^{-1} = C'C,$$

for some invertible C as well. Hence, we can re-write (2.2) as

$$\begin{aligned} & Q'C'CQ - Q'a_0'a_0Q\left(Q'a_0'a_0C^{-1}\left(C'\right)^{-1}a_0'a_0Q\right)^{-1}Q'a_0'a_0Q \\ = \quad & Q'C'\left[I - \left(C'\right)^{-1}a_0'a_0Q\left(Q'a_0'a_0C^{-1}\left(C'\right)^{-1}a_0'a_0Q\right)^{-1}Q'a_0'a_0C^{-1}\right]CQ. \end{aligned} \tag{2.3}$$

Define

$$H = \left(C'\right)^{-1}a_0'a_0Q.$$

Hence, (2.3) could be re-written as

$$Q'C'\left(I - H\left(H'H\right)^{-1}H'\right)CQ.$$

The above matrix is positive semi-definite if $I - H\left(H'H\right)^{-1}H'$ is positive semi-definite. $I - H\left(H'H\right)^{-1}H'$ is idempotent and, consequently, positive semi-definite.

Thus, the first part in the theorem is proved. If $a_N' a_N \xrightarrow{p} a_0' a_0 = \Omega^{-1}$, then the asymptotic variance becomes

$$\left(Q'\Omega^{-1}Q\right)^{-1} Q'\Omega^{-1}\Omega\Omega^{-1}Q \left(Q'\Omega^{-1}Q\right)^{-1} = \left(Q'\Omega^{-1}Q\right)^{-1}.$$

Hence, the second part in the theorem is proved. *Q.E.D.*

The intuition of this theorem is as follows. Note that

$$\Omega = E\left\{f\left(X_1, \beta_0\right) f\left(X_1, \beta_0\right)'\right\},$$

which represents the variation of each moment condition used in estimation. Hence, the efficient GMM estimator is attained when each moment condition in the objective function is weighted by the corresponding inverse of its variance. Consequently, the more information embedded in a moment condition, the higher weight is assigned to it.

Chapter 3

Application of GMM in Linear Models

In the last chapter, I have established certain asymptotic properties of the GMM estimator based on i.i.d observations. I focus on the application of GMM estimator for linear models in this chapter. Linear regression model is probably the most widely used tool in many empirical fields. In addition, GMM estimation technique in linear regression models is probably the most widely used tool in econometrics. In fact, when researchers apply GMM estimator in linear regression models, it is often called the *Instrumental Variable Estimator* (or IV Estimator). The rest of this section is organized as follows. Section 3.1 lays out the basic linear model framework and illustrates the motivations of why GMM estimator is widely used in econometrics. Section 3.2 derives the efficient GMM estimator in a linear model. Section 3.3 gives an example of its application.

3.1 Framework and Motivations

Suppose a researcher is interested in analyzing the relationship between a response variable Y and a vector of covariates X. He/she has collected a random sample of N i.i.d observations. More specifically, suppose the data are observed by the researcher but not collected through some experiment. The information of the i^{th} observation is denoted by $W_i = (Y_i,\, X_i')'$ $(i = 1, 2, ..., N)$, where Y_i is a random scalar (response), $X_i = (X_{1i}, X_{2i}, ..., X_{ki})'$ is a random $k \times 1$ vector (covariates). In many economic applications, Y_i often represents some decision of an economic agent, X_i often contains demographic information about the agent and other relevant variables which may also be decision outcomes of the same agent or even other economic agents. For example, Y_i can be the number of stores a brand is running in a geographical area, X_i contains the unit price set by the brand and its other characteristics. In addition, X_i could also contain the number of stores running by other brands in the same geographical area. Suppose the researcher believes Y_i can be represented by a linear function of X_i as follows

$$Y_i = X_i'\beta + U_i, \tag{3.1}$$

where β is a $k \times 1$ constant parameter vector of interest and U_i is a random scalar, containing all other factors that affect Y_i but are not observed by the researcher. Suppose X_i contains an intercept term, then the researcher can make the following assumption without any cost

$$E(U_i) = 0. \tag{3.2}$$

In addition, due to the nature of observational data, the researcher also believes X_i and U_i are correlated. Under (3.2), this implies

$$E(X_i U_i) \neq 0. \tag{3.3}$$

(3.3) indicates at least one covariates in X_i is correlated with U_i. Suppose only X_{2i} is correlated with U_i. In the econometrics literature, X_{2i} is called *endogenous*; since $E(X_{ji}U_i) = 0$ for $j \neq 2$, those covariates are usually called *weakly exogenous*. In contrast, strong exogeneity requires $E(U_i|X_i) = 0$. In summary, (3.1) to (3.3) together characterize a linear model that is widely used in econometrics. The key characteristics of the above model is the endogeneity issue as indicated by (3.3). In the following, I will discuss *why* this modeling framework (especially (3.3)) is particularly important and popular in econometrics and *how* the condition in (3.3) emerges in application. I will refer condition (3.3) to "endogeneity" here and there.

3.1.1 Motivations of the Framework

The motivation of such model framework can be summarized into one word: causation. Recall that in section 1.3, I've mentioned the main goal in many classical econometric analyses is causal inference. The key characteristic in causal inference is to control for all other relevant factors. In the above model framework, from the economic agent's perspective, Y_i, X_i and U_i are real value variables instead of random variables. Hence, (3.1) implies Y_i is a deterministic function of X_i and U_i. Following the example given at the beginning of section 3.1, from the perspective of the brand, the number of stores is determined once the brand knows its X_i and U_i. This economic theoretical argument underneath the above statistical model is largely neglected in the field. To my best knowledge, few researchers explicitly state this argument in their works. However, it is shown below this is crucial in understanding the above model framework. For ease of illustration, suppose X_i is not a vector but a scalar. Hence the coefficient β can be represented (from the perspective of the agent) as

$$\beta = \frac{\partial Y_i\left(X_i, U_i\right)}{\partial X_i}.$$

Loosely speaking, β measures the units of Y_i changes in response to a unit change in X_i holding all other relevant factors constant. Therefore, from the agent's perspective, the coefficient β is the "causal effect" of X_i on Y_i. Note that from the agent's perspective, there is no randomness in the above model. However, from the researcher's perspective, $\{U_i, i = 1,2,...,N\}$ are nearly always assumed to be random variables coming from a common distribution because they are not observed. Hence, it is fair to say that all randomness arguments in the above model are drawn from the perspective of the researcher due to his/her inability of observing U_i. Yet, how do these arguments lead to the formation of condition (3.3)? I will discuss this issue later in the next sub-section.

In summary, the above model framework is suitable for causal inference in many economics problems. Here, I briefly present another popular modeling framework in statistics and econometrics, as well as compare it to the above framework. As shown later, this framework is more suitable for non-causal inference in many cases. Suppose a researcher is still interested in studying the relationship between Y_i and X_i. But the main focus is no longer the causal effect of X_i on Y_i, instead he/she only focuses on the correlation between Y_i and X_i. In other words, the researcher now is interested in the following question: if X_i is observed to increase one unit, how many units does Y_i will be observed to change on average. In this case, the following model might be useful.

$$Y_i = E\left(Y_i | X_i\right) + e_i \tag{3.4}$$

In (3.4), by construction, we have $E\left(e_i | X_i\right) = 0$. Intuitively, the relationship between X_i and Y_i is completely captured by the conditional expectation. The "leftovers" e_i represents any other random shocks to Y_i that is uncorrelated with X_i. Furthermore, let's assume

$$E\left(Y_i | X_i\right) = X_i'\delta. \tag{3.5}$$

The above model framework, (3.4) to (3.5), is a popular linear regression model in many areas of statistics and some areas of econometrics. The OLS estimator of δ is consistent. The fundamental difference between two modeling framework lies in the decomposition of the response variable. In the model of (3.4) to (3.5), Y_i is decomposed as the complete relationship with X_i and a pure random shock that is uncorrelated with X_i. However, in model (3.1) to (3.3), Y_i is decomposed as the causal relationship with X_i and other relevant factors which might be correlated with X_i. Mathematically

$$E\left(X_iU_i\right) \neq 0 \implies E\left(U_i|X_i\right) \neq 0,$$

therefore

$$E\left(Y_i|X_i\right) \neq X_i'\beta.$$

Suppose U_i is positively correlated with X_i and Y_i. Then intuitively speaking, the complete effect (represented by δ) of a unit change of X_i on Y_i comprises two components: 1, the causal effect of X_i on Y_i, represented by β; 2, the indirect effect of X_i on Y_i through U_i, which is positive by assumption. Hence, in this case we have $\delta > \beta$. Therefore, if we are interested in β, the estimate obtained from running the OLS regression based on the data will overestimate the causal effect. In addition, the model framework (3.1) to (3.3) is often backed up by some economic argument which is from the perspective of some economic agent.

3.1.2 Potential Reasons for Endogeneity

1. *Missing Covariates*. In many economic applications, the response variable Y_i (or decision variable) is often determined by some observed variables X_i and other unobserved variables. For example, a salesperson's sales performance is affected by the salary paid to him and his effort. In general, we cannot observe a salesperson's effort (denoted by U_i). In many cases, we cannot even measure the effort. However, it is generally believed that effort is positively correlated with the salary (reflected by $E\left(X_iU_i\right) \neq 0$). If we want to investigate whether an increase of salary causes higher sales performance, the correlation between salary and effort needs to be considered. Otherwise one may overestimate the causal effect of salary. In an extreme case, if the causal effect of salary on sales performance is nearly zero while effort is highly positively correlated with salary and sales performance, running an OLS regression of sales performance on salary yields a significantly positive salary coefficient. In such case, the significant positive salary coefficient is largely driven by the fact that effort is highly positively correlated with salary and sales performance. As a result, one may be misled to an impression that higher salary causes higher sales performance. Hence, based on the misleading OLS regression result, a firm's manager may raise salary in order to increase sales. However, if one can use the IV regression (introduced later) to estimate the true causal effect of salary on sales performance, the manager may apply another lower-cost policy to stimulate sales. In fact, if the manager also realizes the driving force for sales performance is effort, he may seek to re-design a contract that pays the same amount of salary as before but is more effective in terms of stimulating an employee's effort, i.e.

appropriate design of bonus and commission. This research area in economics is often referred to “contract theory”.

2. *Simultaneity*. This is a jargon heavily used in the economics and econometrics literature. In short, simultaneity means the covariate X_i is chosen optimally by some economic agent (partially) according to Y_i. For example, Y_i represents the salary level of automobile production workers in a city and X_i denotes the number of automobile firms in the city. An important economic question is whether higher competition (reflected by larger number of firms) leads to higher salary level. Intuitively and also suggested by many economic models, higher competition should lead to higher salary level due to the following reason. Employees have larger bargaining power against the firm when competition is more severe because they have more alternatives to choose. From the perspective of the local government, if this economic intuition is true, they may implement some policy to attract more investors to the city. Suppose a policy maker collects a random sample of $\{Y_i, X_i, i = 1, 2, ..., N\}$ from N different cities. If he runs an OLS regression of Y_i on X_i, it is very likely the estimated coefficient of X_i is not significantly positive (even negative in many cases). Based on such result, the policy maker may be misled to believe that encouraging more investors is unable to increase the local salary level. However, the policy maker may neglect the fact that firms choices of entering a city or not largely depend on the salary level of the city. Intuitively, a firm is more likely to enter a city if the local salary level is low. How does this relate to the unobservable U_i? It is intuitive and suggested by some economic theories that the larger the population of a city, the cheaper its labor cost is. Hence, a firm is more likely to enter a city if the local population is larger. Suppose the policy maker does not collect the population data, then the unobserved U_i contains the i^{th} city’s population. In such case, X_i and U_i is positively correlated, i.e. $E(X_i U_i) > 0$. However, Y_i and U_i is negatively correlated. Thus, a negative OLS estimated coefficient of X_i on Y_i may be largely driven by the fact that U_i is positive correlated with X_i but negatively correlated with Y_i. Based on the above argument, one can treat simultaneity issue as a special case of missing covariates. For this example, some researchers argue that the negative OLS estimated coefficient of X_i is generated by the *reverse causality* which states that it is actually a lower salary causes a higher number of firms in a city. From my opinion, the arguments of reverse causality is the same as those of simultaneity in this case.

3. *Measurement Error*. The endogeneity issue characterized by (3.3) can also generated in the case of measurement error. Suppose the researcher is interested in analyzing the causal effect of X_i on Y_i and the true model is

$$Y_i = \alpha + X_i \beta + \varepsilon_i.$$

Further assume X_i is independent of ε_i. However, the researcher cannot observe X_i directly but a proxy for it

$$\tilde{X}_i = X_i + \xi_i.$$

Hence the regression model faced by the researcher is

$$Y_i = \alpha + \tilde{X}_i\beta + U_i, \text{ where} U_i = -\xi_i\beta + \varepsilon_i.$$

Therefore $\tilde{X}_i$ is endogenous, because it is correlated with U_i through ξ_i.

3.1.3 Potential Difficulty of Applying MLE

For the linear regression model characterized by (3.1) to (3.3), the potential difficulty of applying MLE to estimate β arises from the specification of the conditional distribution of U_i given X_i. For ease of illustration, assume the conditional distribution depends on X_i only through $E\left(U_i|X_i\right)$. Further assume X_i is a scalar. If we assume

$$E\left(U_i|X_i\right) = X_i\gamma,$$

then we have

$$E\left(Y_i|X_i\right) = X_i\left(\beta+\gamma\right).$$

Therefore, the parameters β and γ cannot be separately identified from the observed data. However, if we assume

$$E\left(U_i|X_i\right) = X_i^2\tau,$$

it is easy to verify the parameters β and τ can be separately identified from the data. However, such ability of identification is often criticized by the fact that it is only due to a specific functional form assumption. From this simple example, we can see that it is not easy to specify a convincing conditional distribution of U_i given X_i in a real world application.

3.2 GMM Estimation

As discussed above, the information embedded in $\{Y_i, X_i\}$ is usually not enough to consistently estimate the causal parameters. Some additional information may be crucial in identifying the causal parameters. The instrumental variables discussed below play such a role. Suppose Z_i is a $l \times 1$ vector and satisfies

$$E\left(Z_iU_i\right) = 0. \tag{3.6}$$

Condition (3.6) indicates the instruments Z_i are weakly exogenous. Now the observed data set is augmented to be

$$\left\{Y_i, X_i', Z_i', i = 1, 2, ..., N\right\}.$$

Based on the moment conditions in (3.6), according to the definition of GMM estimator in section 1.4, the population model can be defined through:

$$E\left(Z_i\left(Y_i - X_i'\beta\right)\right) = 0. \tag{3.7}$$

Note that (3.7) implies a system of l equations with k unknown parameters. The rest of the section discusses about the identification issue and introduces the GMM estimator based on (3.7). In addition, the efficient GMM estimator is also discussed.

3.2.1 Identification

According to assumption 2.1.1, the population model (3.7) is identifiable *if and only if* there exists a unique $\beta_0 \in R^k$, such that

$$E\left(Z_i\left(Y_i - X_i'\beta_0\right)\right) = 0.$$

Re-arranging (3.7), we have

$$E\left(Z_iX_i'\right)\beta = E\left(Z_iY_i\right). \tag{3.8}$$

Note that (3.8) is a l-equation linear system with k unknown parameters. The identification condition requires there is an unique solution to (3.8). According to linear algebra, this is equivalent to the following condition

$$rank\left\{E\left(Z_iX_i'\right)\right\} = k. \tag{3.9}$$

(3.9) is often referred to the "*rank condition*" in the econometrics literature. Hence the population model defined by (3.7) is identified *if and only if* the rank condition (3.9) is satisfied. Hence, a vector Z_i is said to be "valid instruments" if it satisfies the exogeneity condition (3.7) and the rank condition (3.9). An immediate implication of (3.9) is $l \geq k$, since $rank\{E\left(Z_iX_i'\right)\} \leq min\left(l,k\right)$. In the econometrics literature, the necessary condition $l \geq k$ is often referred to "*order condition*". It means one needs to have at least as many exogenous instruments as those endogenous regressors in order to identify the causal parameters. Given the rank condition is satisfied, when $l = k$, the model is said to be *exactly identified*; while if $l > k$, it is said to be *overidentified*.

In order to gain some insights for the rank condition, consider the following example. Suppose $X_i = (1, X_{1i})'$, where X_{1i} is a scalar random variable. Suppose $Z_i = (1, Z_{1i})'$, where Z_{1i} is also a scalar random variable. Note that the regressor "1" is exogenous due to (3.2). Therefore, there is only one endogenous variable X_{1i} and one exogenous instrument Z_{1i}. The rank condition implies

$$\begin{aligned} & det\left\{E\left(Z_i X_i'\right)\right\} \\ = \; & det\left\{\left(\begin{array}{cc} 1 & E\left(X_{1i}\right) \\ E\left(Z_{1i}\right) & E\left(Z_{1i}X_{1i}\right) \end{array}\right)\right\} \\ = \; & E\left(Z_{1i}X_{1i}\right) - E\left(X_{1i}\right)E\left(Z_{1i}\right) \neq 0, \end{aligned}$$

which means the exogenous instrument needs to be correlated with the endogenous regressor.

3.2.2 GMM Estimator

Based on the population model (3.7), the GMM objective function (according to definition 1.1) in the linear model situation ((3.1) to (3.3)) is

$$d(b) = \left(a_N \cdot N^{-1} \cdot \Sigma_{i=1}^N Z_i \left(Y_i - X_i' b\right)\right)' \left(a_N \cdot N^{-1} \cdot \Sigma_{i=1}^N Z_i \left(Y_i - X_i' b\right)\right), \qquad (3.10)$$

where a_N is a $l \times l$ weighting matrix satisfying the conditions listed in section 1.4. The GMM estimator is the solution to the following equation

$$\frac{\partial d(b)}{\partial b} = 0.$$

In this case, it is given by

$$\hat{\beta}_N^{GMM} = \left(\Sigma_{i=1}^N X_i Z_i' \left(a_N' a_N\right) \Sigma_{i=1}^N Z_i X_i'\right)^{-1} \Sigma_{i=1}^N X_i Z_i' \left(a_N' a_N\right) \Sigma_{i=1}^N Z_i Y_i. \qquad (3.11)$$

According to theorem 2.2, the asymptotic variance of $\hat{\beta}_N^{GMM}$, after proper scaling, is given by

$$V = \left(Q' a_0' a_0 Q\right)^{-1} Q' a_0' a_0 \Omega a_0' a_0 Q \left(Q' a_0' a_0 Q\right)^{-1},$$

where

$$\begin{aligned}
Q &= E\left\{ \frac{\partial Z_i\left(Y_i - X_i' b\right)}{\partial b'} \bigg|_{b=\beta_0} \right\}, \\
\Omega &= E\left\{ Z_i\left(Y_i - X_i'\beta_0\right)\left(Y_i - X_i'\beta_0\right)' Z_i' \right\} \\
&= E\left\{ Z_i U_i U_i Z_i' \right\}, \\
a_N &\overset{p}{\to} a_0.
\end{aligned}$$

In the linear model case, we have

$$\begin{aligned}
Q &= E\left(Z_i X_i'\right), \\
\Omega &= E\left(U_i^2 Z_i Z_i'\right).
\end{aligned}$$

Their natural consistent estimators are given by

$$\begin{aligned}
\hat{Q} &= N^{-1}\Sigma_{i=1}^N Z_i X_i', \\
\hat{\Omega} &= N^{-1}\Sigma_{i=1}^N \hat{U}_i^2 Z_i Z_i',
\end{aligned}$$

where $\hat{U}_i = Y_i - X_i'\hat{\beta}_N^{GMM}$.

3.2.3 Efficient GMM Estimator (General Case)

According to theorem 2.3, the lower bound of the asymptotic variance for $\hat{\beta}_N^{GMM}$ under the linear model is achieved when

$$a_N' a_N \overset{p}{\to} \Omega^{-1} = \left(E\left(U_i^2 Z_i Z_i'\right)\right)^{-1}.$$

The following two step procedure illustrates how to obtain an efficient GMM estimator under the linear model setup.

Step 1. Set $a_N' a_N = I_l$, where I_l is the $l \times l$ identity matrix. Obtain the corresponding GMM estimate, say $\tilde{\beta}_N^{GMM}$, which is inefficient but consistent. Use $\tilde{\beta}_N^{GMM}$ to construct a consistent estimator of Ω, which is given by

$$\begin{aligned}
\hat{\Omega} &= N^{-1}\Sigma_{i=1}^N \hat{U}_i^2 Z_i Z_i', \text{ where} \\
\hat{U}_i &= Y_i - X_i'\tilde{\beta}_N^{GMM}.
\end{aligned}$$

Step 2. Set $a_N' a_N = \hat{\Omega}^{-1}$(obtained from step 1) and substitute back into (3.11), yielding the efficient GMM estimator as

$$\begin{aligned} & \hat{\beta}_N^{GMM*} \\ = & \left(\Sigma_{i=1}^N X_i Z_i' \left(\Sigma_{i=1}^N \hat{U}_i^2 Z_i Z_i' \right)^{-1} \Sigma_{i=1}^N Z_i X_i' \right)^{-1} \\ & \times \Sigma_{i=1}^N X_i Z_i' \left(\Sigma_{i=1}^N \hat{U}_i^2 Z_i Z_i' \right)^{-1} \Sigma_{i=1}^N Z_i Y_i. \end{aligned}$$

3.2.4 Efficient GMM Estimator (Conditional Homoscedasticity)

If we assume the conditional variance of U_i given Z_i is constant (which is also called conditional homoscedasticity)

$$E\left(U_i^2 \middle| Z_i\right) = \sigma^2.$$

Then we have

$$\begin{aligned} \Omega &= E\left(U_i^2 Z_i Z_i'\right) \\ &= E\left[E\left(U_i^2 Z_i Z_i' \middle| Z_i\right)\right] \\ &= E\left[E\left(U_i^2 \middle| Z_i\right) Z_i Z_i'\right] \\ &= \sigma^2 E\left(Z_i Z_i'\right). \end{aligned}$$

A natural consistent estimator of $E\left(Z_i Z_i'\right)$ is $N^{-1}\Sigma_{i=1}^N Z_i Z_i'$. Note that $\hat{\beta}_N^{GMM}$ defined in (3.11) is invariant to $a_N' a_N$ up to a constant term. Hence, in order to obtain the efficient GMM estimator in this case, we can set

$$a_N' a_N = \left(\Sigma_{i=1}^N Z_i Z_i'\right)^{-1}. \tag{3.12}$$

Substituting (3.12) into (3.11), the efficient GMM estimator, also known as the Two-Stage-Least-Square (2SLS) estimator, is given by

$$\hat{\beta}_N^{2SLS} = \left(\Sigma_{i=1}^N X_i Z_i' \left(\Sigma_{i=1}^N Z_i Z_i' \right)^{-1} \Sigma_{i=1}^N Z_i X_i' \right)^{-1} \Sigma_{i=1}^N X_i Z_i' \left(\Sigma_{i=1}^N Z_i Z_i' \right)^{-1} \Sigma_{i=1}^N Z_i Y_i.$$

Denote $X = (X_1, X_2, ..., X_N)'$, $Z = (Z_1, Z_2, ..., Z_N)'$ and $Y = (Y_1, Y_2, ..., Y_N)'$. The 2SLS estimator can be expressed as

$$\hat{\beta}_N^{2SLS} = \left(X'Z \left(Z'Z \right)^{-1} Z'X \right)^{-1} X'Z \left(Z'Z \right)^{-1} Z'Y. \tag{3.13}$$

The reason for the name “Two-Stage-Least-Square” lies in the fact that (3.13) can be obtained through the following two regression steps.

Step 1. Regress X on Z. In other words, project the matrix of regressors X orthogonally onto the space spanned by the instruments Z to obtain

$$\tilde{X} = Z\left(Z'Z\right)^{-1}Z'X.$$

Step 2. Obtain $\hat{\beta}_N^{2SLS}$ by running an OLS regression of Y on $\tilde{X}$, which gives

$$\hat{\beta}_N^{2SLS} = \left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'Y. \qquad (3.14)$$

It is easy to verify (3.14) is equivalent to (3.13).

3.3 Example

In this section, an example is given to illustrate the use of GMM estimator in the linear model. Based on the example, I also discuss the potential difficulties in applying GMM estimators in the above framework. Suppose a researcher is interested to see whether education really helps to improve a person’s wage level, after controlling for the effects of ability. Denote Y_i as the annual wage for the i^{th} person in the researcher’s random sample, X_i as education measured by number of years in school, D_i as ability which is unobserved by the researcher. Suppose we may explain the wage based on the following linear model

$$Y_i \quad = \quad \alpha + \beta \cdot X_i + \delta \cdot D_i + u_i,$$

where $E\left(u_i\right) = 0$, u_i is uncorrelated with both X_i and D_i while β is the parameter of interest. Since the researcher does not observe D_i, the linear model he/she is facing should be

$$Y_i = \tilde{\alpha} + \beta \cdot X_i + \varepsilon_i, \qquad (3.15)$$

where $\tilde{\alpha} = \alpha + \delta \cdot E\left(D_i\right)$ and $\varepsilon_i = \delta \cdot D_i - \delta \cdot E\left(D_i\right) + u_i$. Hence

$$E\left(\varepsilon_i\right) = 0. \qquad (3.16)$$

Since the number of years in school is generally correlated with a person’s ability, we have

$$E\left(X_i\varepsilon_i\right) \neq 0. \tag{3.17}$$

Therefore, the linear model ((3.15) to (3.17)) faced by the researcher belongs to the class of linear models characterized by (3.1) to (3.3). Hence running an OLS regression of Y_i on X_i does not provide a consistent estimate for β. In fact, the level of education is likely to be positively correlated with ability, and ability is also probably positively correlated with wage, the usual OLS estimate of β should be inflated. This is a classical problem in labor economics. Historically, researchers have proposed many instruments under different situations in order to identify the education effect. Here I will discuss two instruments: the quarter of birth of the i^{th} person and the education levels of the i^{th} person's parents. The reason I choose to discuss these two instruments is that they reflect two different kinds of difficulties when applying the GMM technique developed in this section.

To be a valid instrument, the quarter of birth/parent's education should be correlated with a person's education, but not correlated with a person's ability and any other unobservable factors that affect wage level. For the quarter of birth instrument, it is easy to rationalize that it is uncorrelated with a person's ability and other unobservable factors. However, it is not easy to rationalize why it is correlated with a person's number of years in school. Fortunately, the data can provide us evidence to see whether quarter of birth is correlated with number of years in school. In fact, a side-by-side boxplot of number of years in school against four quarters can help verification. Actually, Angrist and Krueger's (1991) show that to some extent, these two variables are indeed correlated though not strongly correlated. In contrast, for the parent's education instrument, it is easy to rationalize that it should be correlated (or even strongly correlated) with a person's education. However, it is hard to rationalize the parent's education is not correlated with a person's ability and any other unobserved factors that influence wage. Unfortunately, this problem is not testable from the data. If one wants to apply such instrument, he/she can only assume the parent's education is exogenous.

The above discussion shows the embarrassment when applying IV regressions. It is relatively easy to find an instrument that is correlated with the endogenous regressor while its exogeneity is hard to rationalize. Moreover, the exogeneity requirement of the instrument is not testable from the data. In contrast, it is relatively easy to find an instrument that is exogenous (uncorrelated with the unobserved term) while its correlation with the endogenous regressor is usually small. Fortunately, the correlation between an instrument and the endogenous regressors can be estimated from the data. In summary, to find a good instrument in reality is not an easy task.

Chapter 4

Simulation Studies

In this section, a series of simulation studies are conducted to investigate the performance of GMM estimator under various conditions. All the simulation studies in this section are based on the linear model introduced in Section 3. Specifically, consider the following linear model:

$$Y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \varepsilon,$$

$$Z = \delta \cdot X_1 + u,\ W = \gamma \cdot X_1 + e,$$

where

$$\begin{pmatrix} X_1 \\ X_2 \\ \varepsilon \\ u \\ e \end{pmatrix} \sim N(\mu, \Sigma),\ \mu = \begin{pmatrix} \mu_{x_1} \\ \mu_{x_2} \\ 0 \\ 0 \\ 0 \end{pmatrix},\ \Sigma = \begin{pmatrix} 1 & \rho_{x_1x_2} & \sigma_\varepsilon \rho_{x_1\varepsilon} & \rho_{x_1u} & \rho_{x_1e} \\ & 1 & \sigma_\varepsilon \rho_{x_2\varepsilon} & \rho_{x_2u} & \rho_{x_2e} \\ & & \sigma_\varepsilon^2 & \sigma_\varepsilon \rho_{\varepsilon u} & \sigma_\varepsilon \rho_{\varepsilon e} \\ & & & 1 & \rho_{ue} \\ & & & & 1 \end{pmatrix}.$$

Y, X_1, X_2, W and Z are random scalars which are assumed to be observable; while ε, u and e are random scalars which are assumed to be unobservable. All parameters are pre-specified and data sets are simulated based on their values. Parameters α, β_1 and β_2 will be estimated by GMM using simulated data sets. In addition, X_1 is assumed to be endogenous and X_2 is assumed to be exogenous. Both Z and W are valid instruments for X_1. In other words, the parameters δ, γ, μ and Σ are chosen such that the following moment conditions are satisfied:

$$E(X_1\varepsilon) \neq 0,\ E(X_2\varepsilon) = 0,$$

$$E(Z\varepsilon) = 0,\ E(W\varepsilon) = 0, \text{(Instrument Exogeneity)} \tag{4.1}$$

$$E(WX_1) \neq 0,\ E(ZX_1) \neq 0 \text{(Instrument Relevance).} \tag{4.2}$$

Such moment conditions imply we are not “free” to choose values of all parameters. The restrictions imposed on the parameters are listed below.

$$\rho_{x_1\varepsilon} \neq 0, \, \rho_{x_2\varepsilon} = 0, \, \rho_{\varepsilon u} = -\delta\rho_{x_1\varepsilon}, \, \rho_{\varepsilon e} = -\gamma\rho_{x_1\varepsilon},$$

$$\delta\left(1+\mu_{x_1}^2\right) + \rho_{x_1 u} \neq 0, \, \gamma\left(1+\mu_{x_1}^2\right) + \rho_{x_1 e} \neq 0.$$

In some studies described later, I will use models with more than two instruments. In those cases, Z, W, u and e could be viewed as random vectors while δ and γ are vectors containing fixed constants correspondingly. Hence, in all simulation studies, a population distribution is indexed by the following collection of parameters

$$\{\mu, \Sigma, \delta, \gamma, \alpha, \beta_1, \beta_2\}. \tag{4.3}$$

In addition, I assume conditional homoscedasticity of ε. Specifically

$$E\left(\varepsilon^2 | Z, W\right) = \sigma_\varepsilon^2.$$

Hence, in all simulation studies below, the GMM estimator is the 2SLS estimator which has a close form expression and hence can be easily obtained.

4.1 Study 1 - Robustness to Mixed Populations

The cornerstone of GMM estimator is based on a set moment conditions that characterize the data generating mechanism. The moment conditions do not completely specify the joint distribution. Under some regularity conditions, the desired properties of GMM estimator, like consistency and asymptotic normality, hold as long as the moment conditions are correctly specified. Study 1 is designed to investigate the performance of GMM estimator when the observed data come from a mixture - two different underlying distributions. Hence, there are two “parts” of the observed data in this study. For example, 90% of the data is generated by distribution A and the rest of the data is generated by another distribution B. Specifically, the variables $\{X_1, X_2, \varepsilon, u, e\}$ are generated from distribution A and B respectively, based on which (together with α, β_1 and β_2) the variable Y is generated accordingly. The mixture of the observed data is characterized in two dimensions. The first one relates to the degree of mixture, which is represented by the relative proportions of distribution A and B in generating the data. The second one relates to the “difference” between distributions A and B, which is represented by the difference in the means of regressors between two distributions. The *objective* of this study is to investigate how bias and efficiency of the GMM estimator changes in response to changes in the above two dimensions.

The distributions generating part A and B are the same for all components in (4.3) except for μ_{x_1} and μ_{x_2}. The values of the parameters are chosen as follows:

Para.	Value	Para.	Value	Para.	Value	Para.	Value
α	1	$\rho_{x_1 x_2}$	0.1	$\rho_{x_2 u}$	0.2	$\mu^A_{x_1}$	1
β_1	2	$\rho_{x_1 \varepsilon}$	0.5	$\rho_{x_2 e}$	0.2	$\mu^A_{x_2}$	1
β_2	3	$\rho_{x_1 u}$	0.2	$\rho_{\varepsilon u}$	-0.5	$\mu^B_{x_1}$	$\mu^A_{x_1} + \mu_0$
δ	1	$\rho_{x_1 e}$	0.2	$\rho_{\varepsilon e}$	-0.5	$\mu^B_{x_2}$	$\mu^A_{x_2} + \mu_0$
γ	1	$\rho_{x_2 \varepsilon}$	0	ρ_{ue}	0.2	σ^A_{ε}	1

Table 4.1: Distributional parameter values

where μ_0 varies in different simulated scenarios and captures the difference between distribution A and B. In addition, I set $\sigma^B_{\varepsilon} = \mu_0 + 1$. By doing so, the structural part of Y (represented $\alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2$) does not dominate the random component ε.

As discussed before, the simulation study is designed mainly in two dimensions: (i) the degree of mixture characterized by a parameter p_B, which is the proportion of data coming from distribution B; (ii) the difference in means of X_1 and X_2 between distribution A and B, denoted by μ_0. In addition, I also vary the sample size in the simulation study. In other words, the objective of this study is to examine the bias and efficiency of the GMM estimator when p_B and μ_0 change. Every dimension has three distinct values, resulting in 27 different scenarios in total. For each scenario, I performed 1000 Monte Carlo simulations. For each combination of parameter values in each simulation, I first generated n vectors of $\{X_1, X_2, \varepsilon, u, e\}$ based on which I then generated n vectors of $\{Y, Z, W\}$. Then the GMM estimator is estimated based on the formula of 2SLS estimator in (3.14). The table below shows the values these three dimensions take.

n	100	500	1000
p_B	0.1	0.2	0.3
μ_0 $(\mu^B_{x_1} - \mu^A_{x_1})$	3	5	10

Table 4.2: Parameter values for study 1

Although there are 27 different scenarios, I only present the results for the following six situations.

			$\hat{\beta}_1$ ($\beta_1 = 2$)		$\hat{\beta}_2$ ($\beta_2 = 3$)	
n	p_B	μ_0	Mean	SD	Mean	SD
100	0	—	1.9987	0.1156	2.9962	0.1016
100	0.3	10	1.9906	1.5075	3.0483	5.7763
100	0.3	5	2.0100	0.5839	3.0167	1.8389
1000	0.3	10	1.9776	0.4914	3.0370	1.9044
1000	0.3	5	2.0001	0.1738	2.9929	0.5570
100	0.1	10	2.0134	0.9291	2.9537	3.8323
100	0.1	5	2.0128	0.3839	2.9804	1.3637

Table 4.3: Results of study 1

The means of $\hat{\beta}_1$ and $\hat{\beta}_2$ in all scenarios are very close to their true values respectively. It indicates the bias of GMM estimator in this case is small even when the sample size is small (e.g. 100). Hence, the first conclusion from this simulation study is that mixed population (characterized in this study) has little effect on the bias of GMM estimator.

However, the table above shows the efficiency of GMM estimator highly depends on how the population is mixed. Specifically, holding all other aspects constant, the higher the degree of mixture (reflected by p_B), the higher the variation of the GMM estimator. For example, when $n = 100$ and $\mu_0 = 10$, changing p_B from 0.1 to 0.3, the standard deviations of $\hat{\beta}_1$ (coefficient of endogenous regressor) and $\hat{\beta}_2$ (coefficient of exogenous regressor) are increased by 62% and 51% respectively. In addition, holding all other aspects fixed, the higher the difference of mixed populations (reflected by μ_0), the higher the variation of the GMM estimator. For instance, when $n = 100$ and $p_B = 0.1$, changing μ_0 from 5 to 10, the standard deviations of $\hat{\beta}_1$ and $\hat{\beta}_2$ are increased by 142% and 181% respectively. The following graph visualizes the results.

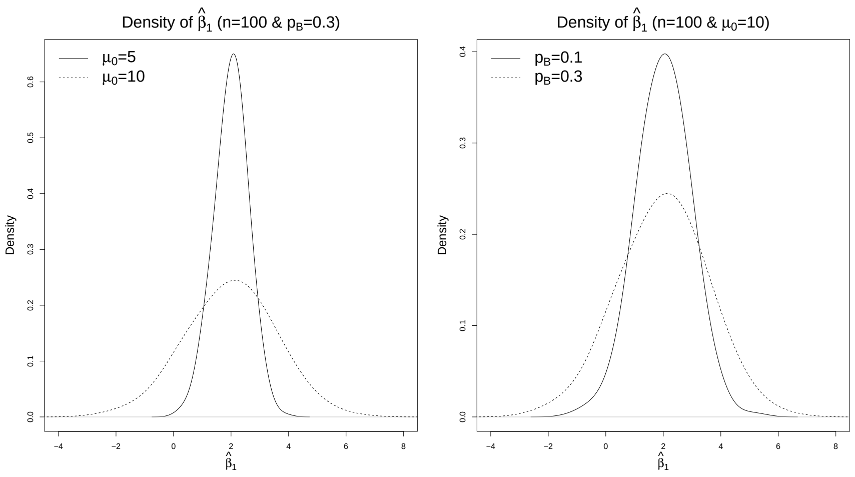

Figure 4.1: Graphical comparison for study 1

Therefore, the second conclusion of this study is that: when the degree of mixture is higher or the difference of the mixed population is larger, the efficiency of the GMM estimator declines significantly. The practical implication of this study is non-trivial. Suppose a researcher is interested in the causal effect of advertisement on sales. Therefore X_1 in the study can represent the annual investment on advertisement of a firm while X_2 represents all other sales-relevant factors. The variation of investment on advertisement of different firms varies a lot in the market, i.e. large firms, particularly internationally reputed firms, spend millions of dollars for advertising while medium or small firms spend significantly less on it. Hence, a random sample of firms collected by the researcher is likely to contain many large, medium and small firms. Suppose in an extreme case, for some reason, the researcher intends to collect a sample with size 100 that consists of 70% small firms (reflected by distribution A in the study) and 30% large firms (reflected by distribution B). Suppose μ_0 is 10 (in millions of dollars) and the true causal effect β_1 is 2. Suggested by the simulation results above, before collecting the sample and performing the analysis, the researcher has 25% (or 75%) chance to underestimate (or overestimate) the advertisement effect by 50% (the chance of $\hat{\beta}_1 = 1$ or $\hat{\beta}_1 = 3$). What is worse is that the researcher has a 10% chance to obtain a negative advertisement effect (the chance of $\hat{\beta}_1 < 0$).

In summary, this simulation study shows that when the sample is mixed with two parts coming from different populations, the bias of the GMM estimator is nearly unaffected but the efficiency decreases significantly.

4.2 Study 2 - Robustness to Outliers

In this study, I will investigate the performance of GMM estimator when a small proportion of the data are contaminated. By contamination, it means a small proportion of the data is generated from another mechanism as opposed to the mechanism that generates the majority of the data. Contaminated data happens for different reasons in practice, i.e. incorrect record. The way I defined contaminated data is as follows: a random disturbance term is added to the response after it was generated by the true underlying model. Suppose I generate 100 observations from a specified distribution, a random disturbance term is then added to the last response. As a result, the last observation is contaminated. For the perspective of data composition, study 1 and study 2 are similar to each other in that the observed data is not generated from a common mechanism. The difference between two studies lies in the following. The mixture feature of the observed data comes from different sources. In study 1, the mixture feature comes from the question itself under investigation. In study 1, the reason of mixture comes from the fact that some firms are of large scale while many other firms are of medium or small scale, reflected by the difference in population means of X_1. As a result, larger firms tend to have a higher value of Y. However, in study 2, the source of mixture is out of the question under investigation. In other words, it can often be characterized by “mistakes”, either made by human being or machine. Since the mixture feature comes from different sources, the characteristics of mixture in study 2 is also different from study 1. The contamination proportion p_C is usually very low (e.g. 1%). The strength of contamination is usually not small, i.e. the mean of the random contamination disturbance μ_C is non-trivial. For example, the data collector may accidentally input the value “1000” instead of “100”. The difference between these two studies can also be seen mathematically. For study 1, those exceptionally large value of Y is attributed to the corresponding large value of X_1; for study 2, it is attributed to a large value of the intercept α.

The *objective* of this study is to investigate the performance of GMM estimator under different situations of p_C and μ_C. We use the parameter values of distribution A in study 1 to generate the majority (uncontaminated) of data in this study. The random contamination disturbance is specified as a normal random variable with mean μ_C and unit variance. The simulation study is designed through two dimensions: (i) the proportion of contaminated data (p_C); (ii) the mean of the random contamination disturbance, μ_C. In addition, I also vary the sample size n. For each scenario, I performed 1000 Monte Carlo simulations. The table below shows the values these three dimensions take.

Sample Size (n)	100	500	1000
Contamination Proportion (p_C)	0.01	0.05	0.1
Mean of Contamination (μ_C)	±5	±10	±50

Table 4.4: Parameter values of study 2

So there are 54 $(3 \times 3 \times 6)$ scenarios in total. Some results are reported below. The table reports the mean and standard deviation of $\hat{\beta}_1$ and $\hat{\beta}_2$ for seven combinations of the parameter values.

			$\hat{\beta}_1$ ($\beta_1 = 2$)		$\hat{\beta}_2$ ($\beta_2 = 3$)	
n	p_C	μ_C	Mean	SD	Mean	SD
100	0	0	1.9987	0.1156	2.9962	0.1016
100	0.01	50	2.0029	0.6263	2.9968	0.5023
100	0.01	-50	2.0349	0.6188	2.9842	0.5197
100	0.05	50	1.9402	1.2865	3.0668	1.1255
100	0.05	-50	2.0119	1.3095	2.9652	1.1198
100	0.05	10	2.0006	0.2942	3.0105	0.2452
100	0.05	-10	1.9965	0.2832	3.0011	0.2623

Table 4.5: Results of study 2

The reason for reporting the results where $\mu_C = \pm 50$ and $n = 100$ lies in the intuition that the distortions of the parameter estimates should be largest in these two cases if there are any. From the table above, the bias of GMM estimator is small in all scenarios reflected by the closeness between mean of the estimates and the true value. In terms of the variation of the estimates, when the contamination proportion p_C is low (e.g. 1%) or the contamination strength μ_C is low (e.g. 10), the standard deviations of $\hat{\beta}_1$ and $\hat{\beta}_2$ are relatively small. However, when both contamination proportion and strength are high (e.g. 5% and 50 respectively), the standard deviations of the estimators are high compared to their means. For example, when $n = 100$, $p_C = 5\%$ and $\mu_C = 50$, the standard deviation of $\hat{\beta}_1$ is about 66% of its mean. Therefore, it is suggested by the results that the efficiency of GMM estimator decreased significantly when both contamination proportion and strength are high. In addition, I found no systematic effect of the sign for contamination strength. The following graph also illustrates the results.

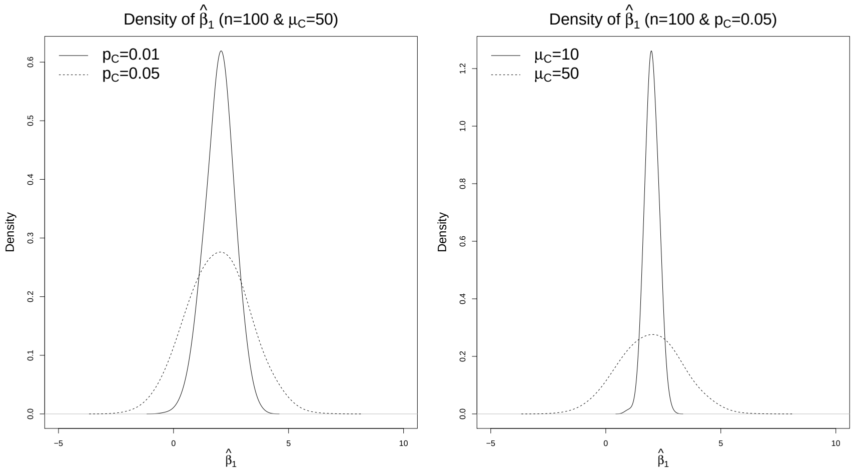

Figure 4.2: Graphical comparison for study 2

In summary, the conclusions of this simulation study are as follows. In terms of bias, the GMM estimator is robust to both contamination proportion and strength. In terms of efficiency, the GMM estimator is relatively robust to either contamination proportion or strength. However, its efficiency decreases significantly when both contamination proportion and strength are high.

4.3 Study 3 - Weak Instruments

Recall that in section 3.3, I discussed the potential difficulties of applying GMM estimator in reality through an example. The two requirements for an instrument to be valid are hard to satisfy simultaneously, especially given that the exogeneity requirement (4.1) is untestable from the data. In this study, I will study the performance of GMM estimator when the exogenous instruments are only weakly correlated with the endogenous regressor. Recall that to be valid instruments, Z and W must satisfy (4.2):

$$E\left(WX_1\right) \neq 0,\ E\left(ZX_1\right) \neq 0.$$

Under our model setup, we have

$$\begin{aligned}
E\left(ZX_1\right) &= E\left[\left(\delta X_1 + u\right)X_1\right] = \delta\left(1+\mu_{x_1}^2\right) + \rho_{x_1 u},\\
E\left(WX_1\right) &= E\left[\left(\gamma X_1 + e\right)X_1\right] = \gamma\left(1+\mu_{x_1}^2\right) + \rho_{x_1 e}.
\end{aligned}$$

In this study, I chose $\rho_{x_1 u} = \rho_{x_1 e} = 0$ and $\mu_{x_1} = 1$. Hence, by varying δ and γ, I could set the correlation between the instruments and X_1 to be arbitrarily small. For simplicity, I let $\delta = \gamma$ in this study. Other parameters that govern the joint distribution of the data are chosen the same as those in study 1. Therefore, the *objective* of this simulation study is to investigate the performance of GMM estimator, both bias and efficiency, in response to δ which determines the correlation between the instruments and the endogenous regressor.

The simulation study is designed through two dimensions: δ and sample size (n). The table below shows the values these two dimensions take.

Sample Size (n)	50	100	500	1000	—
δ	0.01	0.05	0.1	0.15	0.2

Table 4.6: Parameter values of study 3

The results for $\hat{\beta}_1$ and $\hat{\beta}_2$ are summarized in the following table.

		$\hat{\beta}_1$ ($\beta_1 = 2$)				$\hat{\beta}_2$ ($\beta_2 = 3$)			
		GMM Est.		OLS Est.		GMM Est.		OLS Est.	
n	δ	Mean	SD	Mean	SD	Mean	SD	Mean	SD
100	1	2.0067	0.1329	2.5048	0.0899	2.9957	0.1033	2.9471	0.0901
1000	1	1.9979	0.0398	2.5039	0.0261	2.9988	0.0317	2.9488	0.0269
100	0.01	2.5271	1.3493	2.5081	0.0892	2.9449	0.2197	2.9462	0.0887
100	0.05	2.4368	1.2778	2.5036	0.0932	2.9666	0.2103	2.9551	0.0893
100	0.1	2.2980	1.3055	2.5082	0.0896	2.9746	0.1912	2.9511	0.0857
100	0.2	2.0352	1.7101	2.5028	0.0895	2.9968	0.2458	2.9522	0.0886
1000	0.05	2.2068	1.4474	2.5054	0.0275	2.9801	0.1829	2.9507	0.0281
1000	0.1	1.9894	0.3497	2.5052	0.0277	3.0007	0.0514	2.9499	0.0286

Table 4.7: Results of Study 3

Not surprisingly, the standard deviations of the GMM estimator in all scenarios are larger than those of the OLS estimator. In fact, this is a general result although I did not provide a proof in earlier chapters. The intuition can be explained as follows. Remember the GMM estimator in these simulation studies is the 2SLS estimator. Hence, the regressors X_1 and X_2 are projected onto the instruments Z and W in the first stage. The projection is then used as regressors to estimate the parameters β_1 and β_2 in the second stage. Therefore, not all information (or variation) embedded in the regressors is used in the estimation. Intuitively, only the information that can be absorbed by the instruments is used in estimating the parameters. In contrast, the OLS estimator only involves regressing the response on the original regressors X_1 and X_2. Hence, it is not hard to rationalize that the variation of the estimator is larger in GMM than OLS. When the instruments are not weak (e.g. $\delta = 1$), the bias of GMM estimator of both β_1 and β_2 are negligible. In contrast, in such case, the bias of OLS estimator is non-trivial for β_1 (roughly 25% larger than the true value) but very small bias for β_2 (roughly 1.6% larger smaller than the true value).

The above table shows that when sample size is small (e.g. 100) and the instruments are weakly correlated with X_1 (e.g. $\delta = 0.01$ or 0.05), the bias of GMM estimator is roughly the same as that of the OLS estimator. In addition, the effect of weak correlation between the instruments and the endogenous regressor spills over to the estimate of β_2 (the coefficient of the exogenous regressor X_2), making its GMM estimator slightly biased. The following graph visualize the comparison.

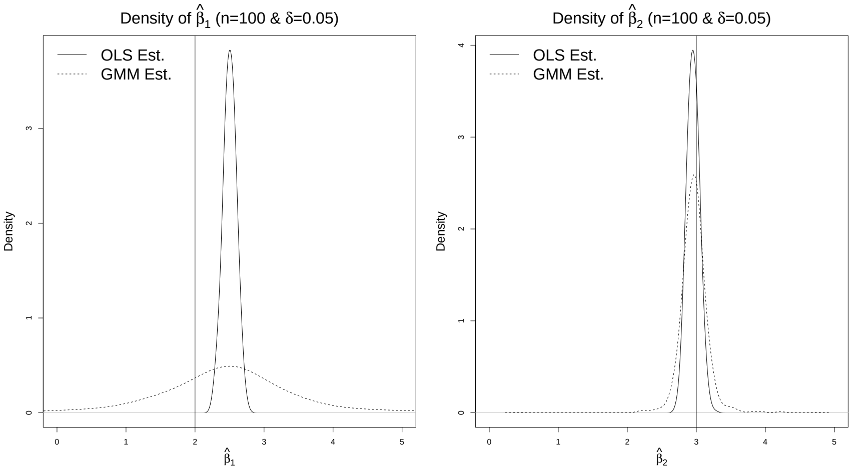

Figure 4.3: Comparison of OLS and GMM estimators

Therefore, when sample size is not large and the instruments are only weakly correlated with the endogenous regressors, the GMM estimator is moderately biased and has relatively large standard deviation. In such cases, the OLS estimator is not much worse in terms of bias.

The results also show that when δ is very small, i.e. 0.05, increasing sample size from 100 to 1000 helps to reduce the bias but not the standard deviation. Even though sample size is 1000 (and $\delta = 0.05$), the GMM estimator still has non-trivial bias, i.e. 10% higher than the true value. On the other hand, when δ is only moderately small, i.e. 0.1, increasing the sample size from 100 to 1000 significantly reduces the bias and standard deviation. In fact, the bias of the GMM estimator when $\delta = 0.1$ and $n = 1000$ is nearly the same as that when $\delta = 1$ and $n = 1000$. The following graph depicts these results.

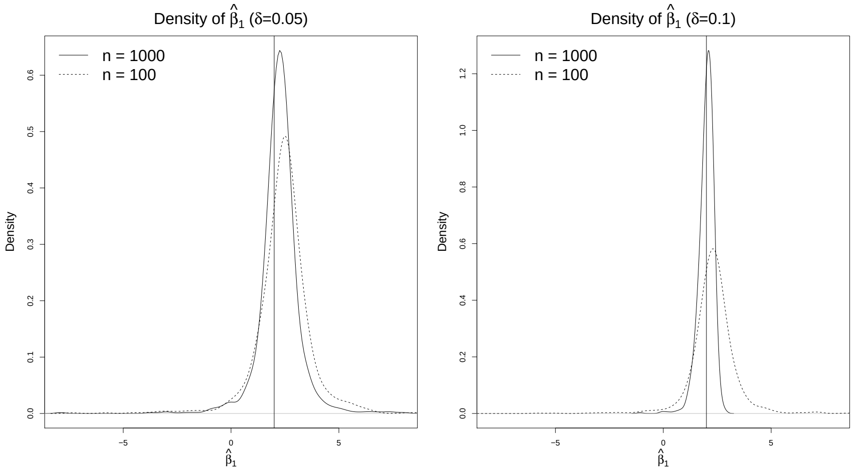

Figure 4.4: Comparison of GMM estimators under different correlations for varied sample size

In addition, the result also shows that for a given sample size, i.e. 100, the bias reduces when the instruments are more correlated with X_1. However when sample size is 100, standard deviations of the GMM estimator stay basically the same when δ increases. The following plot shows this result.

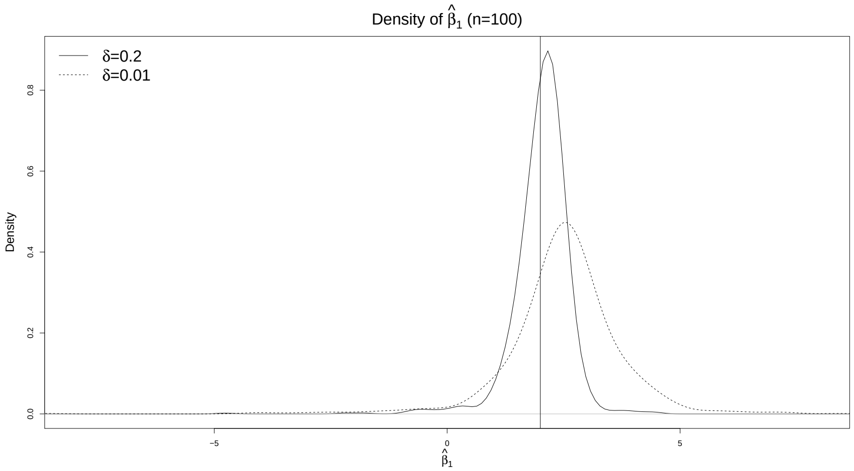

Figure 4.5: Comparison of GMM estimators under different correlations for a given sample size

In summary, when the instrument is only weakly correlated with the endogenous regressor and the sample size is not large, the GMM estimator of the endogenous regressor's coefficient has non-trivial bias and large variation. In such case, the GMM estimator of the exogenous regressor's coefficient is also slightly affected in terms of bias. In addition, it requires a very large sample size, i.e. larger than 1000, to reduce the bias and variation of the GMM estimator for the endogenous regressor's coefficient when the instruments are weak.

4.4 Study 4 - Mixed Strong and Weak Instruments

In the previous study, I investigate the performance of GMM estimator when both instruments are weak and linear functions of the endogenous regressor X_1. In this study, I will examine the performance of GMM estimator when there are strong and weak instruments. Specifically, there is one strong instrument (Z) and two weak instruments in this study. In addition, one of the weak instrument (W) is still a linear function of X_1 but the other one (Q) is a quadratic function of X_1. In the previous study, I show that weak relationship between the endogenous regressors and the instruments may yield undesired results of GMM estimators, i.e. non-trivial bias. In this study, the *objective* is twofold: (1) whether weak instruments still bias the estimator with the presence of a strong instrument; (2) how an additional weak instrument affect the GMM estimator's performance with the presence of strong and weak instruments.

4.4.1 Specification of Quadratic Instrument

As discussed before, I make an instrument Q that is quadratically related to X_1:

$$Q = aX_1^2 + bX_1 + \xi, \text{ where } E(\xi) = 0, Var(\xi) = 1.$$

In order to ensure Q to be exogenous which means

$$E(Q\varepsilon) = E\left(\left(aX_1^2 + bX_1 + \xi\right)\varepsilon\right) = 0,$$

the following restrictions are imposed

$$\begin{aligned} & E(\xi\varepsilon) = -\left[aE\left(X_1^2\varepsilon\right) + bE(X_1\varepsilon)\right] \\ \Rightarrow \quad & \rho_{\varepsilon\xi} = -\left[a\left(1+\mu_{x_1}^2\right) + b\rho_{x_1\varepsilon}\right]. \end{aligned}$$

I also make $E(\xi X_1) = 0$. Hence another restriction (instrument relevance) for Q to be a valid instrument is

$$E(QX_1) = aE\left(X_1^3\right) + bE\left(X_1^2\right) \neq 0.$$

Since $\mu_{x_1} = 1$ and $\sigma_{x_1}^2 = 1$, this leads to the requirement of

$$E(QX_1) = 4a + 2b \neq 0.$$

In this study, I assume $a = 0.5$. Hence, the strength of the relationship between Q and X_1, defined as $E(QX_1)$, is affected only through b. For better illustration, we can re-express it as

$$E(QX_1) = 2(1+b) = 2\tilde{b}, \text{ where } \tilde{b} = 1 + b.$$

4.4.2 Simulation Design and Results

In this case, the measurement of relationship between the linear instrument and the endogenous variable is the same as before. The study is designed in three dimensions: (i) sample size, n; (ii) δ_W, which describes the weak correlation between the linear instrument (W) and X_1 while the strong instruments (Z) take value $\delta_Z = 1$ in this study; (iii) $\tilde{b}$, which describes the weak relation between quadratic instrument Q and X_1. The table below shows the values these three dimensions take.

n	100	500	1000
δ_W	0.01	0.05	0.1
$\tilde{b}$	0.01	0.05	0.1

Table 4.8: Parameter values of study 4

The following table reports some of the results for $\hat{\beta}_1$.

$\hat{\beta}_1$ ($\beta_1 = 2$)			Only Strong Linear Inst. (Z)	Strong & Weak Linear Inst. (Z & W)	All Inst. (Z, W & Q)	OLS
n	δ_W	$\tilde{b}$	Mean	Mean	Mean	Mean
100	0.01	0.1	1.9975	1.9997	1.9425	2.1016
			(0.1460)	(0.1449)	(0.2062)	(0.1019)
100	0.05	0.05	1.9986	2.0005	1.9699	2.1010
			(0.1520)	(0.1503)	(0.2035)	(0.1023)
100	0.05	0.1	2.0021	2.0039	1.9468	2.1029
			(0.1456)	(0.1446)	(0.1855)	(0.1009)
100	0.1	0.01	1.9983	2.0005	1.9945	2.0999
			(0.1467)	(0.1453)	(0.1965)	(0.1050)
1000	0.01	0.1	1.9987	1.9990	1.9396	2.1015
			(0.0449)	(0.0449)	(0.0613)	(0.0340)
1000	0.05	0.05	1.9962	1.9965	1.9664	2.0996
			(0.0465)	(0.0464)	(0.0654)	(0.0327)
1000	0.1	0.01	2.0014	2.0017	1.9927	2.1008
			(0.0446)	(0.0446)	(0.0625)	(0.0319)

Table 4.9: Results of $\hat{\beta}_1$ for study 4. Standard errors are reported in the brackets.

It is shown from the above table that the GMM estimators are basically unbiased and have relatively small standard deviation (compared to their means) when only the strong linear instrument is used. It also shows a general pattern that adding a weak linear instrument has virtually no effect on the bias and variation of the GMM estimator with the presence of a strong linear instrument. However, regardless of the sample size, when the weak linear instrument W is weakly correlated with X_1 (e.g. $\delta_W = 0.05$ or 0.01), adding a weak quadratic instrument Q increases the bias and standard deviation by roughly 2.5% and 30% respectively. The following graph visualizes the comparison.

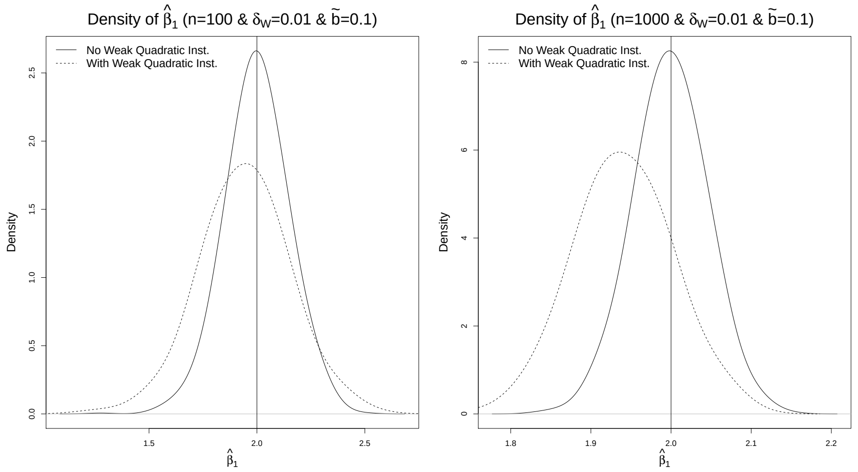

Figure 4.6: Comparisons of GMM estimator with and without a weak quadratic instrument

The first conclusion of this study is: with the presence of a strong linear instrument, adding a weak linear instrument does not affect the GMM estimator's performance but adding a non-linear weak instrument increases the bias and reduce accuracy. What's interesting from table 4.9 is that when the added quadratic weak instrument is extremely weak (e.g. $\tilde{b} = 0.01$), the variation of the GMM estimator increases but the bias is virtually unaffected. Another pattern from table 4.9 is that the standard deviations in all cases are small compared to those in table 4.7 where there are only two weak linear instruments. Hence, the second conclusion is the presence of a strong instrument reduces the variation significantly. The following table reports some of the results for $\hat{\beta}_2$.

$\hat{\beta}_2$ ($\beta_2 = 2$)			Only Strong Linear Inst. (Z)	Strong & Weak Linear Inst. (Z & W)	All Inst. (Z, W & Q)	OLS
n	δ_W	$\tilde{b}$	Mean	Mean	Mean	Mean
100	0.01	0.1	2.9984 (0.1100)	2.9982 (0.1100)	3.0037 (0.1131)	2.9876 (0.1087)
100	0.05	0.05	3.0039 (0.1068)	3.0037 (0.1067)	3.0063 (0.1084)	2.9950 (0.1053)
100	0.05	0.1	2.9972 (0.1029)	2.9970 (0.1029)	3.0030 (0.1045)	2.9871 (0.1018)
100	0.1	0.01	3.0031 (0.1085)	3.0028 (0.1084)	3.0031 (0.1107)	2.9927 (0.1074)
1000	0.01	0.1	3.0007 (0.0309)	3.0007 (0.0309)	3.0067 (0.0315)	2.9903 (0.0306)
1000	0.05	0.05	3.0007 (0.0313)	3.0006 (0.0313)	3.0036 (0.0317)	2.9903 (0.0311)
1000	0.1	0.01	3.0009 (0.0319)	3.0009 (0.0319)	3.0018 (0.0326)	2.9908 (0.0316)

Table 4.10: Results of $\hat{\beta}_2$ for study 4. Standard errors are reported in the brackets.

Table 4.10 shows that the bias and variation of $\hat{\beta}_2$ are very small in all cases. In contrast to the results in table 4.7, the bias of $\hat{\beta}_2$ is slightly increased when there are only weak instruments. Therefore the last conclusion of this study is the GMM estimator of the exogenous regressor's coefficient has little bias and small variation when there exists a strong instrument.

Chapter 5

Summary

Since Hansen's (1982) original paper on GMM estimator, it gained its popularity rapidly in recent years not only in econometrics but also other literatures (e.g. epidemiology). A vast volume of published papers in economics are related to the application of GMM estimator. Its popularity is partially attributed to the tempting theoretical asymptotic results and the ease of application in reality. Another reason lies in the fact that many economic models ultimately imply a set of moment conditions that can be used as the cornerstone of GMM estimation. However, the GMM estimator has its own weakness. A major concern is the so-called weak instrument which is just weakly correlated with the endogenous regressors. In addition, the validity of the moment conditions used in estimation is usually not testable from the observed data. Another concern bothers many practitioners is its performance when the sample size is small.

This thesis reviews the theoretical development of the GMM estimator and conducts several simulation studies to examine its performance under different situations. Specifically, I derive the complete proofs of consistency and asymptotic normality for the GMM estimator under i.i.d data structure. I then review the application of GMM estimator in linear models. Specifically, I emphasize the motivations for model formulations in econometrics, based on which the suitability of GMM estimator in such framework is illustrated. On the other hand, I also explain the potential difficulties of applying GMM estimator through an example. The use of GMM estimator is, however, not restricted to linear models. In fact, it is also widely used in the discrete choice models (described in section 1.2.1), i.e. Steven Berry *et al* (1995) provides one application of GMM estimator in discrete choice models with endogenous regressor and aggregate level data. Paul Gustafson *et al* (2008) provides the application of instrumental variables in generalized linear models in the context of epidemiological studies. In fact, the GMM technique is heavily used in econometrics when endogeneity is a key feature of the model.

In recognition of its potential weakness, the simulation studies are designed to investigate the performance of GMM estimator under different scenarios. The first two studies are related to a common fact that the observed data does not come from a single population. A general conclusion from these two studies is that under mixed populations the bias of GMM estimator is trivial but its efficiency is sensitive to various aspects of the mixture. The last two studies involves the presence of weak instruments. A general conclusion is that weak instruments dramatically increase the bias and variation of the GMM estimator even though the sample size is moderately large.

In summary, the GMM estimator has its own theoretical desired properties but caution is warned for its practical applications.

Bibliography

Amemiya, T. (1985). *Advanced Econometrics*. Cambridge, MA: Harvard University Press.

Angrist, J. D., and Krueger, A. B. (1991). Does Compulsory School Attendance Affect Schooling and Earnings. *Quarterly Journal of Economics*, **106**: 979-1014.

Berry, S., Levinsohn, J., and Pakes, A. (1995). Automobile Prices in Market Equilibrium. *Econometrica*, **63**: 841-890.

Bound, J., and David, A. J. (2000). Do Compulsory Schooling Attendance Laws Alone Explain the Association between Quarter of Birth and Eamings. *Research in Labor Economics*, **19**: 83-108.

Bound, J., Jaeger, D. A., and Baker, R. (1995). Problems With Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variables Is Weak. *Journal of the American Statistical Association*, **90**: 443-450.

Buse, A. (1992). The Bias of Instrumental Variable Estimators. *Econometrica*, **60**: 173-180.

Card, D. E. (1999). The Causal Effect of Education on Earnings. *Handbook of Labor Economics*, **3**: 1801-1863.

Donald, S. G., and Newey, W. K. (2001). Choosing the Number of Instruments. *Econometrica*, **69**: 1161-1191.

Hall, A. R., Rudebusch, G. D., and Wilcox, D. W. (1996). Judging Instrument Relevance in Instrumental Variables Estimation. *International Economic Review*, **37**: 283-289.

Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, **50**: 1029-1054.

Heckman, J. J. (2000). Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective. *Quarterly Journal of Economics*, **115**: 45-97.

Johnston, K. M., Gustafson, P., Levy, A. R., and Grootendorst, P. (2008). Use of instrumental variables in the analysis of generalized linear models in the presence of unmeasured confounding with applications to epidemiological research. *Statistics in Medicine*, **27**:1539-1556.

Stock, J. H., Wright, J. H., and Yogo, M. (2002). A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments. *Journal of Business & Economic Statistics*, **20**: 518-529.

Staiger, D., and Stock, J. H. (1997). Instrumental Variables Regression With Weak Instruments. *Econometrica*, **65**: 557-586.

Stock, J. H., and Wright, J. H. (2000). GMM With Weak Identification. *Econometrica*, **68**: 1055–1096.

Wang, J., and Zivot, E. (1998). Inference on Structural Parameters in Instrumental Variables Regression With Weak Instruments. *Econometrica*, **66**: 1389-1404.