Approaches to modelling heterogeneity in longitudinal studies

by

Xiaomei Li

A thesis
submitted to the Victoria University of Wellington
in partial fulfilment of the
requirements for the degree of
Master of Science
in Statistics and Operations Research.

Victoria University of Wellington
2011

Abstract

This thesis is about estimation bias of longitudinal data when there is correlation between the explanatory variable and the individual effect. In our study, we firstly introduce what is longitudinal data, then we introduce the commonly used estimation methods for the general linear model: the least squares method and maximum likelihood method. We apply these estimation methods to three simple general models which are commonly used to analyse longitudinal data. Secondly, we use frequentist and Bayesian analysis to explore the estimation bias theoretically and empirically, with an emphasis on the heterogeneity bias. This bias occurs where random effect estimation is used to analyse data with nonzero correlation between explanatory variables and the individual effect. We then empirically compare the estimated value with the true value. In this way, we demonstrate and verify the theoretical formulation which can be used to determine the size of the bias [Mundlak, 1978]. In order to avoid the estimation bias, the fixed effect estimation should be used to get the better solution under nonzero correlation situation. The Hausman test is used to confirm this. However, the bias not only occurs when we use frequentist analysis, but also exist by using the Bayesian estimation of random effect model. Finally, we follow the Mundlak [1978] idea, then define the special Bayesian model which can be used as Hausman test and as a comparable model. We also prove that it is best fit model among the random effect, fixed effect and pooled model if there is correlation between explanatory variables and individual effect. Throughout this thesis, we illustrate this ideas using examples based on real and simulated data.

Acknowledgements

I sincerely thank my supervisor Dr. Richard Arnold for all the guidance and help. During my study, he not only gave me his patience and support, but also taught me how to study and learn in unknown areas. So he really gave me an indication to become a real statistical analyst.

I would also like to thank Dr. Ivy Liu for the useful notes on longitudinal study and Dr. Dong Wong and Teo Her Guan for their useful suggestions and comments. Also, I am grateful for all the assistance and support from all of my lecturers, A/Prof. Megan Clark, Prof. Estate Khmaladze, Dr. John Haywood, Dr. Yuichi Hirose, etc and the staff of the School of Mathematics, Statistics and Operations Research, Victoria University of Wellington.

Lastly, I would like to thank all of my family and friends for their encouragement, especially my mother for helping me take care of my first baby, without her help, I couldn't have completed my thesis on time.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Longitudinal Study

A longitudinal study is a study that involves repeated observations of the same individual over time (or over different locations, etc.). Cross sectional study is a study that involves observations of a population or a sample that are made at one single time point. Because longitudinal studies are the repeated observation on the same individual, it may be thought as a repeated cross section. Therefore, longitudinal data analysis can be applied in various of field.

Longitudinal data is used in a wide range of fields: economics, biology, public health, business, social sciences, education, etc. Econometricians call it panel data. Longitudinal data consist of repeated observations of an outcome variable y on a set of individuals. These individuals may be people, animals, plants, businesses, field plots etc. And usually are a sample from a population. The repeated measurements are often taken at different times, however they may instead be at different locations (e.g. within an agricultural trial, the unit being experimental plots) or other form of replicates within an experiment. For most data described as longitudinal the replicates have a unique ordering (such as ordering in time), but longitudinal analysis methods may also be applied to unordered data, and are frequently used in spatial analyses such as small area estimation.

Methods exist for the analysis of continuous, discrete (count) and binary longitudinal data. If a longitudinal data set has all individuals have the same number of observations, taken at time points which are also the same for all individuals, this data set is called balanced longitudinal data. Otherwise it is called unbalanced longitudinal data. In this thesis we concentrate on balanced longitudinal data with continuous outcome variable, and will

assume underlying normality whenever a distributional assumption is required.

1.2 Objective of the Thesis

Our basic model for longitudinal data therefore is

$$y_{it} = x_{it}^T \boldsymbol{\beta} + \varepsilon_{it} \tag{1.1}$$

where y_{it} is the observed value of the outcome variable for individual i at time t. A set of K explanatory variables x_{it} are also observed at the same time, and these are associated with a set of K coefficients $\boldsymbol{\beta}$, which are to be estimated. For each observation we have a disturbance/error term ε_{it}. These error terms have mean zero, are independent **between** individuals, but may be correlated **within** individuals. If it is correlated **within** individuals, one way to define the error term is

$$\varepsilon_{it} = \alpha_i + u_{it} \quad i = 1, \cdots, N \quad t = 1, \cdots, T \tag{1.2}$$

where α_i is called individual effect and random variable u_{it} is independent and identically distributed (iid) with mean zero and variance σ_u^2. In many data sets of interest, individuals are unlike one another; that is, they are heterogeneous. The α_i is used to model the individual behaviour, so if it is within the model, the model is called heterogeneous otherwise it is called homogeneous model.

A fundamental aim of longitudinal analysis is to characterise **within** and **between** individual variation. Variation within individuals is accounted for by any **time dependent** covariates x_{it} and the error term ϵ_{it}. Variation between individuals is accounted for by the full set of covariates x_{it} and also by individual effects which may modelled as fixed effect or random effect. There are two common assumptions made about the individual effect α_i, fixed effect assumption and random effect assumption. If the individual effect α_i are correlated with the explanatory variables or are treated as fixed, we should use fixed effects model. If the individual effect α_i are treated as draws from an unknown population or as a random effect, we should use random effects model. But if we use the random effect estimation to the case that the explanatory variable X is correlated with the individual effect α, the estimator is biased in estimating $\boldsymbol{\beta}$. This is called heterogeneity bias. Figure 1.1 and 1.2 show these two cases graphically as examples: with and without correlation between

explanatory variable and individual effects, respectively. Figure 1.1 shows the intercept of individuals are step increasing and Figure 1.2 shows the intercepts for individuals are independent although they start at same place.

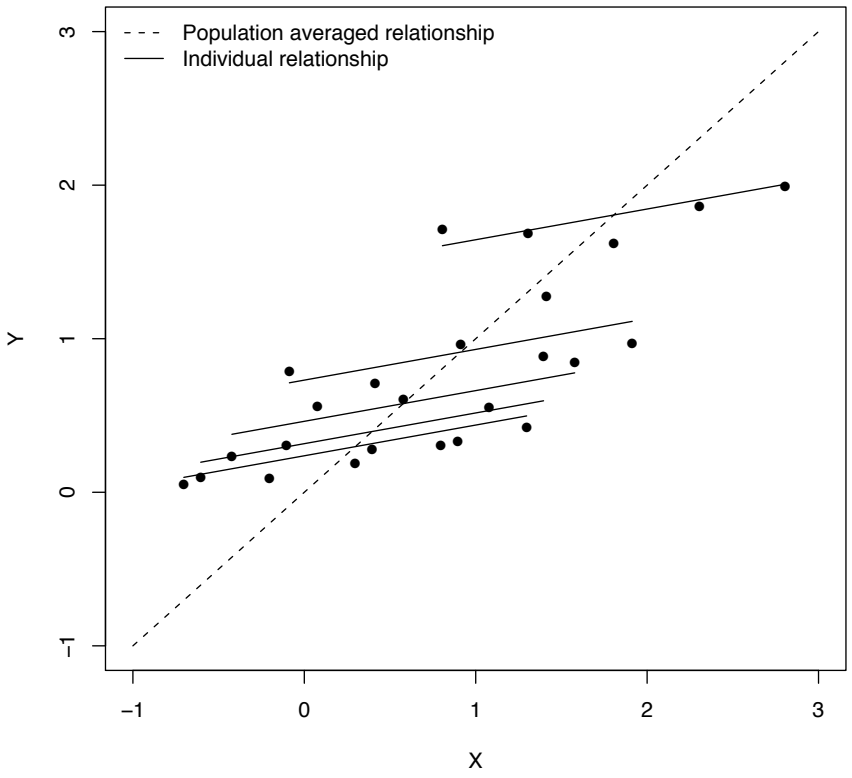

Figure 1.1: Figures of Correlated x and individual effect α

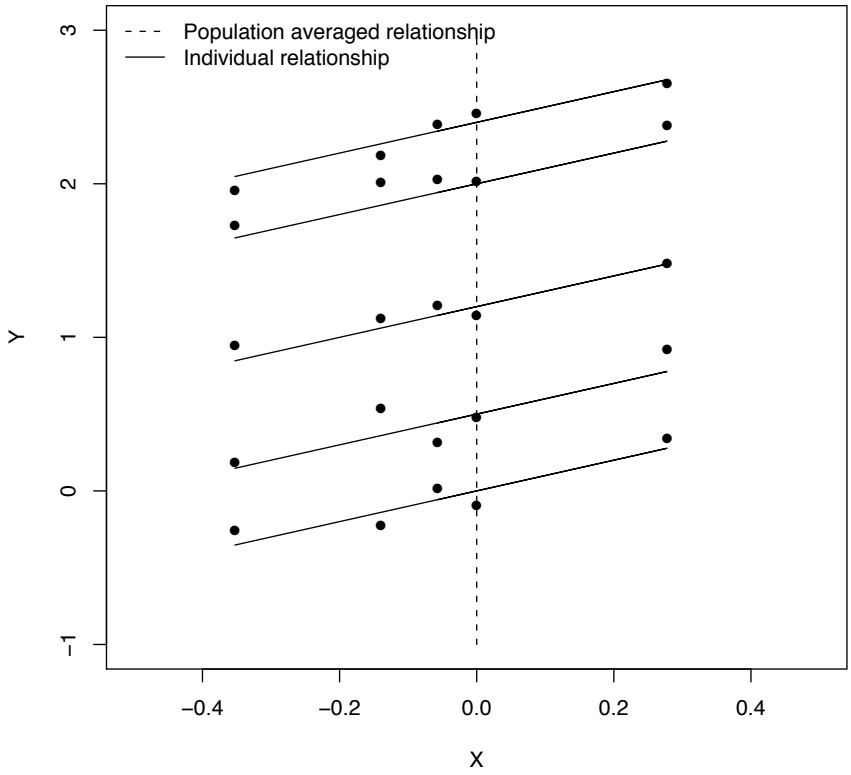

Figure 1.2: Figures of uncorrelated x and individual effect α with same start point for each individual

Hsiao [2003] gave more elaboration on when the heterogeneity bias exist. Rosenzweig and Schultz [1983] showed the problem of heterogeneity bias in a study of child health production and the demand for child health inputs. Also several researchers have indicated the gender wage and earning differences have attempted to heterogeneity bias using as PSID data [Johnson and DiNardo, 2007]. However, these studies leave some important question unanswered. Most of important findings from these studies are identify the heterogeneity bias, then use instrumental variable estimator (see Chapter 5 for detail) to obtain unbiased estimates. Mundlak [1978] produced a formulation for the model with correlation between the individual effects and the explanatory variables to theoretically prove there is biased. In our thesis, we empirically prove the bias exist and then look into the bias, eg. how the degree of the correlation affects the bias and at what stage the correlation exits but is insufficient to produce the bias, etc.

There are wide variety of analytical methods developed by econometricians and statisticians for modelling the behaviour of longitudinal data. Approaches to random effect and fixed effect estimation with longitudinal data fall broadly into three categories:

- Least squares based methods – including ordinary and weighted least squares;
- Likelihood based methods – including maximum likelihood (ML) and restricted maximum likelihood (REML);
- Bayesian hierarchical approaches.

Social scientists prefer the random effect estimation based on likelihood method (refer to Diggle et al. [2002]), the reason for that might be health and social scientists have designed experiments with randomisation that may break down the correlation between the explanatory variable and individual effect (or it may be because social scientists unknown this is a problem).

Econometricians prefer to use the least squares based methods to obtain the fixed effect estimation (Johnson and DiNardo [2007], Wooldridge [2009] and Verbeek [2004] discussed this case a lot). That might be because econometricians are more concerned with causation than association: may be more concerned with correct quantitative values for $\boldsymbol{\beta}$. Thus, when the explanatory variable X is correlated with the individual effect α, the fixed effect estimation is preferred, because the random effect estimation has bias in estimating $\boldsymbol{\beta}$.

In this thesis we:

- Demonstrate the occurrence of bias in two cases:
  - Omitted variables bias case
  - Heterogeneity bias case.
- We show theoretically and empirically (via simulation) how and when these bias occur in each case.
- We discuss the standard Hausman test [**?**] for deciding if a Random Effects or Fixed Effects estimation is more appropriate.
- We investigate Bayesian alternative formulations of this problem to determine whether such solutions also suffer from the same bias.
- We investigate the model selection criterion under frequentist and Bayesian approaches to see whether they produce the same conclusion.

1.3 Overview of the Thesis

This thesis is organized as follows:

- In Chapter 2, we describe the two commonly used analysis methods, least squares based method, likelihood based method and their extensions.
- In Chapter 3, we describe three simple types of longitudinal data models and the Hausman test which can be used to compare the fixed effect estimation and random effect estimation. Also, we prove that the compound symmetry model and random intercept model with iid error term are equivalent models.
- In Chapter 4, we program the different longitudinal datasets generator in R, ie. random effect data with iid or AR(1) or compound symmetry structure, fixed effect data and pooled data (cross sectional data) with different variance-covariance structures, and also program the data generator for the correlated explanatory variable and individual effect case.
- In Chapter 5, we use a simple case to see the bias in omitted variables model and provide a theoretical and empirical investigation via simulation. Then we use the same strategy to investigate the heterogeneity bias when there is correlation between explanatory variable and individual effect. Finally, we use the Hausman test we describe in Chapter 3 to compare the fixed effect estimator and random effect estimator under two cases, with correlation case and without correlation case. And we introduce instrumental variable estimator as an alternative way to obtain the unbiased estimator when correlation exits.
- In Chapter 6, we empirically investigate whether random effect estimator has the same bias under correlation assumption by using Bayesian approach. Then we describe a full Bayesian formulation of the Hausman test under Bayesian approach to test whether the random effect estimator or the fixed effect estimator is more appropriate under correlation case.
- In Chapter 7, we describe two model selection criteria, AIC based on likelihood approach and DIC based on Bayesian approach. Then we compare the results obtained from these two methods with the result from the Hausman test and the full Bayesian formulation by using the real data and simulated data.

- In Chapter 8, we conclude with the discussion, and present directions of future research of our study.

Throughout the thesis, we will illustrate the ideas using examples based on real data and simulated data. Statistical programming software R and WinBugs are used to model and analyse data (the functions are built in or programmed by using R).

1.4 Notation

Before we analyse the longitudinal data, let's define some notation:

$$y_{it}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T,$$

$$x_{it}^{k}, \quad i = 1, \ldots, N, \quad t = 1, \ldots, T, \quad k = 1, \ldots, K$$

where i is the observation dimension, t is the time dimension and k is explanatory variable dimension; since we restrict the study on the balanced panels, thus the total number of observations is NT. y is the value of the dependent variable and x is value of the explanatory variable.

$$\mathbf{y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix}, \quad X_i = \begin{bmatrix} x_{i1}^1 & x_{i1}^2 & \ldots & x_{i1}^K \\ x_{i2}^1 & x_{i2}^2 & \ldots & x_{i2}^K \\ \vdots & \vdots & \ddots & \vdots \\ x_{iT}^1 & x_{iT}^2 & \ldots & x_{iT}^K \end{bmatrix} \tag{1.3}$$

Often the longitudinal data are in vector as

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_N \end{bmatrix}, \quad X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_N \end{bmatrix} \tag{1.4}$$

where y is $NT \times 1$ and X is $NT \times K$.

A general longitudinal data standard linear model is written as

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{1.5}$$

Where

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix}$$

is the coefficient matrix for explanatory variables. And the error term can be written as

$$\boldsymbol{\varepsilon}_i = \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{iT} \end{bmatrix} \text{ and } \boldsymbol{\varepsilon} = \begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_N \end{bmatrix} \tag{1.6}$$

where $\boldsymbol{\varepsilon}$ is $NT \times 1$ vector.

Chapter 2

Estimating Methods for General Linear Models

In this chapter, we introduce the standard estimation methods for longitudinal data, ie. the ordinary least squares method, the generalized least squares method and the maximum likelihood method. These methods are commonly used to estimate the parameters in the general linear model. We use them throughout the thesis.

2.1 Least Squares

One of the powerful estimating method in statistics is least squares which assumes that the best fit model that has the minimal sum of the deviations squared from a given set of data. The least squares approach is based on the criterion which makes as small as possible the sum squared errors between the data and fitted model. And this method is widely used by scientists and mathematician in early times.

2.1.1 Ordinary Least Squares

In statistics and econometrics, ordinary least squares (OLS) is a method based on least square theory which can be applied on a linear regression model. The OLS method minimizes the sum of squared differences between the observed value and prediction value obtained from the linear approximation. In order to use the OLS method to obtain a meaningful result, there are a few assumptions to be made on the linear regression model. Assume a simple linear regression model (in this thesis means standard linear regression) with independent

error term is

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.1}$$

then the OLS assumptions based on this model which might be made are (from [Hayashi, 2000])

- OLS-A1: Strict exogeneity.

$$E[\boldsymbol{\varepsilon}|X] = 0$$

- OLS-A2: No multicollinearity. The regressors in X must all be linearly independent.
- OLS-A3: Spherical errors, ie. homoscedastic and uncorrelated error $\boldsymbol{\varepsilon}$

$$\text{Var}(\boldsymbol{\varepsilon}|X) = \sigma^2 I$$

- OLS-A4: Normality

$$\boldsymbol{\varepsilon} \sim \text{N}(0, \sigma^2 I)$$

All but the OLS-A4 assumption are necessary assumptions for the OLS method. But under this assumption that the errors are normally distributed, OLS can be derived as a maximum likelihood estimator (see proof in section 2.3). The OLS estimator is valid when the regressors are exogenous and there is no multicollinearity, also when the errors are homoscedastic and serially uncorrelated.

Now we can derive the OLS estimate for $\widehat{\boldsymbol{\beta}}$ in Eq.(2.1) and define a vector of residuals $\mathbf{e}$ as

$$\mathbf{e} = \mathbf{y} - X\widehat{\boldsymbol{\beta}}$$

Apply the least squares principle to choose $\widehat{\boldsymbol{\beta}}$ to minimize the residual sum of squares (RSS), $\mathbf{e}^T\mathbf{e}$

$$\begin{aligned} RSS &= (\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta}) \\ &= \mathbf{y}^T\mathbf{y} - \boldsymbol{\beta}^T X^T\mathbf{y} - \mathbf{y}^T X\boldsymbol{\beta} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta} \\ &= \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T X^T\mathbf{y} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta} \end{aligned}$$

The first derivative of RSS gives

$$\frac{\partial RSS}{\partial \boldsymbol{\beta}} = -2X^T\mathbf{y} + 2X^T X\boldsymbol{\beta} = 0$$

So the estimator $\widehat{\boldsymbol{\beta}}$ is

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y} \tag{2.2}$$

The $\widehat{\boldsymbol{\beta}}$ is unbiased, since

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

(see Section 3.1 for the proof).

The variance covariance matrix of $\widehat{\boldsymbol{\beta}}$ is

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2 (X^T X)^{-1} \tag{2.3}$$

and an unbiased estimator of σ^2 is

$$\widehat{\sigma^2} = \frac{\mathbf{e}^T \mathbf{e}}{N - K}$$

where N is the number of the observations and K is the number of variables (see Section 3.1 for the proof).

2.1.2 Generalized Least Squares Estimation

In statistics, generalized least squares (GLS) is applied when the variances of the observations are heteroscedastic (not equal), or when there is a certain degree of correlation between the observations. This means the OLS-A3 assumption is violated. The model is changed to

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.4}$$

where $E[\boldsymbol{\varepsilon}|X] = 0$ and $\mathrm{Var}(\boldsymbol{\varepsilon}|X) = \sigma^2\Omega$ where it is assumed that the variance of Y given X is a known matrix Ω which is positive definite and its inverse is positive definite as well. Thus, we are able to find a non-singular matrix P that has the following equation

$$\Omega^{-1} = P^T P \tag{2.5}$$

Now we let $\tilde{\mathbf{y}} = P\mathbf{y}$, $\tilde{X} = PX$ and $\tilde{\boldsymbol{\varepsilon}} = P\varepsilon$, then we apply OLS estimator to the new equation

$$\tilde{\mathbf{y}} = \tilde{X}\boldsymbol{\beta} + \tilde{\boldsymbol{\varepsilon}} \tag{2.6}$$

This is because if $\tilde{\varepsilon} = P\varepsilon$ for $\varepsilon \sim N(0, \sigma^2\Omega)$ the

$$\begin{aligned}
\text{Var}(\tilde{\varepsilon}) &= \text{Var}(P\varepsilon) \\
&= P\text{Var}(\varepsilon)P^T \\
&= P\sigma^2\Omega P^T \\
&= \sigma^2 P(P^T P)^{-1}P^T \\
&= \sigma^2 PP^{-1}(P^T)^{-1}P^T \\
&= \sigma^2 I
\end{aligned}$$

which indicates these are homoscedastic errors. Then we can apply OLS to Eq.(2.6), so that we obtain

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{\mathbf{y}} \\
&= [(PX)^T(PX)]^{-1}(PX)^T(P\mathbf{y}) \\
&= (X^T P^T PX)^{-1}X^T P^T P\mathbf{y}
\end{aligned}$$

Then the generalized least squares (GLS) estimator by using Eq.(2.5), we have

$$\widehat{\boldsymbol{\beta}}_{GLS} = (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}\mathbf{y} \tag{2.7}$$

$$\begin{aligned}
\text{Var}(\widehat{\boldsymbol{\beta}}_{GLS}) &= \sigma^2(\tilde{X}^T\tilde{X})^{-1} \\
&= \sigma^2(X^T\Omega^{-1}X)^{-1}
\end{aligned} \tag{2.8}$$

The unbiased estimator of the unknown σ^2 in Eq.(2.8) is

$$\begin{aligned}
\widehat{\sigma^2} &= \frac{(\tilde{\mathbf{y}} - \tilde{X}\widehat{\boldsymbol{\beta}}_{GLS})^T(\tilde{\mathbf{y}} - \tilde{X}\widehat{\boldsymbol{\beta}}_{GLS})}{N-K} \\
&= \frac{(\mathbf{y} - X\widehat{\boldsymbol{\beta}}_{GLS})^T\Omega^{-1}(\mathbf{y} - X\widehat{\boldsymbol{\beta}}_{GLS})}{N-K}
\end{aligned} \tag{2.9}$$

So GLS estimation is equivalent to OLS estimation of the transformed data by using a non-singular matrix (Johnson and DiNardo [2007] give more details of the proof). There are number of choices of disturbances or error term structure which are listed below.

Covariance structure

Suppose we have T observations for a given individual i:

$$\mathbf{y}_i = X_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$$

The vector of disturbances or error term

$$\boldsymbol{\varepsilon}_i = [\varepsilon_{i1}, \cdots, \varepsilon_{iT}]^T$$

has a $T \times T$ (symmetric, positive definite) variance-covariance matrix as

$$\text{Var}[\boldsymbol{\varepsilon}_i] = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdot & \sigma_{1T} \\ \sigma_{12} & \sigma_2^2 & \cdot & \sigma_{2T} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1T} & \sigma_{2T} & \cdot & \sigma_T^2 \end{bmatrix} \tag{2.10}$$

Then there are $\frac{1}{2}T(T+1)$ possible covariance parameters with the variance-covariance matrix, since this matrix is symmetric. Also a variance-covariance matrix can have following possible structures:

- **Independence** has a single parameter σ^2:

$$\text{Var}[\boldsymbol{\varepsilon}_i] = \sigma^2 \begin{bmatrix} 1 & 0 & \cdot & 0 \\ 0 & 1 & \cdot & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdot & 1 \end{bmatrix} \tag{2.11}$$

- **Compound Symmetry** has a fixed correlation between all observations, regardless of lag: (two parameters σ^2, ρ):

$$\text{Var}[\boldsymbol{\varepsilon}_i] = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & \cdot & \rho \\ \rho & 1 & \rho & \cdot & \rho \\ \vdots & \vdots & \ddots & \vdots & \\ \rho & \rho & \rho & \cdot & 1 \end{bmatrix} \tag{2.12}$$

- **First-Order Autoregressive AR(1)** has two parameters σ^2, ρ. We define the error terms

as

$$\varepsilon_{it} = \rho\varepsilon_{i,t-1} + u_{it} \tag{2.13}$$

where u_{it} has following distribution

$$u_{it} \sim \text{iid}(0, \sigma^2)$$

Note: iid means independent and identically distributed.

$$\text{Var}\left[\boldsymbol{\varepsilon}_i\right] = \sigma^2 \begin{bmatrix} 1 & \rho & \rho^2 & \cdot & \rho^{T-1} \\ \rho & 1 & \rho & \cdot & \rho^{T-2} \\ \vdots & \vdots & \ddots & \vdots & \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \cdot & 1 \end{bmatrix} \tag{2.14}$$

There are many other possibilities, eg, unstructured covariance, power law covariance structure, exponential covariance structure, etc. In this thesis, we only consider the independence, AR(1) and compound symmetry covariance structure. The covariance structures are introduced here follow by Arnold and Liu [2004].

2.1.3 Feasible Generalized Least Squares Estimation

GLS assumes the Ω matrix is known. However, in practice, the true variance covariance is not known directly. Then the feasible generalized least squares (FGLS) estimator is introduced, where unknown parameters in Eq.(2.4) $\sigma^2\Omega$ can be replaced by unbiased estimates V which can be obtained by using the OLS estimator [Johnson and DiNardo, 2007]. Recall Eq.(2.2) for homoscedastic errors

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$$

and we can obtain the variance covariance matrix V by using OLS method to calculate the residuals for each individual. Then we use the estimates of the residuals as the diagonal elements of matrix to construct the variance covariance matrix V. Now we replace the unknown $\sigma^2\Omega$ in Eq.(2.7). The estimator $\widehat{\boldsymbol{\beta}}_{FGLS}$ can be written as

$$\widehat{\boldsymbol{\beta}}_{FGLS} = (X^T V^{-1} X)^{-1} X^T V^{-1} \mathbf{y}$$

The variance of $\widehat{\beta}_{FGLS}$ is

$$\widehat{\text{Var}}(\widehat{\boldsymbol{\beta}}_{FGLS}) = \sigma^2(X^T V^{-1} X)^{-1}$$

This method is called the FGLS method, the estimator obtain by this method is called the FGLS estimator (some context written as FGLSE), $\widehat{\boldsymbol{\beta}}_{FGLS}$.

2.2 Weighted Least Squares Estimation

Recall the OLS-A3 assumption (homoscedastic errors), if this assumption is unsatisfied, the OLS estimation method can't be used. In such cases, we can use the GLS estimation instead of OLS estimation as in section 2.1.2. In this section, we suppose the variances of the observed values are unequal (heteroscedastic) but all the off-diagonal entries are 0, so that there are no correlations exist among the observed values. Then the weighted least squares (WLS) estimate is introduced to solve this problem. This assumption can be written as

$$\text{Var}(\boldsymbol{\varepsilon}) = \sigma^2 V$$

Then the distribution of **y** is

$$Y \sim \text{N}(X\beta, \sigma^2 V)$$

The weighted least squares method use a symmetric weight matrix W, then apply the least squares principle to choose a $\widehat{\beta}_W$ to minimise

$$\text{RSS}_W = (y - X\beta)^T W (y - X\beta)$$

Follow the same procedure as in section 2.1.1 , $\widehat{\beta}_W$ can be expressed as

$$\widehat{\beta}_W = (X^T W X)^{-1} X^T W y \tag{2.15}$$

then the variance of $\widehat{\beta}_W$ is

$$\text{Var}(\widehat{\beta}_W) = \sigma^2 (X^T W X)^{-1} \tag{2.16}$$

Since $E(Y) = X\boldsymbol{\beta}$, the weighted least squares estimator is unbiased, whatever the choice of W. Diggle et al. [2002] gave two example of choice of W:

- If $W = I$, the identity matrix, the WLS estimator is identical to the OLS estimator

$$\widehat{\beta}_I = (X^T X)^{-1} X^T y$$

and

$$Var(\widehat{\beta}_I) = \sigma^2 (X^T X)^{-1}.$$

- If $W = V^{-1}$, the estimator becomes

$$\widehat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$$

with

$$Var(\hat{\beta}) = \sigma^2 (X^T V^{-1} X)^{-1}$$

Note that for empirical data, the appropriate **W** may be unknown and must be estimated by using Feasible Generalized Least Squares (FGLS) estimation. The weighted least squares estimation is unlike least squares, it gives each term a weight, so that takes the influence of each observation into account.

2.3 Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) is another popular statistical method which is used for fitting a statistical model to data and providing estimates for the parameters in the model.

Definition 2.1. *Let $y = [Y_1, ..., Y_N]^T$ denote a random vector and let the joint probability density function of the Y_i's be*

$$f(y; \theta)$$

*which depends on the vector of parameters $\theta = [\theta, \cdots, \theta_k, \cdots, \theta_K]^T$. The **likelihood function** $L(\theta; y)$ is algebraically the same as the joint probability density function $f(y; \theta)$ but the change in notation reflects a shift of emphasis from the random variables y, with θ fixed, to the parameters θ with y fixed. Since L is defined in terms of the random vector y, it is itself a random variable. The **maximum likelihood estimator** of θ is the value $\hat{\theta}$*

which maximizes the likelihood function, that is

$$L(\hat{\theta}) \geq L(\theta; y) \text{ for all } \theta$$

[Dobson, 2002]

Since $f(y;\theta) > 0$ and the logarithm is a monotonic function on the positive real line, $\hat{\theta}$ is also the value which maximizes the log-likelihood function

$$\ell(\theta; y) = \log L(\theta; y)$$

The estimator $\widehat{\theta}$ can be obtained by differentiating the log-likelihood function with respect to each element θ_k of θ and solving the simultaneous equations

$$\frac{\partial \ell(\theta; y)}{\partial \theta_k} = 0 \text{ for } k = 1, ..., K$$

To check the solutions do return the maximum $\ell(\theta; y)$, we have to verify the second derivative with respect to each element θ_k of θ is negative.

$$\left[\frac{\partial^2 \ell(\theta; y)}{\partial \theta_k^2}\right]_{\theta_k = \hat{\theta}_k} < 0 \text{ for } k = 1, ..., K$$

Here the $\widehat{\theta}$ that maximizes $\ell(\theta; y)$ will also maximize $L(\theta; y)$. The derivative of $\ell(\theta; y)$ is known as the score, $s(\theta; y)$.

Property 2.1. *The major properties of MLEs are large-sample, or asymptotic, ones. They hold under fairly general conditions.*

1 *Consistency*

$$\lim_{N \to \infty} (\hat{\theta}) = \theta$$

2 *Asymptotic normality*

$$\hat{\theta} \overset{a}{\sim} N(\theta, I^{-1}(\theta))$$

This states that the asymptotic distribution of $\hat{\theta}$ is normal with mean θ and variance

given by the inverse of I(θ). I(θ) is the information matrix and is defined as

$$I(\theta) = E\left[\left(\frac{\partial \ell}{\partial \theta}\right)\left(\frac{\partial \ell}{\partial \theta}\right)^T\right]$$

3 *Asymptotic efficiency*

If $\hat{\theta}$ is the MLE of a single parameter θ, the previous property means that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \sigma^2)$$

for some finite constant σ^2. The MLE has minimum variance in the class of consistent, asymptotically normal estimators.

4 *Invariance.*

If $\hat{\theta}$ is the MLE of θ and g(θ) is a continuous function of θ, then g($\hat{\theta}$) is the MLE of g(θ).

[Johnson and DiNardo, 2007]

2.3.1 MLE of simple regression model

By using the definition and properties of MLE, a simple regression model

$$y = X\beta + \varepsilon \tag{2.17}$$

with

$$\varepsilon \sim N(0, \sigma^2 \mathrm{I})$$

can be written in the normal density form for ε if the sample size is N is

$$f(\varepsilon; \beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} e^{-(\frac{1}{2\sigma^2})(\varepsilon^T \varepsilon)}$$

The log-likelihood function is

$$\begin{aligned} \ell(\beta, \sigma^2; y, X) &= -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\varepsilon^T\varepsilon \\ &= -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) \end{aligned} \tag{2.18}$$

In order to get the estimators $\widehat{\beta}$ and $\widehat{\sigma^2}$, we take the partial derivatives of Eq.(2.18)

$$\frac{\partial \ell}{\partial \beta} = -\frac{1}{\sigma^2}(-X^T y + X^T X\beta) \tag{2.19}$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}(y - X\beta)^T(y - X\beta) \tag{2.20}$$

Then we set Eq.(2.19) and (2.20) are equal to zero, the MLEs are

$$\widehat{\beta} = (X^T X)^{-1} X^T y \tag{2.21}$$

and

$$\widehat{\sigma^2} = \frac{1}{n}(y - X\beta)^T(y - X\beta) = \frac{e^T e}{N} \tag{2.22}$$

The MLE $\widehat{\beta}$ is the OLS estimator and $\widehat{\sigma^2}$ is $\frac{e^T e}{N}$. The LS theory gives

$$E(\frac{e^T e}{N - K}) = \sigma^2 \tag{2.23}$$

where N is the number of observation and K is the number of variables in the model (see Chapter 3 for details of proof). Thus

$$E(\widehat{\sigma^2}) = \frac{\sigma^2(N - K)}{N} \tag{2.24}$$

So that $\widehat{\sigma^2}$ is biased for σ^2, but $\widehat{\beta}$ is unbiased for β.

The MLE $\widehat{\beta}$ is identical to OLS estimator from Eq. (2.2). From the properties of the MLE, we can infer that the OLS estimator is consistent (unbiased) and is asymptotically efficient (with minimum variance) if the normality assumption is satisfied.

Now we can extend the above knowledge into the linear model with non-spherical (heteroscedastic) disturbances.

$$\mathbf{y} = X\boldsymbol{\beta} + \varepsilon \text{ with } \varepsilon \sim N(0, \sigma^2\Omega)$$

where Ω is a positive define matrix of order N. The normal density for ε is

$$f(\varepsilon) = (2\pi)^{-\frac{N}{2}} |\sigma^2\Omega|^{-\frac{1}{2}} e^{-\frac{1}{2\sigma^2}\varepsilon^T(\Omega)^{-1}\varepsilon}$$

We can rewrite this as

$$f(\varepsilon) = (2\pi)^{-\frac{N}{2}}\sigma^{-N}|\Omega|^{-\frac{1}{2}}e^{-\frac{1}{2\sigma^2}\varepsilon^T(\Omega)^{-1}\varepsilon}$$

since $|\sigma^2\Omega|$=$\sigma^{2N}|\Omega|$.

The log-likelihood function is

$$\ell = -\frac{N}{2}\log 2\pi - \frac{N}{2}\log\,\sigma^2 - \frac{1}{2}\log\,|\Omega| - \frac{1}{2\sigma^2}(y - X\beta)^T\Omega^{-1}(y - X\beta) \tag{2.25}$$

By differentiating with respect to β and σ^2, we have

$$\frac{\partial \ell}{\partial \beta} = -\frac{1}{\sigma^2}(-X^T\Omega^{-1}y + X^T\Omega^{-1}X\beta)$$

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)^T\Omega^{-1}(y - X\beta)$$

Then the ML estimators for non-spherical disturbances model are

$$\widehat{\beta} = (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}y \tag{2.26}$$

$$\widehat{\sigma^2} = \frac{1}{N}(y - X\widehat{\beta})^T\Omega^{-1}(y - X\widehat{\beta}) \tag{2.27}$$

The MLE $\widehat{\beta}$ for non-spherical disturbances model is the same as the GLS estimator $\widehat{\beta}_{\text{GLS}}$. There are more complex models, Frees [2004] described this by assuming $\mathbf{y} \sim N(X\boldsymbol{\beta}, \nu)$ where $\nu = V(\tau) = V_i$.

The log likelihood of a single individual is

$$\ell_i(\beta, \tau) = -\frac{1}{2}\left(T_i\log(2\pi) + \log|V_i(\tau)| + (\mathbf{y}_i - X_i\beta)^T V_i(\tau)^{-1}(\mathbf{y}_i - X_i\beta)\right) \tag{2.28}$$

Then the full data log likelihood is

$$L(\beta, \tau) = \sum_{i=1}^{N}\ell_i(\beta, \tau)$$

Then the estimators of β and τ can be obtained by maximize the $L(\beta, \tau)$, take the first derivatives with respect to β and τ, and set the equations equal to zero. We

have

$$\begin{aligned}\frac{\partial}{\partial\beta}L(\beta,\tau) &= \sum_{i=1}^{N}\frac{\partial}{\partial\beta}\ell_i(\beta,\tau)\\ &= -\frac{1}{2}\sum_{i=1}^{N}\frac{\partial}{\partial\beta}(\mathbf{y}_i - X_i\beta)^T V_i(\tau)^{-1}(\mathbf{y}_i - X_i\beta)\\ &= \sum_{i=1}^{N}X_i^T V_i(\tau)^{-1}(\mathbf{y}_i - X_i\beta).\end{aligned}$$

Therefore, the estimator of β is

$$\beta_{MLE} = \left(\sum_{i=1}^{N}X_i^T V_i(\tau)^{-1}X_i\right)\sum_{i=1}^{N}X_i^T V_i(\tau)^{-1}y_i$$

Hence, the fixed covariance τ of MLE is the same as the GLS estimation [Frees, 2004].

2.4 Restricted Maximum Likelihood Estimation

In statistics, the restricted maximum likelihood (REML) approach is a way of estimating variance components (see Patterson and Thompson [1971] for definition) which was introduced by Patterson and Thompson [1971]. In contrast to the maximum likelihood estimation introduced earlier, REML can produce unbiased estimates of variance and covariance parameters.

Recall the general linear model Eq.(2.17) with independent errors

$$y = X\beta + \varepsilon$$

with

$$\varepsilon \sim N(0, \sigma^2\mathbf{I})$$

Then the distribution of Y is

$$Y \sim N(X\boldsymbol{\beta}, \sigma^2 I). \tag{2.29}$$

In this case, Y is an N dimensional vector with known covariate values, and X is $N \times K$ matrix, where K is the number of variables of β . It also assumes that all observations are

independent. So the maximum likelihood estimator for σ^2 is Eq.(2.22)

$$\widehat{\sigma^2} = \frac{1}{N}(y - X\beta)^T(y - X\beta) \tag{2.30}$$

which is biased downward by $\frac{N-K}{N}$. See Eq. (2.24) for detail.

Now we introduce REML estimator which is defined by Diggle et al. [2002] as a maximum likelihood estimator based on a linearly transformed set of data $Y^* = M^T Y$ where M is $N \times (N - K)$ matrix. So the distribution of Y^* does not depend on $\boldsymbol{\beta}$, then it follows a normal distribution with mean zero and variance covariance matrix $\sigma^2 M^T M$. Now using the MLE method on the transformed data, the estimated σ^2 is

$$\widehat{\sigma^2} = \frac{1}{N-K}(y - X\beta)^T(y - X\beta) \tag{2.31}$$

which is unbiased for σ^2.

Note: one way to achieve this is by taking M to be a subset of $N-P$ linearly independent columns from the matrix P which converts Y to OLS residuals,

$$M = I - X(X^T X)^{-1} X^T. \tag{2.32}$$

Therefore,

$$\widehat{\sigma^2} = \frac{(\mathbf{Y} - X(X^T X)^{-1} X^T \mathbf{Y})^T (\mathbf{Y} - X(X^T X)^{-1} X^T \mathbf{Y})}{N - K}$$

which is the mean squared error, unbiased for σ^2. It used as the estimator for the residual variance in linear regression. There are more details of proof from [Diggle et al., 2002].

In summary, maximum likelihood and REML estimators will often give very similar estimates for β. But when K is relatively large, the result of ML and REML may differ. However, when they do differ substantially, REML estimators should be unbiased or less biased [Diggle et al., 2002]. It has been recommended by some authors (eg. [Arnold and Liu, 2004]), that the REML can be used to select an appropriate variance covariance structure; ML then be used to select significant terms in a longitudinal model by using AIC (see definition of AIC in section 7.2); and then finally model estimates be estimated using REML estimation in the selected model.

Remark: In this thesis, we consider model where $n >> p$, so the distinction between ML and REML estimators is not important. We therefore do not consider REML estimation.

Chapter 3

Models for Longitudinal Data

In this chapter, we apply the methods from the general linear model presented in chapter 2 to the case of longitudinal data. We now introduce a general longitudinal data standard linear model that can be written in three different forms:

- Full data form:

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{3.1}$$

  where the dimensions of parameters are $\mathbf{y}$ is $NT \times 1$, X is $NT \times K$, $\boldsymbol{\beta}$ is $K \times 1$ and $\boldsymbol{\varepsilon}$ is NT $\times$ 1.

- Vector or individual form:

$$\mathbf{y}_i = X_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i \quad i = 1, \cdots, N \tag{3.2}$$

  where the dimensions of parameters are $\mathbf{y}_i$ is T $\times$ 1, X_i is T $\times$ K, $\boldsymbol{\beta}$ is $K \times 1$ and $\boldsymbol{\varepsilon}_i$ is T $\times$ 1.

- Scalar form:

$$y_{it} = \mathbf{x}_{it}^T\boldsymbol{\beta} + \varepsilon_{it} \quad i = 1, \cdots, N \quad t = 1, \cdots, T \tag{3.3}$$

  where the dimensions of parameters are y_{it} is 1×1, $\mathbf{x}_{it}^T$ is $1 \times K$ and is known, $\boldsymbol{\beta}$ is K $\times$ 1 and ε_{it} is 1×1.

where i is the individual specification and t is the observation specification. There are different assumptions can be made on the error term structure of this general model.

If we assume that $\boldsymbol{\varepsilon} \sim$ iid $(0, \sigma^2\mathrm{I})$, that means we ignore the correlation structure of the data for a given individual, observations within individual are assumed to be uncorrelated;

and across individuals and time, the errors are homoscedastic. Then the simplest model has been defined.

If we don't ignore the relationship between observations for a given individual and time, then one way to define the error term is

$$\varepsilon_{it} = \alpha_i + u_{it} \quad i = 1, \cdots, N \quad t = 1, \cdots, T \tag{3.4}$$

The first term of the right side of Eq.(3.4) is called an individual unobserved effect. It varies across individuals, but is constant across time; this part may or may not be correlated with the explanatory variables X_i. These two assumptions can be made about the individual unobserved effect correspond to two different models:

- Random effects model: $\alpha_i \sim$ iid $(0, \sigma_\alpha^2)$ and are uncorrelated with X_i
- Fixed effects model: α_i are constant over time and may be correlated with X_i.

The second term of the right side of Eq.(3.4), u_{it} is assumed to be a random disturbance, that is uncorrelated with X_{it} (although the u_{it} may be correlated amongst themselves for a particular individual). It varies independently across individuals and may vary across time. Note: In this chapter, we state three models which are following Johnson and DiNardo [2007].

3.1 Pooled Model

Longitudinal data can be modelled as a pooled model. The pooled model stacks data over individuals and time. The estimator derived based on this model is called the pooled estimator.

The pooled estimator can be derived in two ways based on the model

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon}. \tag{3.5}$$

3.1.1 Ordinary Least Squares (OLS) Estimation for pooled model

The least squares principle is used to find an estimate $\widehat{\boldsymbol{\beta}}$ to minimize the residual sum of squares (RSS), $\mathbf{e}^T\mathbf{e}$ where $\mathbf{e} = \mathbf{y} - X\widehat{\boldsymbol{\beta}}$

$$\begin{aligned} RSS &= (\mathbf{y} - X\widehat{\boldsymbol{\beta}})^T(\mathbf{y} - X\widehat{\boldsymbol{\beta}}) \\ &= \mathbf{y}^T\mathbf{y} - \widehat{\boldsymbol{\beta}}^T X^T\mathbf{y} - \mathbf{y}^T X\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\beta}}^T X^T X\widehat{\boldsymbol{\beta}} \\ &= \mathbf{y}^T\mathbf{y} - 2\widehat{\boldsymbol{\beta}}^T X^T\mathbf{y} + \widehat{\boldsymbol{\beta}}^T X^T X\widehat{\boldsymbol{\beta}} \end{aligned}$$

The first derivative of RSS gives

$$\frac{\partial RSS}{\partial \boldsymbol{\beta}} = -2X^T\mathbf{y} + 2X^T X\boldsymbol{\beta} = 0$$

Estimation of $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1}X^T\mathbf{y} \tag{3.6}$$

substituting for $\mathbf{y}$ gives

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1}X^T(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{\beta} + (X^T X)^{-1}X^T\boldsymbol{\epsilon}$$

from which

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (X^T X)^{-1}X^T\boldsymbol{\epsilon}$$

Taking the expectation,

$$\mathrm{E}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (X^T X)^{-1}X^T\mathrm{E}(\boldsymbol{\epsilon}) = 0$$

giving

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta} \tag{3.7}$$

Thus, under the assumptions of this model, the LS coefficients are unbiased estimates of the $\boldsymbol{\beta}$ parameters. The variance-covariance matrix of the OLS estimates is established as

$$\begin{aligned}
\text{Var}(\widehat{\boldsymbol{\beta}}) &= \text{E}[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T] \\
&= \text{E}[(X^TX)^{-1}X^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^TX(X^TX)^{-1}] \\
&= (X^TX)^{-1}X^T\text{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T]X(X^TX)^{-1} \\
&= (X^TX)^{-1}X^T\sigma^2IX(X^TX)^{-1} \\
&= \sigma^2(X^TX)^{-1}
\end{aligned}$$

So

$$\text{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(X^TX)^{-1} \tag{3.8}$$

Estimation of σ^2 :

The residuals from the OLS regression can be expressed as

$$\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} = \mathbf{y} - X\widehat{\boldsymbol{\beta}} = \mathbf{y} - X(X^TX)^{-1}X^T\mathbf{y} = M\mathbf{y}$$

The symmetric matrix M is defined as:

$$M = I - X(X^TX)^{-1}X^T$$

M is idempotent matrix, we have $MM = M$ and $MX = 0$.

Proof: To prove M is idempotent matrix, then

$$\begin{aligned}
MM &= (I - X(X^TX)^{-1}X^T)(I - X(X^TX)^{-1}X^T) \\
&= I - X(X^TX)^{-1}X^T - X(X^TX)^{-1}X^T + X(X^TX)^{-1}X^TX(X^TX)^{-1}X^T \\
&= I - X(X^TX)^{-1}X^T - X(X^TX)^{-1}X^T + X(X^TX)^{-1}X^T \\
&= I - X(X^TX)^{-1}X^T \\
&= M
\end{aligned}$$

Also we have

$$\begin{aligned} MX &= (I - X(X^TX)^{-1}X^T)X \\ &= X - X(X^TX)^{-1}X^TX \\ &= X - X = 0 \end{aligned}$$

Then we can write $\mathbf{e} = M\mathbf{y} = M(X\boldsymbol{\beta} + \boldsymbol{\epsilon}) = M\boldsymbol{\epsilon}$. And we utilize the fact that the trace of a scalar is the scalar. Thus,

$$\begin{aligned} \mathrm{E}(\mathbf{e}^T\mathbf{e}) &= E(\boldsymbol{\epsilon}^T M^T M\boldsymbol{\epsilon}) \\ &= \mathrm{E}(\boldsymbol{\epsilon}^T M\boldsymbol{\epsilon}) \\ &= \mathrm{E}[\mathrm{tr}(\boldsymbol{\epsilon}^T M\boldsymbol{\epsilon})] \\ &= \mathrm{E}[\mathrm{tr}(\boldsymbol{\epsilon}^{\boldsymbol{T}}\boldsymbol{\epsilon} M)] \\ &= \mathrm{tr}(\mathrm{E}[(\boldsymbol{\epsilon}^{\boldsymbol{T}}\boldsymbol{\epsilon} M]) \\ &= \mathrm{tr}(\sigma^2\mathrm{E}[M]) \\ &= \sigma^2\mathrm{tr}(M) \\ &= \sigma^2\mathrm{tr}I - \sigma^2\mathrm{tr}[X(X^TX)^{-1}X^T] \\ &= \sigma^2\mathrm{tr}I - \sigma^2\mathrm{tr}[(X^TX)^{-1}(XX^T)] \\ &= \sigma^2(NT - K) \end{aligned}$$

So

$$\sigma^2 = \frac{\mathrm{E}(\mathbf{e}^T\mathbf{e})}{NT - K}$$

Thus, the unbiased estimator of σ^2 is

$$\widehat{\sigma}^2 = \frac{\mathbf{e}^T\mathbf{e}}{NT - K} \tag{3.9}$$

3.1.2 Maximum Likelihood Estimation for Pooled Model

In Chapter 2, we derived the maximum likelihood estimators in Eq. (2.21) and Eq. (2.22) by assuming the $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \sigma^2\mathrm{I})$. Now we apply the same theory on pooled model. We have

$$\widehat{\boldsymbol{\beta}}_{MLE} = (X^TX)^{-1}X^T\mathbf{y}$$

The $\widehat{\beta}_{MLE}$ is the same as the OLS estimator $\widehat{\beta}$. And

$$\widehat{\sigma}^2_{MLE} = \frac{1}{NT}(\mathbf{y} - X\widehat{\boldsymbol{\beta}}_{MLE})^T(\mathbf{y} - X\widehat{\boldsymbol{\beta}}_{MLE}) = \frac{\mathbf{e}^T\mathbf{e}}{NT}$$

where **e**=**y**-$X\widehat{\boldsymbol{\beta}}_{MLE}$ and is the OLS residual vector. The Ordinary Least Squares estimation assumes only $\boldsymbol{\varepsilon} \sim$ iid $(0, \sigma^2\mathrm{I})$ and gives

$$E(\frac{\mathbf{e}^T\mathbf{e}}{NT - K}) = \sigma^2$$

This property is still true for normal errors. Thus,

$$E(\widehat{\sigma^2}_{MLE}) = \frac{\sigma^2(NT - K)}{NT}$$

So that $\widehat{\sigma}^2_{MLE}$ is biased for σ^2, although $\widehat{\beta}_{MLE}$ is unbiased for β. And an unbiased estimator by using MLE method is

$$\widehat{\sigma^2} = \frac{NT}{NT - K}\widehat{\sigma}^2_{MLE}$$

3.2 Fixed Effects (FE) Model

The fixed effects model is defined as

$$y_{it} = \mathbf{x}_{it}^T\boldsymbol{\beta} + \alpha_i + u_{it} \quad i = 1, \cdots, N \quad t = 1, \cdots, T \tag{3.10}$$

where time invariant individual unobserved effect α_i is a fixed parameter which is representing the effects of those variables is constant over time and may be correlated with X_i.

Note: Hsiao [2003] gives an alternative and equivalent formulation of Eq.(3.10) is to introduce a "mean intercept," μ, so that

$$y_{it} = \mu + \mathbf{x}_{it}^T\boldsymbol{\beta} + \alpha_i + u_{it} \quad i = 1, \cdots, N \quad t = 1, \cdots, T \tag{3.11}$$

Because both μ and α_i are fixed constants, without additional restriction, they are not separately identifiable or estimable. One way to identify μ and α_i is to introduce the restriction $\sum_{i=1}^N \alpha_i$. Then the individual effect α_i represents the deviation of the i^{th} individual from the common mean μ. But both formulation lead to the same least-squares estimator [Hsiao, 2003]. In this thesis, we only concentrate on the formulation Eq.(3.10).

In vector form, Eq. (3.10) can be written as

$$\mathbf{y}_i = X_i\boldsymbol{\beta} + \mathbf{1}_T\alpha_i + \mathbf{u}_i \quad i = 1, \cdots, N \tag{3.12}$$

where $\mathbf{y}_i = [y_{i1}, \cdots, y_{iT}]^T$, $X_i = [\mathbf{x}_{i1}, \cdots, \mathbf{x}_{iT}]^T$, $\mathbf{u}_i = [u_{i1}, \cdots, u_{iT}]^T$ and $\mathbf{1}_T$ is a $T \times 1$ dummy vector. Note: By using the FE model, the estimator $\boldsymbol{\beta}$ only can calculate the time variant variables. The time invariant variables are not estimable, thus X_i is a subset of variables which are time varying.

We make two assumptions on the fixed effect model Eq. (3.10)

A1.

$$\mathrm{E}(\mathbf{u}_i|X_i, \alpha_i) = 0$$

A2.

$$\mathrm{Var}(\mathbf{u}_i|X_i, \alpha_i) = \sigma^2\mathbf{I}_T$$

The assumption A1 means the disturbance term $\mathbf{u}_i$ are uncorrelated with explanatory variable X_i over time, that means all the explanatory variables are strictly exogenous.

The assumption A2 is the common homoscedastic assumption, under this assumption, the OLS estimation of model (3.12) is unbiased.

There are three approaches to the fixed effect estimator that we introduce below.

3.2.1 Within Estimation for fixed effect model

In the first approach, we calculate the mean values of the variables in the observations for each given individual. The mean of y_i averaged over time is

$$\begin{aligned}
\overline{y_i} &= \frac{1}{T}\mathbf{1}^T\mathbf{y}_i \\
&= \frac{1}{T}\mathbf{1}^T(X_i^T\boldsymbol{\beta} + \alpha_i + \mathbf{u}_i) \\
&= \overline{X_i}^T\boldsymbol{\beta} + \alpha_i + \overline{\mathbf{u}_i} \quad i = 1, \cdots, N
\end{aligned}$$

the matrices are defined same as Eq.(3.12) and subtracted from Eq.(3.10) for that individual gives

$$y_{it} - \overline{y_i} = (X_{it} - \overline{X_i})^T\boldsymbol{\beta} + (u_{it} - \overline{u_i}) \quad i = 1, \cdots, N$$

Applying the OLS method, we obtain the estimator

$$\widehat{\boldsymbol{\beta}}_W = \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{X}_{it} - \overline{X}_i)(\mathbf{X}_{it} - \overline{X}_i)^T\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{X}_{it} - \overline{X}_i)(y_{it} - \overline{y}_i)\right] \tag{3.13}$$

We also can write the model (3.10) in full data form

$$\mathbf{y} = X\boldsymbol{\beta} + Z\boldsymbol{\alpha} + \mathbf{u} \tag{3.14}$$

where $\mathbf{y} = (\mathbf{y}_1^T, \cdots, \mathbf{y}_N^T)^T$, $\mathbf{u} = (\mathbf{u}_1^T, \cdots, \mathbf{u}_N^T)^T$ are $NT \times 1$ vectors and

$$Z = \begin{bmatrix} \mathbf{1}_{T\times 1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{1}_{T\times 1} \end{bmatrix}_{NT\times T}$$

and $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_N)^T$. Z can be written as

$$Z = I_N \otimes \mathbf{1}_T$$

where I_N is an identity matrix of dimension N, $\mathbf{1}_T$ is a vector of ones of dimension T and $\otimes$ denotes Kronecker product (which is an production on two matrices resulting in a block matrix.). Then we can derive

$$\begin{aligned} ZZ^T &= \begin{bmatrix} \mathbf{1}\mathbf{1}^T & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{1}\mathbf{1}^T \end{bmatrix} \\ &= \begin{bmatrix} J_T & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & J_T \end{bmatrix}_N \\ &= I_N \otimes J_T \end{aligned}$$

$$
\begin{aligned}
(Z^TZ)^{-1} &= \begin{bmatrix} \mathbf{1}_{T\times 1}^T\mathbf{1}_{T\times 1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{1}_{T\times 1}^T\mathbf{1}_{T\times 1} \end{bmatrix}^{-1} \\
&= \frac{1}{T}\begin{bmatrix} 1 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 1 \end{bmatrix}_T \\
&= \frac{1}{T}I_T
\end{aligned}
$$

where $J_T = \mathbf{1}\mathbf{1}^T$. Thus

$$
\begin{aligned}
P &= Z(Z^TZ)^{-1}Z^T \\
&= \frac{1}{T}I_N \otimes J_T \\
&= I_N \otimes \overline{J}_T
\end{aligned}
$$

Let $Q = I_{NT} - P$, the matrix P and Q have the following properties:

- Idempotent: $P^T = P$ and $P^2 = P$; $Q^2 = Q$.

  Proof: P is idempotent matrix, then

$$
\begin{aligned}
P^T &= [Z(Z^TZ)^{-1}Z^T]^T \\
&= (Z^T)^T\left[(Z^TZ)^{-1}\right]^T(Z)^T \\
&= (Z^T)^T(Z^TZ)^{-1}(Z)^T \\
&= P
\end{aligned}
$$

  So $P^T = P$.

$$P^2 = Z(Z^TZ)^{-1}Z^TZ(Z^TZ)^{-1}Z^T = Z(Z^TZ)^{-1}Z^T = P$$

  So $P^2 = P$

$$
\begin{aligned}
QQ &= (I-P)(I-P) \\
&= I - P - P + PP \\
&= I - P \\
&= Q
\end{aligned}
$$

So $QQ = Q$.

- P and Q are orthogonal: $PQ = 0$ and Q and Z are orthogonal: $QZ = 0$

  Proof:

  $$PQ = P(I - P) = P - P^2 = P - P = 0$$

  So $PQ = 0$.

  $$QZ = (I - P)Z = Z - PZ = Z - Z(Z^T Z)^{-1} Z^T Z = Z - Z = 0$$

  So $QZ = 0$

- Additive identity: $P + Q = I_{NT}$.

  Proof: Since we defined $Q = I - P$, thus $P + Q = I_{NT}$.

We now transform $\mathbf{y}$ using Q; by premultiplying Eq.(3.14) by Q

$$\begin{aligned} Q\mathbf{y} &= QX\boldsymbol{\beta} + QZ\boldsymbol{\alpha} + Q\mathbf{u} \\ &= QX\boldsymbol{\beta} + \mathbf{0} + Q\mathbf{u} \end{aligned}$$

This transformation eliminates the time invariant variable α and any time independent variables. Now we rewrite in standard linear model form

$$\widetilde{\mathbf{y}} = \widetilde{X}\boldsymbol{\beta} + \widetilde{\mathbf{u}} \tag{3.15}$$

where

$$\begin{aligned} \widetilde{\mathbf{y}} = Q\mathbf{y} &= (I_{NT} - \frac{1}{T} ZZ^T)\mathbf{y} \\ &= \mathbf{y} - \frac{1}{T} ZZ^T \mathbf{y} \\ &= \mathbf{y} - \overline{\mathbf{y}} \end{aligned}$$

with

$$\overline{\mathbf{y}} = \begin{bmatrix} \overline{y}_1\mathbf{1} \\ \overline{y}_2\mathbf{1} \\ \vdots \\ \overline{y}_N\mathbf{1} \end{bmatrix}$$

and $\widetilde{\mathbf{u}} = Q\mathbf{u}$ and $\widetilde{X} = QX$ are in the same form. These can be interpreted as individual deviations which measuring are the difference between individuals and its individual mean over time.

By using the least squares theory, we can derive the following estimator:

$$\begin{aligned} \widetilde{\boldsymbol{\beta}}_W &= (\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T\widetilde{\mathbf{y}} && (3.16) \\ &= [(QX)^T(QX)]^{-1}(QX)^T(Q\mathbf{y}) \\ &= (X^TQX)^{-1}X^TQ\mathbf{y} && (3.17) \end{aligned}$$

$$\begin{aligned} \mathrm{E}(\widetilde{\boldsymbol{\beta}}_W) &= (X^TQX)^{-1}X^TQX\boldsymbol{\beta} \\ &= \boldsymbol{\beta} && (3.18) \end{aligned}$$

So $\widetilde{\boldsymbol{\beta}}_W$ is an unbiased estimator for $\boldsymbol{\beta}$ often referred to as the the **within** or **fixed effects** estimator, with

$$\widetilde{\boldsymbol{\beta}}_W = \widehat{\boldsymbol{\beta}}_{FE} \tag{3.19}$$

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}_{FE}) = \sigma^2(X^TQX)^{-1} = \sigma^2(\widetilde{X}^T\widetilde{X})^{-1} \tag{3.20}$$

Let

$$\widehat{\mathbf{u}}_W = \widetilde{\mathbf{y}} - \widetilde{X}\widetilde{\boldsymbol{\beta}}_W$$

Then we have the estimator of σ_u^2

$$\widehat{\sigma^2}_W = \frac{\widehat{\mathbf{u}}_W^T\widehat{\mathbf{u}}_W}{NT - K} \tag{3.21}$$

However, since the transformation matrix eliminated the time invariant variable α_i where $i = 1, \cdots, N$, there is a loss of N degrees of freedom. Therefore, we have to adjust the denominator of Eq.(3.21) . The correct calculation should be NT observations minus N means

and K parameters. Thus, the correct unbiased estimator of $\widehat{\sigma^2}_u$ is

$$(\widehat{\sigma^2}_u)_{FE} = \frac{\widehat{\mathbf{u}}_W^T \widehat{\mathbf{u}}_W}{NT - N - K} = \widehat{\sigma^2}_W \frac{NT - K}{NT - N - K} \tag{3.22}$$

The $\boldsymbol{\alpha}$ can be estimated as

$$\widehat{\boldsymbol{\alpha}}_W = \overline{\mathbf{y}} - \overline{X}\widetilde{\boldsymbol{\beta}}_W$$

This is known as the within estimator. The within estimator is only one possible fixed effects estimator.

Note: In fact, model (3.15) does not satisfy the OLS assumption, because after transformation the residual $\tilde{\mathbf{u}} = Q\mathbf{u}$ are no longer uncorrelated

$$\mathrm{E}[(Q\mathbf{u})(Q\mathbf{u})^T] = \sigma^2 Q \neq \sigma^2 I.$$

So we introduce the OLS estimation below.

3.2.2 OLS Estimation for fixed effect model

The alternative way to derive the fixed effect estimator by using OLS method with a partitioned design matrix $W = \begin{bmatrix} X & Z \end{bmatrix}$ and an augmented parameter vector $\boldsymbol{\gamma} = \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix}$ In order to use this method, we rewrite the Eq.(3.14) as

$$\mathbf{y} = \begin{bmatrix} X & Z \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\alpha} \end{bmatrix} + \mathbf{u} = W\boldsymbol{\gamma} + \mathbf{u} \tag{3.23}$$

By using the Least squares theory, the estimator of $\boldsymbol{\gamma}$ is

$$\widehat{\boldsymbol{\gamma}} = \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\alpha}} \end{bmatrix} = (W^T W)^{-1} W^T \mathbf{y} \tag{3.24}$$

Eq.(3.24) will give the same estimate of $\widehat{\boldsymbol{\beta}}$ as Eq.(3.17). To prove this we have to use blockwise matrix inversion method which is given in Appendix A.

Proof:

Since $W = \begin{bmatrix} X & Z \end{bmatrix}$, then $W^T = \begin{bmatrix} X^T \\ Z^T \end{bmatrix}$.

$$W^T W = \begin{bmatrix} X^T X & X^T Z \\ Z^T X & Z^T Z \end{bmatrix}$$

$$\begin{aligned} (W^T W)^{-1} &= \begin{bmatrix} 0 & 0 \\ 0 & (Z^T Z)^{-1} \end{bmatrix} + \\ &\quad \begin{bmatrix} I_{P\times P} \\ -(Z^T Z)^{-1} Z^T X \end{bmatrix} (X^T X - X^T Z (Z^T Z)^{-1} Z^T X)^{-1} \\ &\quad \begin{bmatrix} I_{P\times P} & -X^T Z (Z^T Z)^{-1} \end{bmatrix} \\ &= \begin{bmatrix} A & B \\ C & D \end{bmatrix} \end{aligned}$$

where

$$\begin{aligned} A &= (X^T X - X^T Z (Z^T Z)^{-1} Z^T X)^{-1} \\ &= (X^T (I - Z (Z^T Z)^{-1} Z^T) X)^{-1} \\ &= (X^T Q X)^{-1} \end{aligned}$$

$$\begin{aligned} B &= -X^T Z (Z^T Z)^{-1} (X^T X - X^T Z (Z^T Z)^{-1} Z^T X)^{-1} \\ &= -X^T Z (Z^T Z)^{-1} (X^T (I - Z (Z^T Z)^{-1} Z^T) X)^{-1} \\ &= -X^T Z (Z^T Z)^{-1} (X^T Q X)^{-1} \end{aligned}$$

$$\begin{aligned} C &= -(Z^T Z)^{-1} Z^T X (X^T X - X^T Z (Z^T Z)^{-1} Z^T X)^{-1} \\ &= -(Z^T Z)^{-1} Z^T X (X^T (I - Z (Z^T Z)^{-1} Z^T) X)^{-1} \\ &= -(Z^T Z)^{-1} Z^T X (X^T Q X)^{-1} \end{aligned}$$

$$
\begin{aligned}
D &= (Z^TZ)^{-1} + (Z^TZ)^{-1}Z^TXX^TZ(Z^TZ)^{-1}(X^TX - X^TZ(Z^TZ)^{-1}Z^TX)^{-1} \\
&= (Z^TZ)^{-1} + (Z^TZ)^{-1}Z^TXX^TZ(Z^TZ)^{-1}(X^T(I - Z(Z^TZ)^{-1}Z^T)X)^{-1} \\
&= (Z^TZ)^{-1} + (Z^TZ)^{-1}Z^TXX^TZ(Z^TZ)^{-1}(X^TQX)^{-1}
\end{aligned}
$$

Now Eq.(3.24) can be written as

$$
\widehat{\boldsymbol{\gamma}} = \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\boldsymbol{\alpha}} \end{bmatrix} = (W^TW)^{-1}W^T\mathbf{y} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} X^T \\ Z^T \end{bmatrix} \mathbf{y}
$$

Thus, the OLS estimate of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}_{OLS} &= [AX^T + BZ^T]\mathbf{y} \\
&= (X^TQX)^{-1}(X^T - X^TZ(Z^TZ)^{-1}Z^T)\mathbf{y} \\
&= (X^TQX)^{-1}X^T(I - Z(Z^TZ)^{-1}Z^T)\mathbf{y} \\
&= (X^TQX)^{-1}X^TQ\mathbf{y} \qquad (3.25)
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\boldsymbol{\alpha}}_{OLS} &= [CX^T + DZ^T]\mathbf{y} \\
&= -(Z^TZ)^{-1}Z^TX(X^TQX)^{-1}X^T + \qquad (3.26) \\
&\quad (Z^TZ)^{-1}Z^T + (Z^TZ)^{-1}Z^TXX^TZ(Z^TZ)^{-1}Z^T(X^TQX)^{-1}\mathbf{y} \\
&= (Z^TZ)^{-1}Z^T[-X(X^TQX)^{-1}X^T + I + XX^TP(X^TQX)^{-1}]\mathbf{y} \\
&= (Z^TZ)^{-1}Z^T[I - (X^TQX)^{-1}(-XX^TP + XX^T)]\mathbf{y} \\
&= (Z^TZ)^{-1}Z^T\mathbf{y} - (Z^TZ)^{-1}Z^TX(X^TQX)^{-1}X^T(I - P)\mathbf{y}
\end{aligned}
$$

Since $\boldsymbol{\alpha} = Z\boldsymbol{\alpha}_{OLS} = \begin{bmatrix} \alpha_1\mathbf{1} \\ \vdots \\ \alpha_N\mathbf{1} \end{bmatrix}$ Therefore

$$
\begin{aligned}
\widehat{\boldsymbol{\alpha}} \qquad (3.27) \\
&= \overline{\mathbf{y}} - \overline{\mathbf{X}}(X^TQX)^{-1}X^TQ\mathbf{y} \\
&= \overline{\mathbf{y}} - \overline{\mathbf{X}}\widehat{\boldsymbol{\beta}}_{OLS} \qquad (3.28)
\end{aligned}
$$

Where $Q = I - Z(Z^TZ)^{-1}Z^T = I - P$,

$$\overline{\mathbf{y}} = (Z^TZ)^{-1}Z^T\mathbf{y} = \frac{1}{T}Z^T\mathbf{y} = \begin{bmatrix} \overline{y}_1\mathbf{1} \\ \overline{y}_2\mathbf{1} \\ \vdots \\ \overline{y}_N\mathbf{1} \end{bmatrix}$$

and

$$\overline{\mathbf{X}} = (Z^TZ)^{-1}Z^T\mathbf{X} = \frac{1}{T}Z^T\mathbf{X} = \begin{bmatrix} \overline{X}_1\mathbf{1} \\ \overline{X}_2\mathbf{1} \\ \vdots \\ \overline{X}_N\mathbf{1} \end{bmatrix}$$

So the OLS estimator gives the same formulation as the within group estimator.

3.2.3 Maximum Likelihood Estimation for fixed effects model

The maximum likelihood estimator of fixed effects model (3.23) is the same as perform MLE on a simplest model by assume $\mathbf{u}_i \sim \mathrm{N}(0, \sigma_u^2)$. Thus the MLE are same as OLS estimator, recall Eq.(3.25):

$$\widehat{\boldsymbol{\beta}}_{MLE} = (X^TX)^{-1}X^T\mathbf{y}$$

and the estimator of $\widehat{\sigma^2}_u$ follows by the definition is:

$$\widehat{\sigma^2}_{MLE} = \frac{1}{NT}(\mathbf{y} - X\boldsymbol{\beta})^T(\mathbf{y} - X\boldsymbol{\beta}) = \frac{\mathbf{e}^T\mathbf{e}}{NT}$$

where $\mathbf{e} = \mathbf{y} - X\widehat{\boldsymbol{\beta}}_{MLE}$ and is the residual vector refer to section 2.3. The ordinary least squares estimation gives

$$\mathrm{E}(\frac{\mathbf{e}^T\mathbf{e}}{NT - N - K}) = \sigma^2$$

Thus

$$\mathrm{E}(\widehat{\sigma^2}_{MLE}) = \frac{\sigma^2(NT - N - K)}{NT}$$

So that $\widehat{\sigma^2}_{MLE}$ is biased for σ^2, although $\widehat{\beta}_{MLE}$ is unbiased for β. The adjusted estimator is

$$\widehat{\sigma^2} = \frac{NT}{NT - N - K}\widehat{\sigma}^2_{MLE}$$

Note: if $\mathbf{u}_i \sim \mathrm{N}(0, D)$ where D is diagonal, the MLE estimator follows by Eq.(2.2) is

$$\widehat{\boldsymbol{\beta}}_{WLS} = (X^T D^{-1} X)^{-1} X^T D^{-1} \mathbf{y}$$

If D is not diagonal, then the MLE estimator follows Eq.(2.7) is

$$\widehat{\boldsymbol{\beta}}_{GLS} = (X^T V^{-1} X)^{-1} X^T V^{-1} \mathbf{y}$$

The proofs of WLS and GLS is similar to OLS case.

3.3 Random Effects (RE) Model

The major difference between random effects model and the fixed effects model is the individual unobserved effect α_i is random and uncorrelated with X_i. The model follows Eq.(3.3) and Eq. (3.4) structure, it can be defined most generally as

- Scalar form:

$$y_{it} = \mathbf{x}_{it}^T \boldsymbol{\beta} + \mathbf{Z}_{it}^T \boldsymbol{\alpha}_i + u_{it} \quad i = 1, \cdots, N \quad t = 1, \cdots, T$$

  where the dimensions of parameters are y_{it} is 1×1, $\mathbf{x}_{it}^T$ is $1 \times K$, $\boldsymbol{\beta}$ is $K \times 1$, $\mathbf{Z}_{it}^T$ is $1 \times P$, $\boldsymbol{\alpha}_i$ is $P \times 1$ and u_{it} is 1×1. $\mathbf{x}_{it}^T$ and $\mathbf{Z}_{it}^T$ are known covariates.

- Vector or individual form:

$$\mathbf{y}_i = X_i \boldsymbol{\beta} + Z_i \boldsymbol{\alpha}_i + \mathbf{u}_i \quad i = 1, \cdots, N \tag{3.29}$$

  where the dimensions of parameters are $\mathbf{y}_i$ is $T \times 1$, X_i is $T \times K$, $\boldsymbol{\beta}$ is $K \times 1$, Z_i is $T \times P$, $\boldsymbol{\alpha}_i$ is $P \times 1$ and $\mathbf{u}_i$ is T $\times$ 1.

- Full data form:

$$\mathbf{y} = X\boldsymbol{\beta} + Z\boldsymbol{\alpha} + \mathbf{u} \tag{3.30}$$

  where the dimensions of parameters are $\mathbf{y}$ is $NT \times 1$, X is $NT \times K$, $\boldsymbol{\beta}$ is $K \times 1$, Z is $NT \times NP$ and Z can be written as $Z = I_N \otimes \mathbf{1}_T$, $\boldsymbol{\alpha}$ is N$P \times 1$ where $\boldsymbol{\alpha}^T = [\boldsymbol{\alpha}_1^T, \boldsymbol{\alpha}_2^T, \cdots, \boldsymbol{\alpha}_N^T]$ ($\boldsymbol{\alpha}_i$ is P$\times$1) and $\mathbf{u}$ is NT $\times$ 1.

There are three components in random effect model, eg, in model (3.29): the fixed effect term $X_i\boldsymbol{\beta}$ with K variables and the random effect term $Z_i\boldsymbol{\alpha}_i$ with P explanatory variables;

the last term is a vector of residual components. This is sometimes called mixed effects model because it includes both fixed effects ($\boldsymbol{\beta}$) and individual random effects ($\boldsymbol{\alpha}_i$).

Now we assume $\mathbf{u}_i \sim$ iid $(0, R_i)$ for some covariance matrix R; the structure of R encodes the correlation among the $\mathbf{u}_{it}$. We have several choices for this structure, for example, it can be independent, AR(1), have a compound symmetry structure or be unstructured (refer to Chapter 2). And R_i is a $(T \times T)$ covariance matrix.

Also, $\boldsymbol{\alpha}_i$ is assumed to be normally distributed as iid $(0, G_i)$ which are independent with $\mathbf{u}_i$. G_i is a $(P \times P)$ covariance matrix with (k, l) element $G_{kl} = G_{lk}$, for simplest model, G_i can be

$$G_i = \begin{bmatrix} \sigma^2_{\alpha_1} & 0 & \cdots & 0 \\ 0 & \sigma^2_{\alpha_2} & \cdots & 0 \\ 0 & 0 & \cdots & \sigma^2_{\alpha_p} \end{bmatrix} \quad \text{for i = 1 } \cdots N.$$

Following the assumptions, conditional on the random effect, $\mathbf{y}_i$ is normally distributed as $N(X_i\boldsymbol{\beta}, Z_i G Z_i^T + R_i)$, the proof is below.

Proof

$$\mathrm{E}(\mathbf{y}_i) = \mathrm{E}(X_i\boldsymbol{\beta} + Z_i\boldsymbol{\alpha}_i + \mathbf{u}_i) = X_i\boldsymbol{\beta}$$

Since

$$\mathrm{E}(X_i\boldsymbol{\beta}) = X_i\boldsymbol{\beta}$$

$$\mathrm{E}(Z_i\boldsymbol{\alpha}_i) = Z_i\mathrm{E}(\boldsymbol{\alpha}_i) = 0$$

$$\mathrm{E}(\mathbf{u}_i) = 0; \quad \mathrm{E}(\alpha_i) = 0$$

$$\begin{aligned} \mathrm{Var}(\mathbf{y}_i) &= \mathrm{Var}(X_i\boldsymbol{\beta} + Z_i\boldsymbol{\alpha}_i + \mathbf{u}_i) \\ &= Z_i\mathrm{Var}(\boldsymbol{\alpha}_i)Z_i^T + \mathrm{Var}(\mathbf{u}_i) \\ &= Z_i G_i Z_i^T + R_i \end{aligned}$$

because

$$\mathrm{Cov}(\boldsymbol{\alpha}_i, \mathbf{u}_i) = 0$$

3.3.1 Two special types of Random Effect Model

Recall the random effect model Eq.(3.29):

$$\mathbf{y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{\alpha}_i + \mathbf{u}_i$$

For longitudinal data, the simplest random effect models are:

- Random Intercept Model

$$\mathbf{y}_i = X_i\boldsymbol{\beta} + \mathbf{1}_T\alpha_i + \mathbf{u}_i \quad i = 1, \cdots, N \tag{3.31}$$

  where α_i is 1×1, $Z_i = \mathbf{1}_T$ where $\mathbf{1}_T$ is a vector of ones of length T. In Eq.(3.31), observations collected on an individual are attributable to a individual"level", or random intercept.

- Random Intercept and Slope Model (Random Trend Model)

$$\mathbf{y}_i = X_i\boldsymbol{\beta} + \alpha_{i0} + Z_i\alpha_{i1} + \mathbf{u}_i \quad i = 1, \cdots, N \tag{3.32}$$

  In Eq.(3.32), each individual has an individual "trend", or follows a random linear trajectory with its own random intercept α_{i0} and slope α_{i1}.

Note: the random effect model we are going to discuss most in this thesis is random intercept model with independent covariance, i.e. $\text{Var}(\mathbf{u}_i) = \sigma_u^2 I_T$ and $\text{Var}(\alpha_i) = \sigma_\alpha^2$.

3.3.2 Generalized Least Squares Estimation for random effect model

Recall the model (3.29)

$$\mathbf{y}_i = X_i\boldsymbol{\beta} + Z_i\boldsymbol{\alpha}_i + \mathbf{u}_i \quad i = 1, \cdots, N$$

$$\mathbf{u}_i \sim \text{iid } (0, R_i)$$

$$\boldsymbol{\alpha}_i \sim \text{iid } (0, G_i)$$

$$\mathbf{u}_1 \cdots \mathbf{u}_N, \ \boldsymbol{\alpha}_1 \cdots \boldsymbol{\alpha}_N \text{ are independent}$$

In this thesis, we are interested in the random intercept model. So we assume R_i to be equal to $\sigma_u^2 I_T$ where I_T is the T dimensional identity matrix. And let G_i to be equal to σ_α^2 and Z_i

equal to $\mathbf{1}_T$. So the model can be written as Eq. (3.31)

$$\mathbf{y}_i = X_i\boldsymbol{\beta} + \mathbf{1}_{T\times 1}\alpha_i + \mathbf{u}_i \quad i = 1, \cdots, N$$

Basically, random effect model can be considered as an extension of the fixed effect model. We need to make another three assumptions based on the fixed effect model assumptions. In addition to

A1.

$$\mathrm{E}(\mathbf{u}_i|X_i, \alpha_i) = 0$$

A2.

$$\mathrm{Var}(\mathbf{u}_i|X_i, \alpha_i) = \sigma^2\mathbf{I}_T$$

We assume:

A3. $\alpha_i \sim$ iid $(0, \sigma_\alpha^2)$

A4. Cov $(\alpha_i, \mathbf{X}_{it}) = 0$

A5. $\mathbf{u}_i|X_i \sim$ iid $(0, \sigma_u^2 I_T + \sigma_{\boldsymbol{\alpha}}^2\mathbf{1}\mathbf{1}^T)$

The assumption A3 assume the unobserved individual effect is randomly independently distributed with mean zero and variance σ_α^2. The assumption A4 assume the α_i can not be correlated with the explanatory variable X_{it}. The assumption A5 assume the α_i and $\mathbf{u}_{it}$ are independent and this is guaranteed by our assumptions

$$\mathbf{u}_i \sim \text{iid } (0, \sigma_u^2 I_T)$$

$$\boldsymbol{\alpha}_i \sim \text{iid } (0, \sigma_\alpha^2)$$

The covariance matrix of the error term can be derived by using the given assumptions listed above. We have

$$\boldsymbol{\varepsilon}_i = \mathbf{1}_T\boldsymbol{\alpha}_i + \mathbf{u}_i,$$

and we can write the error covariance of each individual cross-section unit as

$$V_i = E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^T] = \sigma_u^2 I_T + \sigma_{\boldsymbol{\alpha}}^2\mathbf{1}\mathbf{1}^T \tag{3.33}$$

The proof of this can be found in Section: 3.5. When the data are organized as stacked form, the covariance of the error term for whole data set can be written as

$$\Omega = I_N \otimes V_i = \mathrm{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \begin{bmatrix} V_1 & 0 & \cdots & 0 \\ 0 & V_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V_N \end{bmatrix}$$

where $V_i = \mathrm{E}[\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i^T]$ is a T $\times$ T matrix.

Now we derive the estimator based on model (3.30). The GLS estimator of random effect model is

$$\widehat{\boldsymbol{\beta}}_{RE} = (X^T \widehat{\Omega}^{-1} X)^{-1} X^T \widehat{\Omega}^{-1} y \tag{3.34}$$

The variance of $\widehat{\boldsymbol{\beta}}_{RE}$ is

$$\text{Var } (\widehat{\boldsymbol{\beta}}_{RE}) = (X^T \widehat{\Omega}^{-1} X)^{-1} \tag{3.35}$$

This can be written as

$$\widehat{\boldsymbol{\beta}}_{RE} = \left(\sum_{i=1}^{N} X_i^T V_i^{-1} X_i \right)^{-1} \sum_{i=1}^{N} X_i^T V_i^{-1} \mathbf{y}_i$$

and

$$\text{Var } (\widehat{\boldsymbol{\beta}}_{RE}) = \sigma_u^2 \sum_{i=1}^{N} (X_i^T V_i^{-1} X_i)^{-1} \tag{3.36}$$

The GLS estimation assumes known σ_u^2 and σ_α^2. In fact, we don't know the two parameters in the real data analysis. So we have to estimate these two parameters first, then substitute these two into the GLS estimator in order to obtain the estimation of random effect coefficient. This method is called FGLS ("feasible" GLS) estimation.

Since the within group estimator is an unbiased estimator, we can use the fixed effect model to estimate the residual variance σ_u^2. That's because the fixed effect estimator eliminated the unobserved individual effect already, it would not affect our estimation of σ_u^2. Let the fixed effect residual be

$$\widehat{u}_{it} = (y_{it} - \overline{y}_i) - (\mathbf{X}_{it} - \overline{X}_i)\widehat{\boldsymbol{\beta}}_{FE}$$

where

$$\overline{y_i} = \frac{1}{T} \sum_{t=1}^{T} y_{it} = \frac{1}{T} \mathbf{1}^T \mathbf{y}_i$$

$$\overline{X_i} = \frac{1}{T}\sum_{t=1}^{T} X_{it} = \frac{1}{T}\mathbf{1}^T X_i$$

$$\overline{u_i} = \frac{1}{T}\sum_{t=1}^{T} u_{it} = \frac{1}{T}\mathbf{1}^T \mathbf{u}_i$$

Thus,

$$\widehat{\sigma^2}_u = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} \widehat{u}_{it}^2}{NT - N - K}. \tag{3.37}$$

Then we apply the OLS estimation on model

$$\overline{\mathbf{y}}_i = \overline{X}_i\boldsymbol{\beta} + \boldsymbol{\gamma}_i + \overline{u}_i$$

We define the combined variance

$$\begin{aligned} \sigma_B^2 &= \text{Var}(\alpha_i + \overline{u}_i) \\ &= \text{Var}(\alpha_i) + \text{Var}(\overline{u}_i) \\ &= \sigma_\alpha^2 + \frac{1}{T}\sigma_u^2 \end{aligned}$$

and hence estimates of these variances are related by ($\widehat{\sigma^2}_B$ can be calculated by using Eq.(3.42))

$$\widehat{\sigma^2}_B = \widehat{\sigma^2}_\alpha + \frac{1}{T}\widehat{\sigma^2}_u$$

We combine the $\widehat{\sigma^2}_B$ and $\widehat{\sigma^2}_u$ result, we have

$$\widehat{\sigma^2}_\alpha = \widehat{\sigma^2}_B - \frac{1}{T}\widehat{\sigma^2}_u \tag{3.38}$$

Now we could substitute these two estimators into Eq. (3.33) to use the GLS estimator Eq. (3.34) to get the estimation of $\widehat{\boldsymbol{\beta}}_{RE}$. This estimator is called FGLS estimator.

3.3.3 Between Estimation for random effect model

The between estimator converts all the data into individual averages and performs OLS on the following equation:

$$\overline{y_i} = \overline{X_i}\boldsymbol{\beta} + \alpha_i + \overline{u_i} \quad i = 1, \cdots, N$$

where the *i*th term $\overline{y_i}$ is

$$\overline{y_i} = \frac{1}{T}\sum_{t=1}^{T} y_{it} = \frac{1}{T}\mathbf{1}^T\mathbf{y}_i$$

and $\overline{X}_i$ is defined as the vector of *i*th individual means of the explanatory variables. Let RSS to be the sum of squared residuals, then we have

$$RSS = \sum_{i=1}^{N} \overline{u_i}^T\overline{u_i} = \sum_{i=1}^{N}(\overline{y_i} - \overline{X_i}\boldsymbol{\beta} - \boldsymbol{\alpha}_i)^T(\overline{y_i} - \overline{X_i}\boldsymbol{\beta} - \boldsymbol{\alpha}_i)$$

Take the partial derivatives of RSS with respect to α_i and let it equal to zero, we have

$$\widehat{\alpha} = \overline{\mathbf{y}} - \overline{X}\boldsymbol{\beta}$$

Then take the partial derivatives with respect to $\boldsymbol{\beta}$ and substitute $\widehat{\alpha_i}$ into it and let this equal to zero, we have

$$\widehat{\boldsymbol{\beta}_B} = \left\{\sum_{i=1}^{N}(\overline{X}_i - \overline{X})(\overline{X}_i - \overline{X})^T\right\}^{-1}\left\{\sum_{i=1}^{N}(\overline{X}_i - \overline{X})(\overline{y}_i - \overline{\mathbf{y}})\right\}$$

This estimator is just as the OLS estimator on the cross-sectional equation with α_i as the intercept. The between estimator ignores important information on how the individuals changed over time and is biased if α_i is correlated with $\overline{X_i}$ (see Chapter 7 for more detail).

Now we put this expression in matrix terms. We define a projection matrix P as the transform matrix to get the estimator. The random effect model is

$$\mathbf{y} = X\boldsymbol{\beta} + Z\boldsymbol{\alpha} + \mathbf{u} \tag{3.39}$$

where Z is the matrix of N dummy variables corresponding to each cross-section unit. As we defined $P = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$ in matrix form, this is a symmetric and idempotent matrix. Premultiply the model (3.29) by this matrix transforms the data into the means over time form.

$$P\mathbf{y} = PX\boldsymbol{\beta} + PZ\boldsymbol{\alpha} + P\mathbf{u}$$

where

$$\begin{aligned} P\mathbf{y} &= \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{y} \\ &= \frac{1}{T}\mathbf{Z}\begin{bmatrix}\sum_{t=1}^{T} y_1 & \cdots & \sum_{t=1}^{T} y_N\end{bmatrix}^T \\ &= \mathbf{Z}\begin{bmatrix}\overline{y}_1 & \cdots & \overline{y}_N\end{bmatrix}^T \\ &= \begin{bmatrix}\overline{y}_1\mathbf{1}_T \\ \vdots \\ \overline{y}_N\mathbf{1}_T\end{bmatrix} \\ &= \overline{\mathbf{y}} \end{aligned}$$

and $PX = \overline{X}$, $P\mathbf{u} = \overline{\mathbf{u}}$ and $PZ = 0$. Thus,

$$\overline{\mathbf{y}} = \overline{X}\boldsymbol{\beta} + \overline{\mathbf{u}}$$

By using the least squares theory as above, we have

$$\widehat{\boldsymbol{\beta}}_B = (\overline{X}^T\overline{X})^{-1}\overline{X}^T\overline{\mathbf{y}}$$

This estimator $\widehat{\boldsymbol{\beta}}_B$ is called the between estimator and also can be written as

$$\widehat{\boldsymbol{\beta}}_B = ((X^TP^T)PX)^{-1}(X^TP^T)(P\mathbf{y}) = (X^TPX)^{-1}X^TP\mathbf{y} \tag{3.40}$$

Let

$$\widehat{\mathbf{u}}_B = \overline{\mathbf{y}} - \overline{X}\widehat{\boldsymbol{\beta}_B} \tag{3.41}$$

Then we have the estimator of σ_B^2

$$\widehat{\sigma}_B^2 = \frac{\widehat{\mathbf{u}}_B^T\widehat{\mathbf{u}}_B}{N-K} \tag{3.42}$$

where $\widehat{\mathbf{u}}_B$ is the residual vector obtained by using between estimator.

3.3.4 Maximum Likelihood Estimation for random effect model

To use the maximum likelihood estimation method, we assume the α_i and u_{it} are normally distributed. Thus,

$$\mathbf{y} \sim N(\boldsymbol{\mu}, \Omega) \quad \text{with } \boldsymbol{\mu} = X\boldsymbol{\beta}$$

(See proof from Section 3.5) The probability equation of the random effect model is

$$f(\mathbf{y}|\boldsymbol{\mu},\Omega) = (2\pi)^{-\frac{NT}{2}}|\Omega|^{-\frac{1}{2}}\exp\left[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^T\Omega^{-1}(\mathbf{y}-\boldsymbol{\mu})\right]$$

where $\Omega = I_N \otimes V$. Recall the vector form model (3.31), then the log likelihood function of random effect model is written as

$$\begin{aligned}
\log L &= -\frac{NT}{2}\log 2\pi - \frac{N}{2}\log|V| \\
&\quad -\frac{1}{2}\sum_{i=1}^{N}(\mathbf{y}_i - X_i\boldsymbol{\beta})^T V^{-1}(\mathbf{y}_i - X_i\boldsymbol{\beta}) \\
&= -\frac{NT}{2}\log 2\pi - \frac{N(T-1)}{2}\log\sigma_u^2 - \frac{N}{2}\log(\sigma_u^2 + T\sigma_\alpha^2) \\
&\quad -\frac{1}{2\sigma_u^2}\sum_{i=1}^{N}(\mathbf{y}_i - \mathbf{1}_T\alpha_i - X_i\boldsymbol{\beta})^T Q(\mathbf{y}_i - \mathbf{1}_T\alpha_i - X_i\boldsymbol{\beta}) \qquad (3.43) \\
&\quad -\frac{1}{2(\sigma_u^2 + T\sigma_\alpha^2)}\sum_{i=1}^{N}(\mathbf{y}_i - X_i\boldsymbol{\beta})^T\frac{1}{T}\mathbf{1}\mathbf{1}^T(\mathbf{y}_i - X_i\boldsymbol{\beta}) \\
&= -\frac{NT}{2}\log 2\pi - \frac{N(T-1)}{2}\log\sigma_u^2 - \frac{N}{2}\log(\sigma_u^2 + T\sigma_\alpha^2) \\
&\quad -\frac{1}{2\sigma_u^2}\sum_{i=1}^{N}(\mathbf{y}_i - X_i\boldsymbol{\beta})^T Q(\mathbf{y}_i - X_i\boldsymbol{\beta}) \\
&\quad -\frac{T}{2(\sigma_u^2 + T\sigma_\alpha^2)}\sum_{i=1}^{N}(\overline{y_i} - \overline{X_i}\boldsymbol{\beta})^2 \qquad (3.44)
\end{aligned}$$

where

$$|V| = \sigma_u^{2(T-1)}(\sigma_u^2 + T\sigma_\alpha^2)$$

from Hsiao [2003].

$$V^{-1} = \frac{1}{\sigma_u^2}\left[I_T - \frac{\sigma_u^2}{T\sigma_\alpha^2 + \sigma_u^2}\mathbf{1}\mathbf{1}^T\right] = \frac{1}{\sigma_u^2}\left[Q + \frac{\sigma_u^2}{T\sigma_\alpha^2 + \sigma_u^2}\frac{1}{T}\mathbf{1}\mathbf{1}^T\right]$$

and $Q = I_T - \frac{1}{T}\mathbf{1}\mathbf{1}^T$ as before. The result is obtained from Hsiao [2003].

To get the estimators, we take the first derivative of the log-likelihood of random effect

model and then set each equation is equal to zero.

$$\begin{aligned}\frac{\partial \log L}{\partial \boldsymbol{\beta}} &= \frac{1}{\sigma_u^2}\sum_{i=1}^{N}(\mathbf{y}_i - X_i\boldsymbol{\beta})^T Q X_i \\ &\quad -\frac{1}{\sigma_u^2}\frac{T\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}\sum_{i=1}^{N}(\overline{y_i} - \overline{X_i}\boldsymbol{\beta})\overline{X_i} \\ &= 0\end{aligned}$$

$$\begin{aligned}\frac{\partial \log L}{\partial \sigma_u^2} &= -\frac{N(T-1)}{2\sigma_u^2} - \frac{N}{2(\sigma_u^2 + T\sigma_\alpha^2)} \\ &\quad +\frac{1}{2\sigma_u^4}\sum_{i=1}^{N}(\mathbf{y}_i - X_i\boldsymbol{\beta})^T Q(\mathbf{y}_i - X_i\boldsymbol{\beta}) \\ &\quad +\frac{T}{2(\sigma_u^2 + T\sigma_\alpha^2)^2}\sum_{i=1}^{N}(\overline{y_i} - \overline{X_i}\boldsymbol{\beta})^2 = 0\end{aligned}$$

$$\frac{\partial \log L}{\partial \boldsymbol{\sigma}_\alpha^2} = -\frac{NT}{2(\sigma_u^2 + T\sigma_\alpha^2)} + \frac{T^2}{2(\sigma_u^2 + T\sigma_\alpha^2)^2}\sum_{i=1}^{N}(\overline{y_i} - \overline{X_i}\boldsymbol{\beta})^2 = 0$$

Solve the above partial derivative, we obtain the estimators as

$$\widehat{\boldsymbol{\beta}} = \left\{\sum_{i=1}^{N} X_i^T\left[I_T - \frac{\sigma_u^2}{T\sigma_\alpha^2 + \sigma_u^2}\mathbf{1}\mathbf{1}^T\right]X_i\right\}^{-1}\left\{\sum_{i=1}^{N} X_i^T\left[I_T - \frac{\sigma_u^2}{T\sigma_\alpha^2 + \sigma_u^2}\mathbf{1}\mathbf{1}^T\right]\mathbf{y}_i\right\}$$

$$\widehat{\sigma}_u^2 = \frac{1}{N(T-1)}\sum_{i=1}^{N}(\mathbf{y}_i - \mathbf{1}_T\alpha_i - X_i\boldsymbol{\beta})^T Q(\mathbf{y}_i - \mathbf{1}_T\alpha_i - X_i\boldsymbol{\beta})$$

$$\widehat{\boldsymbol{\sigma}}_\alpha^2 = \frac{1}{N}\sum_{i=1}^{N}(\overline{y_i} - \overline{X_i}\boldsymbol{\beta})^2 - \frac{1}{T}\widehat{\sigma}_u^2$$

Hence, to get the estimation of the maximum likelihood estimator, we can iterate between $\boldsymbol{\beta}$ and σ_u^2, σ_α^2 until convergence.

3.4 Hausman Test

We have described two estimators, random effects (RE) estimator and fixed effects (FE) estimators that have different properties depending on the correlation between unobserved individual effect α_i and the explanatory variable $\mathbf{x}_{it}$ ([Hsiao, 2003]). **?** has named a test called Hausman test for H_0 null hypothesis : α_i and $\mathbf{X}_{it}$ are uncorrelated.

Testing FE vs. RE

We can test whether a fixed or random effects model is appropriate using a Hausman test.

$$H_0 : \alpha_i \perp X_{it}$$

$$H_a : \alpha_i \not\perp X_{it}$$

1. If the effects are uncorrelated with the explanatory variables, the H_0 is true, the random effects (RE) estimator and the fixed effects (FE) estimator is unbiased but the random effects (RE) estimator is the one that should be adopted as it is efficient (random effect estimation has smaller variance than fixed effect estimation).

2. If the effects are correlated with the explanatory variables, the H_a is true, the fixed effects estimator is unbiased and efficient but the random effects estimator is biased (see proof in Chapter 5).

The test statistic is

$$H = (\widehat{\beta}_{RE} - \widehat{\beta}_{FE})^T (\Sigma_{FE} - \Sigma_{RE})^{-1} (\widehat{\beta}_{RE} - \widehat{\beta}_{FE}) \tag{3.45}$$

where Σ_{FE} is the covariance of β_{FE} refer to Eq.(3.20) and Eq.(3.36). Σ_{RE} is the covariance of β_{RE} refer to Eq.(3.34) and Eq.(3.35). This test statistics will be distributed asymptotically as χ^2 with K degrees of freedom under the null hypothesis that the random effects estimator is true where K is the number of β parameters in the model.

3.5 Special Case: Equivalent Model

We note that two important models of interest are equivalent: compound symmetry model and random intercept model with independent error term. A model without random effect, but with compound symmetry variance structure (refer to Eq.(2.12)), we call it compound symmetry model and the random effect model with only random intercept, we call it random intercept model. The proof is below:

Proof of Equivalence

- Compound symmetry model:

$$\mathbf{y}_i = X_i \boldsymbol{\beta} + \mathbf{u}_i \quad i = 1, \cdots, N$$

with

$$\mathbf{u}_i \sim N(0, \sigma^2 V)$$

and

$$V = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

$$\mathrm{E}(\mathbf{y}_i) = \mathrm{E}(X_i\boldsymbol{\beta} + \mathbf{u}_i) = X_i\boldsymbol{\beta}$$

$$\mathrm{Var}(\mathbf{y}_i) = \mathrm{Var}\left[\mathbf{u}_i\right] = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

- Random intercept model with i.i.d $\mathbf{u}_i$:

$$\begin{aligned} \mathbf{y}_i &= X_i\boldsymbol{\beta} + Z_i\boldsymbol{\gamma}_i + \mathbf{u}_i \\ &= X_i\boldsymbol{\beta} + \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \boldsymbol{\gamma}_i + \mathbf{u}_i \\ &= X_i\boldsymbol{\beta} + \mathbf{1}\boldsymbol{\gamma}_i + \mathbf{u}_i \quad i = 1, \cdots, N \end{aligned}$$

with

$$\mathbf{u}_i \sim N(0, \sigma_u^2 I)$$

and

$$\gamma_i \sim N(0, \sigma_\gamma^2)$$

$$\mathrm{E}(\mathbf{y}_i) = \mathrm{E}(X_i\boldsymbol{\beta} + Z_i\boldsymbol{\gamma}_i + \mathbf{u}_i) = X_i\boldsymbol{\beta}; \quad Z_i = \mathbf{1}$$

$$
\begin{aligned}
\text{Var}(\mathbf{y}_i) &= \text{Var}(\mathbf{1}\boldsymbol{\gamma}_i) + \text{Var}(\mathbf{u}_i) \\
&= \mathbf{1}\text{Var}(\boldsymbol{\gamma}_i)\mathbf{1}^T + \sigma_u^2\mathbf{I} \\
&= \sigma_\gamma^2\mathbf{1}\mathbf{1}^T + \sigma_u^2\mathbf{I} \\
&= \sigma_\gamma^2\begin{bmatrix} 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} + \sigma_u^2\begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \\
&= \begin{bmatrix} \sigma_\gamma^2 + \sigma_u^2 & \sigma_\gamma^2 & \cdots & \sigma_\gamma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\gamma^2 & \sigma_\gamma^2 & \cdots & \sigma_\gamma^2 + \sigma_u^2 \end{bmatrix} \\
&= (\sigma_\gamma^2 + \sigma_u^2)\begin{bmatrix} 1 & \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_u^2} & \cdots & \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_u^2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_u^2} & \frac{\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_u^2} & \cdots & 1 \end{bmatrix} \\
&= \sigma^2\begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}
\end{aligned}
$$

where $\sigma^2 = \sigma_\gamma^2 + \sigma_u^2$ and $\rho = \dfrac{\sigma_\gamma^2}{\sigma_\gamma^2 + \sigma_u^2}$

Both models end up with the same distribution of $\mathbf{y}$. So the random intercept model is identical to the compound symmetry model, but is derived from a different perspective.

Chapter 4

Simulation of longitudinal data

In this chapter, we define the simulation functions in R for several common models. For example, random trend model, random intercept model, fixed effects model and pooled model. Within the fixed effect model section, we define two simulation functions: one for the case without correlation between individual effects and explanatory variables and one for the case with such correlation. We also give a special case of random intercept model function which may look like there is correlation between individual effects and explanatory variables. But actually there no such correlation. We then graphically present these models by given the true values for each model, we also indicate the different features of each model. Finally, we define the simulated functions for three simple types of covariance structures, i.e. i.i.d or with serial correlation - AR(1) or compound symmetry. In this section, we assume there is only one explanatory variable. And we only generate the small dataset to demonstrate the features by plot the data for each model with i.i.d as covariance structure only. At the end of this chapter we give the R codes included in this chapter. In the later chapters, we are going to use these functions to generate the data for fixed effects model with correlation and random intercept model without correlation, then we could investigate whether there is estimation bias by using different estimation methods on these two models.

4.1 Random Trend Model

The random effect model for a single covariate X with random slope and intercept, also called the random trend model is defined as:

$$Y_{ij} = \beta_0 + \beta_1 x_{it} + \alpha_{0i} + \alpha_{1i} x_{it} + \varepsilon_{it} \tag{4.1}$$

where $i = 1, \cdots, N$ and $t = 1, \cdots, T$; β_0 and β_1 are fixed effect intercept and slope respectively, and α_{0i} and α_{1i} are random effect intercept and slope for i^{th} individual respectively.

First we randomly sample the N values x_{i1} from a uniform distribution U(a, b) as the first observation of X for each individual.

$$x_{i1} \sim U(a, b)$$

We then let the observations x_{it} (for $t = 2, \cdots, T$) be the order statistics of T-1 draws from the uniform distribution U$(x_{it}, x_{it} + \delta)$ for some δ.

Next, the random distribution ε_{it} is simulated. For independent errors, we have

$$\varepsilon_{it} \sim \text{i.i.d}(0, \sigma_\epsilon^2) \tag{4.2}$$

and we assume this is a Normal distribution with mean 0 and variance σ_ε^2. As noted in chapter 2, it also can be other structure, i.e. AR(1) (first order autoregressive) or CS (compound symmetry). We discuss simulation of such errors later.

The random effect α_{0i} and α_{1i} is multivariate normal distribution with zero mean and covariance G where σ_0^2 for α_{0i} and σ_1^2 for α_{1i}.

$$\begin{bmatrix} \alpha_{0i} \\ \alpha_{1i} \end{bmatrix} \sim \text{MVN}\,(\mathbf{0}, G) \tag{4.3}$$

where $G = \begin{bmatrix} \sigma_0^2 & \sigma_{01} \\ \sigma_{01} & \sigma_1^2 \end{bmatrix}$ and we set $\sigma_{01} = 0$.

To simulate from this distribution, we define the *rmvnorm* function as following (in our case, we always set $\mu_0 = 0$ and $\mu_1 = 0$): Choose the centre of Multivariate (bivariate) Normal distribution at (μ_0,μ_1) with a standard deviation in the (σ_0, σ_1) direction. Define a mean difference vector **u**.

$$\mathbf{u} = \mathbf{x} - \boldsymbol{\mu} = \begin{pmatrix} x_0 - \mu_0 \\ x_1 - \mu_1 \end{pmatrix} = \begin{pmatrix} u_0 \\ u_1 \end{pmatrix}, \quad u_0, u_1 \sim \text{N}(0, 1)$$

then we define a variance matrix **D** and derive **Du** as

$$\mathbf{D} = \begin{pmatrix} \sigma_0 & 0 \\ 0 & \sigma_1 \end{pmatrix}$$

$$\begin{pmatrix} \sigma_0 & 0 \\ 0 & \sigma_1 \end{pmatrix} \mathbf{u} = \begin{pmatrix} \sigma_0 u_0 \\ \sigma_1 u_1 \end{pmatrix} = \mathbf{Du}$$

Now we rotate it with certain angle θ, using matrix

$$\mathbf{R} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

$$\mathbf{RDu} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \sigma_0 u_0 \\ \sigma_1 u_1 \end{pmatrix} \tag{4.4}$$

$$\mathbf{y} = \mu + \mathbf{RDu}$$

Note:

$$\begin{aligned} \text{Var}(\alpha) = \text{Var}(\mathbf{RDu}) &= \mathbf{RD}\text{Var}(\mathbf{u})\mathbf{D}^T\mathbf{R}^T \\ &= \mathbf{RD}I\mathbf{D}^T\mathbf{R}^T \\ &= G \end{aligned}$$

Note, the σ_0 and σ_1 in this function are practical components which are not equal to σ_{0_α} and σ_{1_α} unless $\theta = 0$.

Alternatively, we can simply call the *rmvnorm* function from R built in function to simulate the individual effect.

Finally, we simulate the random effect data by calling the function *sim.RE* with parameters N, T, a, b, δ, $\boldsymbol{\mu}$, G, σ_ε, β_0 and β_1. The true value for these parameters are

$$N = 12,\ T = 5,\ a = 1,\ b = 5,$$

$$\delta = 5,\ \boldsymbol{\mu} = \begin{bmatrix} 0 & 0 \end{bmatrix},\ G = \begin{bmatrix} 100 & 0 \\ 0 & 50 \end{bmatrix},$$

$$\sigma_\varepsilon = 10,\ \beta_0 = 0 \text{ and } \beta_1 = 1.$$

Figure 4.1 shows an example of a random trend model with spherical errors. Data points for a single individual i are connected with lines for $i = 1 \cdots N$. There are 12 simulated individuals and each individual has 5 observations.

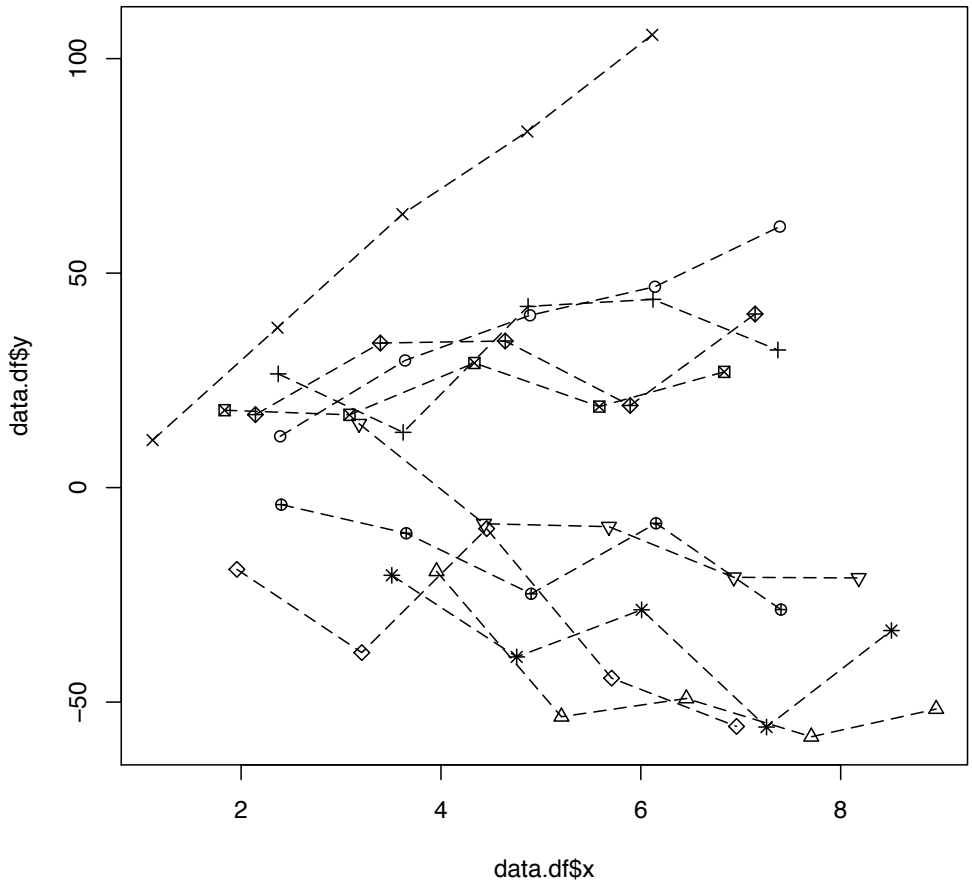

Figure 4.1: Figures of an example of a random trend model with spherical errors data

From Figure 4.1, we can see the most of the individual slopes are positive as most of individuals have upward trend, there are a few individuals have negative slope as a few of them have downwards trend. Especially, the bottom individual has the biggest negative slope.

4.2 Random Intercept Model

The random effect model with random intercept only is a special case of random trend model where $\alpha_{1i} = 0 \quad$ for all $i,\ i = 1, \cdots, N$ and is defined as:

$$Y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_{0i} + \varepsilon_{it} \tag{4.5}$$

where $i = 1, \cdots, N$ and $t = 1, \cdots, T$; β_0 and β_1 are fixed effect intercept and slope respectively, and α_{0i} is the only random effect for i^{th} individual.

In R program, we called the *sim.RE* function. We set the parameter α_{1i} with zero mean and zero variance which means $\sigma_1^2 = 0$. Now, all the α_{1i} for $i = 1, \cdots, N$ are equal to zero. The random trend model simplifies to Eq.(4.5) – the random effect model with random intercept only. We call this the random intercept model. Now we simulate the random intercept data by setting the parameters of *sim.RE* function as

$$N = 10,\ T = 5,\ a = 1,\ b = 5,$$

$$\delta = 5,\ \boldsymbol{\mu} = \begin{bmatrix} 0 & 0 \end{bmatrix},\ G = \begin{bmatrix} 100 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\sigma_\varepsilon = 1,\ \beta_0 = 0 \text{ and } \beta_1 = 1.$$

Figure 4.2 shows an example of a random intercept model with spherical errors. Data points for a single individual i are connected with lines for $i = 1 \cdots N$. There are 10 simulated individuals and each individual has five observation.

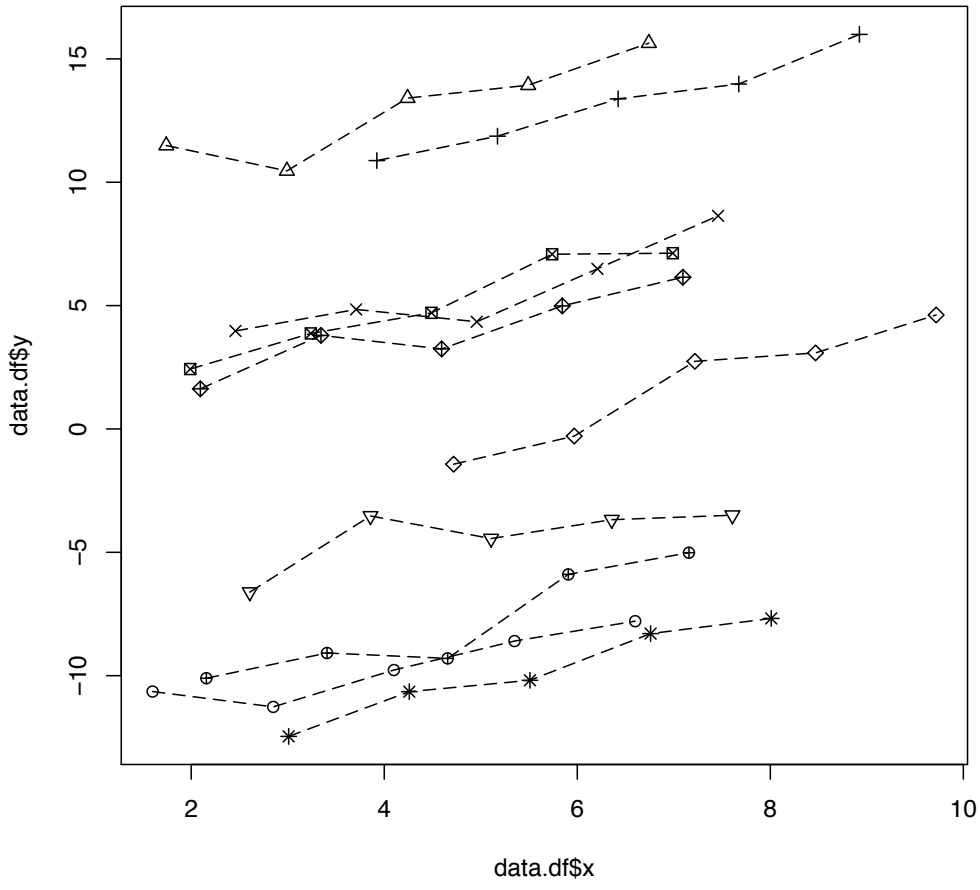

Figure 4.2: Figures of an example of a random intercept model with spherical errors data

From Figure 4.2, we can see all of the individuals have positive slopes and the trends are parallel with different intercept.

4.3 Fixed Effect Model

We have defined the random trend model (or we can call it individual effect model) in Eq.(4.1) with

$$\varepsilon_{it} \sim N(0, \sigma_{\varepsilon}^2)$$

and

$$\begin{bmatrix} \alpha_{0i} \\ \alpha_{1i} \end{bmatrix} \sim N(0, G)$$

For random intercept model, we have $\alpha_{1i} = 0$ and $\alpha_{0i} \sim N(0, \sigma_0^2)$. The fixed effect model is a special case of random intercept model where $\alpha_{1i} = 0$ and α_{0i} is constant over time that may be correlated with $\mathbf{X}_i$ and defined as:

$$Y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_{0i} + \varepsilon_{it} \tag{4.6}$$

where $i = 1 \cdots N$ and $t = 1 \cdots T$; β_0 and β_1 are fixed effect intercept and slope respectively. Note, the random effect model simplifies to the Eq.(4.6) – without any random effect. The fixed effects simulated data set has the same setting as the random intercept model but the definition of α_{0i} is different.

4.3.1 Fixed effects model without correlation

Firstly, we generate the data for the fixed effect model without correlation by setting the parameters as below

$$N = 10,\ T = 5,\ a = -5,\ b = 5,$$

$$\delta = 5,\ \boldsymbol{\mu} = \begin{bmatrix} 0 & 0 \end{bmatrix},\ G = \begin{bmatrix} 20 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\sigma_{\varepsilon} = 1,\ \beta_0 = 0 \text{ and } \beta_1 = 1.$$

In R program, we call the *sim.RE* function and then set α_{1i} equal to zero that means the random slope with zero mean and also set the variance of α_{1i} $\sigma_{1_{\alpha_1}}^2 = 0$ in order to eliminate the random effects slope. We don't have to change the setting of α_{0i}, because α_{0i} is random

and is not correlated with X_i .

The model is Eq.(4.6) with

$$\text{Cov}(x_{it}, \alpha_{0i}) = 0$$

Figure 4.3 shows 10 simulated individuals with constant individual effect.

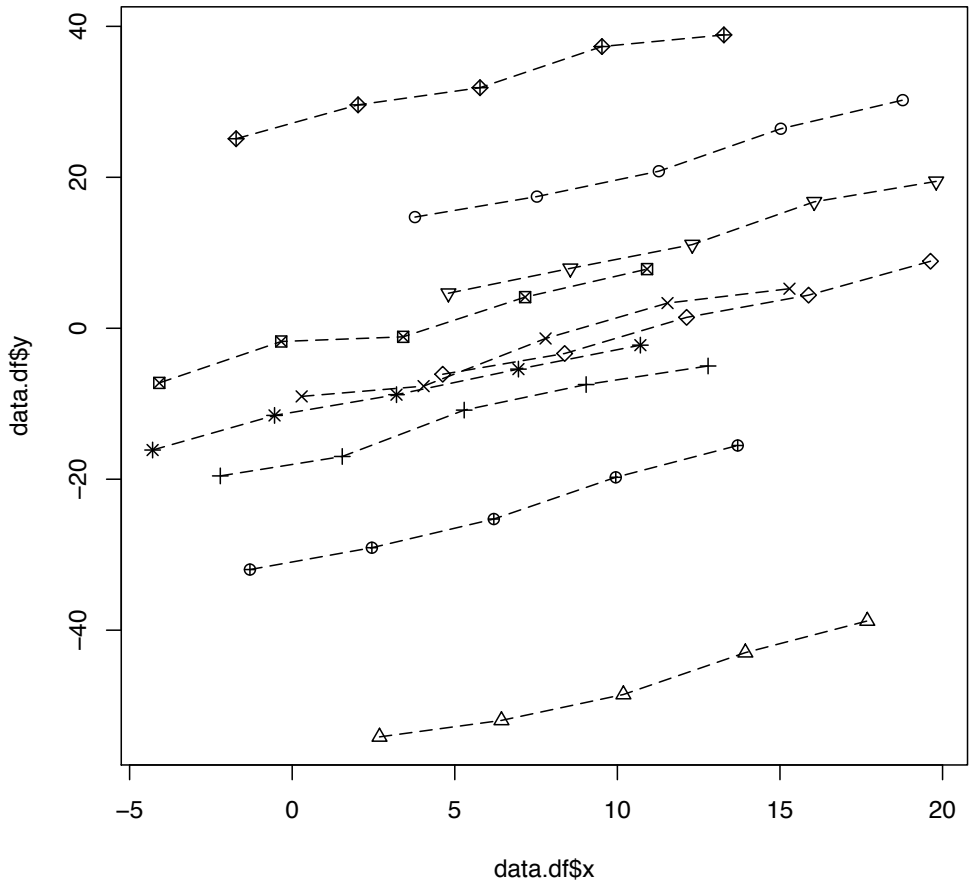

Figure 4.3: Figures of an example of a fixed effect model with $\text{Cov}(x_{it}, \alpha_{0i}) = 0$ data

From Figure 4.3, we can see all of the individuals have positive slopes and the trends are parallel. The data characteristic is hard to distinguish between random intercept model data and fixed effect model data without correlation.

4.3.2 Fixed effect model with correlation

In this section, we generate data of the fixed effects model with correlation. The model is

$$y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_{0i} + \varepsilon_{it} \tag{4.7}$$

with

$$\text{Cov}(X_{it}, \alpha_{0i}) = \rho \neq 0$$

Set the correlation equation by following Mundlak formulation [Mundlak, 1978] as

$$\alpha_{0i} = \overline{\mathbf{X}}_i \rho + w_i \tag{4.8}$$

where

$$w_i \sim \text{N}(0, \sigma_w^2) \tag{4.9}$$

The correlation means the explanatory variable x_{it} is correlated with the individual effect α_{0i}.

Now we define the generator of fixed effects model with correlation in R. Compare with the *sim.RE* function, the only part of program we need to change is the way of generate individual effect α_{0i}. Firstly, we generate N random numbers of w_i to follow Eq.(4.9). Then we calculate $\overline{\mathbf{X}}_i$ and define a constant ρ as the correlation coefficient or degree of the correlation. Finally, we follow the Eq.(4.8), we obtain α_{0i} random generator.

We call this R program function as **sim.cor** with parameters N, T, a, b, δ, ρ, σ_ε, σ_w, β_0 and β_1. Figure 4.4 shows 10 simulated individuals with $\text{Cov}(x_{it}, \alpha_{0i}) \neq 0$ by setting the parameters as

$$N = 10,\ T = 5,\ a = -5,\ b = 5,$$

$$\delta = 10,\ \sigma_\varepsilon = 1,\ \sigma_w = 1,$$

$$\rho = 3,\ \beta_0 = 0 \text{ and } \beta_1 = 1.$$

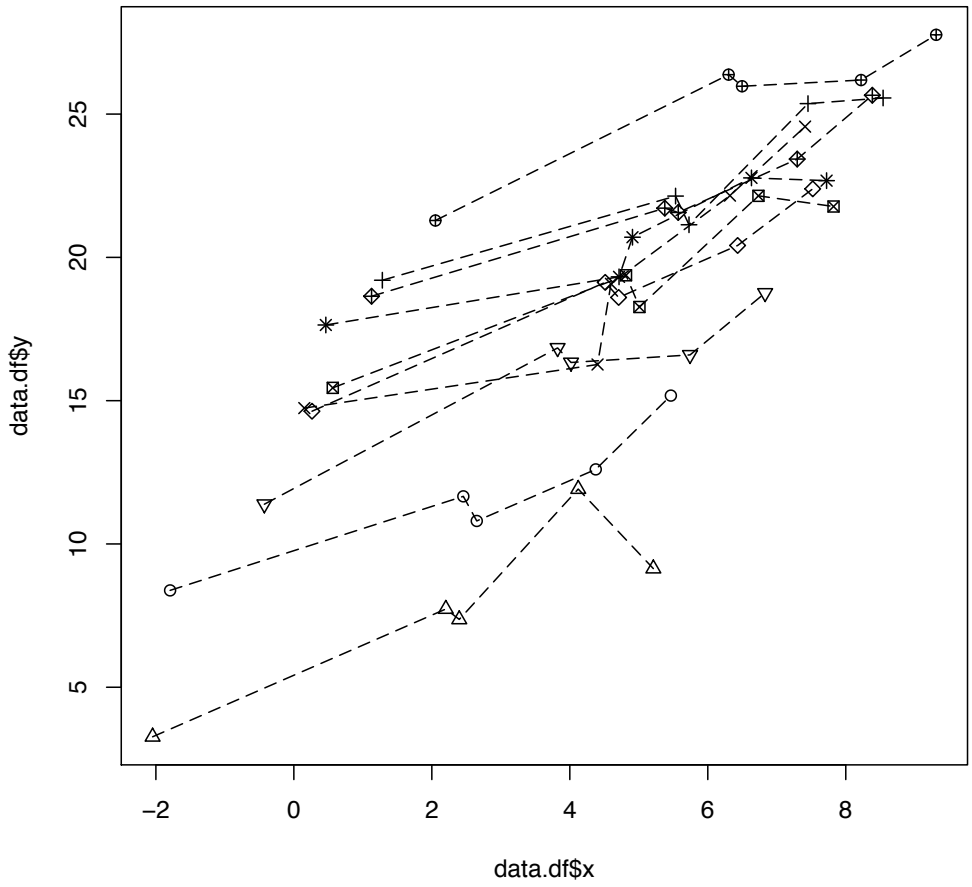

Figure 4.4: Figures of an example of a random intercept model with spherical errors and $\text{Cov}(X, \alpha_0) \neq 0$

From Figure 4.4, we can see all of the individuals have positive slopes and the trends are parallel. The intercepts of each individual are step increasing. This indicate that there is positive correlation between individual effects and covariate X.

4.4 Pooled Model

The pooled model with no individual effects is a special case of fixed effect model where $\alpha_{0i} = 0 \quad$ for all $i,\ i = 1 \cdots N$ and defined as:

$$Y_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it} \tag{4.10}$$

where $i = 1 \cdots N$ and $t = 1 \cdots T$; β_0 and β_1 are fixed effect intercept and slope respectively. Call the function **sim.RE** and we set α_{0i} and α_{1i} to equal to zero that means both random in-

tercept and slope with zero mean and zero variances. Now the random effect model simplifies to Eq.(4.10) – the pooled model without individual effect. Now we simulate the pooled data by setting the parameters of *sim.RE* function as

$$N = 10,\ T = 5,\ a = -5,\ b = 5,$$

$$\delta = 5,\ \boldsymbol{\mu} = \begin{bmatrix} 0 & 0 \end{bmatrix},\ G = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\sigma_\varepsilon = 10,\ \beta_0 = 0 \text{ and } \beta_1 = 1.$$

Figure 4.5 shows 10 simulated individuals with no individual effect and each individual has 5 observations.

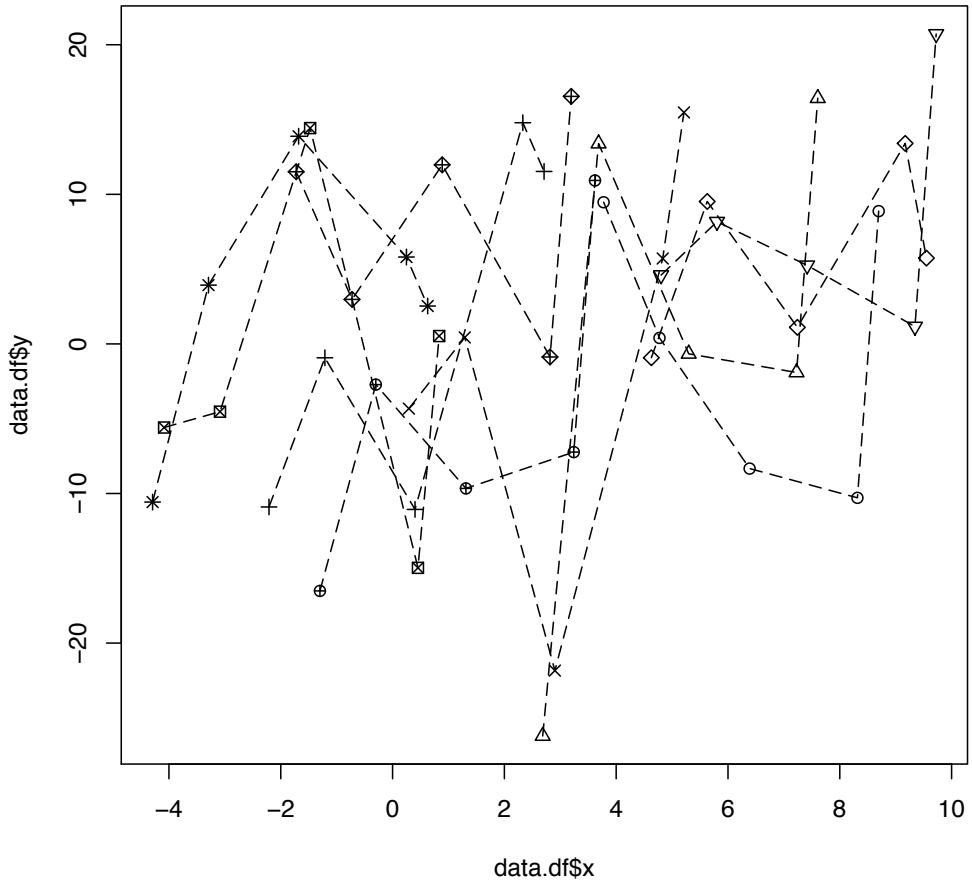

Figure 4.5: Figures of an example of a pooled model with spherical errors data

From Figure 4.5, we can see the individuals distribute randomly. There is a slightly upward trend.

4.5 Error Structure

In this section, we present three simple types of error structures, independent covariance, AR(1) covariance and compound symmetry covariance (refer to Chapter 2). The ε_i is distributed as normal distribution with zero mean and V as covariance. We set $V = \text{Var}(\varepsilon_i)$.

4.5.1 Independence covariance

The independence covariance matrix is the simplest of all covariance matrix. The model with spherical errors have independence covariance. σ^2 is the only parameter in the matrix (refer to Chapter 2), and

$$V = \sigma^2 I$$

with no correlation between observations. The independence matrix is common useful covariance matrices. In this thesis unless individual effects are included, the simulations for different models we introduce in previous sections are all defined with independence matrix. There are other choices for the covariance matrix that we introduce below.

4.5.2 First Order Autoregressive AR(1) covariance

The first order autoregressive processes, AR(1), have the following structure, recall Eq.(2.13):

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t \qquad\qquad (t = 0, \pm 1, ...)$$

where $\{u_t\}$ is a sequence of independent N $(0, \sigma^2)$ random variables and $|\rho| < 1$.
The ε_0 is distributed as N $(0, \sigma^2/(1-\rho^2))$ to proof that we express the AR(1) as

$$\varepsilon_t = \sum_{j=0}^{\infty} \rho^j u_{t-j}$$

$$\begin{aligned} \text{E}[\varepsilon_t] &= E[\sum_{j=0}^{\infty} \rho^j u_{t-j}] \\ &= \sum_{j=0}^{\infty} \rho^j E[u_{t-j}] \\ &= 0 \end{aligned}$$

and

$$\begin{aligned}
\text{Var}[\varepsilon_t] &= \text{Var}[\sum_{j=0}^{\infty} \rho^j u_{t-j}] \\
&= \sum_{j=0}^{\infty} \rho^{2j} \text{Var}[u_{t-j}] \\
&= \sigma^2 \sum_{j=0}^{\infty} \rho^{2j} \\
&= \frac{\sigma^2}{1-\rho^2}
\end{aligned}$$

Since E$[u_t]$ = 0 $\forall$ t and E $[u_s u_t]$ = 0 unless $s = t$, E$[u_t^2]$= Var$[u_t]$=σ^2.
So $\forall$t, ε_t is normal distributed with mean 0 and variance $\frac{\sigma^2}{1-\rho^2}$ by the property of the normal distribution. Then ε_0 is normal distributed with mean 0 and variance $\frac{\sigma^2}{1-\rho^2}$ by the property of the normal distribution.
So AR(1) covariance has two parameters in the matrix, ρ and σ^2. In R, we firstly random generate a number from normal distribution as the zero term in AR(1) process with zero mean and standard deviation $\frac{\sigma}{\sqrt{1-\rho^2}}$ by using *rnorm* function. Then we generate T innovation terms (u_t) by using the same R function $rnorm$ with zero mean and $Seps$ as standard deviation. Now by following Eq.(2.13) we have T+1 terms. In order to get T terms, we have to eliminate the ε_0, finally return $\boldsymbol{\varepsilon}$ as the AR(1) process simulation. For a given individual, the T number of observations follow the AR(1) structure. So we call it *ar1* function which have parameters $T,\ \rho,\ \sigma_\varepsilon$.

4.5.3 Compound Symmetry (CS) covariance

Compound symmetry (CS) is another simple form for the variance-covariance matrix. Eq.(2.12) shows the matrix form of this structure which has two parameters (σ^2 and ρ). The first parameter is the variance of the individuals and is constant across time. The parameter ρ is the correlation between any two observations from the same individual which represents the degree of association of the longitudinal data within individuals, and specifically indicates the proportion of variance in the data attributable to individuals (refer to chapter 2).

Follow the definition of compound symmetry, we have

$$\boldsymbol{\varepsilon}_i \sim CS(\rho, \sigma^2)$$

For each individual, the covariance matrix has σ^2 as the common variance; within the individual pairs of observations have the same ρ. In R, we define the diagonal matrix as $(1, \cdots, 1)$, the off diagonal elements are ρ and we define it in two parts: $lower.tri(corr)$ and $upper.tri(corr)$. Then combine them, we have the covariance matrix for multivariate normal distribution. Now we use *rmvnorm* function to generate $N \times T$ dimension of data. We called the full compound symmetry generator as *cov.cs*.

Alternatively, in Chapter 3, we prove a special case the Random intercept model with iid structure is identical to the simple regression with compound symmetry covariance, it is derive from different perspectives. We can use R to generate random intercept data for the simple regression with compound symmetry covariance model and we call the function as *comp*. For example, the model follows this perspective can be

$$Y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_{0i} + \varepsilon_{it} \tag{4.11}$$

And we generate α_{0i} with zero mean and $\rho * \sigma^2$ variance

$$\alpha_{0i} \sim N(0, \sigma^2_{\alpha_0}) \tag{4.12}$$

then we generate ε_i with zero mean and $(1-\rho) * \sigma^2$ variance

$$\varepsilon_i \sim N(0, \sigma^2_{\varepsilon}) \tag{4.13}$$

where

$$\sigma^2_{\alpha_0} = \rho\sigma^2 \tag{4.14}$$

$$\sigma^2_{\varepsilon} = (1-\rho)\sigma^2 \tag{4.15}$$

Finally, we combine two terms together $\alpha_{0i} + \varepsilon_i$. Then we gather it with other simulated values as the simple regression with compound symmetry covariance model dataset. Therefore, in this section, we introduce the way of generate random numbers with compound symmetry covariance and present a special case to create the compound symmetry structure.

4.6 Special random intercept model without correlation

In this section, we generate data of the random intercept model with same X_{i1}. This kind of data may look like there is correlation between the individual effects and explanatory variables, but actually there is no such correlation. The model is the same as the random intercept model, recall Eq.(4.5)

$$Y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_{0i} + \epsilon_{it} \tag{4.16}$$

But

$$x_{i1} \text{ is same for all individuals}$$

Now we define the generator of this special random intercept model in R. Firstly, we generate N random numbers of ε from normal distribution as $\varepsilon \sim N(0, \sigma_\varepsilon^2)$. Then we generate one random number x_{i1} from a uniform distribution with a and b as $x_{i1} \sim \text{U}(a, b)$, then repeated it N times as the starting points for each individual. We define an increment column vector for each individual, combine the column and row vector using *outer* function to create the design matrix X. Finally we generate N random numbers of α_{0i} from normal distribution as $\alpha_{0i} \sim N(\mu, \sigma_\alpha^2)$.

Follow the random intercept model as Eq.(4.5), we obtained Y_{it}. We call this R program function as **sim.RI.ss** with parameters *N, T, a, b, delta, mu, sigma, Seps, beta0, beta1*. Note: This is different compare with the fixed effects model with correlation. The difference is that we generate X and α_0 separately. So for this special model $\text{Cov}(x_{it}, \alpha_{0i}) = 0$.

Figure 4.6 shows 10 simulated individuals of random intercept model with same same x_{i1} and $\text{Cov}(x_{it}, \alpha_{0i}) = 0$ by setting the parameters as

$$N = 10,\ T = 5,\ a = -5,\ b = 5,$$

$$\delta = 10,\ \mu = 0,\ \sigma_\alpha = 1,$$

$$\sigma_\varepsilon = 1,\ \beta_0 = 0 \text{ and } \beta_1 = 1.$$

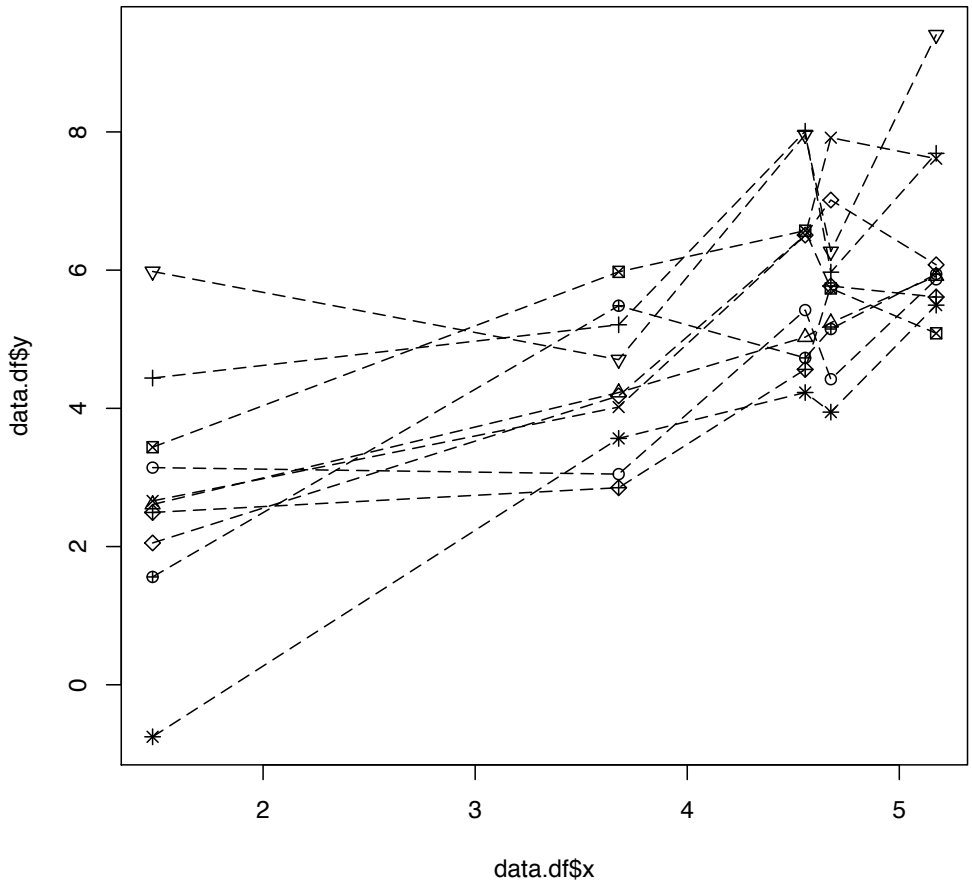

Figure 4.6: Figures of an example of a random intercept model with same x_{i1}

From Figure 4.6, we can easily see all of the individuals have the same x_{i1}. The slops of each individuals are positive and parallel on average.

4.7 R Codes

4.7.1 *sim.RE* function

```
sim.RE<-function(N,T,a,b,delta,mu,G,Seps,beta0,beta1){
            #define the matrix we need
            xM<-matrix(rep(NA,N*T),nrow=N,ncol=T)
            yM<-matrix(rep(NA,N*T),nrow=N,ncol=T)
            idM<-matrix(rep(NA,N*T),nrow=N,ncol=T)
            timeM<-matrix(rep(NA,N*T),nrow=N,ncol=T)
            gammaM0<-matrix(rep(NA,N*T),nrow=N,ncol=T)
            gammaM1<-matrix(rep(NA,N*T),nrow=N,ncol=T)
            #generate X_{i1} as start point for each individual
```

```
            xM[,1]<-runif(N,a,b)
            #generate the individual effect
            gammaM<-rmvnorm(N,mu,G)
            #generate the error term with indep covariance
            epsM<-matrix(rnorm(N*T,0,Seps),nrow=N,ncol=T)
            #define the length of observations for each individual
            D <- delta*seq(from=0, to=1, length=T)
            #follow the model eq. calculate Y_{ij}
            for(i in 1:N){
                for(j in 1:T){
                   xM[i,j]<-xM[i,1]+D[j]
                   yM[i,j]<-beta0+beta1*xM[i,j]
                      +gammaM[i,1]+gammaM[i,2]*xM[i,j]
                      +epsM[i,j]
                   idM[i,j]<-i
                   timeM[i,j]<-j
                   gammaM0[i,]<-gammaM[i,1]
                   gammaM1[i,]<-gammaM[i,2]
                }
        }
            id<-matrix(t(idM),nrow=N*T,ncol=1)
            time<-matrix(t(timeM),nrow=N*T,ncol=1)
            x<-matrix(t(xM),nrow=N*T,ncol=1)
            y<-matrix(t(yM),nrow=N*T,ncol=1)
            gamma0<-matrix(t(gammaM0),nrow=N*T,ncol=1)
            gamma1<-matrix(t(gammaM1),nrow=N*T,ncol=1)
            eps<-matrix(t(epsM),nrow=N*T,ncol=1)
            idtext<-factor(id)
            #combine values as a data frame
            data.df<-data.frame(id=id,idtext=idtext,time=time,
               x=x,y=y,gamma0=gamma0,gamma1=gamma1,eps=eps)

            return(data.df)
}

sim.cor<-function(N,T,a,b,delta,rate,Seps,Weps,beta0,beta1){
              xM<-matrix(rep(NA,N*T),nrow=N,ncol=T)
              yM<-matrix(rep(NA,N*T),nrow=N,ncol=T)
              idM<-matrix(rep(NA,N*T),nrow=N,ncol=T)
              timeM<-matrix(rep(NA,N*T),nrow=N,ncol=T)
              epsM<-matrix(rnorm(N*T,0,Seps),nrow=N,ncol=T)
```

```
        w<-rnorm(N,0,Weps)
        xM[,1]<-runif(N,a,b)
        inc<-delta*c(0,sort(runif(T-1,0,1)))
        xM<-outer(xM[,1],inc,"+")

        gammaM0<-matrix(rep(NA,N*T),nrow=N,ncol=T)
        for(i in 1:N){
           gammaM0[i,1]<-rate*mean(xM[i,])+w[i]
        }
        gammaM0<-outer(gammaM0[,1],rep(1,T))
       for(i in 1:N){
          for(j in 1:T){
           yM[i,j]<-beta0+beta1*xM[i,j]
                    +gammaM0[i,j]+epsM[i,j]
          idM[i,j]<-i
          timeM[i,j]<-j
          }
       }
       id<-matrix(t(idM),nrow=N*T,ncol=1)
       time<-matrix(t(timeM),nrow=N*T,ncol=1)
       x<-matrix(t(xM),nrow=N*T,ncol=1)
       y<-matrix(t(yM),nrow=N*T,ncol=1)
       gamma0<-matrix(t(gammaM0),nrow=N*T,ncol=1)
       eps<-matrix(t(epsM),nrow=N*T,ncol=1)
       idtext<-factor(id)
       data.df<-data.frame(id=id,idtext=idtext,
              time=time,x=x,y=y,gamma0=gamma0,eps=eps)
       return(data.df)
}
```

4.7.2 AR(1)

```
ar1<-function(T,rho,Seps){
        #generate the eps_0
        eps<-rnorm(1,0,Seps/sqrt(1-rho^2))
        #generate the T innovation terms
        innov<-rnorm(T,0,Seps)
        for(j in 1:T){
           #generate T+1 terms
```

```
        eps<-c(eps,rho*eps[j]+innov[j])
         }
         #eliminate the eps_0
         eps<-eps[-1]
         return(eps)
}
```

4.7.3 Compound Symmetry – *cov.cs* function

```
cov.cs<-function(N,sigmaCS,rho,T){
           #diagonal matrix is defined
           corr <- diag(T)
           #off diagonal is defined
           corr[lower.tri(corr)] <- sqrt(rho)
           corr[upper.tri(corr)] <- sqrt(rho)
           #covarianc matrix is defined
           sigmarho <- sqrt(sigmaCS)*corr
           #mean of multivariate normal distribution
           mean <- rep(0,T)
           #generate NT values
           rmvnorm(N,mean,sigmarho)
           }
```

4.7.4 Compound Symmetry – *comp* function

```
comp<-function(N,T,rho,sigma){
         #generate alpha with zero mean and
         #rho*sigma^2 variance
         alpha<-rnorm(N,0,sqrt(rho)*sigma)
         random<-matrix(rep(NA,N*T),nrow=N,ncol=T)
         eps<-matrix(rep(NA,N*T),nrow=N,ncol=T)
         for(i in 1:N){
         for(t in 1:T){
         #generate error term with zero mean and
         #(1-rho)*sigma^2 variance
         eps[i,t]<-rnorm(1,0,sqrt(1-rho)*sigma)
         #combine two terms together as random term
 random[i,t]<-alpha[i]+eps[i,t]
 }
 }
 return(random)
```

```
}

sim.RI.ss<-function(N,T,a,b,delta,mu,sigma,
                    Seps,beta0,beta1){
               xM<-matrix(rep(0,N*T),nrow=N,ncol=T)
               yM<-matrix(rep(NA,N*T),nrow=N,ncol=T)
               idM<-matrix(rep(NA,N*T),nrow=N,ncol=T)
               timeM<-matrix(rep(NA,N*T),nrow=N,ncol=T)
               gammaM0<-matrix(rep(NA,N*T),nrow=N,ncol=T)
               gamma<-rnorm(N,mu,sigma)
               epsM<-matrix(rnorm(N*T,0,Seps),nrow=N,ncol=T)
               xM[,1]<-rep(runif(1,a,b),N)
               inc<-delta*c(0,sort(runif(T-1,0,1)))
               xM<-outer(xM[,1],inc,"+")
               for(i in 1:N){
                   for(j in 1:T){
                       yM[i,j]<-beta0+beta1*xM[i,j]
                                +gamma[i]+epsM[i,j]
                    idM[i,j]<-i
                    timeM[i,j]<-j
                    gammaM0[i,]<-gamma[i]
                }
            }
               id<-matrix(t(idM),nrow=N*T,ncol=1)
               time<-matrix(t(timeM),nrow=N*T,ncol=1)
               x<-matrix(t(xM),nrow=N*T,ncol=1)
               y<-matrix(t(yM),nrow=N*T,ncol=1)
               gamma0<-matrix(t(gammaM0),nrow=N*T,ncol=1)
               eps<-matrix(t(epsM),nrow=N*T,ncol=1)
               idtext<-factor(id)
               data.df<-data.frame(id=id,idtext=idtext,
                     time=time,x=x,y=y,gamma0=gamma0,eps=eps)
               return(data.df)
}
```

Chapter 5

Estimation Bias

As in all statistical modelling there is a risk of model misspecification which may lead to (a) biased estimates of coefficients, and (b) biased estimates of variance components. In this chapter, we firstly show the omitted variables bias which maybe exists when we ignore the variable that is correlated with explanatory variables or maybe it is a determinant variable of the outcome variable and discuss what determines the size of this bias. Then we show theoretically and empirically the heterogeneity bias exists in particular where the correlation exists between the explanatory variables and individual effects. The bias exists under two models, the random effects model and the pooled model. But biases in the coefficient estimates can be tolerable if they are small compared to the standard errors in those coefficients. We also fit a model which use the Mundlak formulation. We empirically prove this model can provide unbiased estimates under the correlation case. Then we investigate estimation bias exists by using simulated data which are generated by using the R functions defined in chapter 4. In this chapter, we also use the Hausman test to compare the fixed effect estimator and random effect estimator, then we found the Hausman test is good to use to decide the best estimation which provide unbiased and efficient estimation.

5.1 Omitted variables bias

Omitted variables bias is due to the correlation between the outcome variables and those variables that should be include in the equation but are not, either because we ignore the variable that is correlated with explanatory variables included in the equation or because of unavailability of data.

To illustrate the omitted variables bias, suppose we have two models, full model Eq.(5.1)

which is the true model and reduced model Eq.(5.2) with X_i only, which is the omitted variable model (incorrect model with bias):

$$Y_i = \beta_0 + \beta_1 X_i + \gamma S_i + \epsilon_i \tag{5.1}$$

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{5.2}$$

where γ and β_0 and β_1 are regression parameters; S and X are random variables and the sample size is n. Now we define Eq.(5.1) and Eq.(5.2) in the matrix form as

$$\text{Full model: } \mathbf{Y} = X_f \boldsymbol{\beta}_f + \boldsymbol{\varepsilon} \tag{5.3}$$

$$\text{Omitted variable model: } \mathbf{Y} = X_r \boldsymbol{\beta}_r + \boldsymbol{\varepsilon} \tag{5.4}$$

$$\mathbf{X}_f = \begin{bmatrix} \mathbf{1} & \mathbf{x} & \mathbf{s} \end{bmatrix}; \quad \mathbf{X}_r = \begin{bmatrix} \mathbf{1} & \mathbf{x} \end{bmatrix}$$

$$\boldsymbol{\beta}_f = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \gamma \end{bmatrix}; \quad \boldsymbol{\beta}_r = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$$

The least squares estimates for both models refer to chapter 2 are given by

$$\widehat{\boldsymbol{\beta}_f} = (\mathbf{X}_f^T \mathbf{X}_f)^{-1} \mathbf{X}_f^T \mathbf{Y} \tag{5.5}$$

$$\widehat{\boldsymbol{\beta}_r} = (\mathbf{X}_r^T \mathbf{X}_r)^{-1} \mathbf{X}_r^T \mathbf{Y} \tag{5.6}$$

And we know that in truth

$$\mathbf{Y} \sim N(\mathbf{X}_f \boldsymbol{\beta}_f, \sigma^2 \mathbf{I})$$

Then using this true model, we can derive the mean and variance of estimator $\boldsymbol{\beta}_f$ and $\boldsymbol{\beta}_r$ by using Eq. (5.5) and Eq. (5.6) are

$$E\left[\widehat{\boldsymbol{\beta}_f}\right] = (\mathbf{X}_f^T \mathbf{X}_f)^{-1} \mathbf{X}_f^T E\left[\mathbf{Y}\right] = (\mathbf{X}_f^T \mathbf{X}_f)^{-1} \mathbf{X}_f^T \mathbf{X}_f \boldsymbol{\beta}_f = \boldsymbol{\beta}_f \tag{5.7}$$

$$\text{Var}(\widehat{\boldsymbol{\beta}_f}) = \hat{\sigma}^2 (\mathbf{X}_f^T \mathbf{X}_f)^{-1} \tag{5.8}$$

ie. $\widehat{\boldsymbol{\beta}_f}$ is unbiased.

For the omitted variable model we have

$$\begin{aligned}
\mathbf{X}_r^T\mathbf{X}_r &= \begin{bmatrix}\mathbf{1}^T \\ \mathbf{x}^T\end{bmatrix}\begin{bmatrix}\mathbf{1} & \mathbf{x}\end{bmatrix} \\
&= \begin{bmatrix} n & \mathbf{1}^T\mathbf{x} \\ \mathbf{x}^T\mathbf{1} & \mathbf{x}^T\mathbf{x}\end{bmatrix} \\
&= \begin{bmatrix} n & \sum X \\ \sum X & \sum X^2\end{bmatrix}
\end{aligned}$$

$$(\mathbf{X}_r^T\mathbf{X}_r)^{-1} = \frac{1}{n\sum X^2 - (\sum X)^2}\begin{bmatrix}\sum X^2 & -\sum X \\ -\sum X & n\end{bmatrix}$$

$$\begin{aligned}
\mathbf{X}_r^T\mathbf{X}_f &= \begin{bmatrix}\mathbf{1}^T \\ \mathbf{x}^T\end{bmatrix}\begin{bmatrix}\mathbf{1} & \mathbf{x} & \mathbf{s}\end{bmatrix} \\
&= \begin{bmatrix}\mathbf{1}^T\mathbf{1} & \mathbf{1}^T\mathbf{x} & \mathbf{1}^T\mathbf{s} \\ \mathbf{x}^T\mathbf{1} & \mathbf{x}^T\mathbf{x} & \mathbf{x}^T\mathbf{s}\end{bmatrix} \\
&= \begin{bmatrix} n & \sum X & \sum S \\ \sum X & \sum X^2 & \sum XS\end{bmatrix}
\end{aligned}$$

therefore,

$$\begin{aligned}
(\mathbf{X}_r^T\mathbf{X}_r)^{-1}\mathbf{X}_r^T\mathbf{X}_f &= \frac{1}{n\sum X^2 - (\sum X)^2}\begin{bmatrix}\sum X^2 & -\sum X \\ -\sum X & n\end{bmatrix}\begin{bmatrix} n & \sum X & \sum S \\ \sum X & \sum X^2 & \sum XS\end{bmatrix} \\
&= \begin{bmatrix} 1 & 0 & \dfrac{\sum X^2\sum S - \sum X\sum XS}{n\sum X^2 - (\sum X)^2} \\ 0 & 1 & \dfrac{n\sum XS - \sum X\sum S}{n\sum X^2 - (\sum X)^2}\end{bmatrix}
\end{aligned}$$

And

$$\bar{S} = \frac{1}{n}\sum S$$

$$\bar{X} = \frac{1}{n}\sum X$$

$$S_{XS} = \sum XS - \frac{1}{n}\sum X\sum S$$

$$S_{XX} = \sum X^2 - \frac{1}{n}(\sum X)^2$$

$$S_{XX}\bar{S} - S_{XS}\bar{X} = \frac{1}{n}\sum S \sum X^2 - \frac{1}{n}\sum X \sum XS$$

So the expectation and variance of estimator $\boldsymbol{\beta}_r$ are

$$\begin{aligned} E\left[\widehat{\boldsymbol{\beta}_r}\right] &= (\mathbf{X}_r^T\mathbf{X}_r)^{-1}\mathbf{X}_r^T E\left[\mathbf{Y}\right] \\ &= (\mathbf{X}_r^T\mathbf{X}_r)^{-1}\mathbf{X}_r^T\mathbf{X}_f\boldsymbol{\beta}_f \\ &= \begin{bmatrix} 1 & 0 & \dfrac{\sum X^2 \sum S - \sum X \sum XS}{n\sum X^2 - (\sum X)^2} \\ 0 & 1 & \dfrac{n\sum XS - \sum X \sum S}{n\sum S^2 - (\sum X)^2} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \gamma \end{bmatrix} \\ &= \begin{bmatrix} \beta_0 + \gamma A \\ \beta_1 + \gamma B \end{bmatrix} \end{aligned} \tag{5.9}$$

$$\begin{aligned} \text{Var}(\widehat{\boldsymbol{\beta}_r}) &= \text{Var}((\mathbf{X}_r^T\mathbf{X}_r)^{-1}X_r\mathbf{Y}) \\ &= (\mathbf{X}_r^T\mathbf{X}_r)^{-1}X_r\text{Var}(\mathbf{Y})(\mathbf{X}_r^T\mathbf{X}_r)^{-1}X_r \\ &= \hat{\sigma}^2(\mathbf{X}_r^T\mathbf{X}_r)^{-1} \end{aligned} \tag{5.10}$$

And let

$$\begin{aligned} A &= \frac{\sum S \sum X^2 - \sum X \sum XS}{n\sum X^2 - (\sum X)^2} \\ &= \frac{nS_{XX}\bar{S} - nS_{XS}\bar{X}}{nS_{XX}} \\ &= \frac{S_{XX}\bar{S} - S_{XS}\bar{X}}{S_{XX}} \end{aligned}$$

$$\begin{aligned} B &= \frac{n\sum XS - \sum X \sum S}{n\sum X^2 - (\sum X)^2} \\ &= \frac{nS_{XS}}{nS_{XX}} \\ &= \frac{S_{XS}}{S_{XX}} \end{aligned}$$

Then we can derive the bias of estimator $\begin{bmatrix} \hat{\beta}_{0r} \\ \hat{\beta}_{1r} \end{bmatrix}$ by using Eq. (5.9) minus the first two

elements of Eq. (5.7), then we have:

$$\begin{aligned}
\text{Bias}\begin{bmatrix}\hat{\beta}_{0r}\\ \hat{\beta}_{1r}\end{bmatrix} &= E[\hat{\boldsymbol{\beta}}_r] - E[\hat{\boldsymbol{\beta}}_f]_{12} \\
&= \begin{bmatrix}\beta_0 + \gamma A\\ \beta_1 + \gamma B\end{bmatrix} - \begin{bmatrix}\beta_0\\ \beta_1\end{bmatrix} \\
&= \begin{bmatrix}\gamma A\\ \gamma B\end{bmatrix} \\
&= \gamma \begin{bmatrix}\dfrac{S_{XX}\bar{S} - S_{XS}\bar{X}}{S_{XX}}\\ \\ \dfrac{S_{XS}}{S_{XX}}\end{bmatrix}
\end{aligned} \tag{5.11}$$

So there is a bias in the coefficient estimate above and the bias can be positive or negative. If $S_{XS} = 0$, that means the covariance of X and S is zero, there is no bias for the slope $\hat{\beta}_1$ estimate; the bias of $\hat{\beta}_0$ is $\bar{S}$. Now we define a simulated dataset to show the bias empirically.

5.1.1 Simulated Example

In this section and following section, we use simulated data and real data to demonstrate when the omitted variable bias exists and how it effects our estimates. We assume S to be a Bernoulli random variable with probability of success p, and X to follow a Beta(a, b) distribution conditional on S with $a(S) = \lambda(S + \frac{1}{2})$ and $b(S) = \lambda(-S + \frac{3}{2})$

$$S \sim \text{Bern}(p)$$

$$X|S \sim \text{Beta}(a, b)$$

Now we can derive the mean and variance for each distribution.

$$\text{E}[S] = \sum_s s\ P(S = s) = p$$

$$\text{E}[S^2] = \sum_s s^2\ P(S = s) = p$$

$$\text{Var}[S] = \text{E}[S^2] - (\text{E}[S])^2 = p - p^2 = p(1 - p)$$

$$\begin{aligned}
\mathrm{E}[X|S] &= \frac{\alpha(s)}{\alpha(s)+\beta(s)} \\
&= \frac{\lambda(s+\frac{1}{2})}{\lambda(s+\frac{1}{2})+\lambda(-s+\frac{3}{2})} \\
&= \tfrac{1}{2}(s+\tfrac{1}{2})
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}[X|S] &= \frac{a(s)b(s)}{(a(s)+b(s))^2(a(s)+b(s)+1)} \\
&= \frac{\lambda(s+\frac{1}{2})\times\lambda(-s+\frac{3}{2})}{(\lambda(s+\frac{1}{2})+\lambda(-s+\frac{3}{2}))^2(\lambda(s+\frac{1}{2})+\lambda(-s+\frac{3}{2})+1)} \\
&= \frac{(s+\frac{1}{2})(\frac{3}{2}-s)}{8\lambda+4} \\
&= \frac{-s^2+s+\frac{3}{4}}{8\lambda+4}
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}[XS] &= \int\sum_s xsP(S=s)f(x|s)\mathrm{d}x \\
&= \sum_s s\int xf(x|s)\mathrm{d}x\ P(S=s) \\
&= \sum_s s\mathrm{E}[X|S]\ P(S=s) \\
&= \frac{3}{4}p
\end{aligned} \tag{5.12}$$

$$\mathrm{E}[X] = \sum_s \mathrm{E}[X|S]\ P(S=s) = \frac{1}{4}+\frac{p}{2} \tag{5.13}$$

$$\mathrm{Cov}[X,S] = \mathrm{E}[XS]-\mathrm{E}[X]\mathrm{E}[S] = \frac{p}{2}(1-p) \tag{5.14}$$

By using the property of Variance

$$\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}[\mathrm{Var}[X|S]]+\mathrm{Var}[\mathrm{E}[X|S]] \\
&= \mathrm{E}[\frac{s-s^2+\frac{3}{4}}{8\lambda+4}]+\mathrm{Var}[\frac{1}{2}(s+\frac{1}{2})] \\
&= \frac{p-p+\frac{3}{4}}{8\lambda+4}+\frac{1}{4}p(1-p) \\
&= \frac{\frac{3}{4}}{8\lambda+4}+\tfrac{1}{4}p(1-p)
\end{aligned} \tag{5.15}$$

Then we have the $\mathrm{E}[X^2]$ is

$$\mathrm{E}[X^2] = \mathrm{Var}[X]+\mathrm{E}[X]^2 \tag{5.16}$$

In Figure 5.1 and Figure 5.2, we demonstrate the situation when there is and isn't bias. In Figure 5.1, we generate n samples for two groups of data ($S = 0$ or 1) with one explanatory variable X with $Cov(X, S) = \rho = \frac{p}{2}(1 - p)$ from above derivation. The two groups of data are generated in R by using function $covXS$ (R code is in section 5.5). We generate S as

$$S \sim \text{Bern}(p)$$

and X as a Beta (a, b) distribution conditional on S with $a = \lambda(S + \frac{1}{2})$ and $b = \lambda(-S + \frac{3}{2})$ The true values in this function are

$$p = 0.5,\ n = 300,\ a = 0,\ b = 1,\ \lambda = 5,$$

$$\sigma_\varepsilon = 0.5,\ \beta_0 = 0,\ \beta_1 = 1,\ \gamma = 3$$

In Figure 5.1 and Figure 5.2, the black lines are true model lines. The red lines are the full model fitted lines by using Eq.(5.1) and the green line is the omitted variable model fitted line by using Eq.(5.2). In Figure 5.1, we can see two full model fitted lines are parallel with small positive slope. The omitted variable model only has one fitted line also can indicate that has positive slope, but it is different from the true slope.

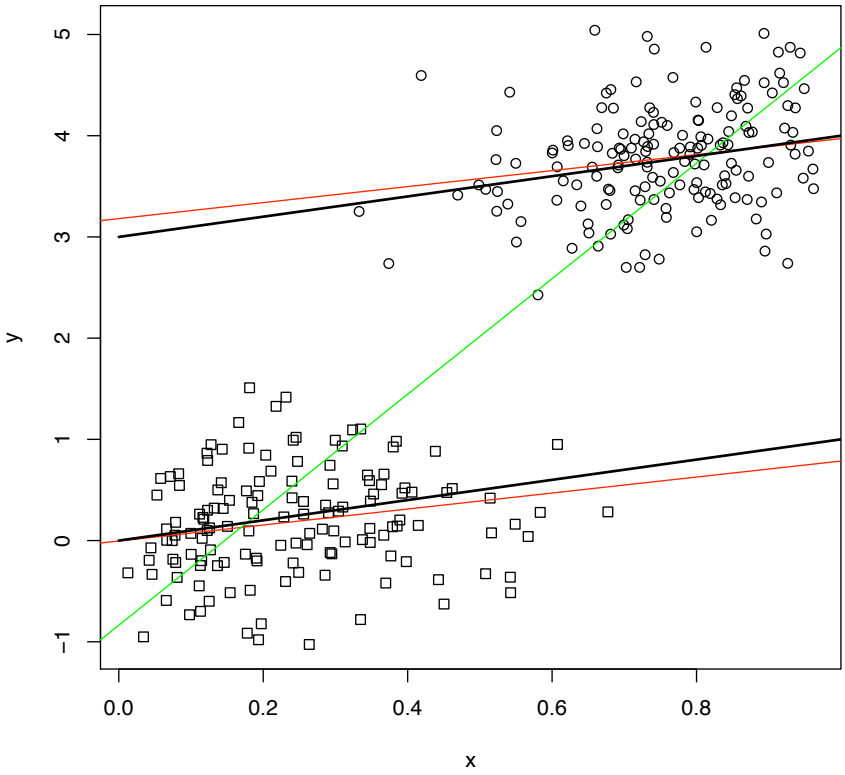

Figure 5.1: Figures of Omitted variable fitting with $Cov(X, S) > 0$

In Figure 5.2, we generate the n samples by using function $indepXS$ (refer to 5.5) without any correlation between S and X, ie. $Cov(X, S) > 0$. Here we generate S and X separately.

$$X \sim \mathrm{N}(a, b)$$

There is no bias. The green line is the omitted variable model fitted line by using Eq.(5.2) and the red lines are the full model fitted lines by using Eq.(5.1). The black lines are the true lines, the data is generated based on Eq.(5.1). From Figure 5.2, we can see there is not much difference among the true slope, the omitted variable model slope and full model slope. So there is no bias. The parameters are used to generate this data are:

$$p = 0.5,\ n = 300\ a = 0\ b = 0.1$$

$$\sigma_\varepsilon = 0.5,\ \beta_0 = 0,\ \beta_1 = 1,\ \gamma = 3$$

Compare with the correlation exists case, the R generation function just modify the way of generate X, in the no correlation case, ie. $X \sim N(a, b)$.

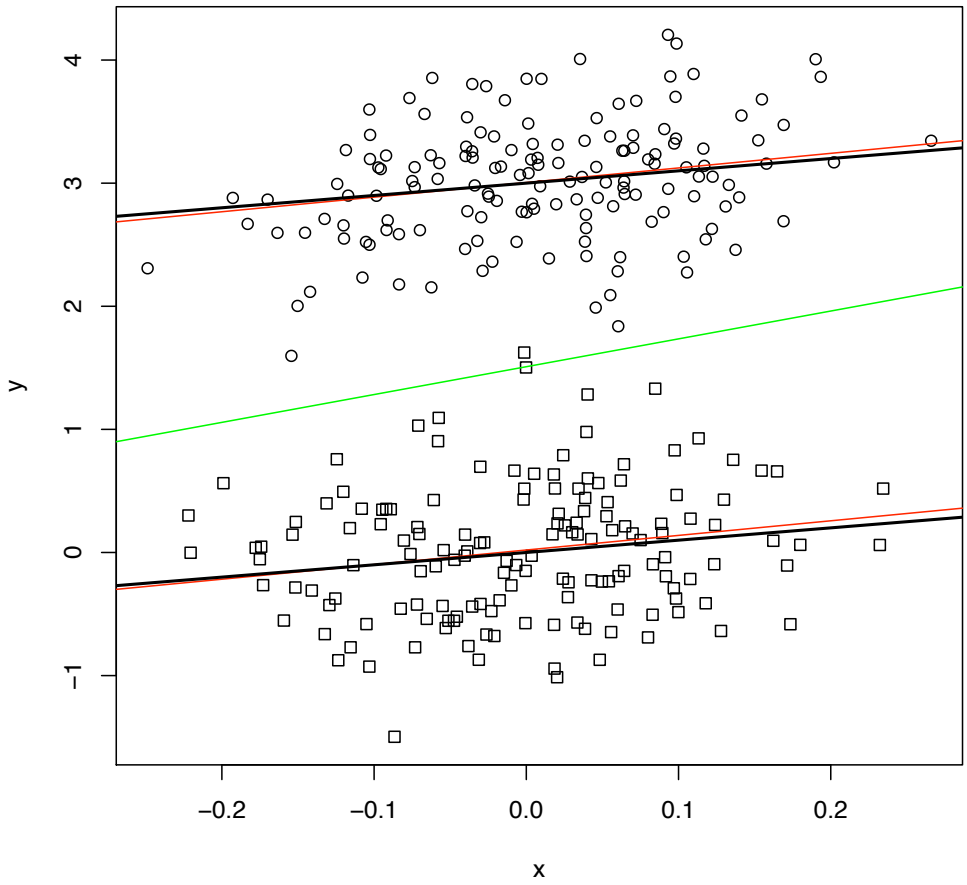

Figure 5.2: Figures of Parallel data fitting with $Cov(X, S) = 0$

Now we define *modelfit* function to follow the above derivations with parameters $n,\ p,\ \lambda, a,\ b,\ \sigma_\varepsilon,\ \beta_0$ which also return the estimates β_0 and β_1 for both full model and omitted variables model (R program of this function can be found in section 5.5). Then we repeat each sample simulation 1000 times and also store the estimates for each replication.

Finally, we calculate the mean and variance of estimator β_1 for full model and omitted variable model by calling the R function *cal* which is defined following the derivations above. The true parameters for each sample are

$$n = 100,\ p = 0.5,\ \lambda = 5,$$

$$a = 0,\ b = 1,\ \sigma_\varepsilon = 0.5,$$

$$\beta_0 = 0,\ \beta_1 = 1,\ \gamma = 3.$$

We compare results of the full model with omitted variables model based on the simulated

estimates and theoretical estimates. Figure 5.3 and Figure 5.4 show the value of the bias for β_0 and β_1 respectively. The bias is horizontal distance between the two vertical lines on the graph. There are two vertical lines for each model. One is simulated estimates and the other is theoretical mean estimates.

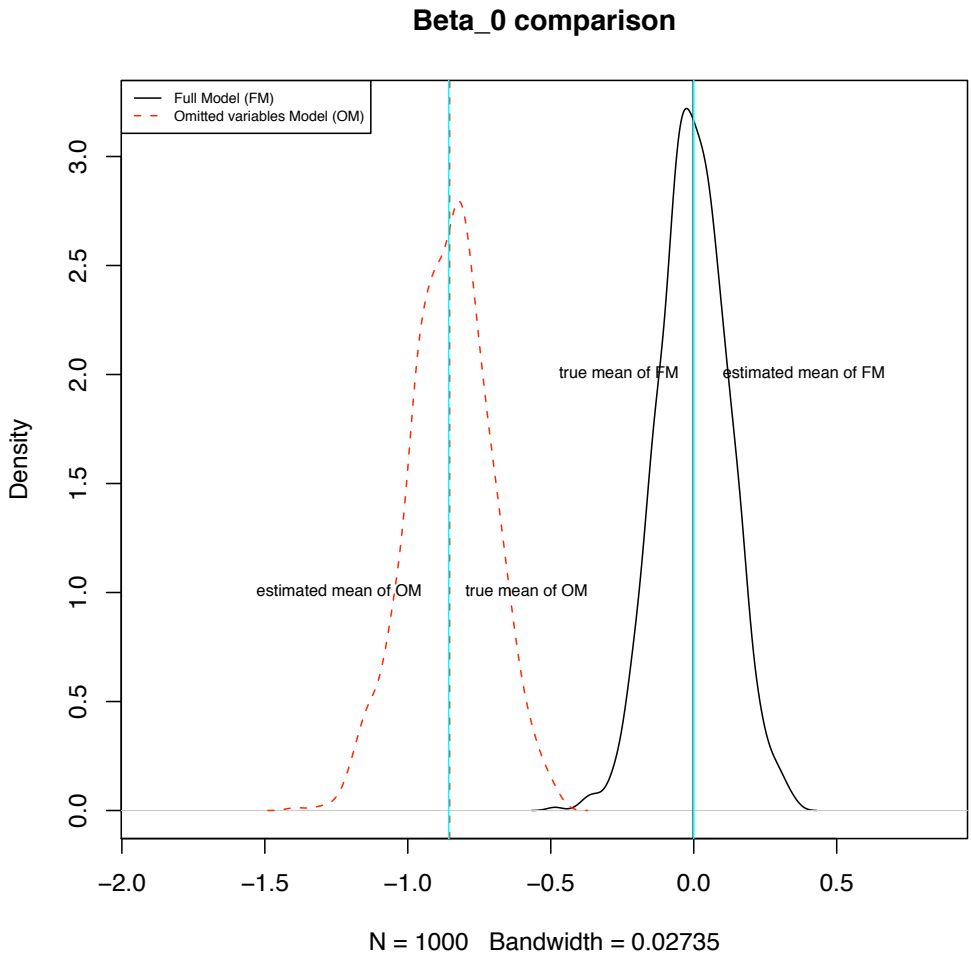

Figure 5.3: Figures of $\hat{\beta}_0$ estimates for both models

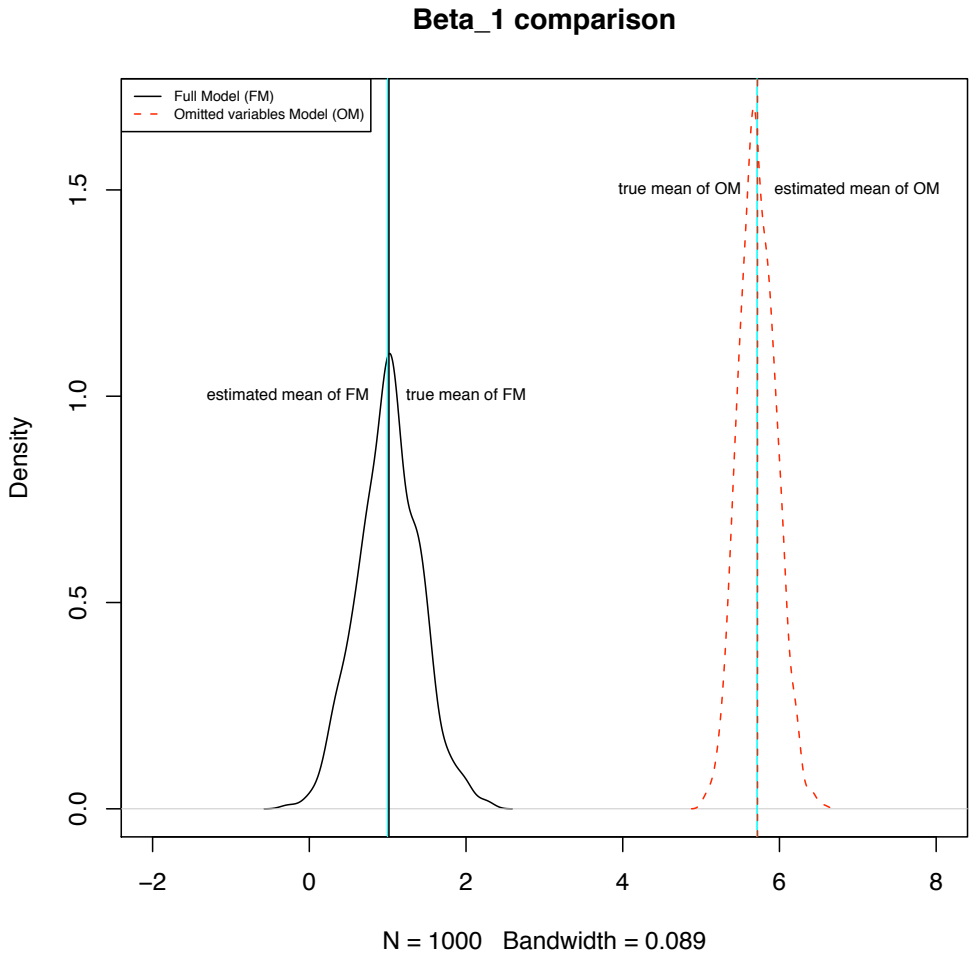

Figure 5.4: Figures of $\hat{\beta}_1$ estimates for both models

The vertical lines on both Figure 5.3 and Figure 5.4 show the estimated mean and theoretical mean for estimates (of intercept and slope) where the true mean for intercept and slope are $\beta_0 = 0$ and $\beta_1 = 1$ respectively. On Figure 5.3, we can see the intercept estimate for full model is merely equal to the true intercept value ($\beta_0 = 0$). But there is bias between the omitted variables model intercept estimate and true intercept value. The bias of intercept is

$$\text{Bias}(\widehat{\beta_0}) = -0.85$$

which is calculated following Eq.(5.11).

From Figure 5.4, we compare the slope estimate between the full model with the true slope where $\beta_1 = 1$, there is no bias between them. The simulated and theoretical estimate of full model is approximately equal to the true slope. Then we compare between the omitted

variables model and the true slope. There is significant bias, the bias is

$$\text{Bias}(\widehat{\beta}_1) = 4.72$$

which is calculated following Eq.(5.11).

Note the theoretical mean is calculated by using *cal* function is based on equations from Eq. (5.12) to Eq. (5.16) and substitute them into Eq. (5.9).

5.1.2 Real Data Example

Girth, Height and Volume for Black Cherry Trees We now demonstrate the effect of omitted variable on a real dataset. The data is sourced from R [Atkinson, 1985]. This data set provides measurements of the girth, height and volume of timber in 31 felled black cherry trees. $Girth$ is measured in inches, $Height$ in ft and $Volume$ in cubic ft. The data is given as

```
> data(trees)
> trees
   Girth Height Volume
1    8.3     70   10.3
2    8.6     65   10.3
3    8.8     63   10.2
4   10.5     72   16.4
5   10.7     81   18.8
6   10.8     83   19.7
7   11.0     66   15.6
8   11.0     75   18.2
.      .      .      .
.      .      .      .
.      .      .      .
```

Then we use function *pairs(trees)* to plot the data shows in Figure 5.5.

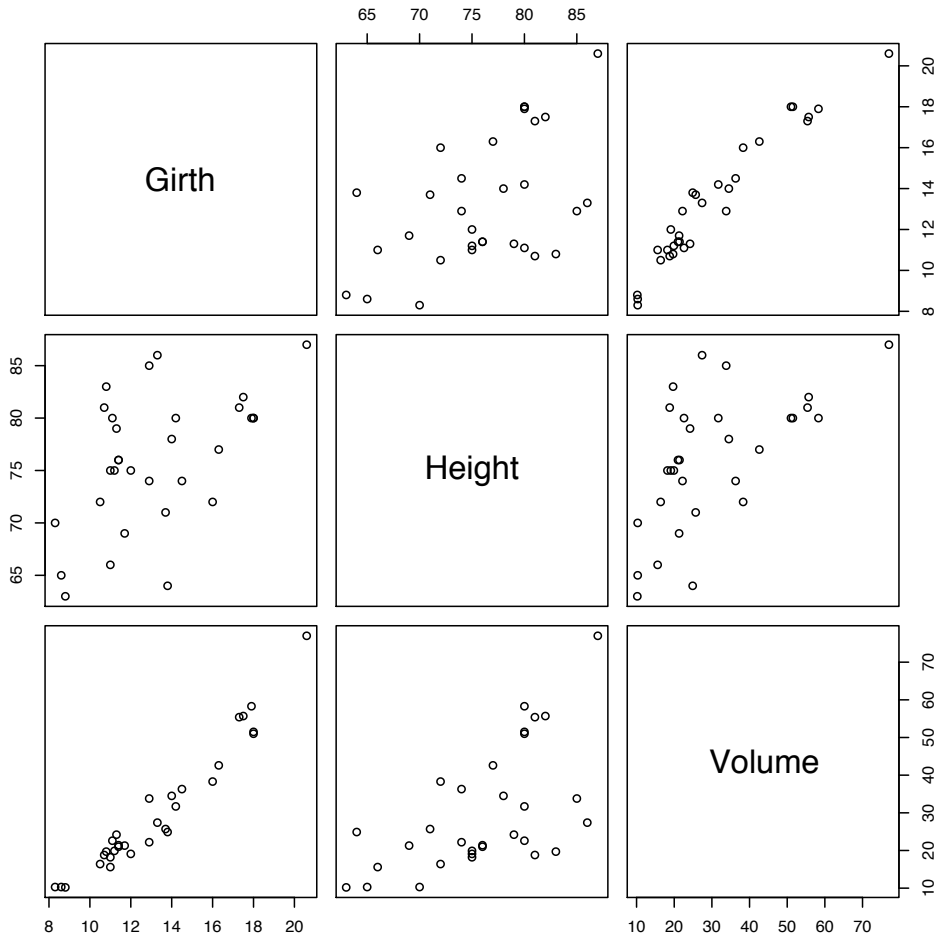

Figure 5.5: Figures of black cherry data

Figure 5.5 shows the correlations between three variables. There is strong correlation between $Girth$ and $Volume$. The correlation between $Girth$ and $Height$ is weaker.

Now we fit five simple model to black cherry trees data. The models in R are defined as

- Model 1 (M1): full model with $Girth$ and $Height$ and their interaction

```
M1 <- lm(Volume ~ Girth*Height, data=trees)
```

- Model 2 (M2): additive model with $Girth$ and $Height$

```
M2 <- lm(Volume ~ Girth+Height, data=trees)
```

- Model 3 (M3): girth only model

```
M3 <- lm(Volume ~ Girth, data=trees)
```

- Model 4 (M4): height only model

```
M4 <- lm(Volume ~ Height, data=trees)
```

- Model 5 (M5): intercept only

```
M5 <- lm(Volume ~ 1, data=trees)
```

The coefficients of the estimates for each model are shown, also with the confidence intervals for each parameter:

Table 5.1: The coefficients of the estimates for Model 1

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	69.40	23.84	2.91	0.01
Girth	-5.86	1.92	-3.05	0.01
Height	-1.30	0.31	-4.19	0.00
Girth:Height	0.13	0.02	5.52	0.00

Table 5.2: The coefficients confidence interval of the estimates for Model 1

	2.5 %	97.5 %
(Intercept)	20.49	118.30
Girth	-9.80	-1.91
Height	-1.93	-0.66
Girth:Height	0.08	0.18

Table 5.3: The coefficients of the estimates for Model 2

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-57.99	8.64	-6.71	0.00
Girth	4.71	0.26	17.82	0.00
Height	0.34	0.13	2.61	0.01

Table 5.4: The coefficients confidence interval of the estimates for Model 2

	2.5 %	97.5 %
(Intercept)	-75.68	-40.29
Girth	4.17	5.25
Height	0.07	0.61

Table 5.5: The coefficients of the estimates for Model 3

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-36.94	3.37	-10.98	0.00
Girth	5.07	0.25	20.48	0.00

Table 5.6: The coefficients confidence interval of the estimates for Model 3

	2.5 %	97.5 %
(Intercept)	-43.83	-30.06
Girth	4.56	5.57

Table 5.7: The coefficients of the estimates for Model 4

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-87.12	29.27	-2.98	0.01
Height	1.54	0.38	4.02	0.00

Table 5.8: The coefficients confidence interval of the estimates for Model 4

	2.5 %	97.5 %
(Intercept)	-146.99	-27.25
Height	0.76	2.33

Table 5.9: The coefficients of the estimates for Model 5

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	30.17	2.95	10.22	0.00

Table 5.10: The coefficients confidence interval of the estimates for Model 5

	2.5 %	97.5 %
(Intercept)	24.14	36.20

In this example, the outcome variable is $Volume$ and the predictors are $Girth$ and $Height$. In order to investigate the omitted variable bias, we assume that the model including these

two predictors (with their interaction) Model 1 is the true model, and define reduced models omitting the interaction Model 2 and simple model omitting $Height$ Model 3.

The estimates of Model 1 are $\beta_{intercept} = 69.4$ and $\beta_{Girth} = -5.86$), the estimates of Model 2 are $\beta_{intercept} = -57.99$ and $\beta_{Girth} = 4.71$) and the estimates of Model 3 are $\beta_{intercept} = -36.94$ and $\beta_{Girth} = 5.07$). The confidence intervals of three models are shown in Figure 5.6 and Figure 5.7 for both intercept and $Girth$ estimates.

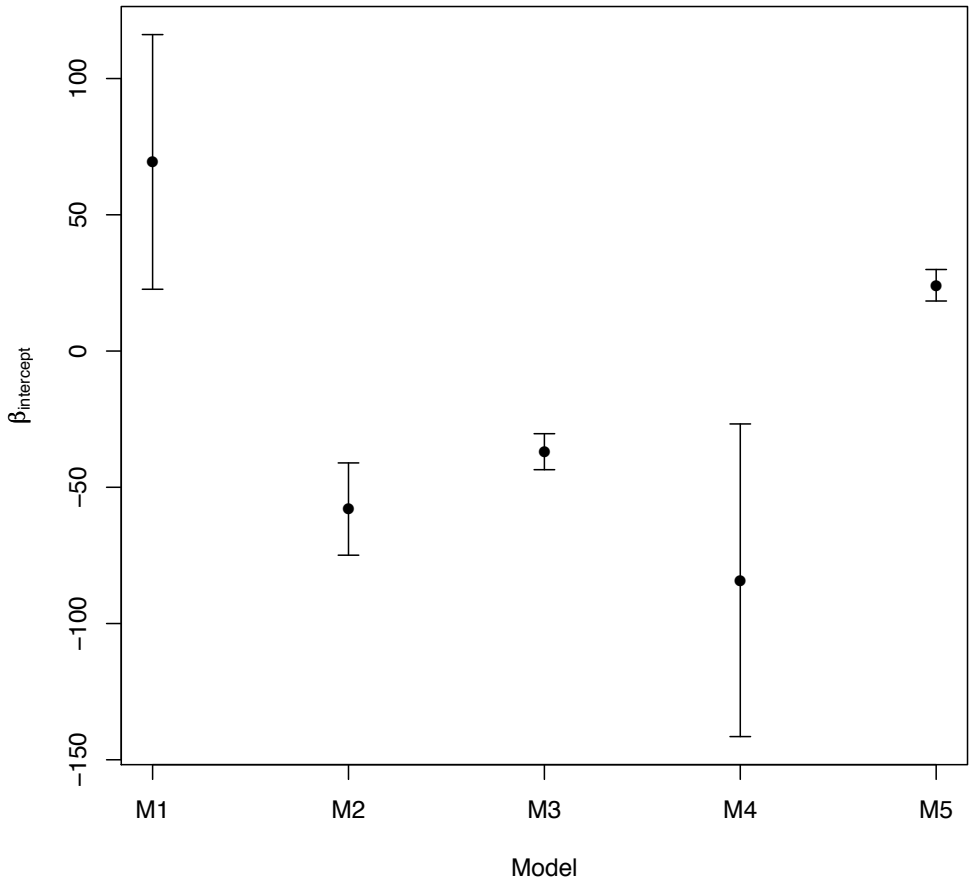

Figure 5.6: Figures of intercept estimate for Model 1, 2, 3, 4 and Model 5

Figure 5.6 shows if Model 1 is the true model, then the reduced models which omit variable $Height$ leads to a significant bias. The middle point on Figure 5.6 is the estimate of intercept for the five fitted models.

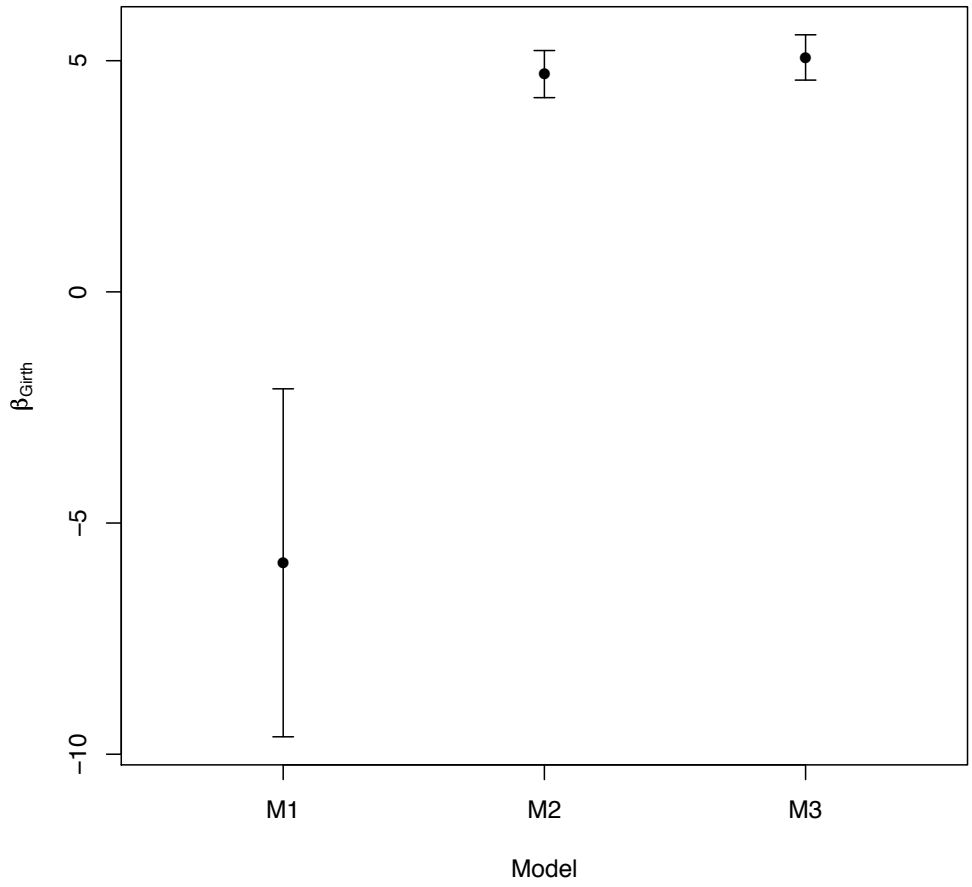

Figure 5.7: Figures of Girth estimate for Model 1, 2 and Model 3

Figure 5.7 shows slightly difference between Model 2 and Model 3. That's because the correlation between the Girth and Height is weaker. But compare with Model 1, there is significant bias. That's because the interaction term has significant effects. So we can conclude the bias can be decided by the degree of correlation between two explanatory variables.

Empirical results are often criticized on the grounds that the researcher has not explicitly recognized the effects of omitted variables that are correlated with the included explanatory variables (the omitted variable $Height$ in the black cherry and S from the simulated data which are correlated with the include variable). So the researcher should be more careful to deal with the effect of the omitted variables.

5.2 Heterogeneity Bias

In order to investigate the random effect estimation bias (heterogeneity bias), we first list the models and their estimators we have introduced in Chapter 3.

- **Pooled model**

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Pooled OLS Estimation Recall Eq. (3.6, 3.7, 3.8 and 3.9) in Chapter 3:

$$\widehat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$$

And

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}) = \widehat{\sigma}^2 (X^T X)^{-1}$$

$$\widehat{\sigma}^2 = \frac{\mathbf{e}^T \mathbf{e}}{NT - K}$$

where

$$\mathbf{e} = \mathbf{y} - X\widehat{\boldsymbol{\beta}}$$

- **Fixed effect model**

$$\mathbf{y} = X\boldsymbol{\beta} + Z\boldsymbol{\alpha} + \mathbf{u}$$

Fixed Effect Estimation Recall Eq. (3.17) in Chapter 3

$$\widehat{\boldsymbol{\beta}}_{FE} = (X^T Q X)^{-1} X^T Q \mathbf{y}$$

where

$$Q = I_{NT} - \frac{1}{T}\mathbf{Z}\mathbf{Z}^T$$

and $\mathbf{Z}$ is a set of N dummy variables (one for each individual).

$$\mathbf{Z} = \begin{bmatrix} \mathbf{1}_{T\times 1} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \mathbf{1}_{T\times 1} \end{bmatrix}_{NT\times N}$$

And Eq. (3.18 and 3.20 and 3.22) in Chapter 3 gives

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}_{FE}) = \boldsymbol{\beta}$$

$$\mathrm{Var}(\widehat{\boldsymbol{\beta}}_{FE}) = \widehat{\sigma}^2 (X^T Q X)^{-1}$$

$$\widehat{\sigma}^2 = \frac{\widehat{\mathbf{u}}^T\widehat{\mathbf{u}}}{NT - N - K}$$

where

$$\widehat{\mathbf{u}} = Q\mathbf{y} - QX\widehat{\boldsymbol{\beta}}$$

- **Random effect model**

$$\mathbf{y} = X\boldsymbol{\beta} + Z\boldsymbol{\alpha} + \mathbf{u}$$

with

$$\boldsymbol{\varepsilon} = Z\boldsymbol{\alpha} + \mathbf{u}$$

Random Effect Estimation Recall Eq. (3.34) in Chapter 3:

$$\widehat{\boldsymbol{\beta}}_{RE} = (X^T\widehat{\Omega}^{-1}X)^{-1}X^T\widehat{\Omega}^{-1}y$$

And Eq. (3.35) in Chapter 3 has

$$\text{Var}(\boldsymbol{\beta}_{RE}) = (X^T\widehat{\Omega}^{-1}X)^{-1}$$

where

$$\Omega = I_N \otimes V = E[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = \begin{bmatrix} V & 0 & \cdots & 0 \\ 0 & V & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V \end{bmatrix}$$

$$V = E[\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_i^T] = \sigma_u^2 I_T + \sigma_{\boldsymbol{\alpha}}^2 \mathbf{1}\mathbf{1}^T.$$

Therefore, the estimator $\widehat{V}$ is

$$\widehat{V} = \widehat{\sigma}_u^2 I_T + \widehat{\sigma}_{\boldsymbol{\alpha}}^2 \mathbf{1}\mathbf{1}^T$$

Recall Eq.(3.37, 3.38 and 3.42), we have

$$\widehat{\sigma^2}_u = \frac{\widehat{\mathbf{u}}_W^T\widehat{\mathbf{u}}_W}{NT - N - K}$$

$$\widehat{\sigma^2}_\alpha = \widehat{\sigma^2}_B - \frac{1}{T}\widehat{\sigma^2}_u$$

$$\widehat{\sigma}_B^2 = \frac{\widehat{\mathbf{u}}_B^T\widehat{\mathbf{u}}_B}{N - K}$$

where

$$\widehat{\mathbf{u}}_W = Q\mathbf{y} - QX\widehat{\boldsymbol{\beta}}_{RE}$$

Also, Eq.(3.40 and 3.41) give

$$\widehat{\boldsymbol{\beta}}_B = (X^T P X)^{-1} X^T P \mathbf{y}$$

$$\widehat{\mathbf{u}}_B = \overline{\mathbf{y}} - \overline{X}\widehat{\boldsymbol{\beta}_B}$$

5.2.1 Theoretical Derivation

From Chapter 3, we know the variance-covariance matrix for random effect model is

$$V = E[\boldsymbol{\epsilon}_i \boldsymbol{\epsilon}_i^T] = \sigma_u^2 I_T + \sigma_{\boldsymbol{\alpha}}^2 \mathbf{1}\mathbf{1}^T$$

Its inverse is

$$V^{-1} = \frac{1}{\sigma_u^2}[I_T - \frac{\sigma_\alpha^2}{\sigma_u^2 + T\sigma_\alpha^2}\mathbf{1}\mathbf{1}^T] \tag{5.17}$$

(see [Hsiao, 2003] for detail). We can easily verify $V^{-1}V = I$ (details can be found in Appendix B). We could rewrite the inverse of V^{-1} as

$$\begin{aligned} V^{-1} &= \frac{1}{\sigma_u^2}[I_T - (\frac{1}{T} - \frac{1}{T}\psi)\mathbf{1}\mathbf{1}^T] \\ &= \frac{1}{\sigma_u^2}[I_T - \frac{1}{T}(1-\psi)\mathbf{1}\mathbf{1}^T] \\ &= \frac{1}{\sigma_u^2}[I_T - \frac{1}{T}\mathbf{1}\mathbf{1}^T + \frac{1}{T}\psi\mathbf{1}\mathbf{1}^T] \\ &= \frac{1}{\sigma_u^2}[Q + \psi P] \end{aligned}$$

where

$$\psi = \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}$$

Now we express the Generalized Least Squares Estimator (GLSE) as a combination of two components, known as the within group estimator and the between group estimator,

$$\begin{aligned} \widehat{\beta}_{GLS} &= [X^T(Q + \psi P)X]^{-1}[X^T(Q + \psi P)\mathbf{y}] \\ &= [W_{XX} + \psi B_{XX}]^{-1}[W_{Xy} + \psi B_{Xy}] \end{aligned}$$

where

$$T_{XX} = X^T X, \quad T_{Xy} = X^T \mathbf{y}$$

$$B_{XX} = X^T P X, \quad B_{Xy} = X^T P \mathbf{y}$$

$$W_{XX} = T_{XX} - B_{XX} = X^T Q X, \quad W_{Xy} = T_{Xy} - B_{Xy} = X^T Q \mathbf{y}$$

We then define Δ and $1 - \Delta$ as

$$\Delta = [W_{XX} + \psi B_{XX}]^{-1} \psi B_{XX}$$

$$1 - \Delta = [W_{XX} + \psi B_{XX}]^{-1} W_{XX}$$

Now the GLS estimator is

$$\widehat{\beta}_{GLS} = \Delta \widehat{\beta}_B + (1 - \Delta) \widehat{\beta}_W$$

where

$$\widehat{\beta}_W = W_{XX}^{-1} W_{Xy}$$

$$\widehat{\beta}_B = B_{XX}^{-1} B_{Xy}$$

If both within group estimator and between group estimator are unbiased, then the GLSE is unbiased estimator as well. But if there is correlation between the unobserved effect α and explanatory variable X, the between group estimator will be biased. That means if the correlation exists between the individual effect and explanatory variable, the random effect estimator will be biased. Mundlak [1978] showed the random effect estimator will be biased if there is such correlation exist. To prove this (refer to Hsiao [2003]), Mundlak assumes that

$$\alpha_i = \overline{\mathbf{X}}_i^T \mathbf{a} + w_i$$

where $w_i \sim N(0, \sigma_w^2)$ and $\mathbf{a}$ is $K \times 1$ matrix and $\overline{\mathbf{X}}_i$ is $K \times 1$ matrix where K is number of explanatory variables have random effect (we assume all the explanatory variables within random effects model have random effect). Now we recall Eq. (3.29) the individual form of random intercept model

$$\mathbf{y}_i = X_i \boldsymbol{\beta} + Z_i \boldsymbol{\alpha}_i + \mathbf{u}_i$$

Then we substitute α_i into the equation to have the new formulation for random intercept model, without assuming α_i and X are uncorrelated.

$$\mathbf{y}_i = X_i\boldsymbol{\beta} + Z_i(\overline{\mathbf{X}}_i^T\mathbf{a} + w_i) + \mathbf{u}_i$$

where $\mathbf{u}_i \sim \mathrm{N}(0, \sigma_u^2 I_T)$. We express the equation as

$$\mathbf{y}_i = X_i\boldsymbol{\beta} + Z_i\overline{\mathbf{X}}_i^T\mathbf{a} + Z_i w_i + \mathbf{u}_i \tag{5.18}$$

We also can write this in stack form

$$\begin{bmatrix}\mathbf{y}_1\\ \vdots \\ \mathbf{y}_N\end{bmatrix} = \begin{bmatrix}\overline{X}_1\\ \vdots \\ \overline{X}_N\end{bmatrix}\boldsymbol{\beta} + \begin{bmatrix}\mathbf{1}\overline{\mathbf{x}}_1^T\\ \vdots \\ \mathbf{1}\overline{\mathbf{x}}_N^T\end{bmatrix}\mathbf{a} + \begin{bmatrix}\mathbf{1}\\ \vdots \\ \mathbf{0}\end{bmatrix} w_1 + \cdots + \begin{bmatrix}\mathbf{0}\\ \vdots \\ \mathbf{1}\end{bmatrix} w_N + \begin{bmatrix}\mathbf{u}_1\\ \vdots \\ \mathbf{u}_N\end{bmatrix}$$

It follows that

$$\mathrm{E}(Z_i w_i + \mathbf{u}_i) = 0$$

The new variance-covariance matrix would be

$$\begin{aligned}\widetilde{V}_{ij} &= \mathrm{E}[(Z_i w_i + \mathbf{u}_i)(Z_j w_j + \mathbf{u}_j)^T]\\ &= \mathrm{E}[\mathbf{u}_i\mathbf{u}_j^T + \mathbf{u}_i Z_i^T w_j + Z w_i \mathbf{u}_j^T + Z_i w_i w_j Z_i^T]\end{aligned}$$

If $i = j$, $\widetilde{V} = \sigma_u^2 I_T + \sigma_w^2 Z_i Z_i^T = \sigma_u^2 I_T + \sigma_w^2 \mathbf{1}\mathbf{1}^T$; If $i \neq j$, $\widetilde{V} = 0$. Using the same method as above we derive the inverse of this matrix is

$$\widetilde{V}^{-1} = \frac{1}{\sigma_u^2}[I_T - \frac{\sigma_w^2}{\sigma_u^2 + T\sigma_w^2}\mathbf{1}\mathbf{1}^T]$$

Now we can write the vector equation as the full data form

$$\mathbf{y} = X\boldsymbol{\beta} + Z\boldsymbol{\alpha} + \mathbf{u}$$

where

$$Z\boldsymbol{\alpha} = PX\mathbf{a} + P\mathbf{W}$$

We assume PX is $NT \times K$, PW is $NT \times 1$ ($\mathbf{W} = \begin{bmatrix} w_1 \\ \vdots \\ w_N \end{bmatrix}$), $P = Z(Z^TZ)^{-1}Z^T$, $PP = P = P^T = P^TP$ and $Q = I - P$ as defined in Chapter 3, and **a** is K $\times$ 1.

And the GLS estimator of random effects model is

$$\widehat{\boldsymbol{\beta}}_{GLS} = \Delta\widehat{\boldsymbol{\beta}}_B + (1-\Delta)\widehat{\boldsymbol{\beta}}_W.$$

We now separately compute the expected value of $\widehat{\boldsymbol{\beta}}_B$ and $\widehat{\boldsymbol{\beta}}_W$. The expectation value for between estimation and within estimation are

$$\begin{aligned}
\mathrm{E}[\widehat{\boldsymbol{\beta}}_B] &= \mathrm{E}[(X^TPX)^{-1}X^TP\mathbf{y} \\
&= \mathrm{E}[(X^TPX)^{-1}X^TP(X\boldsymbol{\beta} + PX\mathbf{a} + P\mathbf{W} + \mathbf{u})] \\
&= \mathrm{E}[(X^TPX)^{-1}X^TPX\boldsymbol{\beta}] + \mathrm{E}[(X^TPX)^{-1}X^TPPX\mathbf{a}] + \\
&\quad \mathrm{E}[(X^TPX)^{-1}X^TPP\mathbf{W}] + \mathrm{E}[(X^TPX)^{-1}X^TP\mathbf{u}] \\
&= \boldsymbol{\beta} + \mathrm{E}[(X^TPX)^{-1}X^TPX\mathbf{a}] \\
&= \boldsymbol{\beta} + \mathbf{a}
\end{aligned}$$

where **a** is a constant numerical vector and by assumptions W is

$$\mathrm{E}[X\mathbf{u}] = 0; \quad \mathrm{E}[X\mathbf{W}] = 0$$

$$\begin{aligned}
\mathrm{E}[\widehat{\boldsymbol{\beta}}_W] &= \mathrm{E}[(X^TQX)^{-1}X^TQ\mathbf{y} \\
&= (X^TQX)^{-1}X^TQ(X\boldsymbol{\beta} + PX\mathbf{a} + P\mathbf{W} + \mathbf{u})] \\
&= \mathrm{E}[(X^TQX)^{-1}X^TQX\boldsymbol{\beta}] + \mathrm{E}[(X^TQX)^{-1}X^TQPX\mathbf{a}] \\
&\quad +\mathrm{E}[(X^TQX)^{-1}X^TQP\mathbf{W}] + \mathrm{E}[(X^TQX)^{-1}X^TQ\mathbf{u}] \\
&= \boldsymbol{\beta} \qquad (5.19)
\end{aligned}$$

So the bias in $\widehat{\boldsymbol{\beta}}_B$ is **a** by using random effect estimation if there is correlation between α and X. Therefore, we can derive the expectation of $\widehat{\beta}_{GLS}$ as

$$\mathrm{E}[\widehat{\beta}_{GLS}] = \Delta(\boldsymbol{\beta} + \mathbf{a}) + (1-\Delta)\boldsymbol{\beta} = \boldsymbol{\beta} + \Delta\mathbf{a} \qquad (5.20)$$

Similarly, if we use the pooled estimation or OLS estimation, the pooled estimation still has bias, to see this, rewrite the pooled estimator as

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (X^T X)^{-1} X^T \mathbf{y} \\
&= (X^T X)^{-1} X^T (X\boldsymbol{\beta} + PX\mathbf{a} + P\mathbf{W} + \mathbf{u}) \\
&= (X^T X)^{-1} X^T X\boldsymbol{\beta} + (X^T X)^{-1} X^T PX\mathbf{a} \\
&\quad + (X^T X)^{-1} X^T P\mathbf{W} + (X^T X)^{-1} X^T \mathbf{u} \\
&= \boldsymbol{\beta} + (X^T X)^{-1} X^T PX\mathbf{a} + (X^T X)^{-1} X^T P\mathbf{W} + (X^T X)^{-1} X^T \mathbf{u}
\end{aligned}$$

then find the expectation of this, we have

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta} + \mathrm{E}[(X^T X)^{-1} X^T PX\mathbf{a}] \tag{5.21}$$

since

$$\mathrm{E}(Xu) = 0 \quad \mathrm{E}(XW) = 0$$

So the bias is $\mathrm{E}[(X^T X)^{-1} X^T PX\mathbf{a}]$ by using the pooled estimation if $a \neq 0$. Although the random effect estimation and pooled estimation give the bias estimation, the fixed effect estimator does provide the unbiased estimator when there is correlation between α and X. To prove this, we can use GLS method and within estimator. We write the model in full data form as

$$\mathbf{y} = \begin{bmatrix} X & PX \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{a} \end{bmatrix} + P\mathbf{W} + \mathbf{u}$$

Let $K = \begin{bmatrix} X & PX \end{bmatrix}$, then $K^T = \begin{bmatrix} X^T \\ X^T P^T \end{bmatrix}$; $\delta = \begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{a} \end{bmatrix}$ and $\varepsilon = P\mathbf{W} + \mathbf{u}$. Then

$$\mathbf{y} = K\delta + \varepsilon \tag{5.22}$$

By apply the GLS method, we have

$$\widehat{\delta} = (K^T K)^{-1} K^T \mathbf{y}$$

Hsiao [2003] shows using the expression of the inverse of a partitioned matrix (refer to Appendix A), we obtain the GLS estimator of $\boldsymbol{\beta}$ and $\mathbf{a}$ as

$$\widehat{\boldsymbol{\beta}}^{*}_{GLS} = \widehat{\boldsymbol{\beta}}_{W}$$

$$\widehat{\mathbf{a}}^{*}_{GLS} = \widehat{\boldsymbol{\beta}}_{B} - \widehat{\boldsymbol{\beta}}_{W}$$

(Details of this proof can be found in Appendix C). Alternatively, we could use the within estimation method to premultiply Eq. (5.18) by Q. We have

$$Q\mathbf{y} = QX\boldsymbol{\beta} + QPX\mathbf{a} + QP\mathbf{W} + Q\mathbf{u}$$

Since

$$QP = (I - P)P = P - P = 0$$

implies $QPX\mathbf{a} = 0$ and $QP\mathbf{W} = 0$. Now we have

$$Q\mathbf{y} = \mathbf{y} = QX\boldsymbol{\beta} + Q\mathbf{u}.$$

We obtain the estimator

$$\widehat{\boldsymbol{\beta}}^{*}_{RE} = \widehat{\boldsymbol{\beta}}_{W}$$

So the within group method and GLS method applied on Mundlak formulation provide unbiased estimates for the fixed effect model.

5.2.2 Simulated Data Deviation

Now we use simulated data to demonstrate heterogeneity bias and the size of the bias for both random effect estimation and pooled estimation, and also demonstrated that the fixed effect estimator is unbiased. We assume our model is

$$Y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_{0i} + \epsilon_{it} \tag{5.23}$$

where $i = 1, \cdots, N$ and $t = 1, \cdots T$,; β_0 and β_1 are fixed effect intercept and slope respectively. We assume $x_{i1} \sim N(0, \sigma_X^2)$. We then let the observations x_{it} (for $t = 2, \cdots, T$) be the order statistics of T-1 draws from the uniform distribution $U(x_{it}, x_{it} + \delta)$ for some δ.

α_{0i} is the only random effect for i^{th} individual, according to the Mundlak formulation

(1978), we have

$$\alpha_{0i} = \overline{\mathbf{X}}_i^T \rho + w_i$$

where ρ is a fixed numerical value and $w_i \sim N(0, \sigma_w^2)$. In R, we generate the α_{0i} without w_i (we generate in next step) by given a constant $\rho = \text{Cov}(X_i, \alpha_i)$ and calculate $\overline{X_i} = \sum_{t=1}^{T} x_{it}$. Finally, we return x_{it} and α_{0i} with individual and time specification. We call this function $newsim.cor$.

Next, we define two functions called $modelfit.mle$, this function is a fitting function which have three models being fitted: random intercept model, fixed effects model and pooled model. We use MLE to fit three different models by call the function lme from $nlme$ package and glm function in R. Note: a -1 is used in the model formula to prevent the default inclusion of an intercept term in the model. Finally, save the slopes for each model. The function is given in section 5.5.

Also, we define a function by using the least square method to calculate the slope of three different models. The estimators are listed at the beginning of this chapter. Then functions $slope.re$, $slope.fe$ and $slope.sr$ are defined based on these equations in R (details of these functions can be found in section 5.5).

We next generate simulated data sets for a fixed set of $\mathbf{X}_i$ values. For each dataset we generate u_i and w_i follow $u_i \sim N(0, \sigma_u^2)$ and $w_i \sim N(0, \sigma_w^2)$, then recall the random intercept model equation Eq. (3.3) to obtain Y_{it}. Combine the $newsim.cor$ function values with Y_{ij}, u_{it} and w_i values as the full data frame. Then we call the functions $modelfit.mle$, $slope.re$, $slope.fe$ and $slope.sr$ to calculate the slope for each model by using two method (LS method and MLE method). Then we define a new function which combine these functions together, this way we could easily call the function at once. We call this combination function *fitting.cor*. Here we assume R replications. That means we replicate function *fitting.cor* R times to get the estimates distribution for each model.

Now using the Eq. (5.20) and Eq. (5.21), we can calculate the bias theoretically as

$$\begin{aligned} \text{Bias Pooled} &= \text{E}[(X^TX)^{-1}X^TPX\boldsymbol{\rho}] \\ &= \boldsymbol{\rho} \times (X^TX)^{-1}X^TPX \end{aligned}$$

and

$$\text{Bias RE} = \Delta\rho$$

where

$$\Delta = [X^T Q X + \psi X^T P X]^{-1} \psi X^T P X$$

Since

$$\psi = \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2},$$

and

$$\begin{aligned} \sigma_\alpha^2 = \text{Var}(\alpha_{0i}) &= \text{Var}(\overline{X_i}\rho + w_i) \\ &= \rho^2 \text{Var}(\overline{X_i}) + \text{Var}(w_i) \\ &= \rho^2 \frac{\sigma_X^2}{T} + \sigma_w^2 \end{aligned}$$

Note: X_i and w_i are independent.

We substitute σ_α^2 into ψ, it becomes

$$\begin{aligned} \psi &= \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2} \\ &= \frac{\sigma_u^2}{\sigma_u^2 + T(\rho^2 \frac{\sigma_X^2}{T} + \sigma_w^2)} \\ &= \frac{\sigma_u^2}{\sigma_u^2 + \rho^2 \sigma_X^2 + T\sigma_w^2} \end{aligned}$$

In R, the calculation function is defined following these equations. Now if we assume the true parameters N, T, ρ, σ_u, σ_X, σ_w, β_0, β_1 and R to be

$$N = 20,\ T = 5,\ \rho = 1,\ \delta = 1,\ \sigma_u = 1,\ \sigma_X = 1,$$

$$\sigma_w = 1,\ \beta_0 = 0,\ \beta_1 = 1 \text{ and } R = 1000$$

The estimates distribution of three models (random effect model (RI), fixed effect model (FE) and pooled model (PL)) are shown in Figure 5.8.

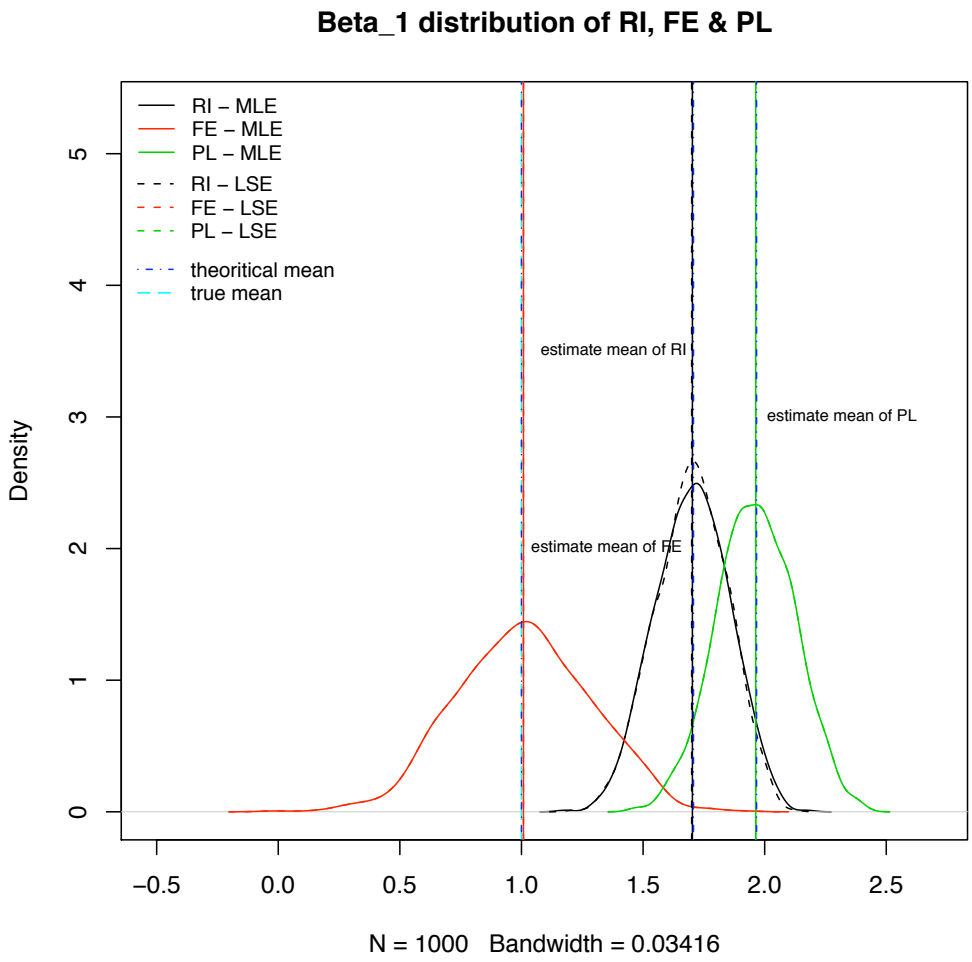

Figure 5.8: Figures of estimates comparison between RI, FE and PL by using LS estimation and MLE with correlation

Figure 5.8 shows the estimates comparison between RI, FE and PL by two different estimation method: least squares estimation (LSE) and maximum likelihood estimation (MLE). From Figure 5.8, we can see both methods give the exactly same distribution for fixed effect model and for pooled model as well, since the lines overlap. There is slightly different between MLE and LSE for random intercept (RI) model. For the estimates of RI model, the dot line (LSE) is slightly flatter than the solid line (MLE) and on the peak of the distribution the LSE is above the MLE. The reason is that by using MLE we iterate to obtain the estimation which is not a linear estimation.

From Figure 5.8, The vertical lines are the estimate mean, theoretical mean and the true mean. The actual value of true mean is $\beta_1 = 1$, and the estimates for each model (MLE and LSE are identical) are $E[\widehat{\beta}_{1_{RI}}] = 1.703$, $E[\widehat{\beta}_{1_{FE}}] = 1.008$ and $E[\widehat{\beta}_{1_{PL}}] = 1.963$. The theoretical mean are approximately equal to the mean estimates and true mean. Since the lines overlap and are the average of fitted values. There is not much difference between the theoretical

mean (from LSE and MLE) and estimated mean (from LSE and MLE) for $\widehat{\beta}_1$ random effect model and also for pooled model as well.

The random effect model and Pooled model have bias on mean estimates. It is calculated by using Eq. (5.20) and Eq. (5.21) for random effect model and pooled model respectively. We can see the biases, the numerical results from simulation match the theoretical formula exactly. The FE model gives the exactly same estimate mean (from LSE and MLE) as the true value. Its estimate mean is calculate by using Eq. (5.19).

Figure 5.9 and 5.10 are shown the estimate variance comparison between RI, FE and PL by using MLE and LSE for $\text{Cov}(X, \alpha) \neq 0$ case.

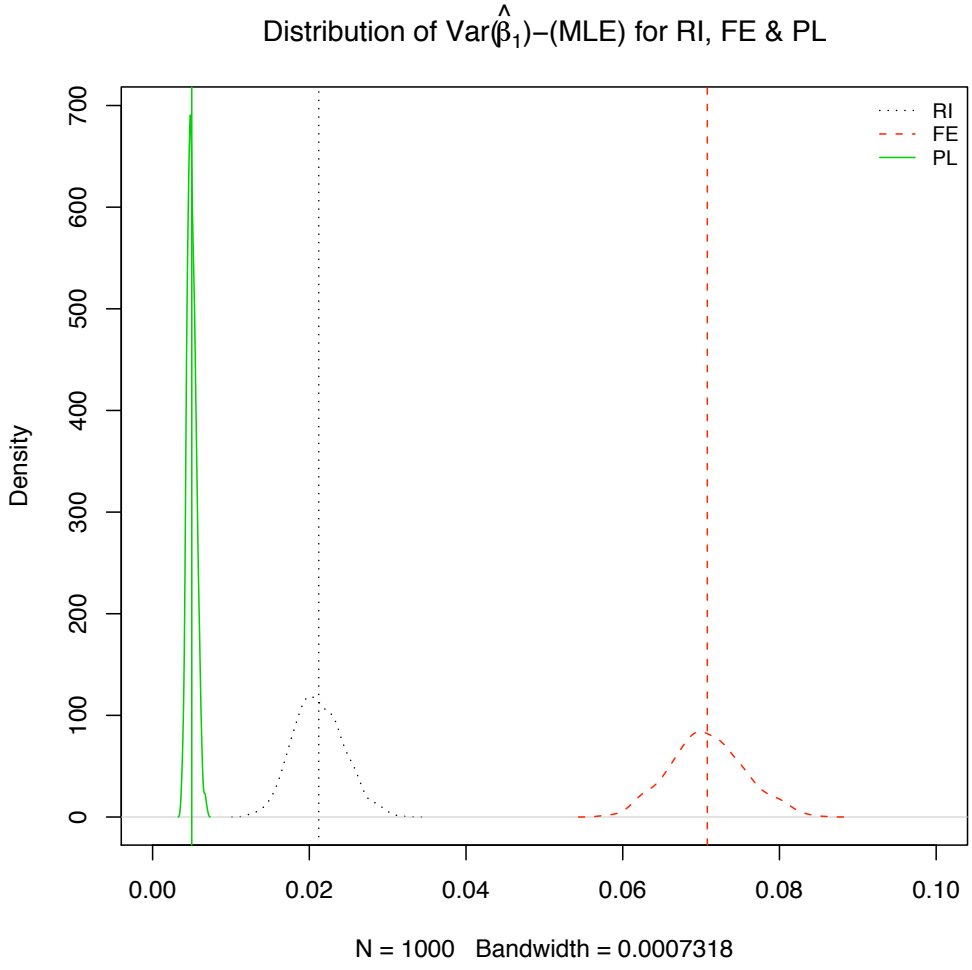

Figure 5.9: Figures of $\text{Var}(\widehat{\beta}_1)$ comparison between RI, FE and PL by using MLE and $\text{Cov}(X, \alpha) \neq 0$

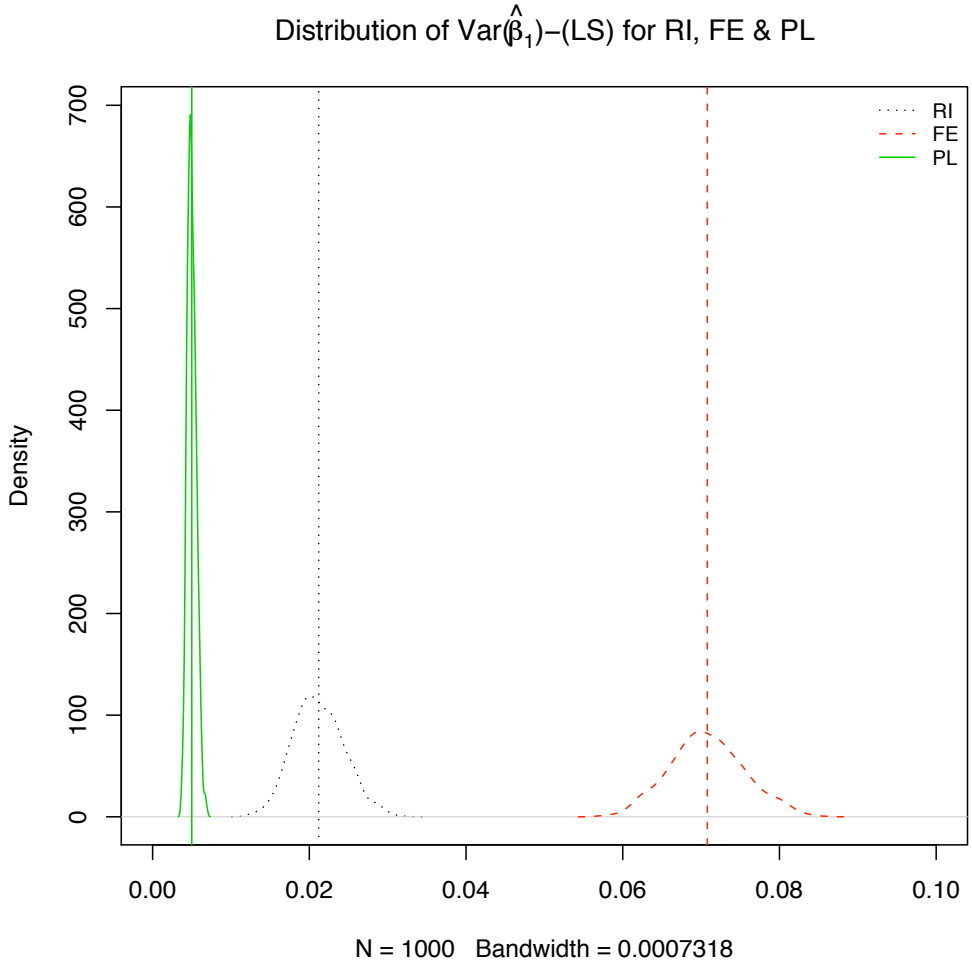

Figure 5.10: Figures of Var$(\widehat{\beta}_1)$ comparison between RI, FE and PL by using LSE and Cov$(X, \alpha) \neq 0$

Since Figure 5.9 and 5.10 are approximately identical, we only comment on Figure 5.9 below. In Figure 5.9, we see the fixed effect estimation has the highest variance, then random effect estimation, then pooled estimation. Because there is a correlation between α_i and X_i, the only appropriate model to fit is the FE model, which has a high variance Var$(\widehat{\beta}_{1_{FE}}) = 0.071$, also the RI model which severely underestimate Var$(\widehat{\beta}_{1_{RI}}) = 0.021$ and even worse the pooled model Var$(\widehat{\beta}_{1_{PL}}) = 0.005$. However, if the variance components are poorly estimated it may lead to inefficient estimation (standard errors overestimated, leading to Type II errors), or unrealistically precise estimation (standard errors underestimated, leading to Type I errors). In the simulation, we have shown the estimated variance for random effect estimate and pooled estimate are unrealistically precise estimation and the Type I error rate is increased.

Therefore, fixed effect estimation should be used as an unbiased and efficient method when there is correlation between the explanatory variable and the individual effect.

When there is no correlation, what would happen to the three models? And which estimation should be used when there is no correlation? Now we can find this answer via simulating the data by setting $rate = 0$, which gives no correlation.

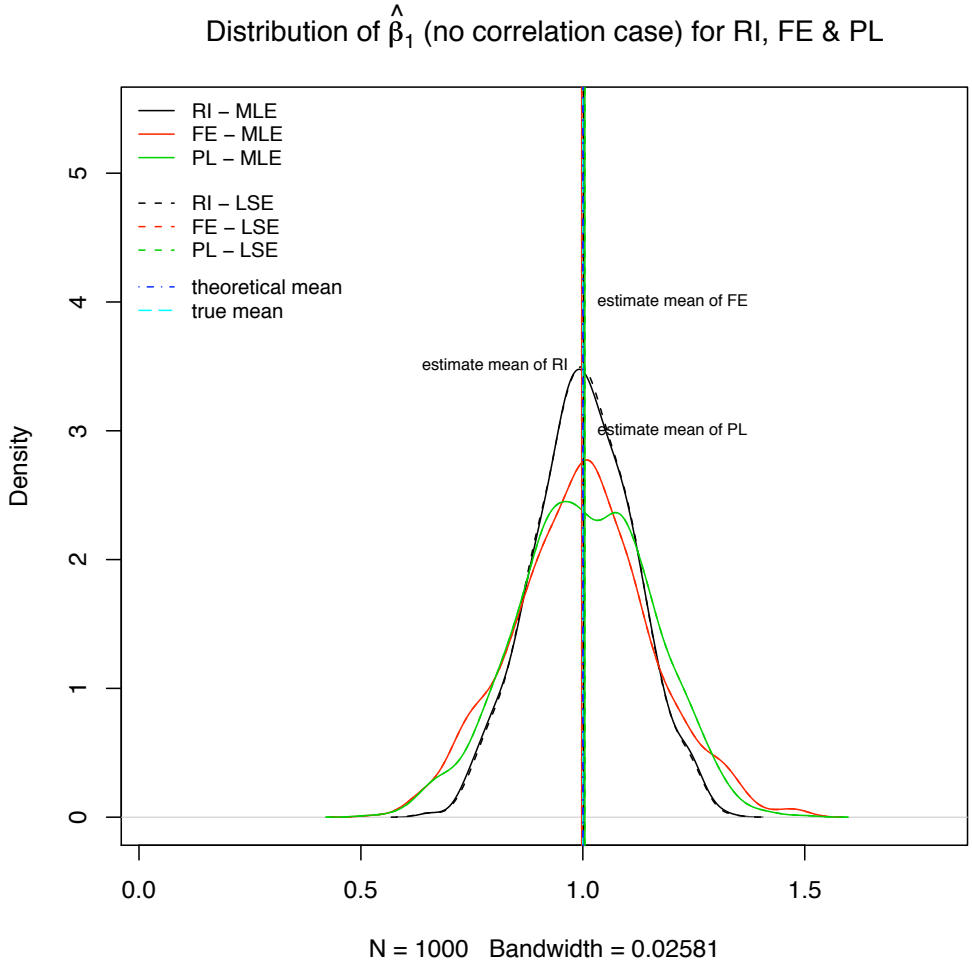

Figure 5.11: Figures of estimates comparison between RI, FE and PL by using LSE and MLE without correlation

Figure 5.11 shows the mean estimates for three model are approximately equal to the true mean, $E[\widehat{\beta}_{1_{RI}}] \approx E[\widehat{\beta}_{1_{FE}}] \approx E[\widehat{\beta}_{1_{PL}}] \approx 1 = E[\beta_{1_{true}}]$. Then the three estimators are all unbiased.

Now we compare the variances for each method. Figure 5.12 and 5.13 show the estimates variance comparison between RI, FE and PL by using MLE and LS when $\text{Cov}(X, \alpha) = 0$.

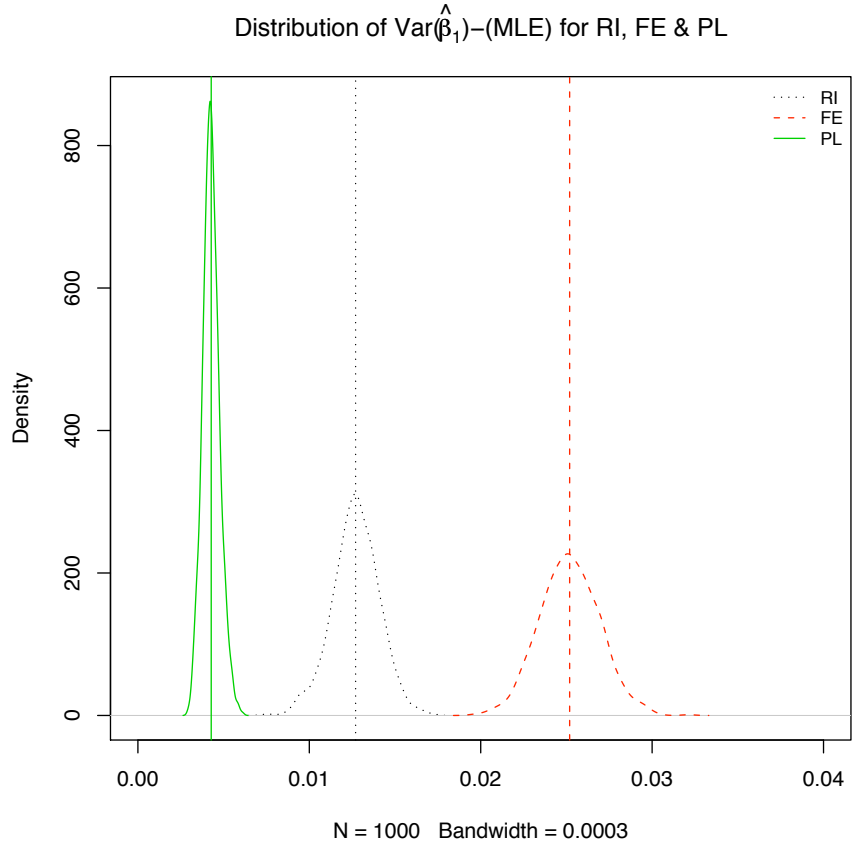

Figure 5.12: Figures of estimates variance comparison between RI, FE and PL by using MLE and Cov$(X, \alpha) = 0$

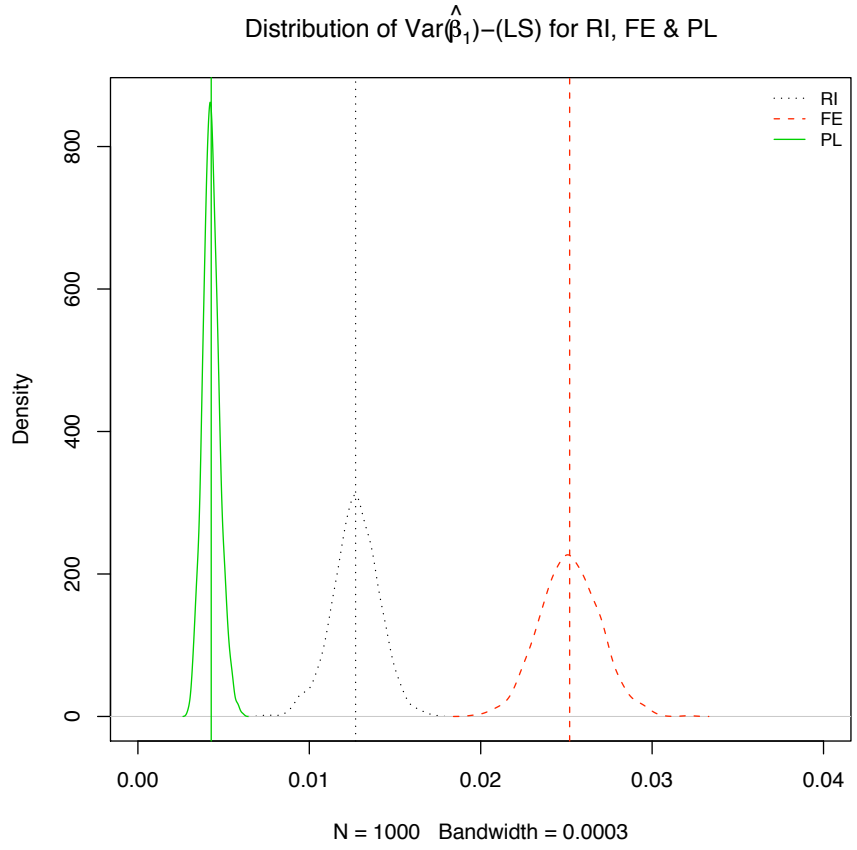

Figure 5.13: Figures of estimates variance comparison between RI, FE and PL by using MLE and Cov$(X, \alpha) = 0$

In both Figure 5.12 and 5.13, they show approximately the same distribution, so we only comment on Figure 5.12 below. In Figure 5.12, we see the fixed effect estimation has the highest variance, then random effect estimation, then pooled estimation. Because there is no correlation between α_i and X_i, the only appropriate model to fit is the RI model, which has variance Var$(\widehat{\beta}_{1_{RI}}) = 0.013$, also the PL model which severely underestimate Var$(\widehat{\beta}_{1_{PL}}) = 0.004$ and FE model which overestimate the variance Var$(\widehat{\beta}_{1_{FE}}) = 0.025$. However, if the variance components are poorly estimated it may lead to inefficient estimation (standard errors overestimated, leading to Type II errors), or unrealistically precise estimation (standard errors underestimated, leading to Type I errors). In this simulation, we have shown the estimated variance for fixed effect estimate is inefficient which gives a loss of statistical power and may lead to Type II errors. Pooled estimate is unrealistically precise estimation and may lead to Type I error.

Hence, the random effect estimation as an unbiased and efficient method should be used when there is no correlation between the explanatory variable and the individual effect. Although, the fixed effect estimator and pooled estimator are all unbiased, the random effect

estimation is the best one should be used in this case.

We summarise the empirical results as: if there are no individual effects, the pooled model gives the best fit; if there are individual effects which are correlated with explanatory variables, the fixed effects model gives the best fit; if there are individual effects, but they are not correlated with explanatory variables, the random intercept model gives the best fit.

5.3 Hausman Test on Model selection

In practice, the true model is unknown, so the question is now to select the best model. In order to decide which is the appropriate model, especially between random effect model and fixed effect model for longitudinal data, we could turn the question to identify whether there is correlation exist between explanatory variables and individual effect. The Hausman test is the common test to use in this case (discussed in section 3.4 and more discussion about Hausman test [**?**] can be found in Cameron and Trivedi [2005]), as it can used to select the unbiased and efficient estimation method. So it can identify the correlation. No correlation means the random effect estimation is unbiased and efficient, while correlation means the fixed effect method is unbiased. In this section, we not only empirically and theoretically prove the correlation exist, but we also demonstrate how the correlation affects the choice of random effect and fixed effect estimation. For the two models, one with and one without correlation, we generate 1000 replicate datasets. We then fit the FE and RI models for each replicate, and compute the Hausman test statistic (by using Eq. (3.45)). Our expectation is the Hausman test is sensitive enough to detect the correlation between explanatory variables and individual effects.

Figure 5.14 and Figure 5.15 show the distribution of Hausman statistic for with correlation and without correlation cases where the degree of the correlation can be expression as $\rho = \text{Cov}(\alpha_i, \overline{X}_i)$.

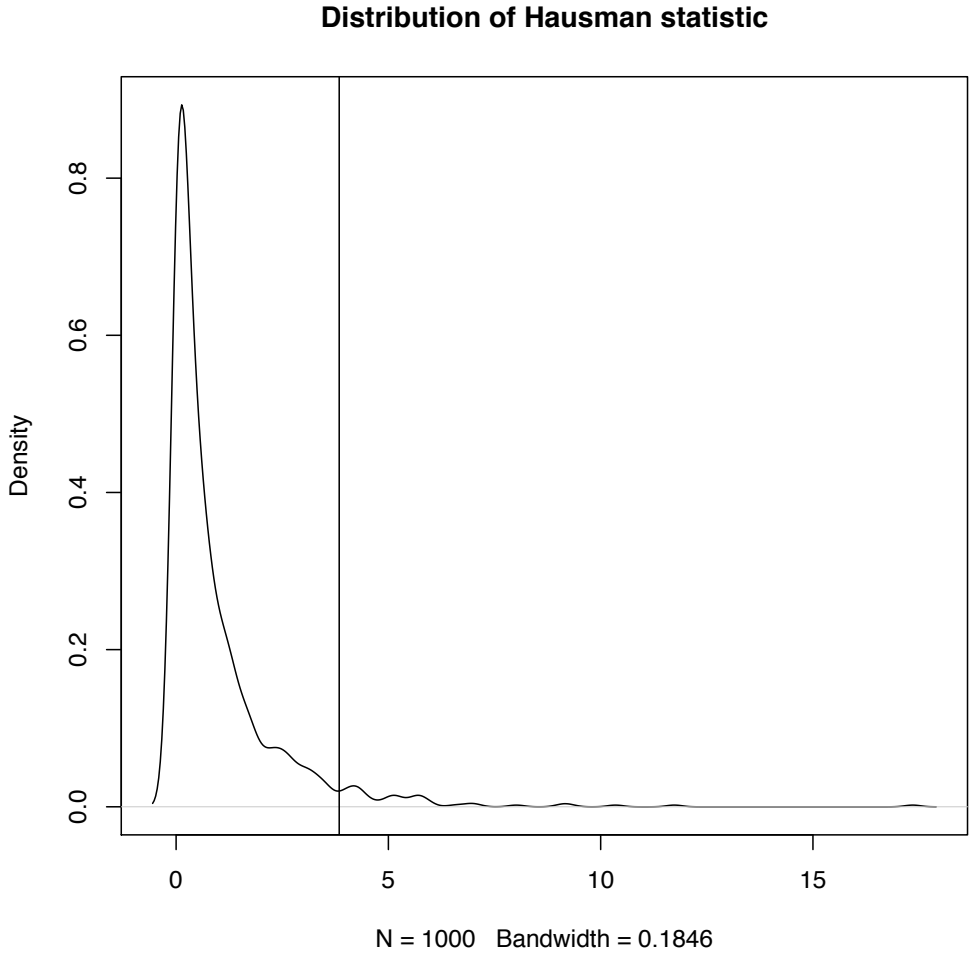

Figure 5.14: Figures of distribution of Hausman statistic - no correlation $\rho = 0$

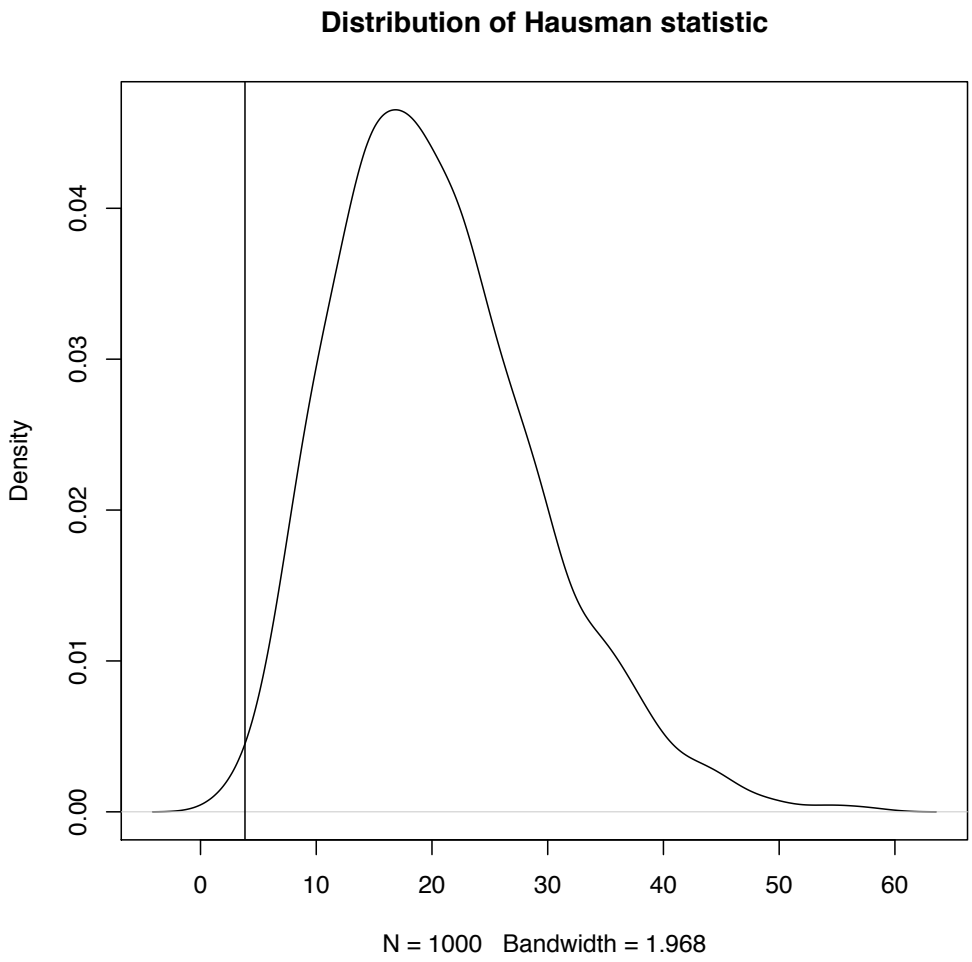

Figure 5.15: Figures of distribution of Hausman statistic - with correlation (ie. $\rho = 1$)

Since we only have one variable in the model, the degrees of freedom of the Hausman statistic is 1, and the corresponding critical value for 5% significance is 3.841. We draw this critical value on Figure 5.14 and Figure 5.15 as a vertical line. Figure 5.14 is the distribution of the Hausman statistic for the case where there is no correlation, the $\rho = 0$ (ρ is a scalar to measure the degree of the correlation). In Figure 5.14, 95.6 % of the distribution is on the left hand side of the vertical line for a 5 % of significance level. The area under the distribution of left hand side is the probability (or proportion) of the acceptance (accept H_0 rate), if this probability is high, that means we accept H_0 a lot for no correlation case. That indicates the random effect estimation should be used. Figure 5.15 shows the opposite situation. The proportion of acceptance is low, since most of the distribution area is on the right hand side of the vertical line. So when there is correlation, we are more likely to reject the H_0 than accept. In this case, we should use fixed effect estimation.

Note:

- Figure 5.14 shows there is the Type I error on right hand side of the vertical line, which 4.4 % of chance we fail to accept H_0.
- Figure 5.15 shows there is Type II error on the left hand side of the vertical line, which caused by 0.4 % we fail to reject H_0.

To investigate how the correlation rate (or the degree of the correlation) effects the proportion of acceptance, we repeat the simulation for each case 100 times with 100 different rate value from 0 to 1 in order. For each rate, we do the same calculation as before, we calculate the test statistic H for single data set and find the proportion of acceptance (calculate by using the number of $H < 3.841$ out of 1000 simulation). So far, for each rate, we have a proportion of acceptance as its probability (the highest probability is 95 % and lowest is 0%). We show the ρ vs. proportion of acceptance on Figure 5.16.

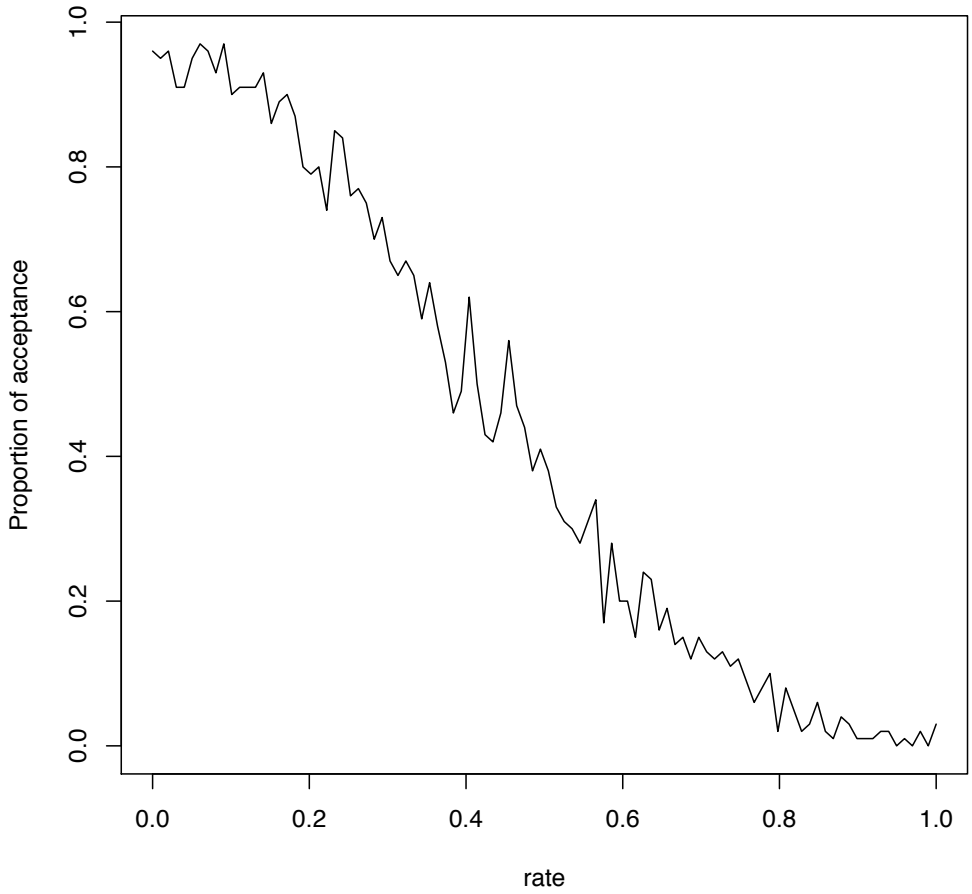

Figure 5.16: Figures of rate vs. proportion of acceptance

Figure 5.16 shows when the ρ is increasing, the proportion of acceptance is decreasing. Although the proportions fluctuate, it shows a consistent downward trend. It also suggest

that for $\rho < 0.1$ that random effect estimator is still ok and for even small correlation the Hausman test is sensitive enough to detect this correlation. Now we use simulated data and real data to demonstrate how Hausman test works. In the following examples, we only use likelihood approach to estimate, because we have proved least squares based approach produces the similar estimates as likelihood based approach for RE model and produces exactly same estimates as likelihood based approach for FE and PL model. We also fit a new model called Mundlak formulation model (MF Model), because we proved it is unbiased and is a special case of the RE model. Hence, the MF estimates should be unbiased for both correlation and none correlation case.

5.3.1 Simulated Data Example

The simulation data sets we use in this section is generated by using the R functions defined in Chapter 4. There are two types of datasets, one has no the correlation between individual effect α and explanatory variable X_{ij} called **RINOCOR** which is randomly generated by using *sim.RE* function and setting the true parameters for **RINOCOR** dataset as

$$N = 20,\ T = 4,\ a = -10,\ b = 10,$$

$$\delta = 3,\ \boldsymbol{\mu} = \begin{bmatrix} 0 & 0 \end{bmatrix},\ G = \begin{bmatrix} 0.6 & 0 \\ 0 & 0 \end{bmatrix},$$

$$\sigma_\varepsilon = 1,\ \beta_0 = 0 \text{ and } \beta_1 = 1.$$

The other is with such correlation called **RICOR** which is randomly generated by using *sim.cor* function and setting the true parameters for **RICOR** dataset as

$$N = 20,\ T = 5,\ a = -5,\ b = 5,$$

$$\sigma_w = 1,\ \delta = 1$$

$$\sigma_\varepsilon = 1,\ \beta_0 = 0 \text{ and } \beta_1 = 1.$$

The candidate models are

1. Random intercept model (RE Model):

$$Y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_{0i} + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma_{\varepsilon}^2)$$

$$\alpha_{0i} \sim \mathrm{N}(0, \sigma_{\alpha}^2)$$

Where $i = 1 \cdots N$ and $t = 1 \cdots T$, for $N = 20$ and $T = 4$; β_0 is intercept. In R, we fit RE Model as

```
RE <-lme(fixed = y ~ x, random = ~ 1 | idtext, data=data.df)
```

The estimates for RE Model are obtained by using R for two datasets and are listed as

Table 5.11: Estimates $\widehat{\beta}_k$ of RE model for **RINOCOR** dataset, $k = 0, 1$

	Value	Std.Error	DF	t-value	p-value
(Intercept)	0.072	0.180	59.000	0.400	0.691
x	1.011	0.033	59.000	30.394	0.000

Table 5.12: Estimates Var$(\widehat{\beta}_k)$ of RE model for **RINOCOR** dataset, $k = 0, 1$

	(Intercept)	x
(Intercept)	0.074	-0.018
x	-0.018	0.039

The estimates of **RINOCOR**: σ_{ε} is 1.022 and σ_{α} is 0.620.

Table 5.13: Estimates $\widehat{\beta}_k$ of RE model for **RICOR** dataset, $k = 0, 1$

	Value	Std.Error	DF	t-value	p-value
(Intercept)	0.246	0.272	79.000	0.905	0.368
x	1.594	0.198	79.000	8.061	0.000

Table 5.14: Estimates Var$(\widehat{\beta}_k)$ of RE model for **RICOR** dataset, $k = 0, 1$

	(Intercept)	x
(Intercept)	0.074	-0.018
x	-0.018	0.039

The estimates of **RICOR**: σ_{ε} is 1.063 and σ_{α} is 0.964.

2. Fixed effect model (FE Model)

$$Y_{it} = \beta_1 x_{it} + \alpha_{0i} + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$$

α_{0i} are fixed constant

Where $i = 1 \cdots N$ and $t = 1 \cdots T$, for $N = 20$ and $T = 5$. Here we define the simplest FE model only. No correlation with X_i is assumed, but we may detect this. The correlations case will be introduced by using Mundlak formulation. In R, we fit FE Model as

```
FE <- glm(y ~ x + idtext, data=data.df)
```

The estimates for FE Model are obtained by using R for two datasets and are listed as

Table 5.15: Estimates $\widehat{\beta}_1$ of FE model for **RINOCOR** dataset

	Estimate	Std. Error	t-value	p-value
x	1.016	0.103	9.865	0.000

The estimates of **RINOCOR**: σ_ε is 1.061 and Var($\widehat{\beta}_{1_{FE}}$)= 0.011.

Table 5.16: Estimates $\widehat{\beta}_1$ of FE model for **RICOR** dataset

	Estimate	Std. Error	t-value	p-value
x	1.018	0.312	3.260	0.002

The estimates of **RICOR**: σ_ε is 0.898 and Var($\widehat{\beta}_{1_{FE}}$)= 0.098.

3. Pooled model (PL Model)

$$Y_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$$

Where $i = 1 \cdots N$ and $t = 1 \cdots T$, for $N = 20$ and $T = 4$; β_0 is intercept. In R, we fit PL Model as

```
PL <-glm(y ~ x, data=data.df)
```

The estimates for PL Model are obtained by using R for two datasets and are listed as

Table 5.17: Estimates $\widehat{\beta}_k$ of PL model for **RINOCOR** dataset, $k = 0, 1$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.072	0.133	0.543	0.589
x	1.011	0.025	39.914	0.000

Table 5.18: Estimates Var$(\widehat{\beta}_k)$ of PL model for **RINOCOR** dataset, $k = 0, 1$

	(Intercept)	x
(Intercept)	1.76e-02	-2.31e-04
x	-2.31e-04	6.41e-04

The estimates of **RINOCOR**: σ_ε is 1.401.

Table 5.19: Estimates $\widehat{\beta}_k$ of PL model for **RICOR** dataset, $k = 0, 1$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.116	0.150	0.772	0.442
x	1.881	0.130	14.480	0.000

Table 5.20: Estimates Var$(\widehat{\beta}_k)$ of PL model for **RICOR** dataset, $k = 0, 1$

	(Intercept)	x
(Intercept)	0.022	-0.008
x	-0.008	0.017

The estimates of **RICOR**: σ_ε is 1.896.

4. Mundlak Formulation (MF Model)

$$Y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + \varepsilon_{it}$$

$$\alpha_{0i} = \rho \overline{X}_i + w_i$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$$

$$w_i \sim \mathrm{N}(0, \sigma_w^2)$$

Where $i = 1 \cdots N$ and $t = 1 \cdots T$, for $N = 20$ and $T = 4$; β_0 is intercept; ρ is scalar. In R, we fit MF Model as

```
MF <-lme(fixed = y ~ x+barX, random = ~ 1 | idtext, data=data.df)
```

The estimates for MF Model are obtained by using R for two datasets and are listed as

Table 5.21: Estimates $\widehat{\beta}_k$ and ρ of MF model for **RINOCOR** dataset, $k = 0, 1$

	Value	Std.Error	DF	t-value	p-value
(Intercept)	0.072	0.181	59.000	0.400	0.691
x	1.016	0.103	59.000	9.865	0.000
barX	-0.005	0.109	18.000	-0.049	0.961

Table 5.22: Estimates Var$(\widehat{\beta}_k)$ of MF model for **RINOCOR** dataset, $k = 0, 1$

	(Intercept)	x
(Intercept)	3.26e-02	-2.71e-18
x	-2.71e-18	1.06e-02

The estimates of **RINOCOR**: σ_ε is 1.030 and σ_w is 0.620.

Table 5.23: Estimates $\widehat{\beta}_k$ and a of MF model for **RICOR** dataset, $k = 0, 1$

	Value	Std.Error	DF	t-value	p-value
(Intercept)	0.081	0.268	79.000	0.302	0.764
x	1.018	0.312	79.000	3.260	0.002
barX	0.939	0.394	18.000	2.383	0.028

Table 5.24: Estimates Var$(\widehat{\beta}_k)$ of MF model for **RICOR** dataset, $k = 0, 1$

	(Intercept)	x
(Intercept)	7.16e-02	-2.43e-17
x	-2.43e-17	9.76e-02

The estimates of **RICOR**: σ_ε is 1.007 and σ_w is 0.948.

Now we have to decide which estimator give the unbiased estimation and efficient estimation between random effect estimator and fixed effect estimator. We use Hausman test Eq.(3.4) to compare the random effect estimator and fixed effect estimator.

- H_0: There is no correlation between the individual effect and the explanatory variable;
- H_a: There is correlation between the individual effect and the explanatory variable.

The test statistic is

$$H = (\widehat{\beta}_{RE} - \widehat{\beta}_{FE})^T (\Sigma_{FE} - \Sigma_{RE})^{-1} (\widehat{\beta}_{RE} - \widehat{\beta}_{FE}) \sim \chi_1^2$$

where Σ_{FE} is the covariance of $\widehat{\beta}_{FE}$ and Σ_{RE} is the covariance of $\widehat{\beta}_{RE}$. We use S_{RE}^2 to represent the estimated covariance for the random effect estimator and S_{FE}^2 to represent the estimated covariance for the fixed effect estimator.

In the **RINOCOR** case, the true value of the β_1 is 1. $\widehat{\beta}_{RE} = 1.011$ and SE$(\widehat{\beta}_{RE}) = 0.033$; $\widehat{\beta}_{FE} = 1.015$ and SE$(\widehat{\beta}_{FE}) = 0.103$; $\widehat{\beta}_{PL} = 1.011$ and SE$(\widehat{\beta}_{PL}) = 0.025$ and $\widehat{\beta}_{MF} = 1.016$ and SE$(\widehat{\beta}_{MF}) = 0.103$, all of estimates of β_1 give the close estimates to the true value. That means they are all unbiased. There is not very much difference between fixed effect estimate and random effect estimate. This indicates there is no correlation between the individual effects and the explanatory variable. Now we use Hausman test to apply on this dataset.

In **RINOCOR** case, the test statistic is 0.003 with 1 degree of freedom (we only have one variable here), thus P$(H > 0.003) = 0.959$, so we accept H_0, there is no correlation between X and α at 5% significance level. Hausman test identify the random effect estimator is appropriate estimation to use for **RINOCOR** data.

Note: Our **RINOCOR** data is generated by random effect model, therefore, the Hausman test gives the correct conclusion.

In the **RICOR** case, the true value of the β_1 is 1. $\widehat{\beta}_{RE} = 1.594$ and SE$(\widehat{\beta}_{RE}) = 0.198$; $\widehat{\beta}_{FE} = 1.018$ and SE$(\widehat{\beta}_{FE}) = 0.312$; $\widehat{\beta}_{PL} = 1.881$ and SE$(\widehat{\beta}_{PL}) = 0.130$ and $\widehat{\beta}_{MF} = 1.018$ and SE$(\widehat{\beta}_{MF}) = 0.312$. The fixed effect estimate is close to the true value. The MF estimate is the same as the fixed effect estimate. So the FE estimator and MF estimator give unbiased estimates. RE and PL estimates are different from true value, so they produce biased estimates as proved before. The fixed effect estimate is unbiased and there is a difference between the fixed effect estimate and the random effect estimate. This also indicates there is correlation between the individual effects and the explanatory variable. Now we use Hausman test to confirm this finding.

In **RICOR** case, The test statistic is 4.762 with 1 degree of freedom, thus $\mathrm{P}(H > 4.762) = 0.029$, so we reject H_0 at 5% significance level, there is correlation between X and α. Hausman test confirm that the fixed effect estimator is appropriate estimation to use for **RICOR** data.

Note: Our **RICOR** data is generated by MF model, MF model give the same estimate as the fixed effect model. Therefore, the Hausman test gives the correct conclusion. And we should conclude both MF and FE estimators are appropriate estimation use for **RICOR** data.

5.3.2 Real Data Example

In this section, we apply the random effect estimation and fixed effect estimation on **WAGE** dataset (this dataset can be obtain from Wooldridge [2009]) and compare the estimates. Then we could choose an appropriate estimator and also we use the Hausman test to confirm the result. Therefore, the Hausman test no only can compare the estimator, but also can indicate whether there is correlation between the explanatory variables and the individual effects. The data are sourced from the National Longitudinal Survey held in USA. There are 545 full-time working males who have completed their education by 1980 and follow over the period until 1987. The males in the sample with an age in 1980 ranging from 17 to 23 and entered the labour market recently, with an average of 3 years of experience. The data and specifications we define is the same as in Wooldridge [2009]. Log wages are explained by years of education, years of experience and its square, dummy variables for being a union member, working in the public sector and being married and two racial dummies (the variables in the model are selected same as on Verbeek [2004]). The models we fit are given as

- Random intercept Model (RE)

$$\begin{aligned} \text{lwage}_{it} &= \beta_0 + \beta_1 \text{educ}_{it} + \beta_2 \text{exper}_{it} + \beta_3 \text{expersq}_{it} + \beta_4 \text{union}_{it} \\ &\quad + \beta_5 \text{married}_{it} + \beta_6 \text{black}_{it} + \beta_7 \text{hisp}_{it} \\ &\quad + \beta_8 \text{pub}_{it} + \alpha_i + \varepsilon_{it} \end{aligned} \tag{5.24}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma_\epsilon^2)$$

$$\alpha_i \sim \mathrm{N}(0, \sigma_\alpha^2)$$

where $k = 0, 1 \cdots K$, $i = 1 \cdots N$ and $t = 1 \cdots T$ for $K = 8$, $N = 545$ and $T = 8$. In R, we fit RE Model as

```
RE <- lme(fixed =lwage~educ+black+hisp
          +exper+expersq+union+married+pub,
          random = ~ 1 | nr,na.action=na.omit
          data=males)
```

The estimates for RE Model are obtained by using R and are listed as

Table 5.25: Estimates $\widehat{\beta}_k$ of RE model for **WAGE** dataset, $k = 0, 1 \cdots 8$

	Value	Std.Error	DF	t-value	p-value
(Intercept)	-0.105	0.112	3810.000	-0.931	0.352
educ	0.101	0.009	541.000	11.152	0.000
black	-0.144	0.048	541.000	-2.975	0.003
hisp	0.020	0.043	541.000	0.456	0.648
exper	0.112	0.008	3810.000	13.556	0.000
expersq	-0.004	0.001	3810.000	-6.877	0.000
union	0.106	0.018	3810.000	5.910	0.000
married	0.062	0.017	3810.000	3.695	0.000
pub	0.030	0.036	3810.000	0.832	0.405

Table 5.26: Estimates Var$(\widehat{\beta}_k)$ of RE model for **WAGE** dataset, $k = 0, 1 \cdots 8$

	(Intercept)	educ	black	hisp	exper	expersq	union	married	pub
(Intercept)	1.26e-02	-9.79e-04	-6.21e-04	-1.28e-03	-1.42e-04	5.16e-06	-7.47e-05	1.14e-04	1.40e-04
educ	-9.79e-04	8.21e-05	2.76e-05	7.96e-05	-4.45e-06	5.96e-07	7.22e-07	-8.04e-06	-8.90e-06
black	-6.21e-04	2.76e-05	2.34e-03	3.51e-04	-9.13e-06	2.73e-07	-5.16e-05	6.48e-05	5.64e-06
hisp	-1.28e-03	7.96e-05	3.51e-04	1.88e-03	4.20e-06	-3.55e-07	-1.83e-05	4.40e-06	-1.88e-05
exper	-1.42e-04	-4.45e-06	-9.13e-06	4.20e-06	6.82e-05	-4.65e-06	-9.97e-07	-3.13e-05	-1.50e-05
expersq	5.16e-06	5.96e-07	2.73e-07	-3.55e-07	-4.65e-06	3.49e-07	1.88e-07	1.10e-06	5.14e-07
union	-7.47e-05	7.22e-07	-5.16e-05	-1.83e-05	-9.97e-07	1.88e-07	3.19e-04	-9.65e-06	-4.11e-05
married	1.14e-04	-8.04e-06	6.48e-05	4.40e-06	-3.13e-05	1.10e-06	-9.65e-06	2.82e-04	-1.00e-05
pub	1.40e-04	-8.90e-06	5.64e-06	-1.88e-05	-1.50e-05	5.14e-07	-4.11e-05	-1.00e-05	1.33e-03

The estimates σ_ε is 0.351 and σ_α is 0.332.

- Fixed effects Model (FE Model)

$$\begin{aligned} \text{lwage}_{it} &= \beta_1 \text{exper}_{it} + \beta_2 \text{expersq}_{it} + \beta_3 \text{union}_{it} \\ &\quad + \beta_4 \text{married}_{it} + \beta_5 \text{hisp}_{it} \\ &\quad + \beta_6 \text{pub}_{it} + \alpha_i + \varepsilon_{it} \end{aligned} \tag{5.25}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma_\epsilon^2)$$

where $k = 1 \cdots K$, $i = 1 \cdots N$ and $t = 1 \cdots T$ for $K = 8$, $N = 545$ and $T = 8$. In R, we fit FE Model as

```
FE<-glm(lwage~-1+nr+educ+exper+expersq
        +union+married+black+hisp+pub,
        data=males)
```

The estimates for FE Model are obtained by using R and are listed as

Table 5.27: Estimates $\widehat{\beta}_k$ of FE model for **WAGE** dataset, $k = 2, 3, 4, 5, 8$

	Estimate	Std. Error	t value	Pr(>|t|)
exper	0.116	0.008	13.813	0.000
expersq	-0.004	0.001	-7.083	0.000
union	0.081	0.019	4.204	0.000
married	0.045	0.018	2.463	0.014
pub	0.035	0.039	0.905	0.366

Table 5.28: Estimates Var$(\widehat{\beta}_k)$ of FE model for **WAGE** dataset, $k = 2, 3, 4, 5, 8$

	exper	expersq	union	married	pub
exper	7.11e-05	-4.86e-06	1.12e-08	-3.75e-05	-1.66e-05
expersq	-4.86e-06	3.67e-07	1.32e-07	1.34e-06	5.26e-07
union	1.12e-08	1.32e-07	3.73e-04	-9.01e-06	-3.77e-05
married	-3.75e-05	1.34e-06	-9.01e-06	3.35e-04	-8.42e-06
pub	-1.66e-05	5.26e-07	-3.77e-05	-8.42e-06	1.49e-03

Table 5.29: F-test of FE model for **WAGE** dataset

	Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
NULL			4359	1236.53	
nr	544	664.48	3815	572.05	0.0000
educ	0	0.00	3815	572.05	
exper	1	91.80	3814	480.25	0.0000
expersq	1	6.99	3813	473.26	0.0000
union	1	2.30	3812	470.96	0.0000
married	1	0.76	3811	470.20	0.0134
black	0	0.00	3811	470.20	
hisp	0	0.00	3811	470.20	
pub	1	0.10	3810	470.10	0.3657

Note: Table 5.29 shows no estimated effect of $educ$, $black$ and $hisp$. The fixed effect estimator eliminates the time invariant variables from the model (ie. $educ$, $black$ and $hisp$), so FE Model can't detect time invariant variables. The estimates σ_ε is 0.123

- Pooled Model (PL Model)

$$\begin{aligned} \text{lwage}_{it} &= \beta_0 + \beta_1 \text{educ}_{it} + \beta_2 \text{exper}_{it} + \beta_3 \text{expersq}_{it} + \beta_4 \text{union}_{it} \\ &\quad + \beta_5 \text{married}_{it} + \beta_6 \text{black}_{it} + \beta_7 \text{hisp}_{it} \\ &\quad + \beta_8 \text{pub}_{it} + \varepsilon_{it} \end{aligned} \tag{5.26}$$

$$\varepsilon_{it} \sim \text{N}(0, \sigma_\varepsilon^2)$$

where $k = 0, 1 \cdots K$, $i = 1 \cdots N$ and $t = 1 \cdots T$ for $K = 8$, $N = 545$ and $T = 8$; β_0 is the intercept. In R, we fit PL Model as

```
PL<-glm(lwage~educ+exper+expersq
        +union+married+black+hisp+pub,
        data=males)
```

The estimates for PL Model are obtained by using R and are listed as

Table 5.30: Estimates $\widehat{\beta}_k$ of PL model for **WAGE** dataset, $k = 0, 1 \cdots 8$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.034	0.065	-0.531	0.595
educ	0.099	0.005	21.219	0.000
exper	0.089	0.010	8.807	0.000
expersq	-0.003	0.001	-4.023	0.000
union	0.180	0.017	10.451	0.000
married	0.108	0.016	6.853	0.000
black	-0.144	0.024	-6.104	0.000
hisp	0.016	0.021	0.752	0.452
pub	0.004	0.037	0.095	0.925

Table 5.31: Estimates Var($\widehat{\beta}_k$) of PL model for **WAGE** dataset, $k = 0, 1 \cdots 8$

	(Intercept)	educ	exper	expersq	union	married	black	hisp	pub
(Intercept)	4.18e-03	-2.59e-04	-2.41e-04	1.01e-05	-5.21e-05	1.01e-04	-9.38e-05	-2.96e-04	1.32e-04
educ	-2.59e-04	2.19e-05	-5.63e-06	7.50e-07	1.15e-06	-7.23e-06	4.17e-06	1.75e-05	-7.91e-06
exper	-2.41e-04	-5.63e-06	1.02e-04	-6.86e-06	-7.18e-06	-2.64e-05	-9.06e-06	6.60e-06	-1.62e-05
expersq	1.01e-05	7.50e-07	-6.86e-06	5.01e-07	5.93e-07	8.79e-07	2.53e-07	-5.26e-07	7.26e-07
union	-5.21e-05	1.15e-06	-7.18e-06	5.93e-07	2.96e-04	-1.37e-05	-4.86e-05	-1.69e-05	-6.66e-05
married	1.01e-04	-7.23e-06	-2.64e-05	8.79e-07	-1.37e-05	2.47e-04	5.76e-05	4.39e-06	-1.76e-05
black	-9.38e-05	4.17e-06	-9.06e-06	2.53e-07	-4.86e-05	5.76e-05	5.55e-04	8.34e-05	7.54e-06
hisp	-2.96e-04	1.75e-05	6.60e-06	-5.26e-07	-1.69e-05	4.39e-06	8.34e-05	4.33e-04	-1.89e-05
pub	1.32e-04	-7.91e-06	-1.62e-05	7.26e-07	-6.66e-05	-1.76e-05	7.54e-06	-1.89e-05	1.40e-03

The estimates σ_ε is 0.231.

- Mundlak formulation Model (MF Model)

$$\begin{aligned} \text{lwage}_{it} &= \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_{it} + \beta_3 \text{expersq}_{it} + \beta_4 \text{union}_{it} \\ &\quad + \beta_5 \text{married}_{it} + \beta_6 \text{black}_i + \beta_7 \text{hisp}_i \\ &\quad + \beta_8 \text{pub}_{it} + \alpha_i + \varepsilon_{it} \end{aligned} \tag{5.27}$$

$$\varepsilon_{it} \sim \text{N}(0, \sigma_\varepsilon^2)$$

$$w_i \sim \text{N}(0, \sigma_w^2)$$

$$\begin{aligned} \alpha_i &= \rho_1 \overline{\text{exper}}_i + \rho_2 \overline{\text{expersq}}_i + \rho_3 \overline{\text{union}}_i \\ &\quad + \rho_4 \overline{\text{married}}_i + \rho_5 \overline{\text{pub}}_i + w_i \end{aligned} \tag{5.28}$$

Note: here we wipe out the time invariant variables in Eq. (5.28), because they are time invariant variables, the mean of its variable is just itself .

where $k = 0, 1 \cdots K$, $i = 1 \cdots N$, $p = 1 \cdots P$ and $t = 1 \cdots T$ for $P = 5$, $K = 8$, $N = 545$ and $T = 8$; β_0 is the intercept. In R, we fit MF Model as

```
MF<-lme(fixed = lwage~educ+exper+expersq+union+married
            +black+hisp+pub+barexper+barexpersq+barunion
            +barmarried+barpub, random = ~ 1 | nr, data=males)
```

Note: $barexper$, $barexpersq$, $barunion$, $barmarried$ and $barpub$ are the mean for each variable over time.

The estimates for ML Model are obtained by using R and are listed as

Table 5.32: Estimates $\widehat{\beta}_k$ and ρ_p of ML model for **WAGE** dataset, $k = 0, 1 \cdots 8$ and $p = 1, \cdots, 5$

	Value	Std.Error	DF	t-value	p-value
(Intercept)	0.490	0.221	3810.000	2.217	0.027
educ	0.095	0.011	536.000	8.682	0.000
exper	0.116	0.008	3810.000	13.813	0.000
expersq	-0.004	0.001	3810.000	-7.083	0.000
union	0.081	0.019	3810.000	4.204	0.000
married	0.045	0.018	3810.000	2.463	0.014
black	-0.139	0.049	536.000	-2.845	0.005
hisp	0.005	0.043	536.000	0.128	0.898
pub	0.035	0.039	3810.000	0.905	0.366
barexper	-0.167	0.051	536.000	-3.263	0.001
barexpersq	0.009	0.003	536.000	2.873	0.004
barunion	0.193	0.051	536.000	3.792	0.000
barmarried	0.099	0.045	536.000	2.204	0.028
barpub	-0.091	0.116	536.000	-0.789	0.431

Table 5.33: Estimates Var$(\widehat{\beta}_k)$ of ML model for **WAGE** dataset, $k = 0, 1 \cdots 8$

	(Intercept)	educ	exper	expersq	union	married	black	hisp	pub
(Intercept)	4.89e-02	-1.52e-03	1.20e-17	-6.62e-19	-2.42e-18	-1.27e-18	1.69e-04	-1.56e-03	1.00e-18
educ	-1.52e-03	1.19e-04	-4.08e-19	1.82e-20	-1.46e-20	-6.67e-20	1.77e-06	6.71e-05	4.07e-20
exper	1.20e-17	-4.08e-19	7.11e-05	-4.86e-06	1.12e-08	-3.75e-05	4.51e-21	-4.29e-19	-1.66e-05
expersq	-6.62e-19	1.82e-20	-4.86e-06	3.67e-07	1.32e-07	1.34e-06	2.16e-21	2.54e-20	5.26e-07
union	-2.42e-18	-1.46e-20	1.12e-08	1.32e-07	3.73e-04	-9.01e-06	2.61e-20	2.92e-20	-3.77e-05
married	-1.27e-18	-6.67e-20	-3.75e-05	1.34e-06	-9.01e-06	3.35e-04	-9.89e-20	-1.62e-20	-8.42e-06
black	1.69e-04	1.77e-06	4.51e-21	2.16e-21	2.61e-20	-9.89e-20	2.39e-03	3.57e-04	3.39e-21
hisp	-1.56e-03	6.71e-05	-4.29e-19	2.54e-20	2.92e-20	-1.62e-20	3.57e-04	1.83e-03	2.71e-20
pub	1.00e-18	4.07e-20	-1.66e-05	5.26e-07	-3.77e-05	-8.42e-06	3.39e-21	2.71e-20	1.49e-03

The estimates σ_ε is 0.351 and σ_w is 0.325.

Now we have to decide which estimator give the unbiased estimation and efficient estimation between random effect estimator and fixed effect estimator. In our case, the variables $educ$, $black$ and $hisp$ are cancelled out by using fixed effect model. Thus, we only need to compare the remaining five variables that are time varying in both models. Theoretically, if the zero correlation assumption is satisfied, then random effect estimator and fixed effect estimator should present the same results which are unbiased. The random effect estimator is more efficient (with smaller variance). If there is correlation, the fixed effect estimator produces unbiased estimator, and the random effect estimator is biased. We can see the estimates in Table 5.27, 5.28, 5.25 and 5.26 that they are significant different, eg. the estimates

of $married$ for fixed effect estimator and random effect estimator are

$$\widehat{\beta}_{RE} = 0.062, \quad \text{SE}(\widehat{\beta}_{RE}) = 0.017$$

$$\widehat{\beta}_{FE} = 0.045, \quad \text{SE}(\widehat{\beta}_{FE}) = 0.018$$

we can see that

$$\widehat{\beta}_{RE} \neq \widehat{\beta}_{FE}$$

and

$$\text{SE}(\widehat{\beta}_{RE}) < \text{SE}(\widehat{\beta}_{FE})$$

So this indicates the fixed effect estimator should be used in this case.

Also, we use Hausman test introduced in section 3.4 to do the test in order to confirm the conclusion we made based on the Tables.

- H_0: There is no correlation between the individual effect and the explanatory variable;
- H_a: There is correlation between the individual effect and the explanatory variable.

The test statistic is

$$H = (\widehat{\beta}_{RE} - \widehat{\beta}_{FE})^T (\Sigma_{FE} - \Sigma_{RE})^{-1} (\widehat{\beta}_{RE} - \widehat{\beta}_{FE}) \ \sim \ \chi^2_5$$

where Σ_{FE} is the covariance of $\widehat{\beta}_{FE}$ and Σ_{RE} is the covariance of $\widehat{\beta}_{RE}$. We use S^2_{RE} to represent the estimate covariance for random effect estimator and S^2_{FE} to represent the estimate covariance for fixed effect estimator.

In order to calculate the test statistic, we firstly find the difference in coefficients for time varying covariates between $\widehat{\beta}_{RE}$ and $\widehat{\beta}_{FE}$. Then we calculate the estimated variance covariance difference between S^2_{RE} and S^2_{FE}. Now use the test statistic formula to calculate the Hausman test statistic is 27.10. That means the difference in the coefficients on experience, experience squared and the union, married and public sector dummies are significant or at least one is significant. Under the null hypothesis, the test statistic follows a Chi-squared distribution with 5 degrees of freedom and the critical value for χ^2_5 is 11.07 , so that we have to reject the null hypothesis at 5% significance level since the test statistic is 27.10 (p-value $<$0.001) and $>$ 11.07. So there is correlation between the individual effect and explanatory variables $exper$, $union$ and $married$, etc. Vella and Verbeek [1998] concentrate on the impact of endogenous union status on the wages and consider some complicated estimators

to solve the problem. Johnson and DiNardo [2007] also describe endogenous X: the marital status and year of experiences, $married$ and $exper$ are correlated with the individual effect. These variables may capture other unobserved difference between married and unmarried and year of experiences of workers. In this case, the fixed effect estimator is good to use to eliminate the individual effect in order to avoid the heterogeneity biased estimation.
Note: The Hausman test statistic is calculate based on Σ_{FE} is the covariance of $\widehat{\beta}_{FE}$ and Σ_{RE} is the covariance of $\widehat{\beta}_{RE}$. We use S^2_{RE} to represent the estimate covariance for random effect estimator and S^2_{FE} to represent the estimate covariance for fixed effect estimator. There are more details of covariance calculation described in Wooldridge [2002]. Here we use R to extract the covariance S^2_{FE} and S^2_{RE} which are shown in Table 5.28 and 5.26 respectively from their summary statistics by using $cov.scaled$ (not $cov.unscaled$ which give without dispersion) and $varFix$.

Verbeek [2004] gives a slightly different value of the test statistic, 31.75, obtained by using Stata (Data Analysis and Statistical Software). Because Stata adjusts the covariance of the estimator in order to avoid a negative test statistic value and make sure the difference of covariance between two estimators is the positive definite. Details of how Stata does this are given at http://www.stata.com/.

Caution: There are other sorts of misspecification (ie. simultaneity bias, measurement errors, selection bias, etc.) which may also cause Hausman test rejection, but in this thesis we only concentrate on the correlation between the explanatory variable and the individual effect.

5.4 Instrumental Variable (IV) Estimator

The fixed effect estimator provides unbiased estimates when there is correlation between the individual effects and the explanatory variables which also eliminates the time invariant variables from the model. That is a high price to pay for allowing such correlation. There is an alternative method called Instrumental Variable (IV) estimator which gives unbiased estimates when the explanatory variables are correlated with the individual effects. The instrumental variable (IV) estimator can be seen as in between the random effect estimator and fixed effect estimator. To prove this, we first recall Eq.(3.13).

$$\widehat{\boldsymbol{\beta}}_W = \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{X}_{it}-\overline{X}_i)(\mathbf{X}_{it}-\overline{X}_i)^T\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{X}_{it}-\overline{X}_i)(y_{it}-\overline{y}_i)\right]$$

We could rewrite this equation as

$$\widehat{\boldsymbol{\beta}}_W = \left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{X}_{it} - \overline{X}_i)\mathbf{X}_{it}^T\right]^{-1}\left[\sum_{i=1}^{N}\sum_{t=1}^{T}(\mathbf{X}_{it} - \overline{X}_i)y_{it}\right] \tag{5.29}$$

We could write Eq.(5.29) in full data form

$$\widehat{\boldsymbol{\beta}}_W = (Z^T X)^{-1} Z^T \mathbf{y}$$

This can be interpreted as each explanatory variable is instrumented by its value in deviation from the individual specific mean. That is, X_{it} is instrumented by $Z_{it} = \mathbf{X}_{it} - \overline{X}_i$ [Verbeek, 2004]. Then Z is the instrument variable. The choice of an instrumental variable Z is one that is correlated with the explanatory variable but not with the error term or the individual effects. The IV estimator may also be seen as two stage least squares (2SLS) [Johnson and DiNardo, 2007]:

Stage 1: Regress each of the variables in the X matrix on Z to obtain a matrix of fitted values $\widehat{X}$:

$$\widehat{X} = Z(Z^T Z)^{-1} Z^T X = P_Z X$$

Stage 2: Regress $\mathbf{y}$ on $\widehat{X}$ to obtain the estimated $\boldsymbol{\beta}$ vector

$$\begin{aligned}\widehat{\beta}_{2SLS} &= (\widehat{X}^T\widehat{X})^{-1}\widehat{X}^T\mathbf{y} \\ &= (X^T P_Z X)^{-1}(X^T P_Z \mathbf{y}) \\ &= \widehat{\beta}_{IV}\end{aligned}$$

Thus the IV estimator can be obtained by a two-stage least-squares procedure. The variance-covariance matrix is

$$\text{Var}(\widehat{\beta}_{IV}) = \sigma^2(X^T P_Z X)^{-1}$$

and the error variance may be estimated consistently from

$$\widehat{\sigma}^2 = \frac{(\mathbf{y} - X\widehat{\beta}_{IV})^T(\mathbf{y} - X\widehat{\beta}_{IV})}{N - K}$$

Note: see more details of these from Johnson and DiNardo [2007]. In this thesis, we do not concentrate on the IV estimator, we only describe the method here to let the reader know

there is another option to deal with the correlation between the individual effects and the explanatory variables.

5.5 R codes

5.5.1 Omitted variables bias

```
covXS <- function(prob,n,a,b,lambda,sigmaE,beta0,beta1,gamma){
            s <- rbinom(n, 1, prob)
            alpha <- (s+1/2)*lambda
            beta <- (-s+3/2)*lambda
            u <- rbeta(n,alpha,beta)
            x <- a+(b-a)*u
            eps <- rnorm(n,0,sigmaE)
            y <- beta0+beta1*x+gamma*s+eps
            d.f <- data.frame(x=x,y=y,s=s)
            return(d.f)
          }
indepXS <- function(prob,n,a,b,sigmaE,beta0,beta1,gamma){
              x <- rnorm(n,a,b)
              s <- rbinom(n, 1, prob)
              eps <- rnorm(n,0,sigmaE)
              y <- beta0+beta1*x+gamma*s+eps
              d.f <- data.df<-data.frame(x=x,y=y,s=s,eps=eps)
              return(d.f)
            }
modelfit <- function(prob,n,a,b,lambda,sigmaE,beta0,beta1,gamma){
               s <- rbinom(n,1,prob)
               alpha <- (s+1/2)*lambda
               beta <- (-s+3/2)*lambda
               u <- rbeta(n,alpha,beta)
               x <- a+(b-a)*u
               eps <- rnorm(n,0,sigmaE)
               y <- beta0+beta1*x+gamma*s+eps
               data.df <- data.frame(x=x,y=y,s=s)
               fit <- glm(y~x+s,data=data.df)
               b0 <- summary(fit)$coef[1]
               b1 <- summary(fit)$coef[2]
               bs <- summary(fit)$coef[3]
               b0s <- summary(fit)$cov.scaled[1]
```

```
            b1s <- summary(fit)$cov.scaled[5]
            bss <- summary(fit)$cov.scaled[9]
            disp <- summary(fit)$dispersion
            fitted <- glm(y~x,data=data.df)
            b0r <- summary(fitted)$coef[1]
            b1r <- summary(fitted)$coef[2]
            b0rs <- summary(fitted)$cov.scaled[1]
            b1rs <- summary(fitted)$cov.scaled[4]
            dispr <- summary(fitted)$dispersion
            bbeta <- c(b0,b0s,b1,b1s,bs,bss,b0r,b0rs,
                       b1r,b1rs,disp,dispr)
            return(bbeta)
          }
cal <- function(beta0,beta1,prob,lambda,gamma,a,b){
        Varx <- (3/(4*(8*lambda+4)))*(b-a)^2+(prob*(1-prob)/4)*(b-a)^2
        Exs <- (a+(b-a)*3/4)*prob
        Ex <- (1/4*(b-a)+a)*(1-prob)+(3/4*(b-a)+a)*prob
        Covxs <- Exs-Ex*prob
        Ex2 <- Varx+Ex^2
        A <-(Ex2*prob-Ex*Exs)/Varx
        B <- Covxs/Varx
        Mb0r <- beta0+gamma*A
        Mb1r <- beta1+gamma*B
        M<-c(Mb0r,Mb1r)
        return(M)
      }
```

5.5.2 Heterogeneity Bias

```
newsim.cor <- function(N,T,delta,rate,Xsigma){
                xM <- matrix(rep(NA,N*T),nrow=N,ncol=T)
                idM <- matrix(rep(NA,N*T),nrow=N,ncol=T)
                timeM <- matrix(rep(NA,N*T),nrow=N,ncol=T)
                xM[,1] <- rnorm(N,0,Xsigma)
                inc <- delta*c(0,sort(runif(T-1,0,1)))
                xM <- outer(xM[,1],inc,"+")
                gammaM0 <- matrix(rep(NA,N*T),nrow=N,ncol=T)
                for(i in 1:N){
                   gammaM0[i,1] <- rate*mean(xM[i,])
                }
                gammaM0 <- outer(gammaM0[,1],rep(1,T))
```

```
                for(i in 1:N){
                    for(j in 1:T){
                       idM[i,j] <- i
                    timeM[i,j] <- j
                    }
                }
                id <- matrix(t(idM),nrow=N*T,ncol=1)
                time <- matrix(t(timeM),nrow=N*T,ncol=1)
                x <- matrix(t(xM),nrow=N*T,ncol=1)
                gamma0 <- matrix(t(gammaM0),nrow=N*T,ncol=1)
                idtext <- factor(id)
                data.df <- data.frame(id=id,idtext=idtext,
                           time=time,x=x,gamma0=gamma0)
                return(data.df)
                }
modelfit.mle <- function(data.df){
                  randomint.inc <- lme(fixed = y ~ x,
                  random = ~ 1 | id, data=data.df)
                  slope <- summary(randomint.inc)$coef$fixed[2]
                  fix.iid <- glm(y ~ -1+x+idtext,data=data.df)
                  slope1 <- summary(fix.iid)$coef[1,1]
                  sr.iid <- glm(y ~ x,data=data.df)
                  slope2 <- summary(sr.iid)$coef[2,1]
                  return(c(slope,slope1,slope2))
               }
slope.sr <- function(data.df){
              x <- data.df$x
              y <- data.df$y
              b.sr <- sum((x-mean(x))*(y-mean(y)))/sum((x-mean(x))^2)
              return(b.sr)
        }
slope.fe <- function(data.df,N,T){
              x <- data.df$x
              y <- data.df$y
              id <- data.df$id
              X <- matrix(rep(NA,N),N,1)
              Y <- matrix(rep(NA,N*T),N,1)
              for(i in 1:N){
                 X[i] <- mean(x[id==i])
                 Y[i] <- mean(y[id==i])
```

```
                }
                denom <- sum(x^2)-T*sum(X^2)
                b.fe <- (sum(x*y)-T*sum(X*Y))/(denom)
                return(b.fe)
        }
slope.re <- function(data.df,N,T,rate){
                xx <- matrix(rep(NA,N),N,1)
                ww <- matrix(rep(NA,N),N,1)
                for(i in 1:N){
                xx[i] <- mean(data.df$x[data.df$id==i])
                ww[i] < -mean(data.df$w[data.df$id==i])
        }
                Esigma <- sqrt(var(data.df$eps))
                Xsigma <- sqrt(var(xx))
                Wsigma <- sqrt(var(ww))
                Asigma <- sqrt(Wsigma^2+rate^2*Xsigma^2/T)
                phi <- Esigma^2/(Esigma^2+T*Asigma^2)
                x <- data.df$x
                y <- data.df$y
                id <- data.df$id
                X <- matrix(rep(NA,N),N,1)
                Y <- matrix(rep(NA,N),N,1)
                XM <- matrix(rep(NA,N*T),N,T)
                YM <- matrix(rep(NA,N*T),N,T)
                for(i in 1:N){
                   X[i] <- mean(x[id==i])
                   Y[i] <- mean(y[id==i])
                   for(j in 1:T){
                  XM[i,j] <- X[i]
              YM[i,j] <- Y[i]
               }
                }
                denom1 <- sum(x^2)-T*sum(X^2)
                denom2 <- phi*T*sum((X-mean(x))^2)
                denom3 <- sum(x*y)
                denom4 <- T*sum(X*Y)
                denom5 <- phi*T*sum((X-mean(x))*(Y-mean(y)))
                b.re <- (denom3-denom4+denom5)/(denom1+denom2)
                return(b.re)
   }
```

```
fitting.cor <- function(N,T,data.df,R,beta1,Wsigma,Esigma,rate){
               result <- matrix(rep(NA,3*R),nrow=R,ncol=3)
               slope.re <- matrix(rep(NA,R),nrow=R,ncol=1)
               slope.fe <- matrix(rep(NA,R),nrow=R,ncol=1)
               slope.sr <- matrix(rep(NA,R),nrow=R,ncol=1)
               epsM <- matrix(rep(NA,N*T*R),R,N*T)
               WM <- matrix(rep(NA,N*T*R),R,N*T)
               yM <- matrix(rep(NA,N*T*R),nrow=R,ncol=N*T)
               scor <- NULL
               x <- data.df$x
               gamma0 <- data.df$gamma0
               for(i in 1:R){
                  epsM[i,] <- rnorm(N*T,0,Esigma)
                  WM[i,] <- rep(rnorm(N,0,Wsigma),each=T)
               }
               for(s in 1:R){
                  yM[s,] <- beta1*x+gamma0+epsM[s,]+WM[s,]
                  scor[[s]] <- data.frame(data.df,y=yM[s,],
                  eps=epsM[s,],w=WM[s,])
               slope.sr[s] <- slope.sr(scor[[s]])
               slope.fe[s] <- slope.fe(scor[[s]],N,T)
               result[s,] <- modelfit.mle(scor[[s]])
               slope.re[s] <- slope.re(scor[[s]],N,T,rate)
               }
               beta <- data.frame(result,slope.re,slope.fe,slope.sr)
               return(beta)
               }
```

Chapter 6

Bayesian Estimation

As an alternative to the Hausman test, we can model the entire possible set of dependencies, introducing parameters for possible correlation between individual effects and covariates. We then test for this correlation using hypothesis tests on the parameters. The Bayesian approach provides a natural hierarchical framework for such modelling. In a Bayesian approach, we can fit longitudinal data with random effects model, fixed effects model, pooled model and Mundlak formulation model. In this chapter, we are only interested in how the Bayesian approach works for longitudinal data fitting and show the performance of this approach by using the WinBUGS software. We develop a full Bayesian formulation as an alternative to the Hausman test to do model comparison between random effect model and fixed effect model. To see this, we empirically illustrate the idea using our simulated data and real data *WAGE*.

6.1 Bayesian Analysis

In the Bayesian analysis, prior probability distributions are used to describe the uncertainty of all unknown parameters prior to seeing the data. After observing the data, the posterior distribution provides a summary of the remaining uncertainty of the data which is relevant for parameter estimation. Bayesian analysis can be implemented in WinBUGS [Thomas et al., 2000]. The computational program takes samples from the posterior distribution of the parameter θ given y by using the Markov Chain Monte Carlo (MCMC) method where "Monte Carlo" implies the random sampling. "Markov chain" refers to the method of generating the random samples. There is a sequence of random variables, each variable is conditional on the previous variable in the sequence θ_{t-1}, such a distribution is known as

a Markov chain. MCMC algorithms are constructed in such a way that sufficiently large samples from the Markov chain are equivalent to samples from the required posterior distribution.

6.1.1 Markov Chain Monte Carlo (MCMC)

MCMC methods provide a convenient and generally applicable means of summarising posterior distribution in Bayesian Analysis, and are a useful method for sampling from a complicated distribution. The Metropolis algorithm is one of the widely used and simplest MCMC algorithms. The following describes how the Metropolis algorithm obtains samples from the distribution of a single parameter, but we can easily extend to the multiple parameter simulation.

6.1.2 Metropolis Algorithm

The Metropolis algorithm is defined by McCarthy [2007] as below

> Start with an initial arbitrary value for the parameter θ_0, which is the first value of the Markov chain. We are interested in obtaining subsequent values of θ_t such that they are samples of a random variable with certain probability density function.
>
> A new possible value (θ^*) is generated by drawing it from an arbitrary symmetric probability distribution. This proposal distribution is defined by its probability density function; given the current values θ_t, the probability of drawing the value of θ^* as the possible next value of the Markov chain is equal to $q(\theta^*|\theta_t)$.
>
> Next, the acceptance probability is equal to
>
> $$R(\theta^*|\theta_t) = \min[1, \frac{p(\theta^*)}{p(\theta_t)}]. \tag{6.1}$$
>
> In a Bayesian application, the ratio depends on the posterior probability density function at two different points $p(\theta^*)$ and $p(\theta_t)$. Based on Bayes' rule, these two values are equal to:
>
> $$p(\theta^*) = \frac{\pi(\theta^*)L(\theta^*)}{\int_{\infty}^{-\infty} \pi(\theta)L(\theta)d\theta}$$
>
> and
>
> $$p(\theta_t) = \frac{\pi(\theta_t)L(\theta_t)}{\int_{\infty}^{-\infty} \pi(\theta)L(\theta)d\theta}$$

where π is the prior probability density function and L is the likelihood function. Because both expressions have the same denominator, the ratio of the two values is simply equal to the ratio of the prior probabilities and likelihoods; the integral is not calculated. Therefore, if the ratio of the posterior probability is greater than or equal to 1 (i.e. $p(\theta^*) \geq p(\theta_t)$), then θ^* is chosen as the next value of the Markov chain ($\theta_{t+1} = \theta^*$). If otherwise, $p(\theta^*) < p(\theta_t)$, then θ^* is chosen as the next value of the Markov chain with probability $p(\theta^*)/p(\theta_t)$, and θ_t is chosen otherwise.

Hastings [1970] modified the Metropolis algorithm to permit non-symmetric distributions to be used for generating the new possible values. The new algorithm is called Metropolis-Hastings algorithm. It defines the acceptance probability as equal to

$$R(\theta^*|\theta_t) = \min(1, r) = \min[1, \frac{p(\theta^*) \times q(\theta_t|\theta^*)}{p(\theta_t) \times q(\theta^*|\theta_t)}] \tag{6.2}$$

There is another algorithm called Gibbs sampling that is a special case of Metropolis-Hastings algorithm in which $q()$ is chosen to be the full conditional probability and the R (ratio of the posterior probability) is always equal to 1. We illustrate it by using a bivariate distribution.

- The target density for a bivariate distribution is $p(\theta_1, \theta_2) = p(\theta_1|\theta_2) \times p(\theta_2)$
- We need to propose (θ_1^*, θ_2^*) from $q(\theta_1^*, \theta_2^*|\theta_1, \theta_2)$
- We break q into 2 pieces:
  - propose θ_1^* first from $q_1(\theta_1^*|\theta_1, \theta_2)$
  - then propose θ_2^* from $q_2(\theta_2^*|\theta_1, \theta_2)$

    or propose θ_2^* from $q_2(\theta_2^*|\theta_1^*, \theta_2)$

  where at each step of the chain we only update one parameter, then each new point is selected using a proposal density that along the line is $p(\theta_2|\theta_1)$ or $p(\theta_1|\theta_2)$, this is called full conditonal density.
- The proposal density is $q(\theta_1|\theta_2) = p(\theta_1|\theta_2)$ or $q(\theta_2|\theta_1) = p(\theta_2|\theta_1)$.

$$r = \frac{p(\theta_1^*, \theta_2^*) \times q(\theta_1, \theta_2|\theta_1^*, \theta_2^*)}{p(\theta_1, \theta_2) \times q(\theta_1^*, \theta_2^*|\theta_1, \theta_2)}$$

When we update θ_1, θ_1^* is drawn from $f(\theta_1|\theta_2)$ and $\theta_2^* = \theta_2$. So we have r to become

$$\begin{aligned} r &= \frac{p(\theta_1^*, \theta_2) \times p(\theta_1|\theta_2)}{p(\theta_1, \theta_2) \times p(\theta_1^*|\theta_2)} \\ &= \frac{p(\theta_1^*|\theta_2) \times p(\theta_2) \times p(\theta_1|\theta_2)}{p(\theta_1|\theta_2) \times p(\theta_2) \times p(\theta_1^*|\theta_2)} \\ &= 1 \end{aligned}$$

ie. the ratio $R(\theta^*|\theta_t) = \min(1, r)$= 1 = acceptance probability, that means we always accepted for Gibbs sampling proposals.

Gibbs Sampling

In Gibbs sampling samples are drawn from a multivariate distribution by taking successive samples from the full conditional distribution of each element of the parameter space. This is more straight forward in many cases than sampling from the joint distribution.

When updating an arbitrary parameter θ_j, we fix the other parameters and select θ_j from $\pi(\theta_j|\text{other parameters})$ which is called full conditional distribution. So the proposal density comes directly from the target density.

> **Gibbs Sampler**
>
> Suppose that a sample has distribution depending on a parameter vector $\theta \in \Theta$ of length d. For a joint distribution $\pi(\theta_1, \ldots, \theta_d)$ with full conditionals $\pi_1, \ldots, \pi_d$ where π_j is the distribution of θ_j conditional on $(\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_d)$, the Gibbs sampler simulates successively from all conditionals, modifying one component of θ at a time.
>
> Initialization: Start with an arbitrary value $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_d^{(0)})$
>
> Iteration t: Given $(\theta_1^{(t-1)}, \ldots, \theta_d^{(t-1)})$, generate
>
> 1. θ_1^t according to $\pi_1(\theta_1|\theta_2^{(t-1)}, \ldots, \theta_d^{(t-1)})$,
> 2. θ_2^t according to $\pi_2(\theta_2|\theta_1^t, \theta_3^{(t-1)}, \ldots, \theta_d^{(t-1)})$,
>
> $\vdots$
>
> d. θ_d^t according to $\pi_d(\theta_d|\theta_1^t, \ldots, \theta_{d-1}^t)$
>
> Marin and Robert [2007].

Use of the above algorithms has three main consequences. The first consequence is that θ_{t+1} typically depends on θ_t. The dependence means that each new sample provides a frac-

tion of the information about the posterior distribution compared to an uncorrelated sample. An extremely large number of samples will be needed to obtain a good estimate of the posterior distribution if there is strong correlation between samples.

The second main consequence is that the initial value does influence the Markov chain, so the first part of the chain needs to be discarded as a "burn-in" until the influence of the initial value is no longer apparent.

The third main consequence is that the proposal density may have parameters which need tuning. If step size is too large (taking θ^* a long way from θ_t) proposals may be rejected often, and the chain will stick and will not mix (explore the density) efficiently. If step size is too small proposals may be accepted with high probability. The chain will mix too slowly, ie. it will take a long term to explore the density.

From the theory of Markov chains, we expect the chains to eventually converge to the target distribution. In order to see whether the chain appears to be converged, we can use Gelman-Rubin diagnostics. Gelman and Rubin diagnostics is introduced by Gelman and Rubin [1992] and Brooks and Gelman [1997] provide graphical methods which can be used as a visual inspection to test convergence. This diagnostic is based on analyzing multiple simulated MCMC chains by comparing the variances within each chain and the variance between chains. If within and between chain variation values are both close to 1, that indicate convergence. If there is large deviation between these two variances, that indicates nonconvergence, then we need to run out a longer chain. Alternatively, we could use the trace plot which is a plot of the iteration number against the value of the draw of the parameter at each iteration. By using this plot, we can visually inspect whether the chain gets stuck in certain areas of the parameter space, if so it indicates nonconvergence.

Note: a Markov chain has Optimal performance when the possible values θ^* are not too far from θ_t and with appropriate length of "burn-in".

6.2 WinBUGS Implementation

We implement the models in WinBUGS which use the Gibbs sampling to sample from the full conditional distribution and run two chains for the model from different starting places, to see whether they end up in the same place. We store samples after a 5000 burn-in. We check for convergence by using the Gelman-Rubin diagnostics (bgr diagnostics).

Definition 6.1. *The posterior density of the vector of unobservables θ_A in the model A is*

$$p(\boldsymbol{\theta}_A|\boldsymbol{y}, A) = \frac{p(\boldsymbol{\theta}_A|A)p(\boldsymbol{y}|\boldsymbol{\theta}_A, A)}{p(\boldsymbol{y}|A)} \tag{6.3}$$

The expression in the denominator of Eq.(6.3) is the marginal likelihood. In many circumstances it suffices to know just the shape of the posterior density $p(\boldsymbol{\theta}_A|\boldsymbol{y}, A)$ and it is costly to evaluate $p(\boldsymbol{y}|A)$. In this case it is useful to exploit the fact that

$$p(\boldsymbol{\theta}_A|\boldsymbol{y}, A) \propto (\boldsymbol{\theta}_A|A)p(\boldsymbol{y}|\boldsymbol{\theta}_A, A) \tag{6.4}$$

The expression on the right side of Eq.(6.4) is a kernel of the posterior density, we call it ***kernel posterior density***. *[Geweke, 2005]*

Also we produce the kernel posterior density for the estimators to check their distribution. Here we assume all parameters prior distributions are normal and their variance are inverse gamma distributions.

6.2.1 Simulation Example

The simulation data sets **RINOCOR** and **RICOR** that we use are generated in Chapter 5. The candidate models are

1. Random intercept model (RE Model):

$$Y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_{0i} + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$$

$$\alpha_{0i} \sim \mathrm{N}(0, \sigma_\alpha^2)$$

Prior distributions are

$$\beta_0 \sim \mathrm{N}(0, \sigma_{\beta_0}^2)$$

$$\beta_1 \sim \mathrm{N}(0, \sigma_{\beta_1}^2)$$

$$\frac{1}{\sigma_\varepsilon^2} \sim \mathrm{Gamma}(\nu, \tau)$$

$$\frac{1}{\sigma_\alpha^2} \sim \mathrm{Gamma}(\nu, \tau)$$

where $\sigma_{\beta_0}^2$ and $\sigma_{\beta_1}^2$ are large numbers, because they are uninformative and then we

assume ν and τ are small number. Since the variance of a variable is positive and Gamma is a built in function in WinBUGS that can sample values between 0 and $+\infty$. In order to properly draw from gamma distribution in practice, we assume ν and τ are small number, ie. $\nu = \tau = 0.0001$. Here $i = 1 \cdots N$ and $t = 1 \cdots T$.

2. Fixed effect model (FE Model)

$$Y_{it} = \beta_1 x_{it} + \alpha_{0i} + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$$

Prior distributions are

$$\beta_1 \sim \mathrm{N}(0, \sigma_{\beta_1}^2)$$

Here we define the simplest FE model only. No correlation with X_i is assumed, but we may detect this. The correlations case will be introduced in section 6.10 below.

$$\alpha_{0i} \sim \mathrm{N}(0, \sigma_\alpha^2)$$

$$\frac{1}{\sigma_\varepsilon^2} \sim \mathrm{Gamma}(\nu, \tau)$$

where $\sigma_{\beta_1}^2$ and σ_α^2 are large number and ν and τ are small numbers. Here $i = 1 \cdots N$ and $t = 1 \cdots T$. In FE model, we could include the time independent covariates, but it may cause the nonconvergence problem. The best idea is to exclude the time invariant variables. We compare FE model with the RE model, we can see there is a definite difference on the prior distribution of α_{0i}. For RE model, the inverse variance of α_{0i}, $\frac{1}{\sigma_\alpha^2}$ draws from gamma distribution with smaller numbers of ν and τ. But for FE model, we draw α_{0i} directly from a normal distribution with zero mean and large variance σ_α^2.

3. Pooled model (PL Model)

$$Y_{it} = \beta_0 + \beta_1 x_{it} + \varepsilon_{it}$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$$

Prior distributions are

$$\beta_0 \sim \mathrm{N}(0, \sigma_{\beta_0}^2)$$

$$\beta_1 \sim \mathrm{N}(0, \sigma_{\beta_1}^2)$$

$$\frac{1}{\sigma_{\varepsilon}^2} \sim \text{Gamma}(\nu, \tau)$$

where $\sigma_{\beta_0}^2$ and $\sigma_{\beta_1}^2$ are large numbers and ν and τ are small number. Here $i = 1 \cdots N$ and $t = 1 \cdots T$.

4. The full Bayesian model of Mundlak formulation (MF) is defined as

$$Y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + \varepsilon_{it}$$

$$\alpha_{0i} = \rho \overline{X}_i + w_i$$

$$\varepsilon_{it} \sim \text{N}(0, \sigma_{\varepsilon}^2)$$

$$w_i \sim \text{N}(0, \sigma_w^2)$$

Prior distributions are

$$\beta_0 \sim \text{N}(0, \sigma_{\beta_0}^2)$$

$$\beta_1 \sim \text{N}(0, \sigma_{\beta_1}^2)$$

$$\frac{1}{\sigma_{\varepsilon}^2} \sim \text{Gamma}(\nu, \tau)$$

$$\frac{1}{\sigma_w^2} \sim \text{Gamma}(\nu, \tau)$$

$$\rho \sim \text{N}(0, \sigma_{\rho}^2)$$

where $\sigma_{\beta_0}^2$, $\sigma_{\beta_1}^2$ and σ_{ρ}^2 are large numbers and ν and τ are small numbers. Here $i = 1 \cdots N$ and $t = 1 \cdots T$.

RINOCOR dataset

The following code is used to implement the random intercept model in WinBUGS.

```
model {
   for(i in 1:N){
      for(t in 1:T){
         Y[i,t] ~ dnorm(mu[i,t],tau.e)
         mu[i,t] <- beta0+beta1*X[i,t]+alpha[i]
      }
      alpha[i] ~ dnorm(0,tau.a)
   }
   #prior distribution
```

```
    beta0 ~ dnorm(0.0,0.0001)
    beta1 ~ dnorm(0.0,0.0001)
    tau.e ~ dgamma(0.0001,0.0001)
    tau.a ~ dgamma(0.0001,0.0001)
}
```

The Gelman-Rubin diagnostics are shown on Figure 6.1.

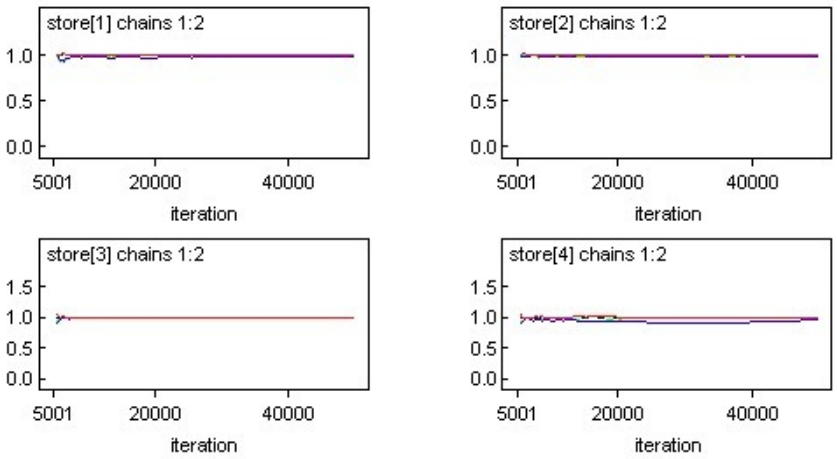

Figure 6.1: Gelman-Rubin diagnostics of β_0, β_1, σ_ε and σ_α for **RINOCOR**

The chain of convergence chains are shown on Figure 6.2.

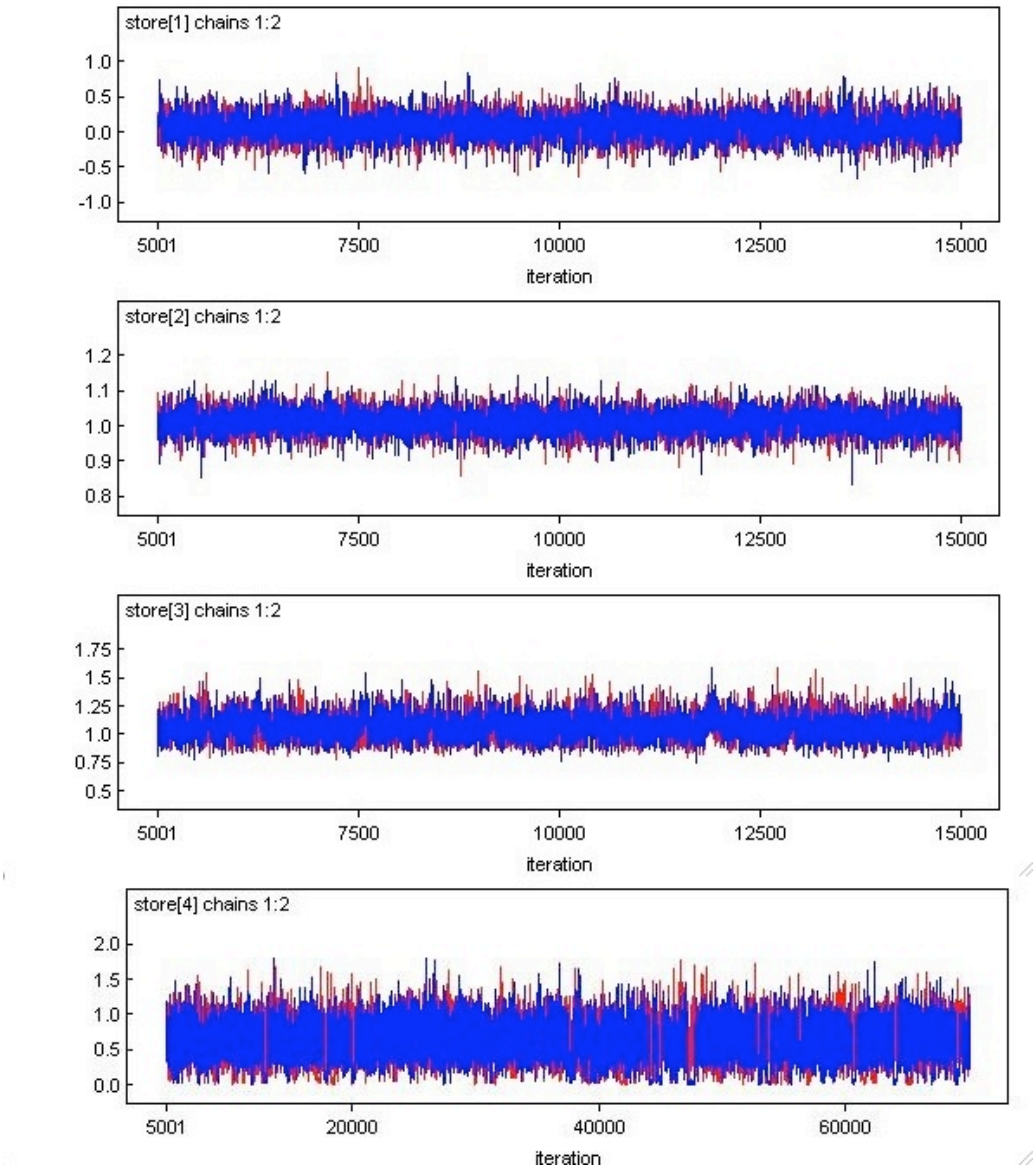

Figure 6.2: Convergence chains of β_0, β_1, σ_ε and σ_α for **RINOCOR**

Figures 6.1 and 6.2 are used to check the convergence. From Figure 6.1, the Gelman-Rubin diagnostics compare within and between chain variation the lines are all close to 1 which indicate the convergence. Figure 6.2 shows two chains for each estimator in RE Model are converging to the same place or converging to its mean within certain standard variance indicate we have appropriate initial value, the length of burn-in and sample size is satisfied. Thus, the result are acceptable.

The posterior distribution of estimators and their estimates are shown on Figure 6.3 and Table 6.1.

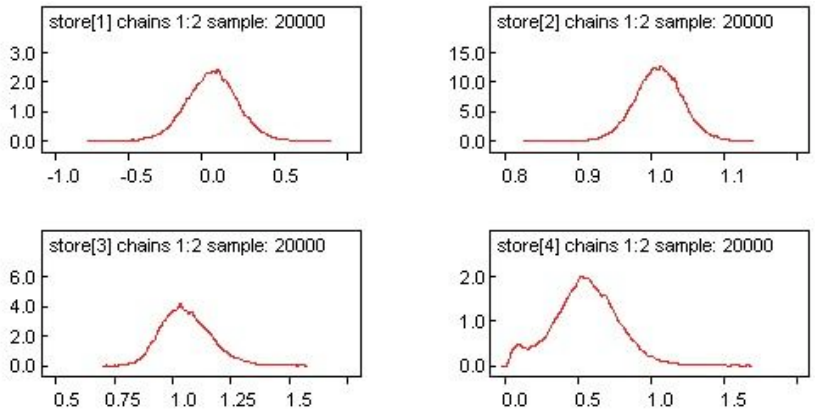

Figure 6.3: Distribution of estimates of β_0, β_1, σ_ε and σ_α for**RINOCOR**

Figure 6.3 shows the kernel density of the posterior distribution for the parameter estimators β_0, β_1, σ_ε and σ_α. The true values, the posterior estimates and their credible intervals are listed as

Table 6.1: Table of estimates of **RINOCOR** for RE Model

	True value	Bayesian Estimate			
		mean	std.err	2.5% Credible Interval	97.5% Credible Interval
β_0	0	0.071	0.178	-0.284	0.424
β_1	1	1.011	0.033	0.945	1.076
σ_ε	1	1.060	0.105	0.880	1.291
σ_α	0.6	0.552	0.226	0.084	0.994

Table 6.1 shows the Bayesian estimates are approximately equal to the true value. Then credible intervals for each estimator show the inclusion of the true value. Therefore, the Bayesian approach perform just as good as frequentist approach.

Implementation for other models see Appendix D and the Gelman-Rubin diagnostics and convergence chains for each model list in Appendix E. The result for other models can be found in Section 6.4. The result of PL Model is acceptable. Because the Gelman-Rubin diagnostics and convergence chains of PL Model are converge. The result of FE Model is acceptable as well. Note, the FE Model we fit is without the intercept term in order to get the convergence of the Gelman-Rubin diagnostics and convergence chains. The intercept term is confounded with α_i, since it is time invariant. So we eliminate intercept term. At this point, we can see WinBUGS can't automatically wipe out the time invariant variable, but R can wipe out the time invariant variable automatically. Thus, when we fit the fixed effect model in WinBUGS, we have to get rid of the time invariant variables.

Note: To detect whether there are correlation between α_i and X_i, we could use BMF model to investigate which is introduced in section 6.3.

RICOR dataset

Now we fit another simulated data set **RICOR**. The programs for four candidate models (time invariant variable in fixed effect model has been removed) of WinBUGS are the same as codes introduced in the **RINOCOR** data example. In WinBUGS, we only need to change the dataset. Implementation codes for four models see Appendix D. The Gelman-Rubin diagnostics and convergence diagnostics for each model list in Appendix E. And the results of four models can be found in Section 6.4. The convergence of the four models are acceptable by the Gelman-Rubin convergence and chains convergence.

6.2.2 Real Data Example

In this section we use WinBUGS to apply to the real data by fitting the random effect model, fixed effect model and pooled model. Three models we fit are given as

1. Random effect Model (RE)

$$\begin{aligned} \text{lwage}_{it} &= \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_{it} + \beta_3 \text{expersq}_{it} + \beta_4 \text{union}_{it} \\ &\quad + \beta_5 \text{married}_{it} + \beta_6 \text{black}_i + \beta_7 \text{hisp}_i \\ &\quad + \beta_8 \text{pub}_{it} + \alpha_i + \varepsilon_{it} \end{aligned} \tag{6.5}$$

$$\varepsilon_{it} \sim \text{N}(0, \sigma_\varepsilon^2)$$

$$\alpha_i \sim \text{N}(0, \sigma_\alpha^2)$$

Prior distributions are

$$\beta_0 \sim \text{N}(0, \sigma_{\beta_0}^2)$$

$$\beta_k \sim \text{N}(0, \sigma_\beta^2)$$

$$\frac{1}{\sigma_\varepsilon^2} \sim \text{Gamma}(\nu, \tau)$$

$$\frac{1}{\sigma_\alpha^2} \sim \text{Gamma}(\nu, \tau)$$

where $\sigma_{\beta_0}^2$ and σ_β^2 are large number and ν and τ are small number. Here $k = 1 \cdots K$, $i = 1 \cdots N$ and $t = 1 \cdots T$ for $K = 8$, $N = 545$ and $T = 8$.

2. Fixed effect Model with time varying covariates only (FE)

$$\begin{aligned} \text{lwage}_{it} &= \beta_1 \text{exper}_{it} + \beta_2 \text{expersq}_{it} + \beta_3 \text{union}_{it} \\ &\quad + \beta_4 \text{married}_{it} + \beta_5 \text{pub}_{it} + \alpha_i + \varepsilon_{it} \end{aligned} \tag{6.6}$$

$$\varepsilon_{it} \sim \text{N}(0, \sigma_\varepsilon^2)$$

Prior distributions are

$$\mu \sim \text{N}(0, \sigma_\mu^2)$$

$$\beta_p \sim \text{N}(0, \sigma_\beta^2)$$

$$\alpha_i \sim \text{N}(0, \sigma_\alpha^2)$$

$$\frac{1}{\sigma_\varepsilon^2} \sim \text{Gamma}(\nu, \tau)$$

where σ_β^2 and σ_α are large number and ν and τ are small number. Here $p = 1 \cdots P$, $i = 1 \cdots N$ and $t = 1 \cdots T$ for $P = 5$, $N = 545$ and $T = 8$.

3. Pooled Model (PL)

$$\begin{aligned} \text{lwage}_{it} &= \beta_0 + \beta_1 \text{educ}_{it} + \beta_2 \text{exper}_{it} + \beta_3 \text{expersq}_{it} + \beta_4 \text{union}_{it} \\ &\quad + \beta_5 \text{married}_{it} + \beta_6 \text{black}_{it} + \beta_7 \text{hisp}_{it} \\ &\quad + \beta_8 \text{pub}_{it} + \varepsilon_{it} \end{aligned} \tag{6.7}$$

$$\varepsilon_{it} \sim \text{N}(0, \sigma_\varepsilon^2)$$

Prior distributions are

$$\beta_0 \sim \text{N}(0, \sigma_{\beta_0}^2)$$

$$\beta_k \sim \text{N}(0, \sigma_\beta^2)$$

$$\frac{1}{\sigma_\varepsilon^2} \sim \text{Gamma}(\nu, \tau)$$

where $\sigma_{\beta_0}^2$ and σ_β^2 are large number and ν and τ are small number. Here $k = 1 \cdots K$, $i = 1 \cdots N$ and $t = 1 \cdots T$ for $K = 8$, $N = 545$ and $T = 8$.

4. The full Bayesian model of Mundlak formulation (MF Model) for **WAGE** data is de-

fined as

$$\begin{aligned} \text{lwage}_{it} &= \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_{it} + \beta_3 \text{expersq}_{it} + \beta_4 \text{union}_{it} \\ &\quad + \beta_5 \text{married}_{it} + \beta_6 \text{black}_i + \beta_7 \text{hisp}_i \\ &\quad + \beta_8 \text{pub}_{it} + \alpha_i + \varepsilon_{it} \end{aligned} \tag{6.8}$$

$$\varepsilon_{it} \sim \text{N}(0, \sigma_\varepsilon^2)$$

$$w_i \sim \text{N}(0, \sigma_w^2)$$

$$\begin{aligned} \alpha_i &= \rho_1 \overline{\text{exper}}_i + \rho_2 \overline{\text{expersq}}_i + \rho_3 \overline{\text{union}}_i \\ &\quad + \rho_4 \overline{\text{married}}_i + \rho_5 \overline{\text{pub}}_i + w_i \end{aligned} \tag{6.9}$$

Note: here we wipe out the time invariant variables in Eq. (5.28), because they may cause the convergence problem.

Prior distributions are

$$\beta_0 \sim \text{N}(0, \sigma_{\beta_0}^2)$$

$$\beta_k \sim \text{N}(0, \sigma_\beta^2)$$

$$\rho_p \sim \text{N}(0, \sigma_\rho^2)$$

$$\frac{1}{\sigma_\varepsilon^2} \sim \text{Gamma}(\nu, \tau)$$

$$\frac{1}{\sigma_w^2} \sim \text{Gamma}(\nu, \tau)$$

where $\sigma_{\beta_0}^2$, σ_β^2 and σ_ρ are large number and ν and τ are small number. Here $k = 1 \cdots K$, $i = 1 \cdots N$, $p = 1 \cdots P$ and $t = 1 \cdots T$ for $P = 5$, $K = 8$, $N = 545$ and $T = 8$.

Here is WinBUGS code for random effect model:

```
model{
for(i in 1:N){
   for(t in 1:T){
      L[i,t] ~ dnorm(mu[i,t],tau.e)
      mu[i,t] <- beta0+beta[1]*S[i,t]+beta[2]*E[i,t]
      +beta[3]*E2[i,t]+beta[4]*U[i,t]+beta[5]*M[i,t]
      +beta[6]*B[i,t]+beta[7]*H[i,t]+beta[8]*P[i,t]+alpha[i]
   }
```

```
    alpha[i] ~ dnorm(0,tau.a)
  }
  #prior distribution
  beta0 ~ dnorm(0.0,0.0001)
  for(k in 1:8){
    beta[k] ~ dnorm(0.0,0.0001)
  }
  tau.e ~ dgamma(0.0001,0.0001)
  tau.a ~ dgamma(0.0001,0.0001)
```

For the other three candidate models, the codes of WinBUGS are able to find in Appendix D. The Gelman-Rubin convergence diagnostics for each model are listed in Appendix E, and the results of four models are acceptable and can be found in Section 6.4. Since the Gelman-Rubin diagnostics and convergence chains show convergence.

6.3 Full Bayesian Formulation

In Chapter 5, we introduced the Mundlak formulation to show that there is bias when we use a random effects model when the explanatory variable are correlated with individual unobserved effect. Empirically, we use the Hausman test to test whether the fixed effect and random effect estimators are significant different. If there is a significantly difference, the fixed effect estimator is unbiased. Now in a Bayesian approach, we follow the Mundlak formulation idea to define a new model called full Bayesian model which is the MF Model under Bayesian approach. Then we fit this model in WinBUGS to obtain the posterior distribution of ρ. If the mean of this distribution is close to zero, then we may conclude there is no correlation between the explanatory variables and the individual unobserved effects. Otherwise, there is a correlation. This model works just as the Hausman test. In this section, we use simulated data and real data to demonstrate how this full Bayesian formulation works as the equivalent to a Hausman test, then we compare this method with AIC, DIC and Hausman test in Chapter 7 to see whether it works just as well as other model selection methods.

6.3.1 Simulated Data

Now we use the simulated data **RINOCOR** and **RICOR** to demonstrate how the full Bayesian model works as Hausman test to select between fixed effect estimator and random effect es-

timator. Recall the full Bayesian model (MF Model) in section 6.2.1

$$Y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + \varepsilon_{it}$$

$$\alpha_{0i} = \rho \overline{X}_i + w_i$$

$$\varepsilon_{it} \sim \mathrm{N}(0, \sigma_\varepsilon^2)$$

$$w_i \sim \mathrm{N}(0, \sigma_w^2)$$

Prior distributions are

$$\beta_0 \sim \mathrm{N}(0, \sigma_{\beta_0}^2)$$

$$\beta_1 \sim \mathrm{N}(0, \sigma_{\beta_1}^2)$$

$$\frac{1}{\sigma_\varepsilon^2} \sim \mathrm{Gamma}(\nu, \tau)$$

$$\frac{1}{\sigma_w^2} \sim \mathrm{Gamma}(\nu, \tau)$$

$$\rho \sim \mathrm{N}(0, \sigma_\rho^2)$$

where $\sigma_{\beta_0}^2$, $\sigma_{\beta_1}^2$ and σ_ρ^2 are large numbers and ν and τ are small numbers. Here $i = 1 \cdots N$ and $t = 1 \cdots T$.

RINOCOR data

Now we apply the full model to **RINOCOR** data by using WINBUGS to fit this full Bayesian model .
The WinBUGS code of this model is

```
model{
   for(i in 1:N){
      for(t in 1:T){
         Y[i,t]~dnorm(mu[i,t],tau.e)
         mu[i,t]<-beta0+beta*X[i,t]+alpha[i]
      }
      alpha[i]<-rho*mean(X[i,])+w[i]
   }
   #prior distribution
   for(i in 1:N){
      w[i]~dnorm(0,tau.w)
   }
```

```
    beta0~dnorm(0.0,0.0001)
    beta~dnorm(0.0,0.0001)
    rho~dnorm(0,0.0001)
    tau.e~dgamma(0.0001,0.0001)
    tau.w~dgamma(0.0001,0.0001)
}
```

In section 6.2.1, we have checked the Gelman-Rubin diagnostics and convergence diagnostics for all estimators show convergence. Then the result we have are all valid. The posterior distribution of ρ is given in Figure 6.4.

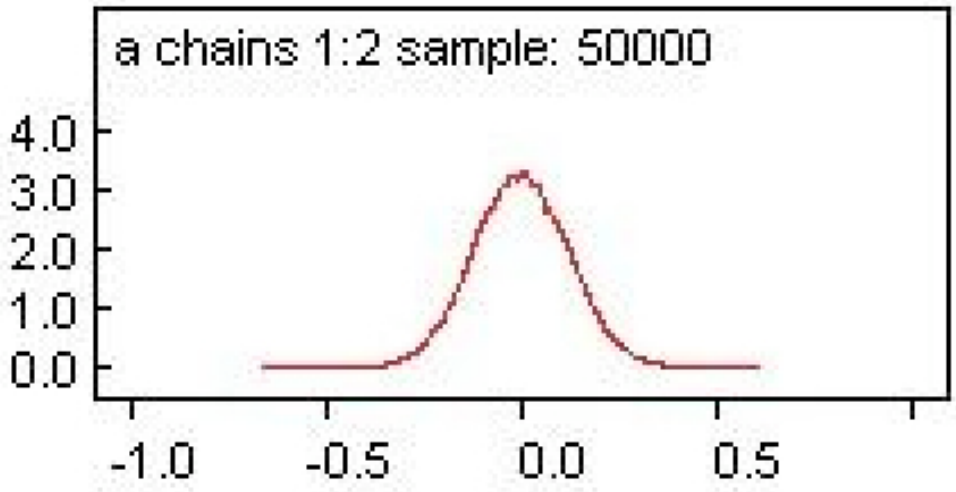

Figure 6.4: Posterior distribution of ρ for **RINOCOR** data

Figure 6.4 shows the estimate mean of ρ is -0.005. The 95 % of posterior credible interval for ρ is (-0.228, 0.217). The true value of ρ is zero which is within the credible interval and the estimate mean of ρ is approximately equal to zero. This indicates there is no correlation between explanatory variable X and individual effect α within **RINOCOR** data. Thus, the random effect estimator should be used for **RINOCOR** data.

RICOR data

The **RICOR** dataset is generated by using the Mundlak formulation, ie. Eq.(4.8). We apply the full model to **RICOR** dataset by using the WINBUGs. The program code is the same as we quote in **RINOCOR** example. The only difference is we use **RICOR** data instead of **RINOCOR** data.

In section 6.2.1, the Gelman-Rubin diagnostics and convergence diagnostics for all estimators show convergence. Then the result we have are all valid. The posterior distribution

of ρ is given in Figure 6.5.

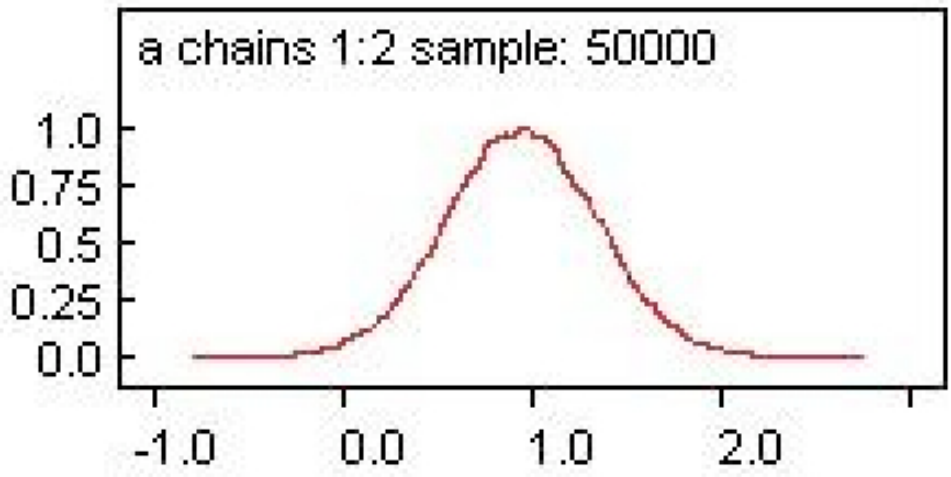

Figure 6.5: Posterior distribution of ρ for **RICOR** data

Figure 6.5 shows the estimate mean of ρ is 0.94. The 95 % posterior credible interval for ρ is (0.135, 1.742). The true value of ρ is 1. Here ρ is the degree of the correlation between the individual effects α_i and the explanatory variable X_i. The estimate of $\rho \neq 0$ indicates there is correlation exists. Therefore, the fixed effects model would produce unbiased estimates for **RICOR** dataset.

6.3.2 Real Data: WAGE

Now we use real data **WAGE** to demonstrate how does this full Bayesian model works as Hausman test. We use the full Bayesian formulation (MF Model) for **WAGE** data which is defined in section 6.2.2. The WinBUGS code of this model is

```
model{
   for(i in 1:N){
      for(t in 1:T){
         L[i,t] ~ dnorm(mu[i,t],tau.e)
         mu[i,t] <- beta0+beta[1]*S[i,t]+beta[2]*E[i,t]
         +beta[3]*E2[i,t]+beta[4]*U[i,t]+beta[5]*M[i,t]
         +beta[6]*B[i,t]+beta[7]*H[i,t]+beta[8]*P[i,t]
         +alpha[i]
      }
      alpha[i] <- rho[1]*mean(E[i,])+rho[2]*mean(E2[i,])
               +rho[3]*mean(U[i,])+rho[4]*mean(M[i,])
               +rho[5]*mean(P[i,])+w[i]
```

```
    }
    for(i in 1:N){
        w[i] ~ dnorm(0.0,tau.w)
    }                          }
    #prior distribution
    for(k in 1:8){
       beta[k] ~ dnorm(0.0,0.0001)
    }
    for(p in 1:8){
       rho[p] ~ dnorm(0.0,0.0001)
    }
    beta0 ~ dnorm(0.0,0.0001)
    tau.e ~ dgamma(0.0001,0.0001)
    tau.w ~ dgamma(0.0001,0.0001)
}
```

We have checked the Gelman-Rubin diagnostics and convergence diagnostics for all estimators show convergence. Then the result we have are all valid. The posterior distribution of ρ is given in Figure 6.6.

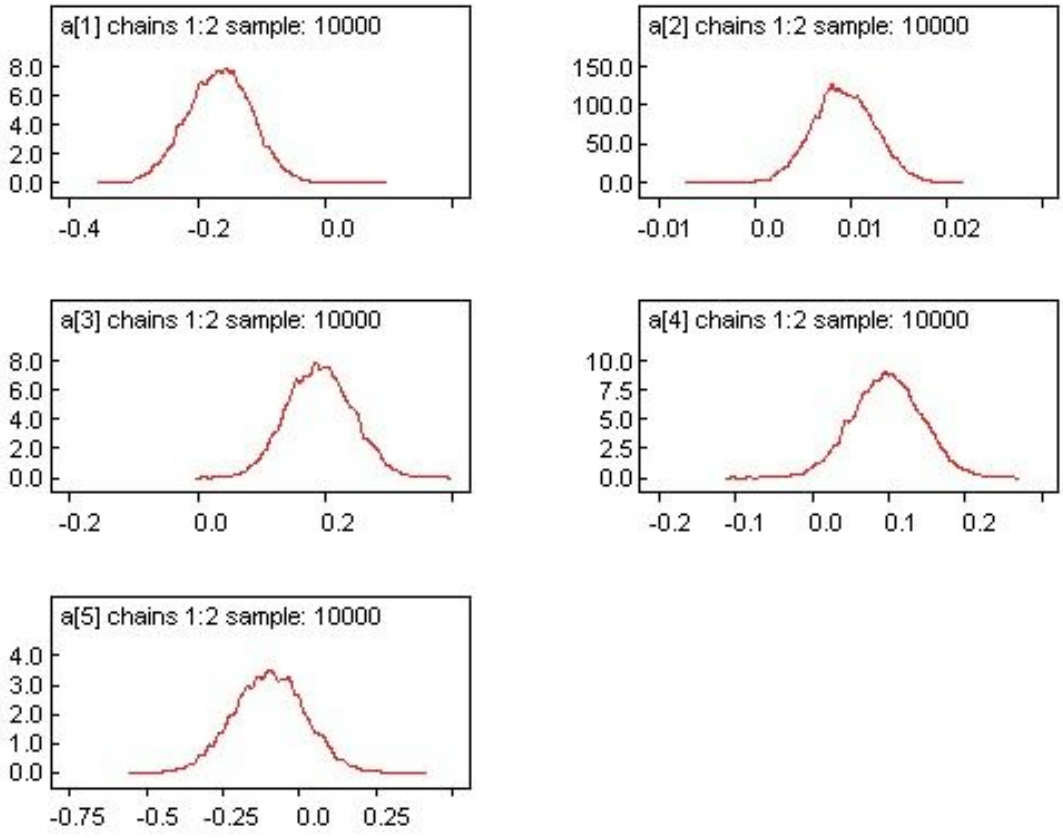

Figure 6.6: Posterior distribution of ρ_p for $exper$, $expersq$, $union$, $married$ and pub respectively in **WAGE** data where $p = 1, 2, 3, 4, 5$

Figure 6.6 shows the estimate mean of ρ_p for $exper$, $expersq$, $union$, $married$ and pub are -0.165, 0.009, 0.190, 0.097 and -0.096 respectively. The 95 % credible intervals for $exper$, $expersq$, $union$, $married$,and pub are (-0.265, -0.063), (0.003, 0.016), (0.089, 0.293), (0.010, 0.186) and (-0.320, 0.127) respectively. These indicate the $\rho_p \neq 0$ for $p = 1, 2, 3, 4$, we say they are different from zero, but ρ_5 (for pub) has zero within the 95 % credible interval. That means there is correlation between individual effect α_i and explanatory variables ($exper$, $union$, $married$ and $expersq$). The random effect estimation is biased. The fixed effect estimator should be used.

Now we define a new equation

$$\begin{aligned} \alpha_i &= \lambda_1 \text{educ}_i + \lambda_2 \overline{\text{exper}}_i + \lambda_3 \overline{\text{expersq}}_i + \lambda_4 \overline{\text{union}}_i \\ &\quad + \lambda_5 \overline{\text{married}}_i + \lambda_6 \text{black}_i + \lambda_7 \text{hisp}_i + \lambda_8 \overline{\text{pub}}_i + w_i \end{aligned} \tag{6.10}$$

where the mean of educ_{it}, black_{it} and hisp_{it} are educ_i, black_i and hisp_i,since they are time independent variables. Then we substitute Eq. (6.10) into Eq. (5.27), then we rearrange it. We have

$$\begin{aligned} \text{lwage}_{it} &= \beta_0 + (\tilde{\beta}_1 + \lambda_1)\text{educ}_i + \beta_2 \text{exper}_{it} + \beta_3 \text{expersq}_{it} + + \beta_4 \text{union}_{it} \\ &\quad + \beta_5 \text{married}_{it} + (\tilde{\beta}_6 + \lambda_6)\text{black}_{it} + (\tilde{\beta}_7 + \lambda_7)\text{hisp}_{it} + \beta_8 \text{pub}_{it} \\ &\quad + \lambda_2 \overline{\text{exper}}_i + \lambda_3 \overline{\text{expersq}}_i + \lambda_4 \overline{\text{union}}_i + \lambda_5 \overline{\text{married}}_i \\ &\quad + \lambda_8 \overline{\text{pub}}_i + w_i + \varepsilon_{it} \end{aligned} \tag{6.11}$$

We can implement both models Eq. (5.27) and Eq. (6.11) in WinBUGS. Then estimates of coefficient for time independent variables in Eq. (5.27) and Eq. (6.11) have the following relationship

$$\tilde{\beta}_p + \lambda_p = \beta_p \quad \text{for } p = 1, 6, 7$$

$$\lambda_p = \rho_p \quad \text{for } p = 2, 3, 4, 5, 8$$

Now we verify this by using real data **WAGE**.

Table 6.2: Table of result comparison from BMF model and BMF model with time independent variables in Mundlak formulation

	educ	black	hisp
BMF	0.095	-0.138	0.008
BMF (λ)	0.553	-0.732	0.398
λ	-0.458	0.593	-0.393
BMF-BMF(λ)	-0.458	0.594	-0.390

Table 6.2 shows the estimates of time independent variables from BMF model and BMF model with time independent variables in Mundlak formulation are the same. But in WinBUGS, we still use the Eq. (5.27), the Eq. (6.11) cause non-convergence. The standard error are overestimated, that lead to inefficient estimation and the Type II error rate is increased.

6.4 Results Comparison

In this section, we compare the results we produce by using computer program R and WinBUGS. Firstly, we compare the simulation datasets **RINOCOR** and **RICOR** results, then we compare the real dataset **WAGE** result to see whether there is difference between R and WinBUGS results for four models. On the Figures, we define RE, FE, PL and MF to represent the MLE estimates results, then we define BRE, BFE, BPL and BMF as Bayesian approach random effects model, fixed effects model, pooled model and full Bayesian model.

6.4.1 Simulation Example

RINOCOR

The result of **RINOCOR** from R and WinBUGS are obtained by two different approaches is showed on Table 6.3.

Table 6.3: Table of estimates comparison of **RINOCOR**

		R	WinBUGS
Model	Estimate	MLE	Bayesian
RE	β_1	1.011	1.011
	$s.e$	0.033	0.033
FE	β_1	1.015	1.017
	$s.e$	0.103	0.103
PL	β_1	1.011	1.011
	$s.e$	0.025	0.026
MF	β_1	1.016	1.015
	$s.e$	0.103	0.108

Figure 6.7 shows the Table 6.3 results graphically.

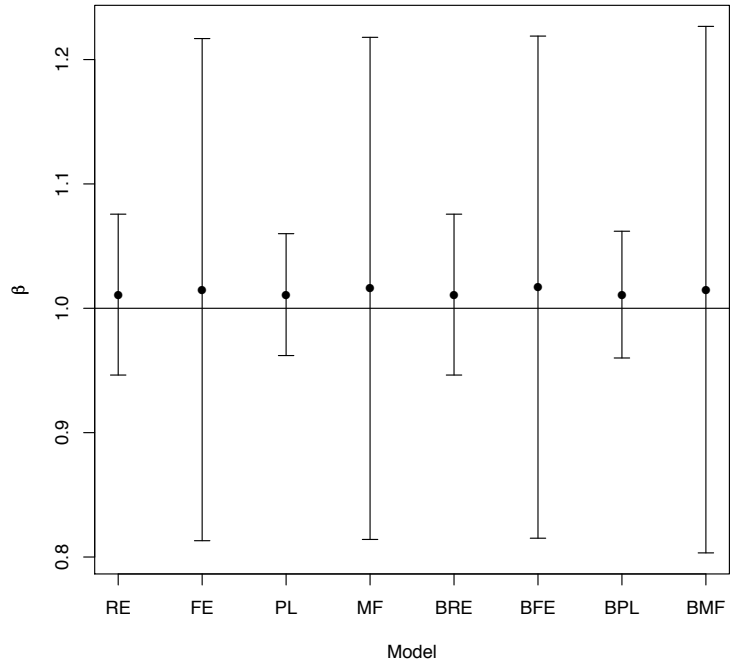

Figure 6.7: Table of estimates comparison of **RINOCOR**

In Table 6.3 and Figure 6.7, we can see there is not much difference among two approaches results for the β_1 estimate and its standard error of RE model, FE model, PL model and MF model. The true value of $\beta_1 = 1$. All of the estimates of β_1 for each model give close estimate. That means all of the model give unbiased estimates. This indicates there is no correlation. In Chapter 5, the Hausman test suggest there is no correlation between the individual effects and the explanatory variable as well. The random effect estimator should be used in this example. The full Bayesian formulation method also confirm this.

RICOR

The result of **RICOR** from R and WinBUGS obtain by different approaches is showed on Table 6.4.

Table 6.4: Table of estimates comparison of **RICOR**

		R	WinBUGS
Model	Estimate	MLE	Bayesian
RE	β_1	1.594	1.589
	$s.e$	0.225	0.282
FE	β_1	1.018	1.019
	$s.e$	0.312	0.321
PL	β_1	1.881	1.880
	$s.e$	0.130	0.131
MF	β_1	1.018	1.018
	$s.e$	0.312	0.318

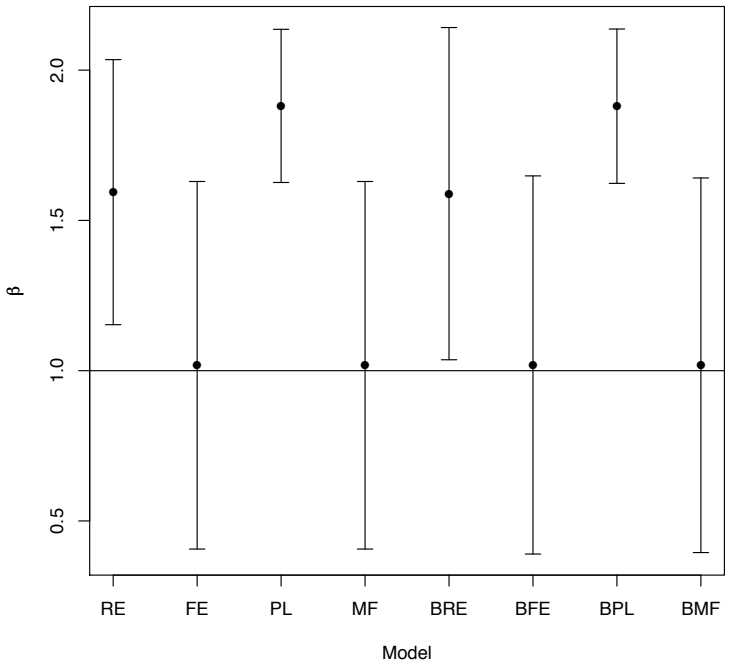

Figure 6.8: Table of estimates comparison of **RICOR**

In Table 6.4 and Figure 6.8, we can see there is not much difference among two approaches result of the β_1 estimate and its standard error. The true value of $\beta_1 = 1$. MF model and FE model have similar estimates close to the true value. The other estimates of β_1 give biased estimates. This indicates there is correlation, since FE estimate and RE estimate are different. In Chapter 5, the Hausman test suggest there is correlation between the

individual effects and the explanatory variable. The fixed effect estimator should be used in this example. The full Bayesian formulation method also confirm this. Since the MF model gives the unbiased estimates similar as FE model's estimate, we could also use MF model.

6.4.2 Real Data Example

WAGE

The result of **WAGE** from R and WinBUGS obtain by two approaches is shown in Table 6.5.

Table 6.5: Table of estimates comparison of **WAGE**

Variables	R				WinBUGS			
	RE	FE	PL	MF	BRE	BFE	BPL	BMF
constant	-0.104	-	-0.034	0.490	-0.111	-	-0.035	0.488
	0.111	-	0.065	0.221	0.112	-	0.064	0.218
education	0.101	-	0.099	0.095	0.102	-	0.099	0.095
	0.009	-	0.005	0.010	0.009	-	0.005	0.011
experience	0.112	0.116	0.089	0.116	0.112	0.117	0.089	0.117
	0.008	0.008	0.010	0.008	0.008	0.008	0.010	0.008
experience^2	-0.0041	-0.0043	-0.0028	-0.0043	-0.0041	-0.0043	-0.0028	-0.0043
	0.0006	0.0006	0.0007	0.0006	0.0006	0.0006	0.0007	0.0006
union member	0.106	0.081	0.180	0.081	0.106	0.082	0.180	0.082
	0.018	0.019	0.017	0.019	0.018	0.019	0.017	0.019
married	0.063	0.045	0.108	0.045	0.063	0.045	0.108	0.045
	0.017	0.018	0.016	0.018	0.017	0.018	0.016	0.018
black	-0.144	-	-0.144	-0.139	-0.140	-	-0.144	-0.138
	0.048	-	0.024	0.049	0.048	-	0.024	0.049
hispanic	0.020	-	0.016	0.005	0.021	-	0.016	0.008
	0.043	-	0.021	0.043	0.044	-	0.021	0.044
public sector	0.030	0.035	0.004	0.035	0.029	0.035	0.004	0.035
	0.036	0.039	0.037	0.039	0.036	0.039	0.038	0.039

Note: the second row in each block is the standard error of the estimator.

We also can graphically present the estimates which is shown on Figure 6.9 – 6.13.

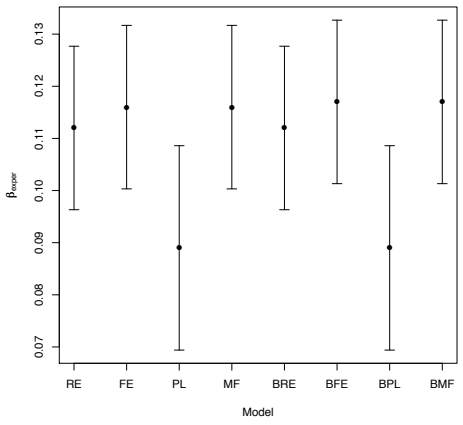

Figure 6.9: Estimates comparison for $exper$

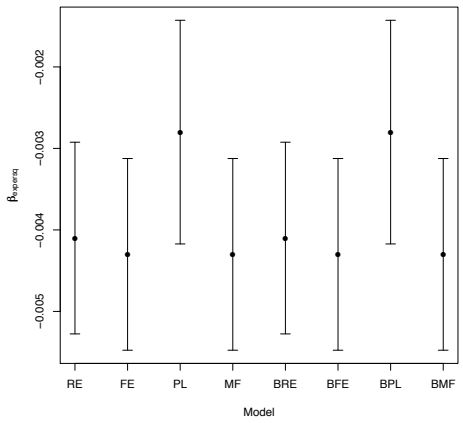

Figure 6.10: Estimates comparison for $expersq$

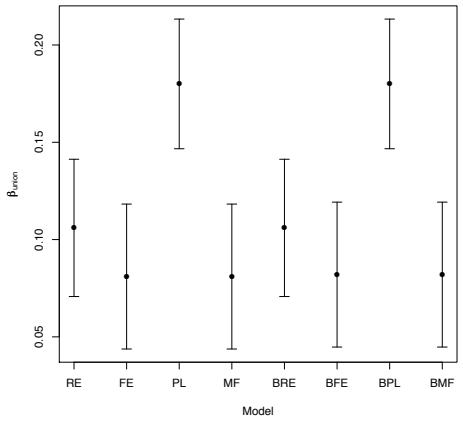

Figure 6.11: Estimates comparison for *union*

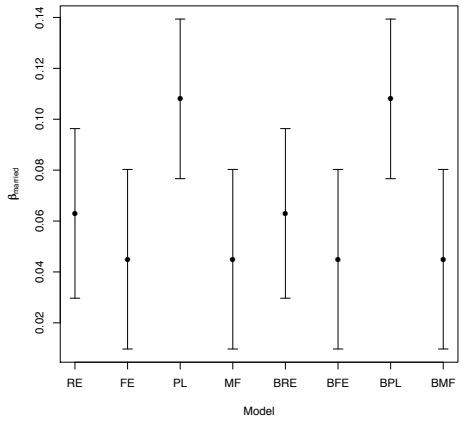

Figure 6.12: Estimates comparison for *married*

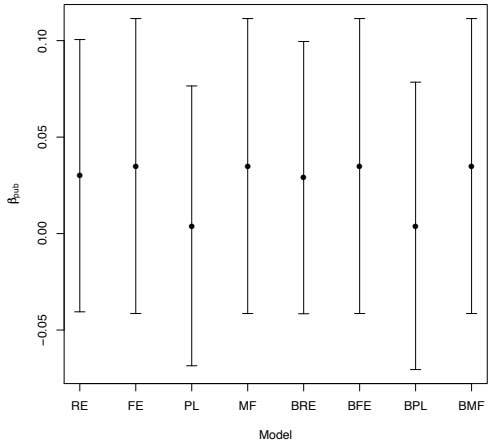

Figure 6.13: Estimates comparison for *pub*

Table 6.5 and Figure 6.9 – 6.13 show there is not much difference between the R estimation and WinBUGS estimates for each model. The FE model and MF model have the similar estimates. In Chapter 5, Hausman test gives the result that the random effects estimator is biased and we should use FE Model for **WAGE** data. The full Bayesian formulation method also suggests the fixed effect estimator is appropriate. The estimates means of ρ_p for $exper$, $union$, $married$, $hisp$ and pub have been calculated in section 6.3 which are -0.165, 0.009, 0.190, 0.097 and -0.096 respectively. The 95 % credible intervals for $exper$, $union$, $married$, $hisp$ and pub are (-0.265, -0.063), (0.003, 0.016), (0.089, 0.293), (0.010, 0.186) and (-0.320, 0.127) respectively. These indicate there is certain correlation between the covariates and individual effect, so the fixed effect estimator must be used in this case. MF model gives similar estimates as fixed effect estimates, it should be another option.

Discussion From the simulation data and real data comparison, we can see there is not much difference between the R and WinBUGs estimates for each model. Both of the computer softwares give the similar results. That means the frequentist and Bayesian statistical analysis give the same estimation results. And Bayesian estimation also suffer the bias when we use the random effect model to fit the data with correlation between explanatory variable and individual effect. The full Bayesian formulation we use in this chapter is just working as good as Hausman test and also it is easy to use compared with the Hausman test. For Hausman test, sometimes it is hard to find the inverse of the differenced covariance of random effect estimator and fixed effect estimator or in some case, the Hausman test statistic may be negative. The full Bayesian formulation do not need to worry about these problem. So the full Bayesian formulation is more easy to use to identify the correlation between the explanatory variables and the individual effects.

Chapter 7

Model Comparison

In Chapter 3, we describe three simple models which are commonly used to fit the longitudinal data. And in Chapter 5 we define a new model follow the Mundlak formulation, then we introduce the Hausman test to compare the random effect estimation and fixed effect estimation. In Chapter 6 we develop a ways to decide whether the random effect model is good or the fixed effect model is good to use for fitting based on Bayesian approach. In this chapter, we describe two commonly used model selection criteria, AIC and DIC from two approaches, and then we compare them with Hausman test and the one we developed in Chapter 6. In order to see how these methods work, we use real data and simulated data to demonstrate these methods.

7.1 Model Comparison

For generalised linear models we have the deviance as a measure of goodness of fit. For longitudinal data model fitting, if the models are nested we can use the deviance to compare the model where nested models means one model is a strict subset of the other. But in this thesis, the models (pooled model, fixed effect model and random effect model) we are comparing are not nested models, so we can't use deviance to do the non-nested model comparison. For non-nested models, we could compare the models by using AIC (was developed by Akaike [1974]) based on likelihood approach and DIC (was introduced by Spiegelhalter et al. [2002]) based on Bayesian approach. Note, for non-nested models for example comparison of models with different correlation structures, (eg. AR(1) or compound symmetry), we can't use the deviance to do the model comparison as well.

7.2 Akaike Information Criterion (AIC)

Akaike's information criterion, developed by Hirotsugu Akaike under the name of "an information criterion" (AIC) in 1971 and proposed in Akaike [1974], is a measure of the goodness of fit of an estimated statistical model and measures the amount of information lost when using a model to approximate reality. This approach measures how well different models approximate reality even though the reality may be unknown. Models that lose the least amount of information will tend to make the best predictions of datasets. Akaike demonstrated a relationship between the expected information content of a model and the log-likelihood at its maximum point.

Definition 7.1. *Based on this criterion, the model should be chosen such that*

$$AIC = 2k - 2\ell_{max} \tag{7.1}$$

where k is the number of parameters in the statistical model, and ℓ_{max} is the maximized value of the likelihood function for the estimated model.

[Demidenko, 2004].

By definition, AIC not only rewards goodness of fit, but also includes a penalty that is an increasing function of the number of estimated parameters. Hence the model with the smaller AIC value is preferred. The AIC methodology attempts to find the model that best explains the data with a minimum of parameters.

In order to use AIC as a tool for model selection, we have to define the rule to assess the difference in AIC values between two compared models. Burnham and Anderson [2002] developed rules of thumb for assessing differences in AIC values between a given model i and the model with the smallest AIC (the best model among those tested). These differences are given by ΔAIC_i (Table 7.1). All models with AIC differences of less than 2 have a substantial level of empirical support, while those within 2-4 might be regarded as the more likely candidates, those within 4-7 substantially less support, and greater than 10 essentially no support.

Table 7.1: Interpretation of the level of AIC values from Burnham and Anderson [2002]

Δ AIC$_i$	Level of Empirical Support of Model i
0–2	Substantial
2–4	Equal likely
4–7	Considerable less
$>$10	Essentially none

Note, AIC is not an overall goodness of fit test, only a comparison of relative goodness of fit.

7.2.1 Real Data Example: MILK data

Now we use **MILK** data to illustrate how AIC works to identify the best model for longitudinal data. The data were from Ms Alison Frensham (from Diggle et al. [2002]). Milk was collected weekly from 79 Australian cows to analyse for its protein content and time is measured in weeks, and the experiment was continued for 19 weeks. The cows were maintained on one of three diets: barley, a mixture of barley and lupins, or lupins alone. The dataset is shown in Figure 7.1.

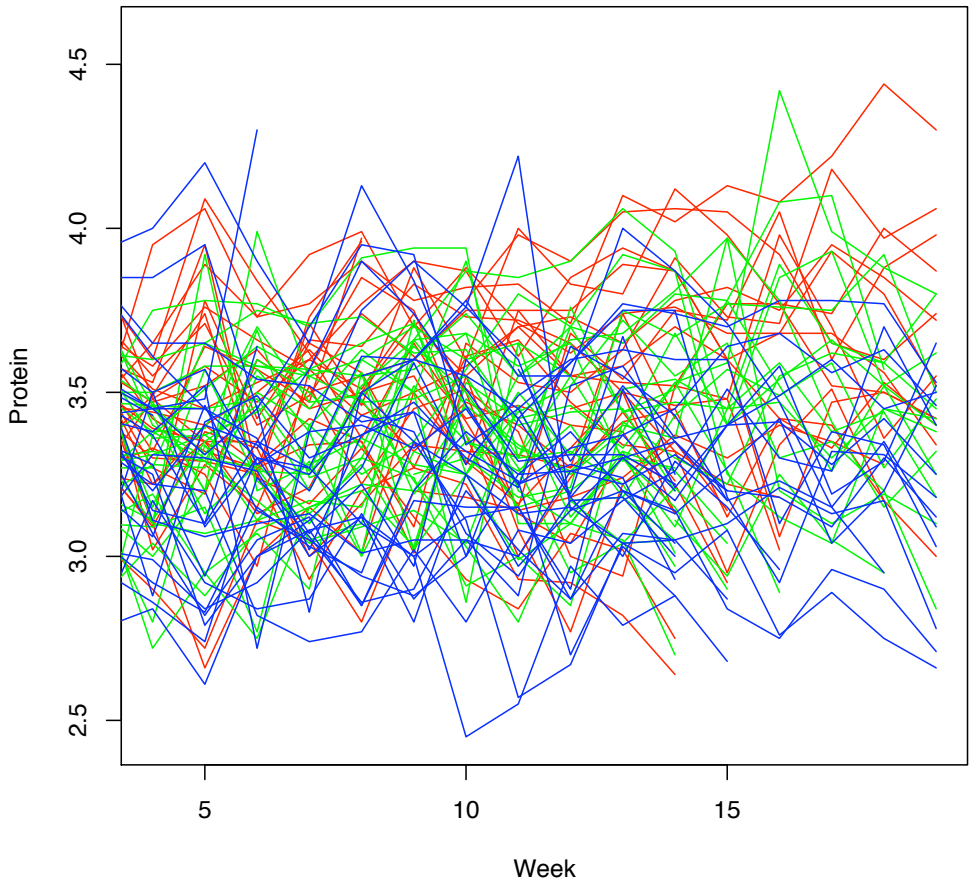

Figure 7.1: Figure of **MILK** data

Note, Figure 7.1 shows the data without the first few observations: a washout period from the previous diet.

Figure 7.1 displays the three subsets of the data corresponding to each of the three diets. The repeated measurements on each cow are joined to accentuate the longitudinal nature of the data set. From the figure, the barley gives higher values than the mixture, which in turn gives higher values than lupins alone.

To do the model comparison, firstly, we have to specify the models: we select the possible models we could fit for **MILK** data with appropriate covariance structure (ie. iid and AR(1)). The six models are defined in R as below

- Model 1: pooled model without any covariate (PLNO)

```
lm1 <- glm(protein ~ 1, data=milk))
```

- Model 2: pooled model (PL)

```
lm2 <- glm(protein ~ diet, data=milk)
```

- Model 3: fixed effect model with iid errors (FE – iid)

```
lm3 <- glm(protein ~ id + diet, data=milk)
```

- Model 4: fixed effect model with AR(1) (FE – AR (1))

```
lm4 <- gls(protein ~ diet, na.action=na.omit,
           correlation=corCAR1(form=~time|id),
           data=milk)
```

  Note, in Model 3 and 4, there are some levels eliminated here, because diet is time invariant variable.

- Model 5: random intercept model with iid errors (RI – iid)

```
lm5 <- lme(fixed = protein ~ diet, random = ~ 1 | id,
           data=milk,na.action=na.omit)
```

- Model 6: random intercept model with AR(1) (RI – AR (1))

```
lm6 <- lme(fixed = protein ~ diet, random = ~ 1 | id,
           corr=corCAR1(form=~time|id),data=milk,
           na.action=na.omit)
```

Secondly we select covariates need to be included, here the *diet* is the only covariate available. Actually time is also available which can be involved. But we don't take time into account here, because we prefer models only involving *diet* as a significant covariate (see Table 7.2). The AICs of six models is shown in Table 7.2.

Table 7.2: Table of AIC comparison for **MILK** data

Model	AIC	ΔAIC
PLNO	846.82	689.00
PL	750.73	592.91
FE-iid	423.00	265.18
FE-AR(1)	158.87	1.05
RI-iid	507.58	249.76
RI-AR(1)	157.82	0.00

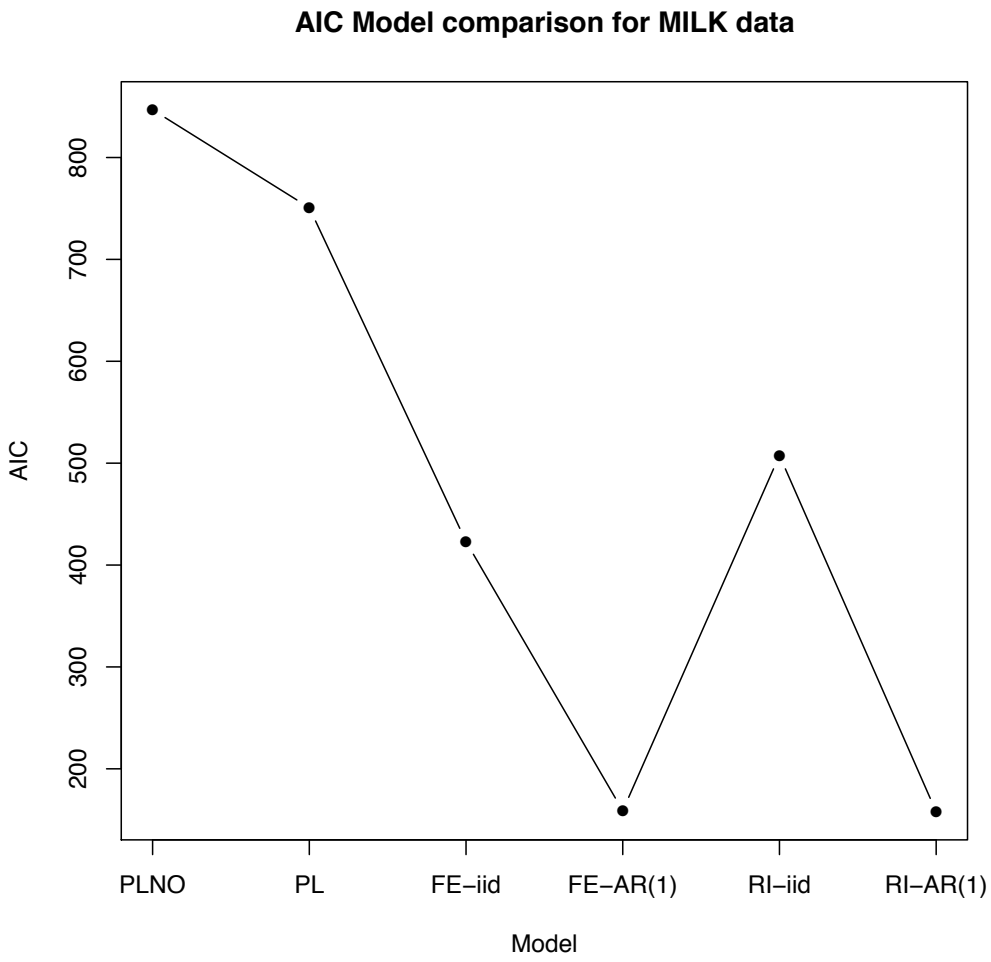

Figure 7.2: Figure of AIC comparison of **MILK** data

From Table 7.2, we first compare the PLNO Model has no covariate with PL Model has *diet* as covariate, we can see the PL Model has smaller AIC than PLNO Model, so the diet is the significant covariate. And then we compare FE – iid Model with FE – AR (1) Model, the FE – AR (1) Model has the smaller AIC, therefore, there is time series correlation. Then we compare RE – iid Model with RI – AR (1) Model, the RI – AR (1) Model has the smaller AIC, therefore, we confirm again there is time series correlation exits. Finally, we compare the RI – AR (1) with FE – AR (1), the RI – AR (1) Model has the smallest AIC. This indicate there is no correlation between *diet* and individual effects. The random effect estimator should be used in this case.

Note, we can't do a Hausman test here: *diet* is a time invariant variable, so we can't fit FE Model including it. Although this case is not good enough, but it shows when the Hausman test does not work, we could use AIC as alternative method.

Then we can use it to calculate Δ AIC, the Δ AIC of FE – AR (1) Model is 1.05 which is less than 2. According to Table 7.1, FE – AR (1) Model and RI – AR (1) Model have

comparable AIC, they are the best models among these. We have tested, and are as good as each other. Since the RI – AR(1) Model have the smallest AIC, we choose Model RI – AR (1) as our best fit model. Either we do the pairwise comparison or do the overall comparison, the conclusion is the same.

Finally, we refit the data by using the model with the smallest AIC based on maximum likelihood method (if the best model is random effect model we use REML method to refit it). So we refit data by using RI – AR (1) Model, the estimates obtained by R are listed below:

Table 7.3: Estimates of RI – AR (1) Model for **MILK** data

	Value	Std.Error	DF	t-value	p-value
(Intercept)	3.55	0.04	1260	97.10	0.0000
dietL	-0.21	0.05	76	-4.08	0.0001
dietM	-0.10	0.05	76	-1.94	0.0562

The correlation $\rho = 0.627$, the error standard deviation is $\sigma_\varepsilon = 0.319$ and the random effect standard deviation is $\sigma_\alpha = 0.108$.

7.3 Deviance Information Criterion (DIC)

[Spiegelhalter et al., 2002] proposed an alternative to AIC, known as the deviance information criterion (DIC) designed for use with Bayesian models where the posterior distributions of the models have been obtained by Markov chain Monte Carlo (MCMC) simulation. And it can be calculated easily in WinBUGS. It has a very similar form to AIC and is given by:

$$\text{DIC} = \widehat{D} + 2p_D = \overline{D} + p_D \tag{7.2}$$

- $\widehat{D}$ is the deviance when using the mean of the posterior distributions for the parameters and is given by

$$\widehat{D} = -2\log(p(y|\overline{\theta})) + C \tag{7.3}$$

  where y are the data, $\overline{\theta}$ are estimated as the posterior mean of the parameter and $p(y|\overline{\theta})$ is the likelihood function. C is a constant, which does not need to be known.

- $\overline{D}$ is the mean deviance over the chain or the posterior mean of the deviance and p_D is

the effective number of estimated parameters and is given by

$$p_D = \bar{D} - \hat{D} \tag{7.4}$$

- p_D is the posterior mean of the deviance minus the deviance of the posterior means. In hierarchical model for normal data, p_D = tr(H) where H is the "hat" matrix that maps the observed data to their fitted values.

The minimum DIC model is the best model, in the same spirit as Akaike's criterion. Very roughly, those DIC values within 5 of the smallest DIC values might be regarded as equivalent models, while those within 5-10 might be regarded as likely candidates and those greater than 10 might be regarded as poor fit models, compared to the best fit model.

The idea is that models with smaller DIC should be preferred to models with larger DIC. Models are penalized both by the value of $\bar{D}$, which favours a good fit, but also (in common with AIC) by the effective number of parameters p_D. Since $\bar{D}$ will decrease as the number of parameters in a model increases, the p_D term compensates for this effect by favouring models with a smaller number of parameters.

The advantage of DIC over other criteria, for Bayesian model selection, is that the DIC is easily calculated from the samples generated by a Markov chain Monte Carlo simulation. AIC requires calculating the likelihood at its maximum over θ, which is not readily available from the MCMC simulation in some computer software like WinBUGS.

7.4 Simulated Data Example

In this section, we use two kinds of dataset **RINOCOR** and **RICOR** to illustrate the two model comparison methods: AIC is based on likelihood approach and DIC is based on Bayesian approach. The four models we are interested are defined in section 5. Note: When we present result graphically, we define RE, FE, PL and MF to represent the four models based on frequentist approach, then we define BRE, BFE, BPL and BMF as Bayesian approach.

7.4.1 RINOCOR data

The AICs and DICs from two approaches for **RINOCOR** data are list in Table 7.4.

Table 7.4: Table of AIC and DIC comparison for **RINOCOR** data

Model	AIC	ΔAIC	DIC	ΔDIC
RE	254.52	3.15	248.66	0.00
FE	251.37	0.00	255.14	6.48
PL	258.00	6.63	258.07	9.41
MF	256.52	5.15	251.85	3.19

We also separately compare the AIC and DIC on Figure 7.3 and Figure 7.4

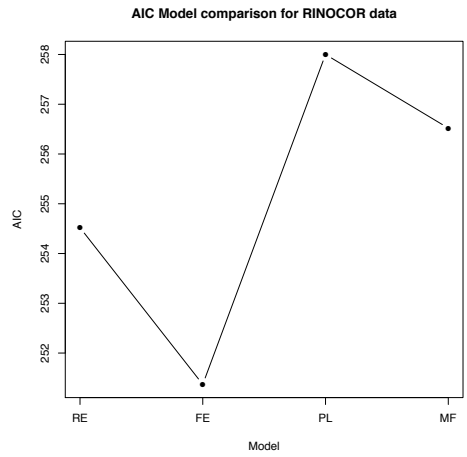

Figure 7.3: AICs comparison for **RINOCOR** data

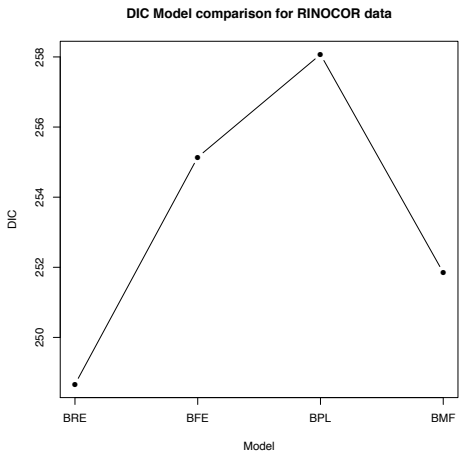

Figure 7.4: DICs comparison for **RINOCOR** data

Figure 7.3 shows the AIC method indicates FE model is the best fit model, since the FE model provides the smallest AIC. So we choose FE model as best fit model based on likelihood approach.

FE model and RE model have similar AICs. Since our true dataset is generated by using RE model, the FE model and RE model give similar AIC indicate there is no correlation between the individual effects and the explanatory variables. So the RE model should be the best model. But AIC method choose the FE model as the best model, maybe the reason is RE model have more parameters then FE model and both models provide unbiased estimates, AIC penalize the model with more parameters.

MF model AIC is slightly bigger than RE model. MF model is a special case of RE model, it suppose to improve the estimates, but with more parameters. Therefore, for no correlation case, both RE and MF model provide unbiased estimates, the RE model would be the best model compare with MF model.

From Figure 7.4, the DIC method indicates that RE model is the best fit model, since it gives smallest DIC. And MF model gives similar DIC. The true dataset does not have

correlation. So the DIC gives the correct result.

In Chapter 5, the Hausman test gives the result that the RE estimator is the unbiased and efficient, so the RE model is chosen in this case.

In chapter 6, we use the full Bayesian formulation method to compare the RE model and FE model, the full Bayesian formulation indicates there is no correlation, so the RE estimator should be used.

Thus, except AIC method, the conclusions from Hausman test, full Bayesian formulation and DIC method are consistent, the RE model is the best fit model. The true dataset has no correlation. So the DIC, Hausman test and full Bayesian formulation give the correct result.

7.4.2 RICOR data

The AICs and DICs from two approaches for **RICOR** data are listed in Table 7.5.

Table 7.5: Table of AIC and DIC comparison for FE data

Model	AIC	Δ AIC	DIC	Δ DIC
RE	321.53	28.08	297.28	2.83
FE	293.45	0.00	296.42	1.97
PL	351.73	58.28	351.77	57.32
MF	317.78	24.33	294.45	0.00

We also separately compare the AIC and DIC on Figure 7.5 and Figure 7.6

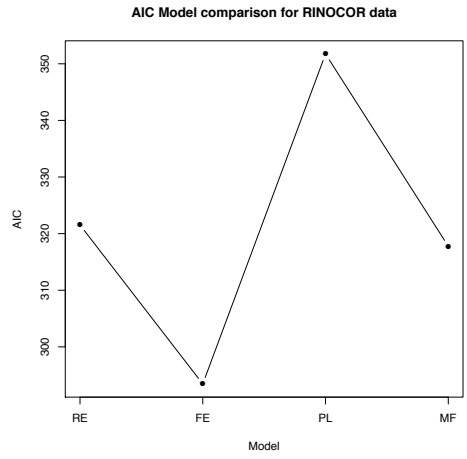

Figure 7.5: AICs comparison for **RICOR** data

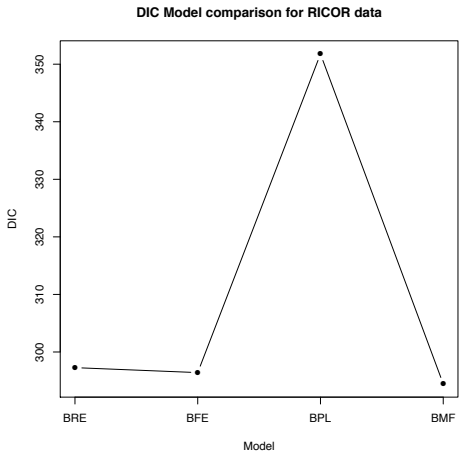

Figure 7.6: DICs comparison for **RICOR** data

The result from Table 7.5 is achieved by using R and WinBUGs. From Figure 7.5, the AIC method indicates FE model is the best model based on likelihood approach, since FE model

has the smallest Δ AIC.

From Figure 7.6, the DIC method indicates RE model, FE model and MF model are all acceptable, however, the MF model provides the smallest DIC, so we choose MF model as our best fit model for **RICOR** data based on Bayesian approach.

In Chapter 5, Hausman test shows the FE estimator is unbiased, so the FE model should be used to fit this data.

In Chapter 6, the full Bayesian formulation method indicates there is correlation, so the FE model is the best fit model.

The MF model we fit is our true model where the **RICOR** data is generated from. The MF model should be used in this case. The AIC, full Bayesian formulation method and Hausman test give the consistent result, there is correlation, the FE model is the best fit model. DIC method seems to give a incorrect result, although it choose MF as best fit model. But similar DICs for RE, FE and MF model indicate there is no correlation. So DIC fails to differentiate the important feature when there is correlation (see Figure 7.6).

7.5 Real Data Example: WAGE data

Although DIC method and AIC method have something in common, we still need to investigate whether they give the same suggestion on model selection for same longitudinal data. Here we use the same real data – **WAGE** data (introduce in Chapter 5) again to demonstrate how AIC and DIC work in model selection. We fit four models that are defined in section 6.2.2.

The AICs and DICs of four models obtained by likelihood method and Bayesian approach respectively are listed in Table 7.6.

Table 7.6: Table of AIC and DIC comparison for **WAGE** data

Model	AIC	ΔAIC	DIC	ΔDIC
RE	4407.88	643.67	3737.53	3.99
FE	3764.21	0.00	3801.61	68.07
PL	5998.42	2234.21	5998.38	2264.88
MF	4390.39	626.18	3733.54	0.00

We also present the AICs and DICs on Figure 7.7 and Figure 7.8.

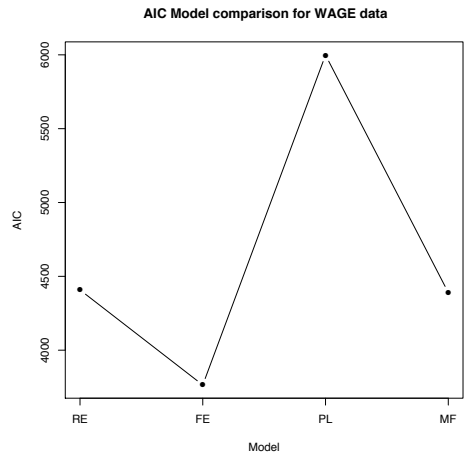

Figure 7.7: AICs comparison for **WAGE** data

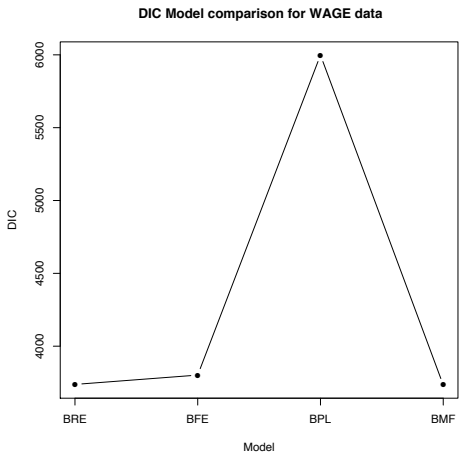

Figure 7.8: DICs comparison for **WAGE** data

Figure 7.7 shows the FE model has the smallest AIC among four models and MF model gives smaller AIC than RE model, this means MF model slightly improve the fitting. FE model gives enormous different AIC to compare with RE model AIC. This indicates there is correlation between the individual effects and the explanatory variables.

Figure 7.8 shows MF model has the smallest DIC, so it is the best model among four models. RE model gives similar DIC as MF model. This indicates there is no correlation.

In Chapter 5, we use Hausman test to compare the RE estimator and FE estimator which gives that we reject the null hypothesis at 5% of significant level (refer to Chapter 5).

In Chapter 6, we use the full Bayesian formulation method to compare the random effect model and fixed effect model. This method suggests that there is correlation, so the best fit model is the FE Model.

Therefore, the results are obtained from AIC based on likelihood method is consistent with the result obtain from Hausman test and the full Bayesian formulation method. The result from DIC may be incorrect.

Summary We summarise the results from the AIC, DIC, Hausman test (H-T) and full Bayesian formulation method (B-MF) for simulated datasets and the **WAGE** data in Table 7.8 and Table 7.7.

Table 7.7: Table of result comparison of four methods for simulated data

True Model	Model Fit	AIC	DIC	H-T	B-MF
RE	RE	✓*	✓	✓	✓
	FE	✓			
	PL			-	-
	BM		✓*	-	-
FE	RE		✓*		
	FE	✓	✓*	✓	✓
	PL			-	-
	BM		✓	-	-

Table 7.8: Table of result comparison of four methods **WAGE**

Model Fit	AIC	DIC	H-T	B-MF
RE		✓*		
FE	✓		✓	✓
PL			-	-
BM		✓	-	-

Note:

✓ means this is the selected model using the relevant criterion.

✓* means this model give a similar values.

− means this combination of model selection criterion was not done.

As we show on the Table 7.7, these four methods not always return the same result. Generally, AIC is not working very well when there is no correlation between the covariates and the individual effects. It seems AIC is more prefer the FE model than RE model. DIC gives incorrect result when there is correlation. e.g, for the **WAGE** data, Hausman test suggests the FE estimator is the best one (explained in detail in chapter 5). And the DIC is given that the RE model as the best fit. Hence we can't rely on DIC as model selection method, instead of this, we could use the full Bayesian formulation method. The full Bayesian formulation method is working very well in both real data case and simulated data cases. This method can indicate whether there is correlation between the covariates and the individual effects or not, by doing this, we can use it as model selection method. Therefore, in order to specify the best model for longitudinal data, do not just rely on a single method when do the model comparison.

Chapter 8

Conclusions

In this thesis, we have explored the estimation bias in the situation where a correlation exists between the explanatory variables and individual effects in longitudinal data. The main objectives of our research is to demonstrate the occurrence of bias in two cases, omitted variable bias by using frequentist statistic and heterogeneity bias, by using frequentist statistics and Bayesian statistics. We state the structure of the longitudinal data used in this thesis and describe two popular analysis estimation methods, least square estimation and maximum likelihood estimation. Then we describe the properties of three simple models random effect model, fixed effect model and pooled model by using these two approaches. The difference of these two approaches for each model have been compared and discussed in Chapter 3. We also proved the bias exists empirically (via simulation), we use the R software to generate the dataset with and without correlation case in Chapter 4. In Chapter 5 and Chapter 6, we used frequentist and Bayesian statistics to show theoretically and empirically the bias exists in particular cases.

In Chapter 5, we considered the omitted variables bias and heterogeneity bias.

- Omitted variables: we showed the bias does exist when we fit a omitted variable model. We proved the bias is affected by the size of the covariance of the explanatory variable and the omitted variable. If this covariance is zero, there is no bias for the coefficient of the explanatory variable. However, the bias for the intercept still existed and the size of this is the mean of the omitted variable. We assume a particular case, then used the R software to simulate the dataset 1000 times, the estimations we got confirmed our finding. Then a real data example is used to confirm the covariance of explanatory variable and the omitted variable does have an effect on estimation.

- Heterogeneity bias: we proved the random effect estimator is unbiased and consistent

compare with fixed effect estimator and pooled estimator when there is no correlation between explanatory variable and individual effect. When we assumed there is such correlation, we had used the formula introduced by Mundlak on 1978 to prove the bias exists and derived the formulation of this bias for random effect estimation. Also, we derived the bias formulation for the pooled estimation. However, when there is correlation, the fixed effect estimation is unbiased. So it is the appropriate estimation which should be used under this situation. Then we used the simulation data and real data to demonstrate this theory is correct by fitting four simple models, random effects model, fixed effects model, pooled model and Mundlak formulation model. The simulated two datasets **RINOCOR** and **RICOR** are generated by using the function defined in Chapter 4. The real data is from the popular National Longitudinal Survey held in USA [Wooldridge, 2009]. And we proved the Mundlak formulation model gives unbiased estimates when there is correlation, although it is a special case of random effects model. The Hausman test which we used also confirmed the fixed effect estimator is more appropriate in this case. Then we use the Hausman test to study how the degree of correlation affects the choice of using fixed effect estimator. Then we found when the correlation is small, the random effect estimator still can be used as unbiased estimator. This conclusion is made based on our empirical result. Finally, we use the simulated data and real data to demonstrate how the Hausman test works as a model comparison method.

In Chapter 6, we firstly discuss the principle idea of the WinBUGS software and we show how the Bayesian estimation works. Then we investigated whether the Bayesian estimation did suffer from the same bias as the random effect estimation in frequentist analysis. Secondly, we used the Mundlak formula to propose a computational approach to check whether there is a correlation between the explanatory variables and the individual effects. Then we used WinBUGS to apply this approach on the real data (*WAGE*) and simulated data generated in Chapter 5. Finally, we compared the estimates from frequentist and Bayesian analysis which showed there is not much different on estimates from both analysis approaches.

In Chapter 7, we used the AIC based on likelihood method and DIC based on Bayesian method to do the model comparison. And then we compare these two method with Hausman test result and the full Bayesian formulation method result which we defined by following Mundlak's formulation. AIC, Hausman test and full Bayesian formulation methods all confirmed the fixed effect model is the best model when there is correlation between the

explanatory variables and the individual effects. But DIC gives different conclusion which has been proved is incorrect. So we couldn't rely on the DIC method.

In this thesis, we demonstrated this conclusion theoretically and empirically. We also investigate the Bayesian approach suffer the same bias when use the random effect model under correlation situation. By using the common model selection method DIC under Bayesian approach, we found the DIC can't produce the correct result. But we developed a full Bayesian model as Hausman test, to compare the fixed effects model and random effects model under no correlation and correlation cases. And we have proved it works as good as Hausman test. Therefore, this thesis study the correlation affects on bias and compare the frequentist approach with Bayesian approach on estimates and model comparison methods. If there is correlation between the explanatory variables and individual effects, the fixed effect estimator is the appropriate estimator which is unbiased and efficient. Otherwise, the random effect estimator is unbiased and efficient. Of course, this conclusion is based on a certain assumption about the correlation between the explanatory variables and the individual effects only and more study are required under the other assumptions (i.e. the error term is correlated with the explanatory variables). In future, we could investigate the Mundlak formulation under other assumptions for both frequentist and Bayesian approaches to see whether it still gives the unbiased estimation under certain assumption.

Appendix A

Blockwise Matrix Inversion

Let a block matrix be $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$. Then the inverse of this block matrix is

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}$$

Now we verify this result.

$$\begin{aligned} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} &= \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix} \\ &= \begin{bmatrix} E & F \\ G & H \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} = I \end{aligned}$$

where

$$\begin{aligned} E &= A(A - BD^{-1}C)^{-1} - BD^{-1}C(A - BD^{-1}C)^{-1} \\ &= (A - BD^{-1}C)^{-1}(A - BD^{-1}C) = I \end{aligned}$$

$$\begin{aligned} F &= -A(A - BD^{-1}C)^{-1}BD^{-1} + BD^{-1} + BD^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \\ &= BD^{-1}[-A(A - BD^{-1}C)^{-1} + I + BD^{-1}C(A - BD^{-1}C)^{-1}] \\ &= BD^{-1}[I - (A - BD^{-1}C)^{-1}(A - BD^{-1}C)] \\ &= BD^{-1}[I - I] = 0 \end{aligned}$$

$$
\begin{aligned}
G &= C(A - BD^{-1}C)^{-1} - DD^{-1}C(A - BD^{-1}C)^{-1} \\
&= (A - BD^{-1}C)^{-1}[C - DD^{-1}C] \\
&= 0
\end{aligned}
$$

$$
\begin{aligned}
H &= -C(A - BD^{-1}C)^{-1}BD^{-1} + DD^{-1} + DD^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \\
&= DD^{-1} - C(A - BD^{-1}C)^{-1}BD^{-1} + DD^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \\
&= I - C(A - BD^{-1}C)^{-1}BD^{-1} + C(A - BD^{-1}C)^{-1}BD^{-1} \\
&= I
\end{aligned}
$$

Therefore, the inverse of the block matrix $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ is

$$
\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}
$$

Appendix B

Verification of Inverse Matrix V^{-1}

To verify the $V^{-1}V = I$, we have to use the Eq. (5.17) from Chapter 5.

$$V^{-1} = \frac{1}{\sigma_u^2}[I_T - \frac{\sigma_\alpha^2}{\sigma_u^2 + T\sigma_\alpha^2}\mathbf{1}\mathbf{1}^T]$$

Then we could rewrite the inverse of V^{-1} as

$$\begin{aligned} V^{-1} &= \frac{1}{\sigma_u^2}[I_T - (\frac{1}{T} - \frac{1}{T}\psi)\mathbf{1}\mathbf{1}^T] \\ &= \frac{1}{\sigma_u^2}[I_T - \frac{1}{T}(1-\psi)\mathbf{1}\mathbf{1}^T] \\ &= \frac{1}{\sigma_u^2}[I_T - \frac{1}{T}\mathbf{1}\mathbf{1}^T + \frac{1}{T}\psi\mathbf{1}\mathbf{1}^T] \\ &= \frac{1}{\sigma_u^2}[Q + \psi P] \end{aligned}$$

where

$$\psi = \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2}$$

We also write V as

$$\begin{aligned} V &= \sigma_u^2[I_T - (\frac{1}{T} - \frac{1}{T}\psi^{-1})\mathbf{1}\mathbf{1}^T] \\ &= \sigma_u^2[I_T - \frac{1}{T}(1-\psi^{-1})\mathbf{1}\mathbf{1}^T] \\ &= \sigma_u^2[I_T - \frac{1}{T}\mathbf{1}\mathbf{1}^T + \frac{1}{T}\psi^{-1}\mathbf{1}\mathbf{1}^T] \\ &= \sigma_u^2[Q + \psi^{-1}P] \end{aligned}$$

So $V^{-1}V$ is

$$V^{-1}V = QQ + \psi PQ + \psi^{-1}PQ + PP = Q + P = I$$

where $QQ = Q$, $PP = P^2 = P$ and $PQ = 0$ from Chapter 3. Therefore, $V^{-1}V = I$.

Appendix C

GLSE for Mundlak Formulation

Refer to section 3.2.2, we use the similar way to prove the GLSE for Mundlak formulation.

C.1 Proof:

Since $K = \begin{bmatrix} X & PX \end{bmatrix}$, then $K^T = \begin{bmatrix} X^T \\ X^T P^T \end{bmatrix}$.

$$K^T K = \begin{bmatrix} X^T X & X^T PX \\ X^T P^T X & X^T P^T PX \end{bmatrix}$$

$$\begin{aligned} (K^T K)^{-1} &= \begin{bmatrix} 0 & 0 \\ 0 & (X^T P^T PX)^{-1} \end{bmatrix} + \\ &\quad \begin{bmatrix} I \\ -(X^T P^T PX)^{-1} X^T P^T X \end{bmatrix} (X^T X - X^T PX (X^T P^T PX)^{-1} X^T P^T X)^{-1} \\ &\quad \begin{bmatrix} I & -X^T PX (X^T P^T PX)^{-1} \end{bmatrix} \\ &= \begin{bmatrix} A & B \\ C & D \end{bmatrix} \end{aligned}$$

where

$$\begin{aligned} A &= (X^TX - X^TPX(X^TP^TPX)^{-1}X^TP^TX)^{-1} \\ &= (X^TX - X^TPX(X^TPX)^{-1}X^TP^TX)^{-1} \\ &= (X^TX - X^TP^TX)^{-1} \\ &= (X^TX - X^TPX)^{-1} \\ &= (X^T(I-P)X)^{-1} \\ &= (X^TQX)^{-1} \end{aligned}$$

$$\begin{aligned} B &= -X^TPX(X^TP^TPX)^{-1}(X^TX - X^TPX(X^TP^TPX)^{-1}X^TP^TX)^{-1} \\ &= -X^TPX(X^TPX)^{-1}(X^TX - X^TPX(X^TPX)^{-1}X^TP^TX)^{-1} \\ &= -(X^TQX)^{-1} \end{aligned}$$

$$\begin{aligned} C &= -(X^TP^TPX)^{-1}X^TP^TX(X^TX - X^TPX(X^TP^TPX)^{-1}X^TP^TX)^{-1} \\ &= -(X^TP^TX)^{-1}X^TP^TX(X^TX - X^TPX(X^TPX)^{-1}X^TP^TX)^{-1} \\ &= -(X^TQX)^{-1} \end{aligned}$$

$$\begin{aligned} D &= (X^TP^TPX)^{-1} + (X^TP^TPX)^{-1}X^TP^TXX^TPX(X^TP^TPX)^{-1} \\ &\quad (X^TX - X^TPX(X^TP^TPX)^{-1}X^TP^TX)^{-1} \\ &= (X^TPX)^{-1} + (X^TP^TX)^{-1}X^TP^TXX^TPX(X^TPX)^{-1} \\ &\quad (X^TX - X^TPX(X^TP^TX)^{-1}X^TP^TX)^{-1} \\ &= (X^TPX)^{-1} + (X^TQX)^{-1} \end{aligned}$$

Now Eq.(5.22) can be written as

$$\widehat{\boldsymbol{\delta}} = \begin{bmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{a}} \end{bmatrix} = (K^TK)^{-1}K^T\mathbf{y} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} X^T \\ X^TP^T \end{bmatrix} \mathbf{y}$$

Thus, the GLS estimate of $\widehat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{a}}$

$$\begin{aligned} \widehat{\boldsymbol{\beta}} &= [AX^T + BX^TP^T]\mathbf{y} \\ &= (X^TQX)^{-1}X^T - (X^TQX)^{-1}X^TP^T)\mathbf{y} \\ &= (X^TQX)^{-1}X^T(I-P)\mathbf{y} \\ &= (X^TQX)^{-1}X^TQ\mathbf{y} \end{aligned}$$

$$
\begin{aligned}
\widehat{\mathbf{a}} &= [CX^T + DX^T P^T]\mathbf{y} \\
&= \left[-(X^T QX)^{-1}X^T + (X^T PX)^{-1}X^T P^T + (X^T QX)^{-1}X^T P^T\right]\mathbf{y} \\
&= (X^T PX)^{-1}X^T P^T\mathbf{y} - \left[(X^T QX)^{-1}X^T (I-P)\right]\mathbf{y} \\
&= (X^T PX)^{-1}X^T P\mathbf{y} - (X^T QX)^{-1}X^T Q\mathbf{y} \\
&= \widehat{\boldsymbol{\beta}}_B - \widehat{\boldsymbol{\beta}}_W
\end{aligned}
$$

Therefore, the GLS for Mundlak formulation gives

$$\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}_W$$

$$\widehat{\mathbf{a}} = \widehat{\boldsymbol{\beta}}_B - \widehat{\boldsymbol{\beta}}_W$$

Appendix D

WINBUGS Codes

D.1 RINOCOR and RICOR

D.1.1 RE Model

```
model {
   for(i in 1:N){
      for(t in 1:T){
         Y[i,t] ~ dnorm(mu[i,t],tau.e)
         mu[i,t] <- beta0+beta1*X[i,t]+alpha[i]
      }
      alpha[i] ~ dnorm(0,tau.a)
   }
   #prior distribution
   beta0 ~ dnorm(0.0,0.0001)
   beta1 ~ dnorm(0.0,0.0001)
   tau.e ~ dgamma(0.0001,0.0001)
   tau.a ~ dgamma(0.0001,0.0001)
}
```

D.1.2 FE Model

```
model{
   for(i in 1:N){
      for(t in 1:T){
```

```
            Y[i,t]~dnorm(mu[i,t],tau.e)
            mu[i,t]<-beta1*X[i,t]+alpha[i]
        }
        alpha[i]~dnorm(0,0.001)
    }
    #prior distribution
    beta1~dnorm(0,0.0001)
    tau.e~dgamma(0.0001,0.0001)
    s.e<-sqrt(1/tau.e)
}
```

D.1.3 PL Model

```
model{
    for(i in 1:N){
        for(t in 1:T){
            Y[i,t]~dnorm(mu[i,t],tau.e)
            mu[i,t]<-beta0+beta1*X[i,t]
        }
    }
    #prior distribution
    beta0~dnorm(0.0,0.0001)
    beta1~dnorm(0,0.0001)
    tau.e~dgamma(0.0001,0.0001)
    s.e<-sqrt(1/tau.e)
}
```

D.1.4 MF Model

```
model{
    for(i in 1:N){
        for(t in 1:T){
            Y[i,t]~dnorm(mu[i,t],tau.e)
            mu[i,t]<-beta0+beta1*X[i,t]+alpha[i]
        }
```

```
      alpha[i]<-rho*mean(X[i,])+w[i]
   }
   #prior distribution
   for(i in 1:N){
      w[i]~dnorm(0,tau.w)
   }
   beta0~dnorm(0.0,0.0001)
   beta1~dnorm(0.0,0.0001)
   rho~dnorm(0,0.0001)
   tau.e~dgamma(0.0001,0.0001)
   tau.w~dgamma(0.0001,0.0001)
}
```

D.2 WAGE

D.2.1 RE Model

```
model{
   for(i in 1:N){
      for(t in 1:T){
         L[i,t]~dnorm(mu[i,t],tau.e)
         mu[i,t]<-beta0+beta[1]*S[i,t]+beta[2]*E[i,t]
         +beta[3]*E2[i,t]+beta[4]*U[i,t]+beta[5]*M[i,t]
         +beta[6]*B[i,t]+beta[7]*H[i,t]+beta[8]*P[i,t]
         +alpha[i]
      }
      alpha[i]~dnorm(0,tau.a)
   }
   #prior distribution
   for(v in 1:8){
      beta[v]~dnorm(0.0,0.0001)
   }
   beta0~dnorm(0.0,0.0001)
   tau.e~dgamma(0.0001,0.0001)
```

```
    tau.a~dgamma(0.0001,0.0001)
}
```

D.2.2 FE Model

```
model{
    for(i in 1:N){
        for(t in 1:T){
            L[i,t]~dnorm(mu[i,t],tau.e)
            mu[i,t]<-beta[1]*E[i,t]+beta[2]*E2[i,t]
            +beta[3]*U[i,t]+beta[4]*M[i,t]+beta[5]*P[i,t]
            +alpha[i]
        }
        alpha[i]~dnorm(0,0.0001)
    }
    #prior distribution
    for(v in 1:5){
        beta[v]~dnorm(0.0,0.0001)
    }
    tau.e~dgamma(0.0001,0.0001)
    s.e<-1/sqrt(tau.e)
}
```

D.2.3 PL Model

```
model{
    for(i in 1:N){
        for(t in 1:T){
            L[i,t]~dnorm(mu[i,t],tau.e)
            mu[i,t]<-beta0+beta[1]*S[i,t]+beta[2]*E[i,t]
            +beta[3]*E2[i,t]+beta[4]*U[i,t]+beta[5]*M[i,t]
            +beta[6]*B[i,t]+beta[7]*H[i,t]+beta[8]*P[i,t]
        }
    }
    #prior distribution
```

```
    for(v in 1:8){
       beta[v]~dnorm(0.0,0.0001)
    }
    beta0~dnorm(0.0,0.0001)
    tau.e~dgamma(0.0001,0.0001)
}
```

D.2.4 MF Model

```
model{
    for(i in 1:N){
       for(t in 1:T){
          L[i,t] ~ dnorm(mu[i,t],tau.e)
          mu[i,t] <- beta0+beta[1]*S[i,t]+beta[2]*E[i,t]
          +beta[3]*E2[i,t]+beta[4]*U[i,t]+beta[5]*M[i,t]
          +beta[6]*B[i,t]+beta[7]*H[i,t]+beta[8]*P[i,t]
          +alpha[i]
       }
       alpha[i] <- rho[1]*mean(E[i,])+rho[2]*mean(E2[i,])
                   +rho[3]*mean(U[i,])+rho[4]*mean(M[i,])
                   +rho[5]*mean(P[i,])+w[i]
    }
    for(i in 1:N){
       w[i] ~ dnorm(0.0,tau.w)
    }                        }
    #prior distribution
    for(k in 1:8){
       beta[k] ~ dnorm(0.0,0.0001)
    }
    for(p in 1:8){
       rho[p] ~ dnorm(0.0,0.0001)
       }
    beta0 ~ dnorm(0.0,0.0001)
    tau.e ~ dgamma(0.0001,0.0001)
```

```
    tau.w ~ dgamma(0.0001,0.0001)
}
```

Appendix E

WINBUGs Output

E.1 RINOCOR

E.1.1 RE Model

The Gelman-Rubin diagnostics for each estimate are

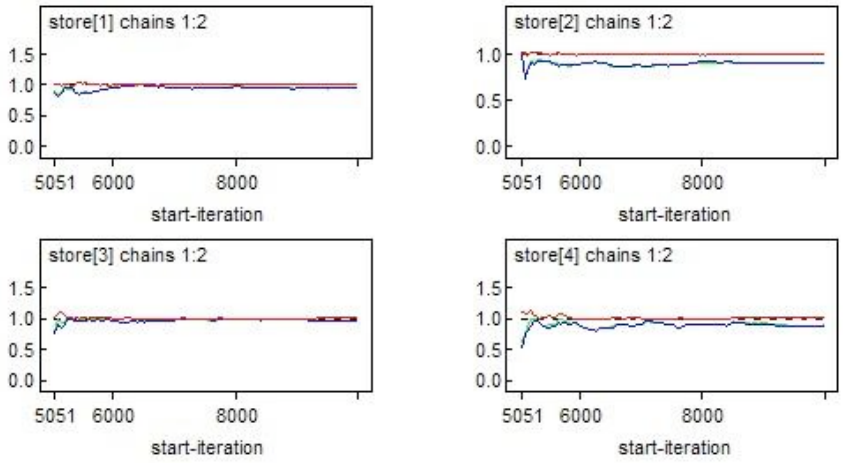

Figure E.1: Gelman-Rubin diagnostics of β_0, β_1, σ_ε and σ_α

The convergence chains for each estimate are

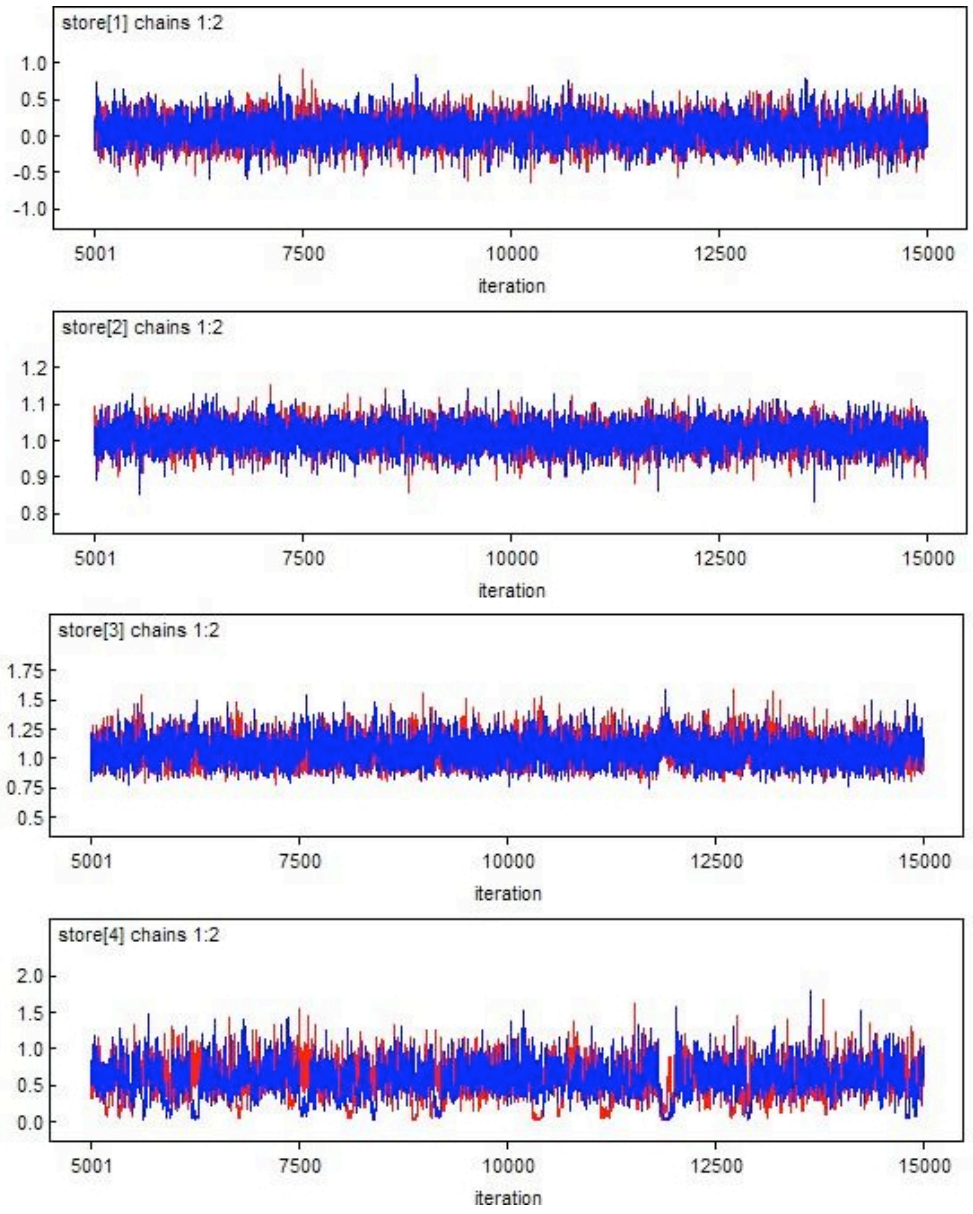

Figure E.2: Convergence chains of β_0, β_1, σ_ε and σ_α

The posterior distribution for each estimate are

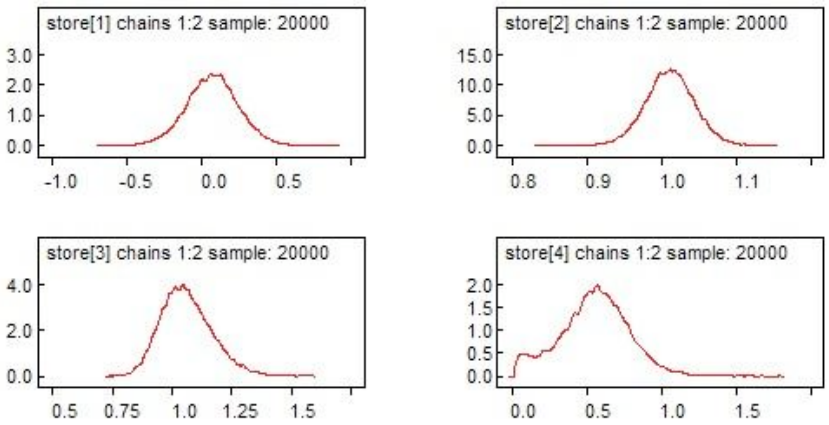

Figure E.3: Posterior distribution of β_0, β_1, σ_ε and σ_α

E.1.2 FE Model

The Gelman-Rubin diagnostics for each estimate are

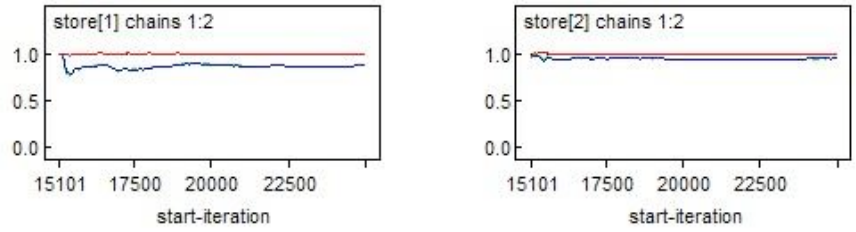

Figure E.4: Gelman-Rubin diagnostics of β_1 and σ_ε

The convergence chains for each estimate are

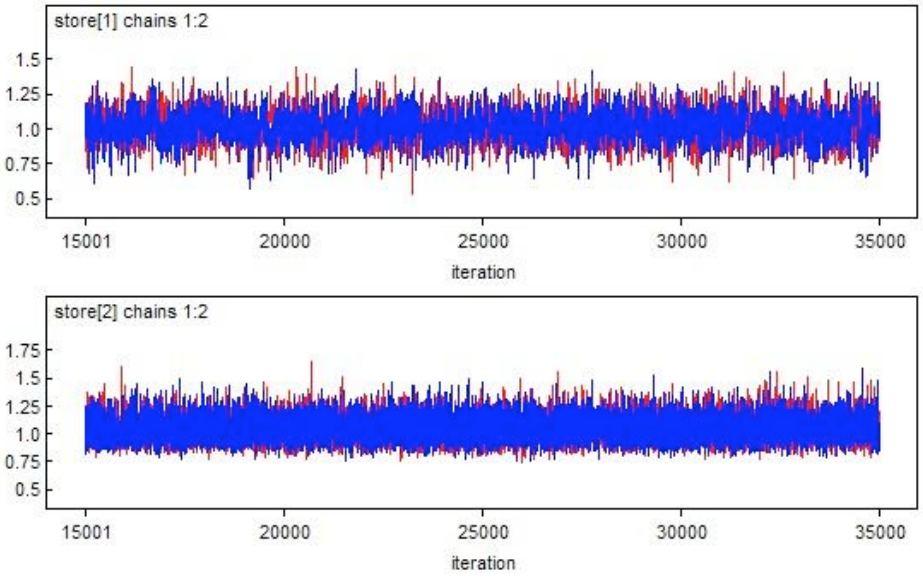

Figure E.5: Convergence chains of β_1 and σ_ε

The posterior distribution for each estimate are

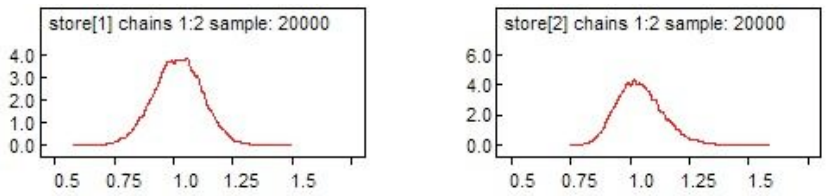

Figure E.6: Posterior distribution of β_1 and σ_ε

E.1.3 PL Model

The Gelman-Rubin diagnostics for each estimate are

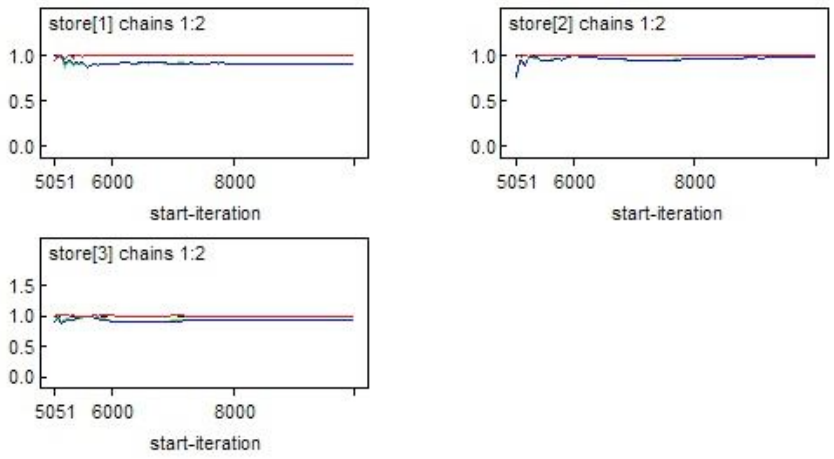

Figure E.7: Gelman-Rubin diagnostics of β_0, β_1 and σ_ε

The convergence chains for each estimate are

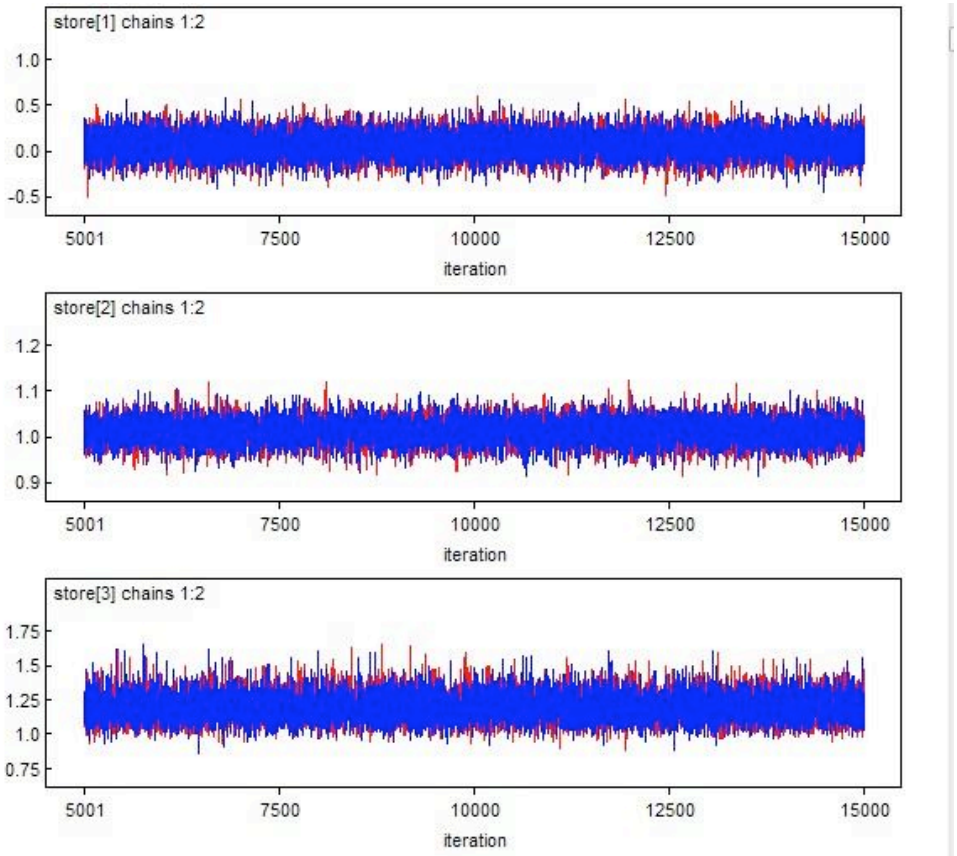

Figure E.8: Convergence chains of β_0, β_1 and σ_ε

The posterior distribution for each estimate are

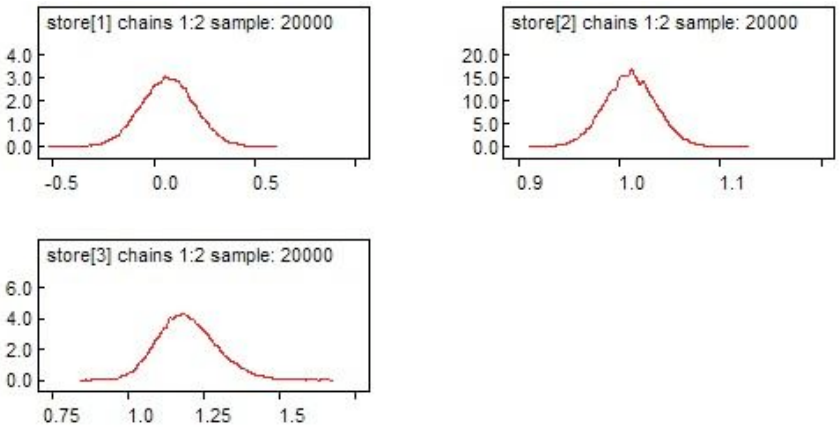

Figure E.9: Posterior distribution of β_0, β_1 and σ_ε

E.1.4 MF Model

The Gelman-Rubin diagnostics for each estimate are

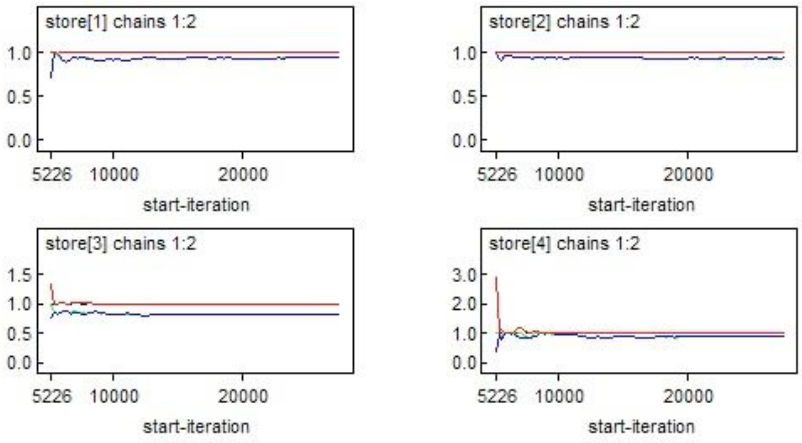

Figure E.10: Gelman-Rubin diagnostics of β_0, β_1, σ_ε and σ_α

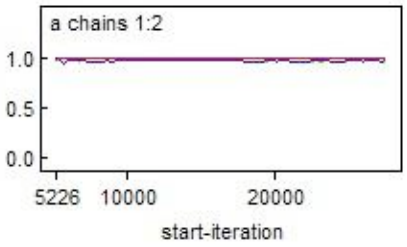

Figure E.11: Gelman-Rubin diagnostics of ρ

The convergence chains for each estimate are

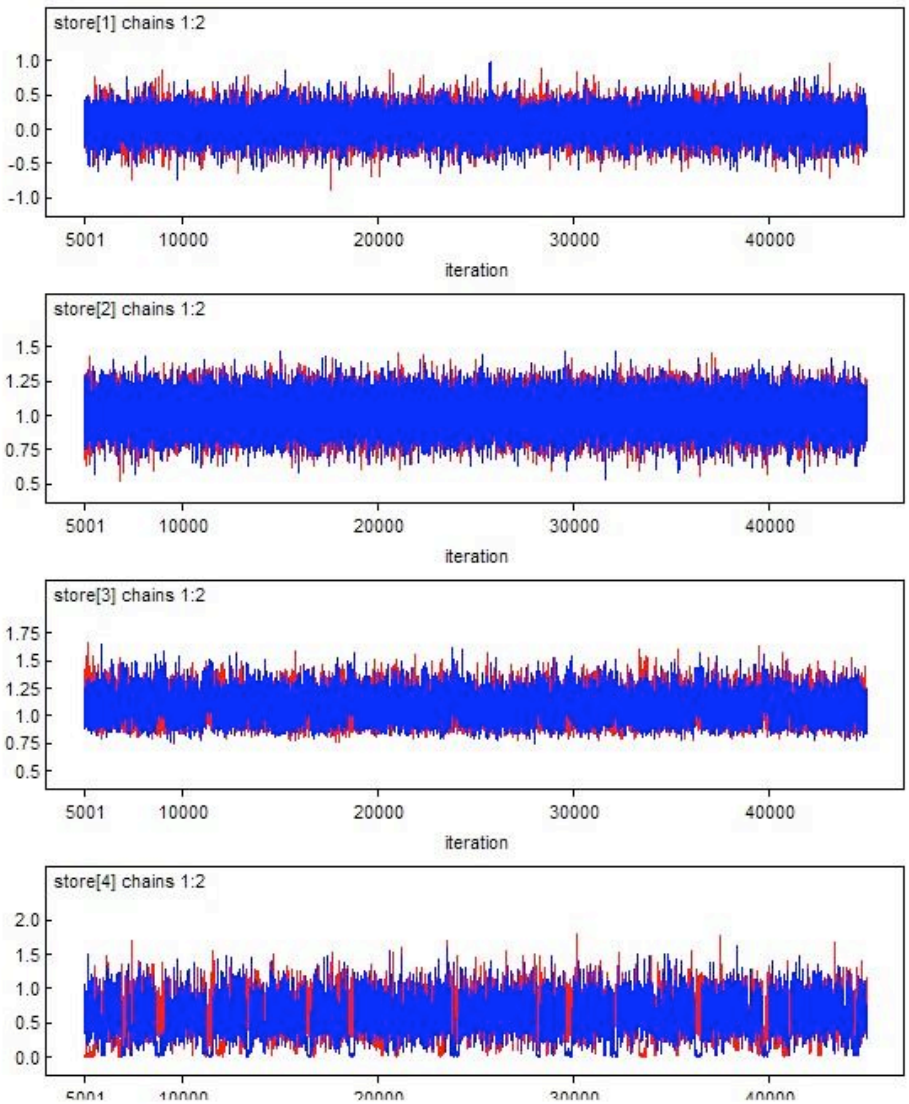

Figure E.12: Convergence chains of β_0, β_1, σ_ε and σ_α

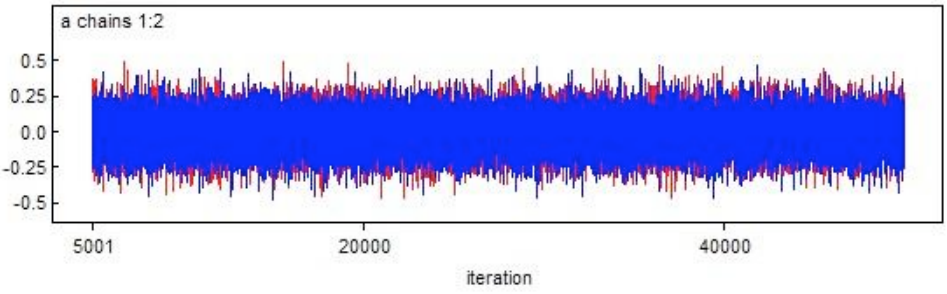

Figure E.13: Convergence chains of ρ

The posterior distribution for each estimate are

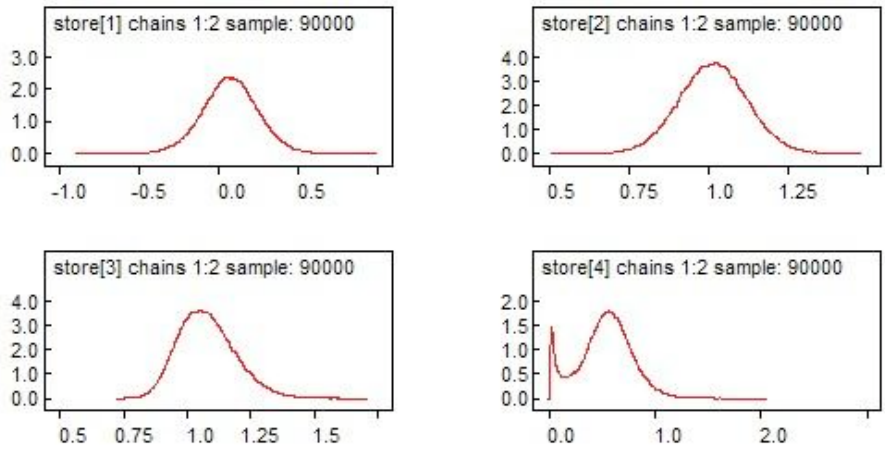

Figure E.14: Posterior distribution of β_0, β_1, σ_ε and σ_α

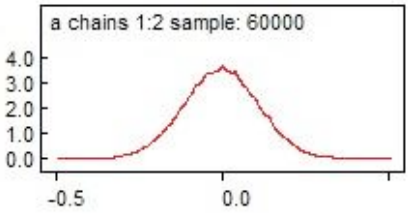

Figure E.15: Posterior distribution of ρ

E.2 RICOR

E.2.1 RE Model

The Gelman-Rubin diagnostics for each estimate are

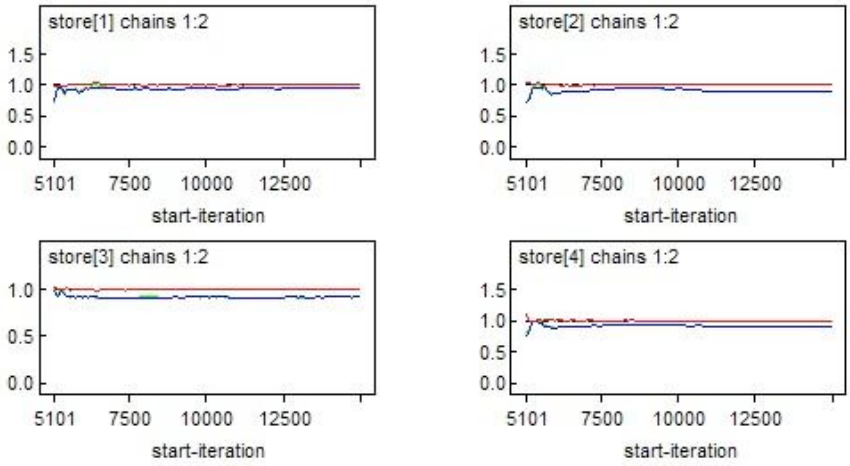

Figure E.16: Gelman-Rubin diagnostics of β_0, β_1, σ_ε and σ_α

The convergence chains for each estimate are

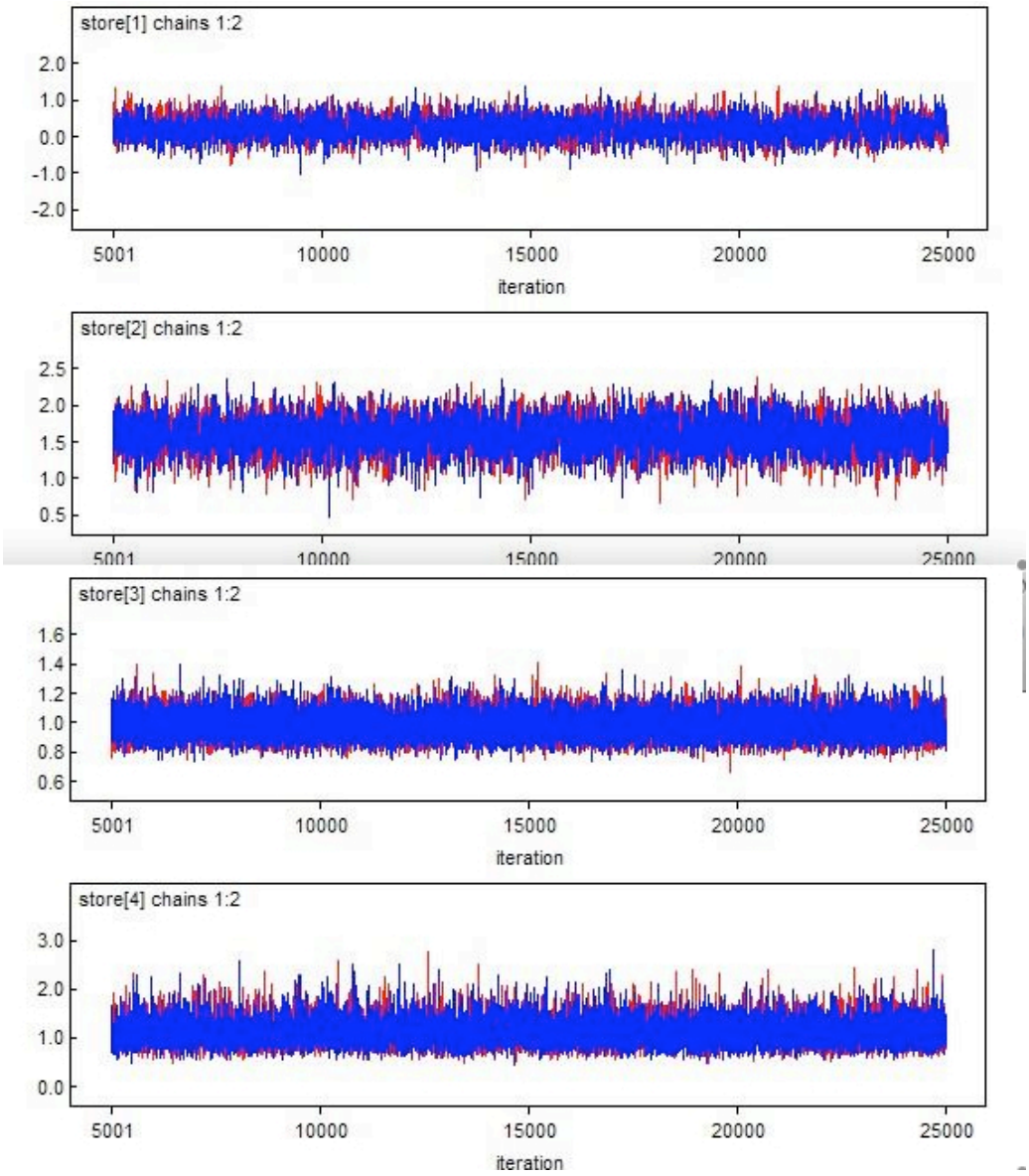

Figure E.17: Convergence chains of β_0, β_1, σ_ε and σ_α

The posterior distribution for each estimate are

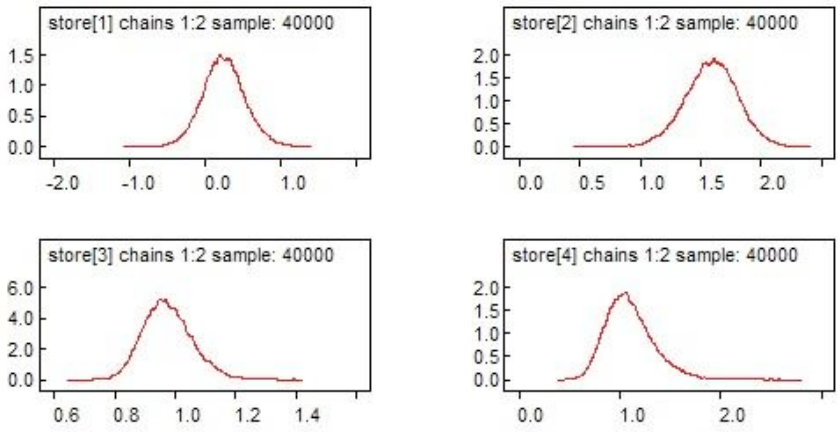

Figure E.18: Posterior distribution of β_0, β_1, σ_ε and σ_α

E.2.2 FE Model

The Gelman-Rubin diagnostics for each estimate are

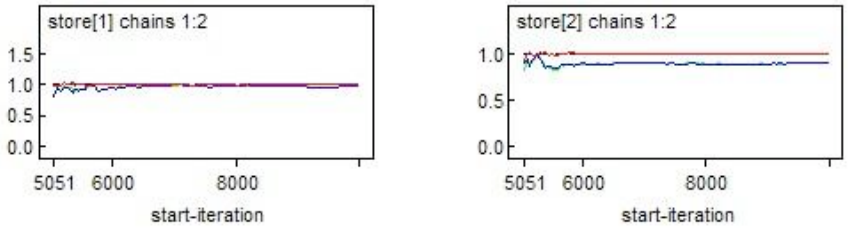

Figure E.19: Gelman-Rubin diagnostics of β_1 and σ_ε

The convergence chains for each estimate are

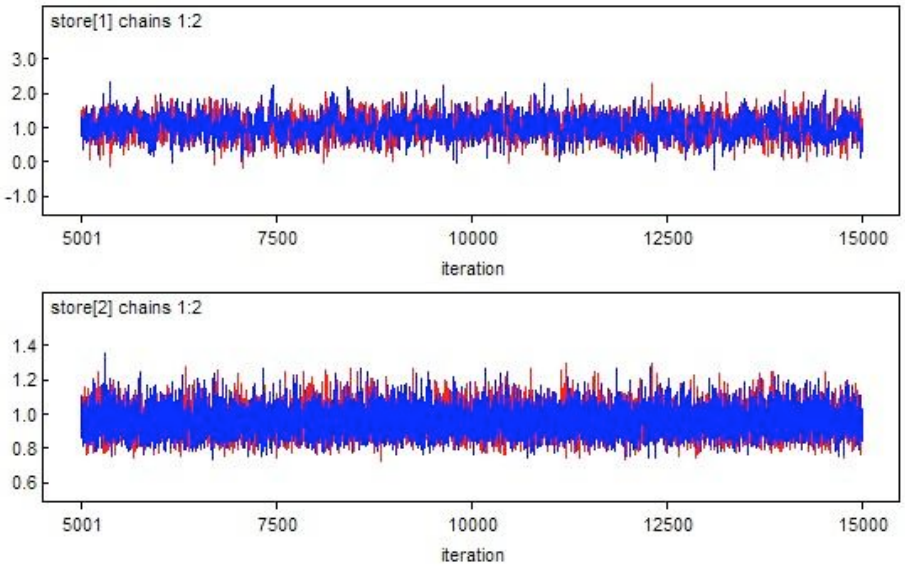

Figure E.20: Convergence chains of β_1 and σ_ε

The posterior distribution for each estimate are

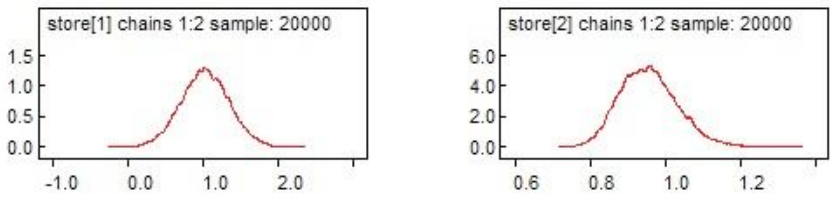

Figure E.21: Posterior distribution of β_1 and σ_ε

E.2.3 PL Model

The Gelman-Rubin diagnostics for each estimate are

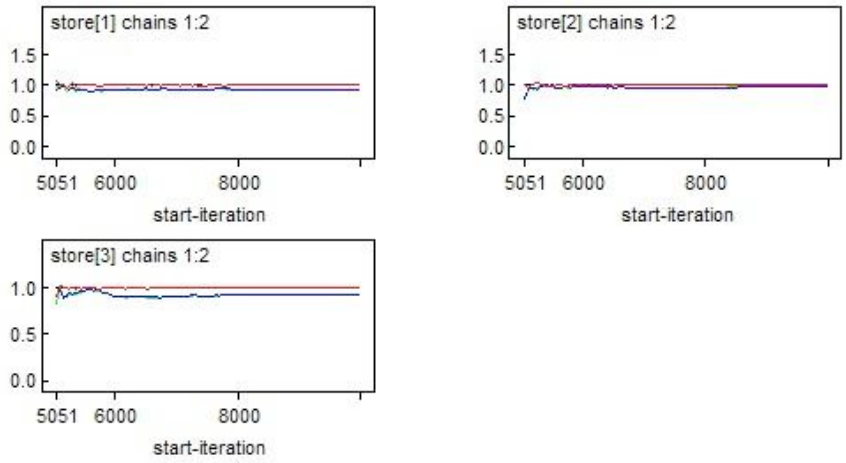

Figure E.22: Gelman-Rubin diagnostics of β_0, β_1 and σ_ε

The convergence chains for each estimate are

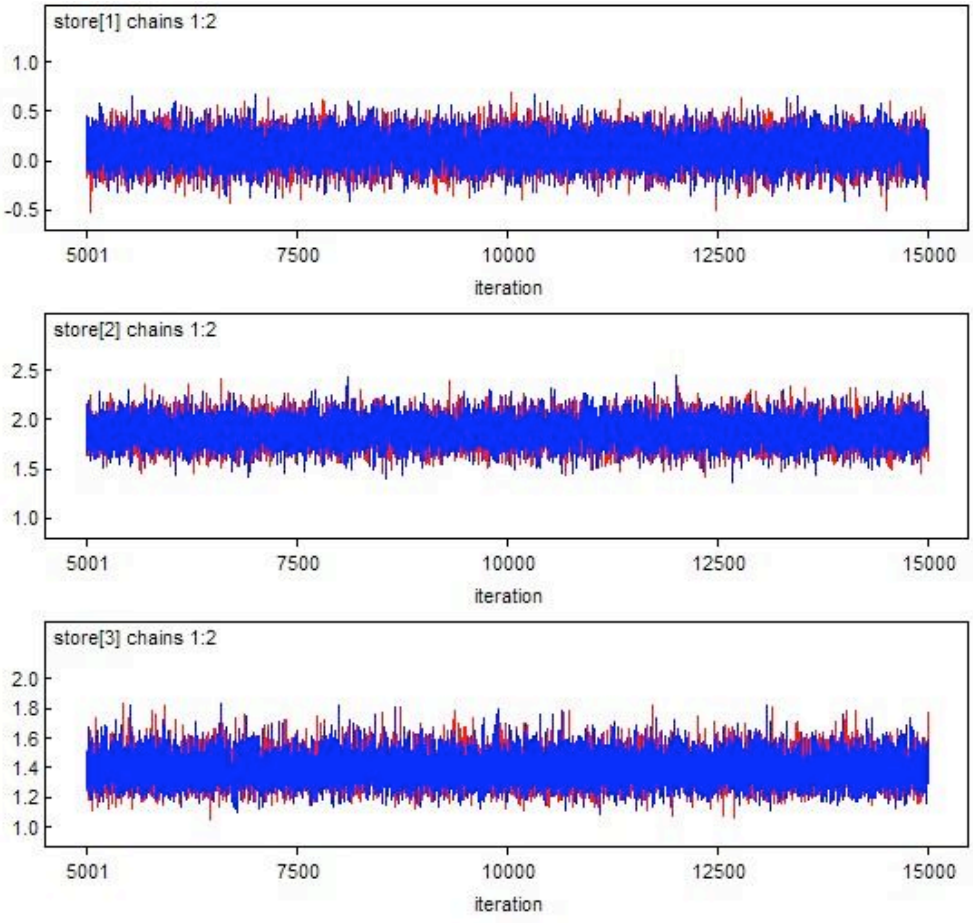

Figure E.23: Convergence chains of β_0, β_1 and σ_ε

The posterior distribution for each estimate are

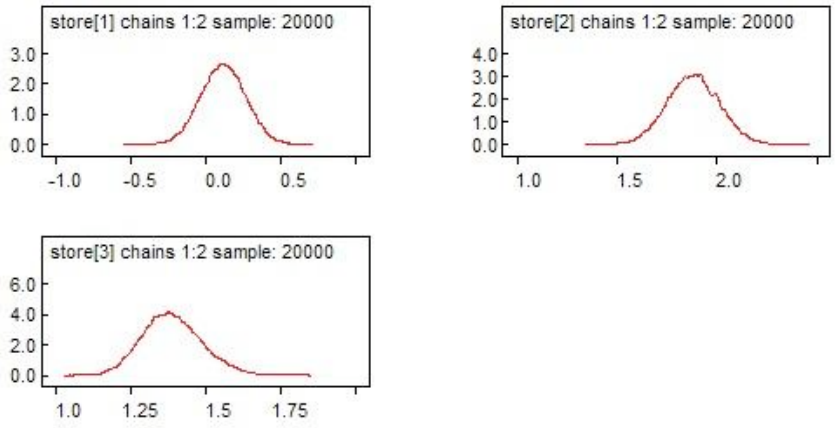

Figure E.24: Posterior distribution of β_0, β_1 and σ_ε

E.2.4 MF Model

The Gelman-Rubin diagnostics for each estimate are

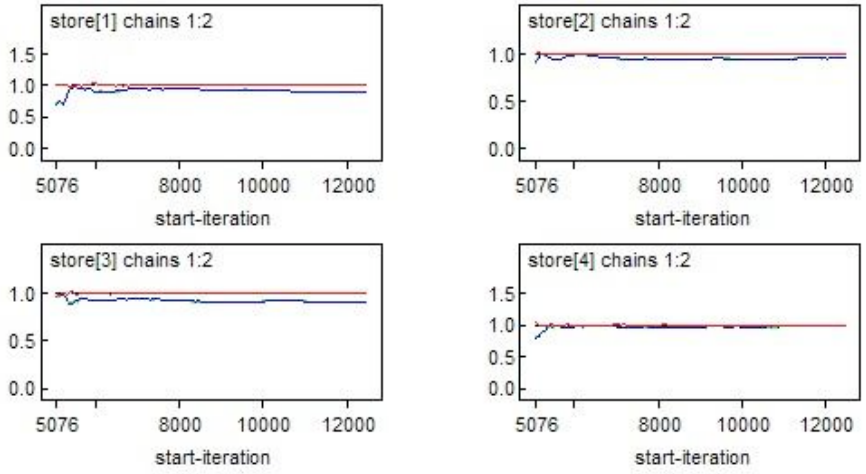

Figure E.25: Gelman-Rubin diagnostics of β_0, β_1, σ_ε and σ_α

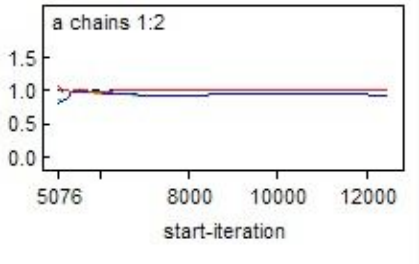

Figure E.26: Gelman-Rubin diagnostics of ρ

The convergence chains for each estimate are

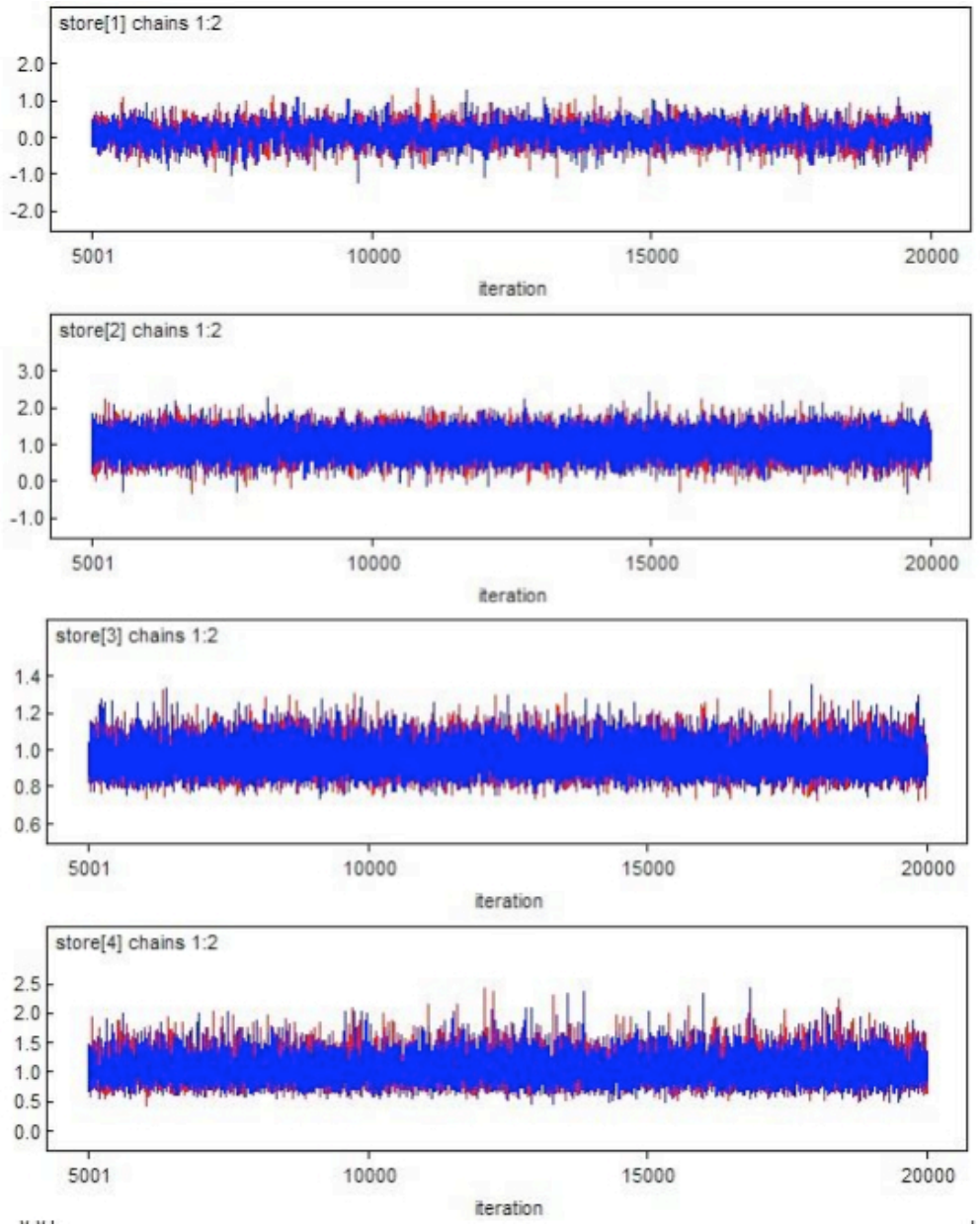

Figure E.27: Convergence chains of β_0, β_1, σ_ε and σ_α

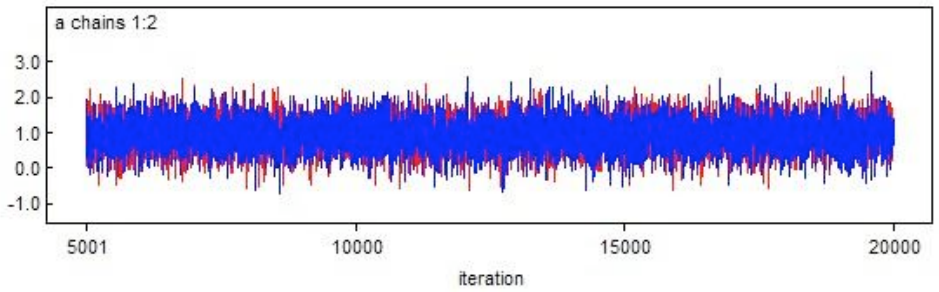

Figure E.28: Convergence chains of ρ

The posterior distribution for each estimate are

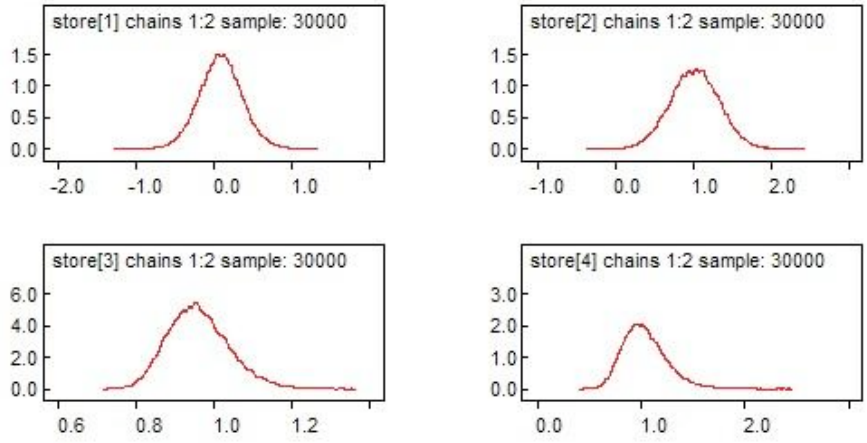

Figure E.29: Posterior distribution of β_0, β_1, σ_ε and σ_α

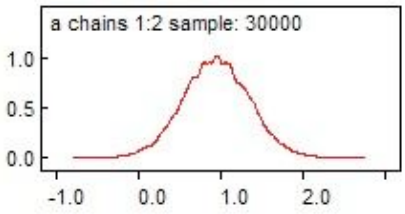

Figure E.30: Posterior distribution of ρ

E.3 WAGE

E.3.1 RE Model

The Gelman-Rubin diagnostics for each estimate are

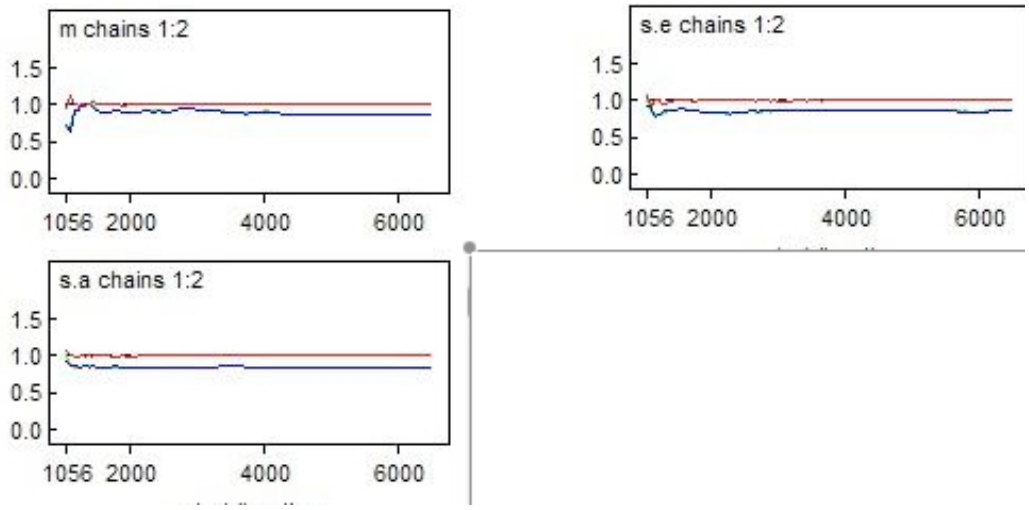

Figure E.31: Gelman-Rubin diagnostics of β_0, σ_ε and σ_α

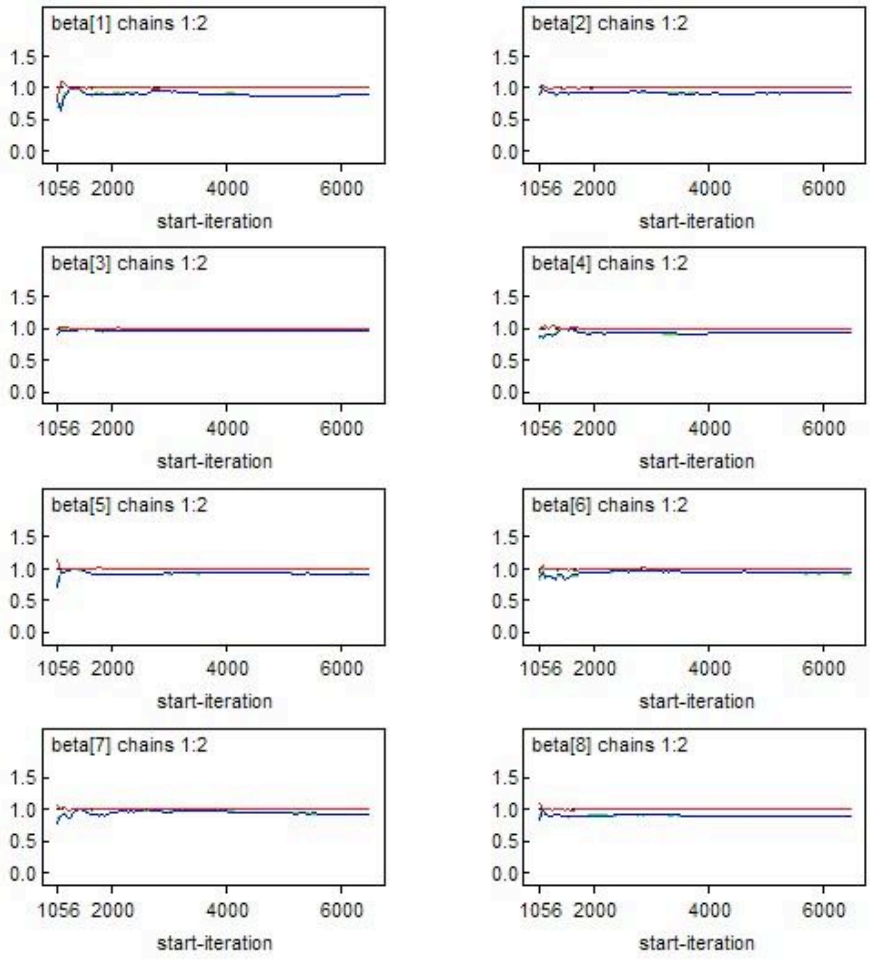

Figure E.32: Gelman-Rubin diagnostics of β_k, $k = 1, \cdots, 8$

The convergence chains for each estimate are

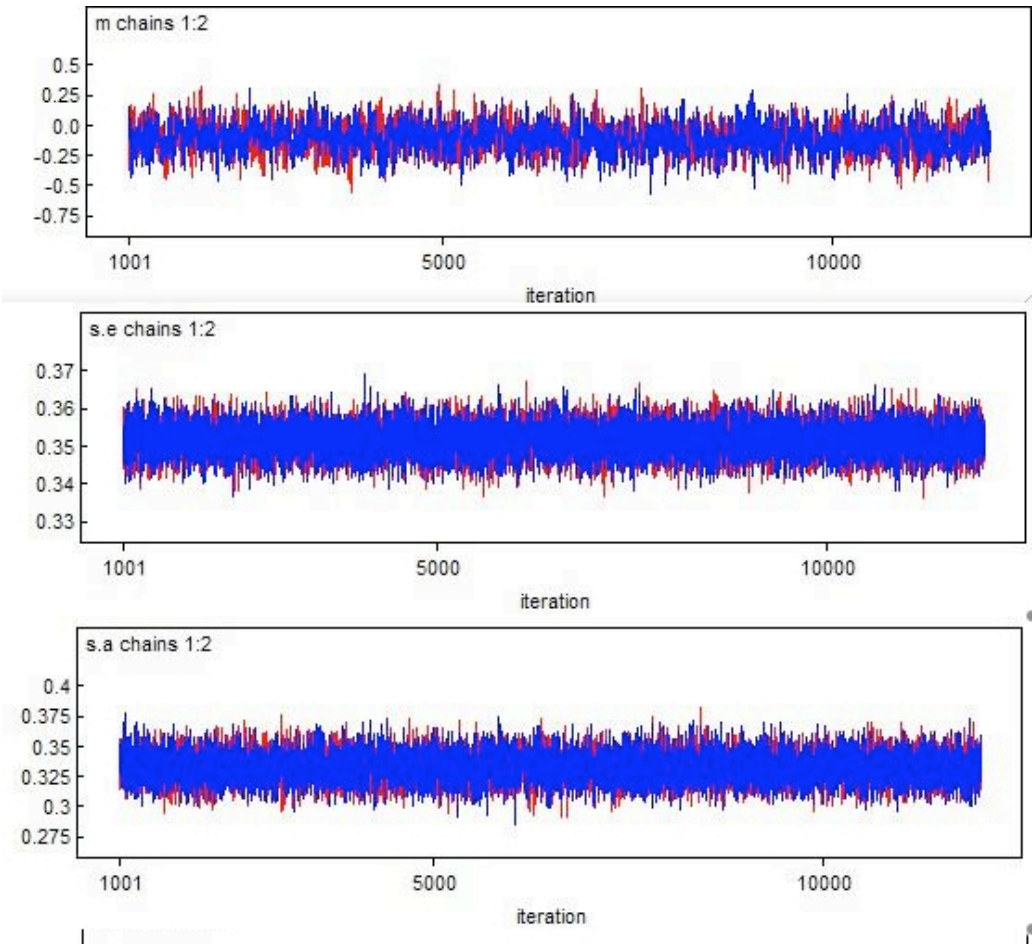

Figure E.33: Convergence chains of β_0, σ_ε and σ_α

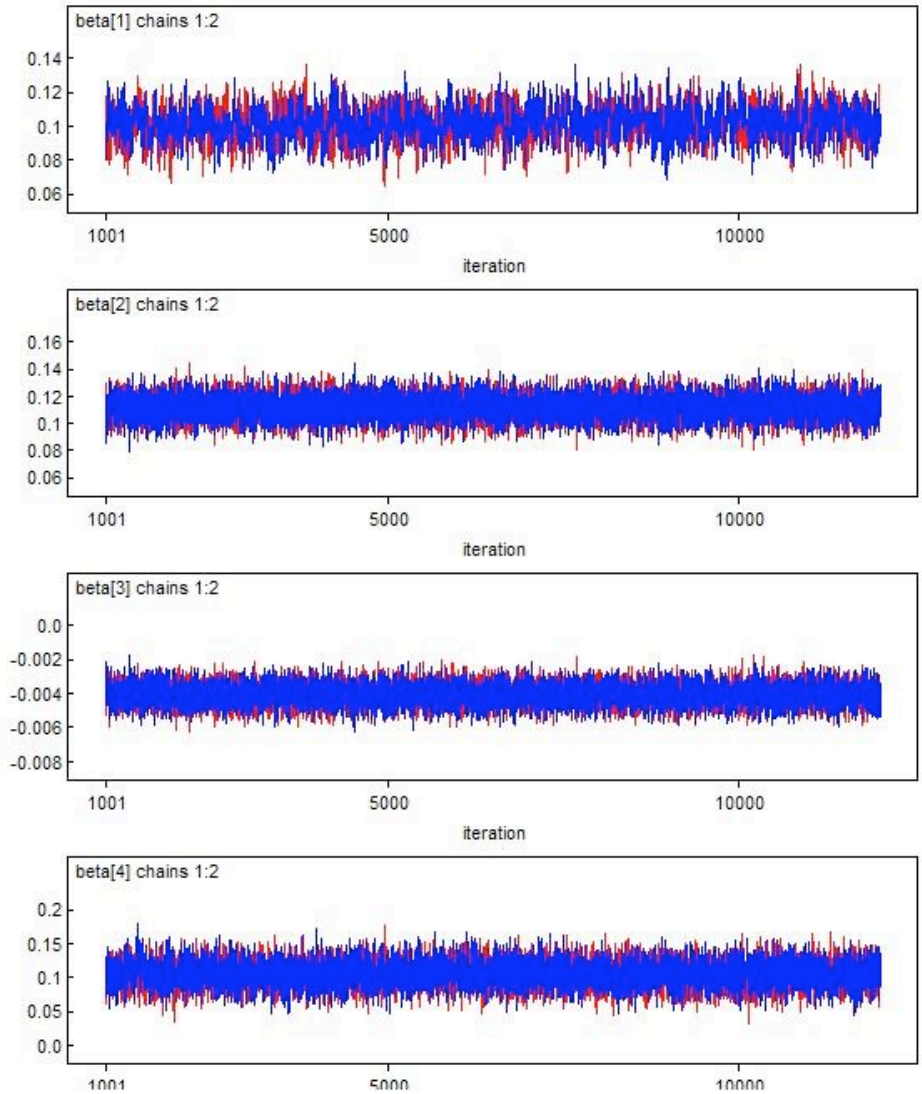

Figure E.34: Convergence chains of β_k, $k = 1, \cdots, 4$

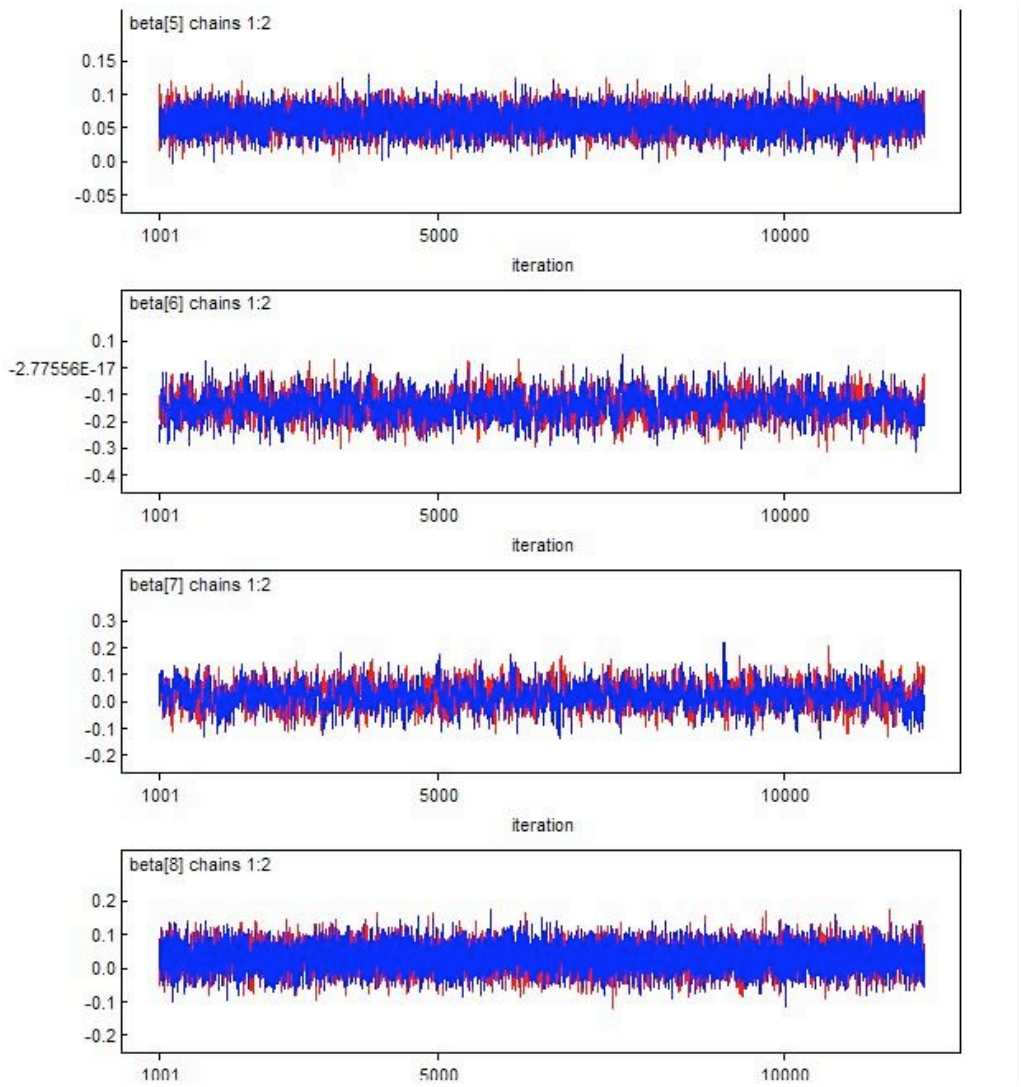

Figure E.35: Convergence chains of β_k, $k = 5, \cdots, 8$

The posterior distribution for each estimate are

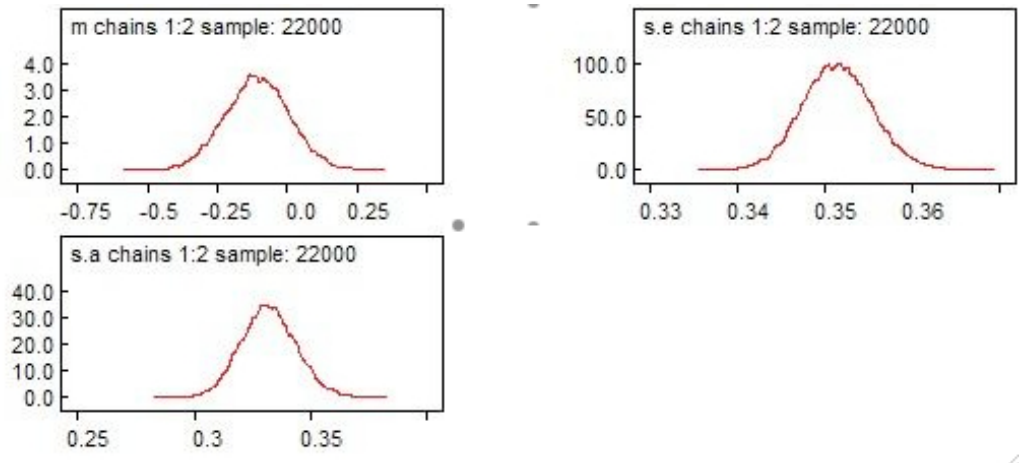

Figure E.36: Posterior distribution of β_0, σ_ε and σ_α

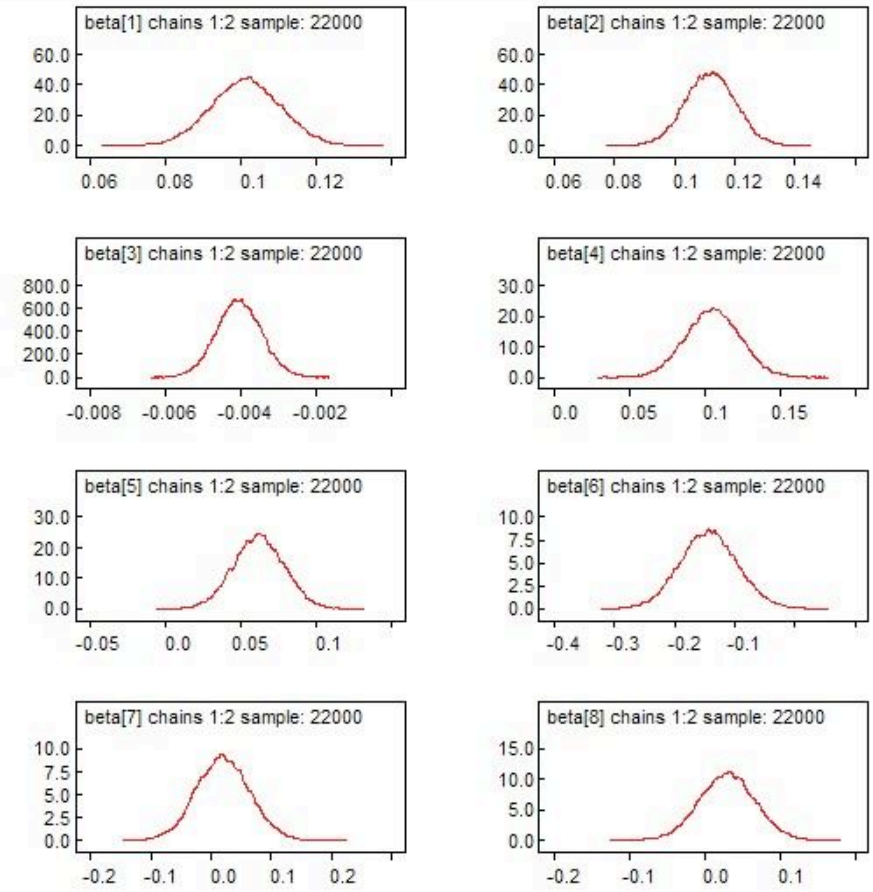

Figure E.37: Posterior distribution of β_k, $k = 1, \cdots, 8$

E.3.2 FE Model

The Gelman-Rubin diagnostics for each estimate are

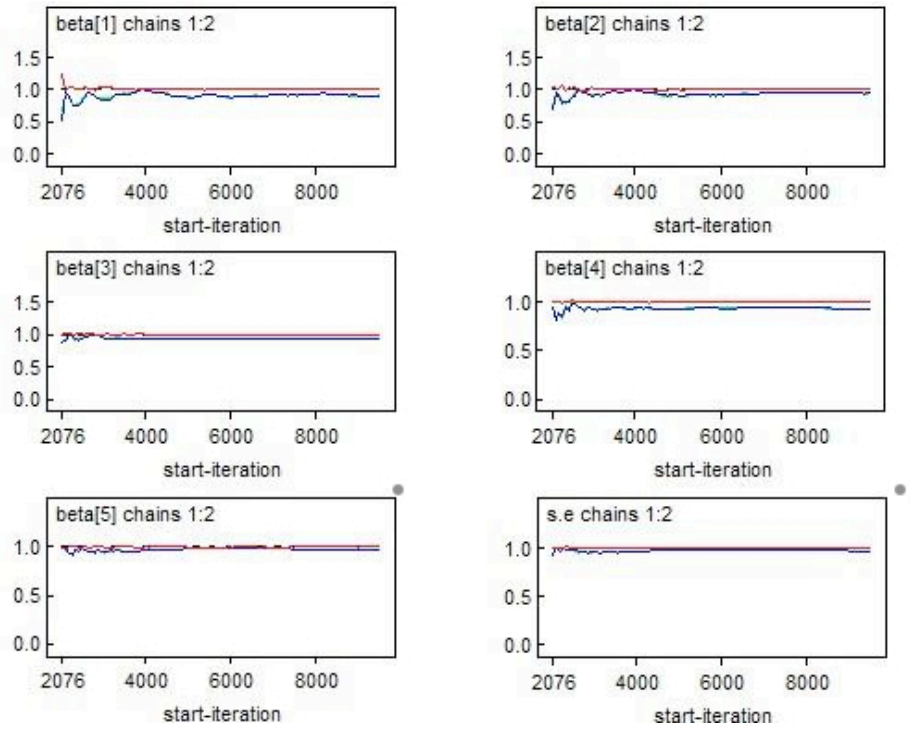

Figure E.38: Gelman-Rubin diagnostics of β_k, $k = 2, 3, 4, 5, 8$ and σ_ε

The convergence chains for each estimate are

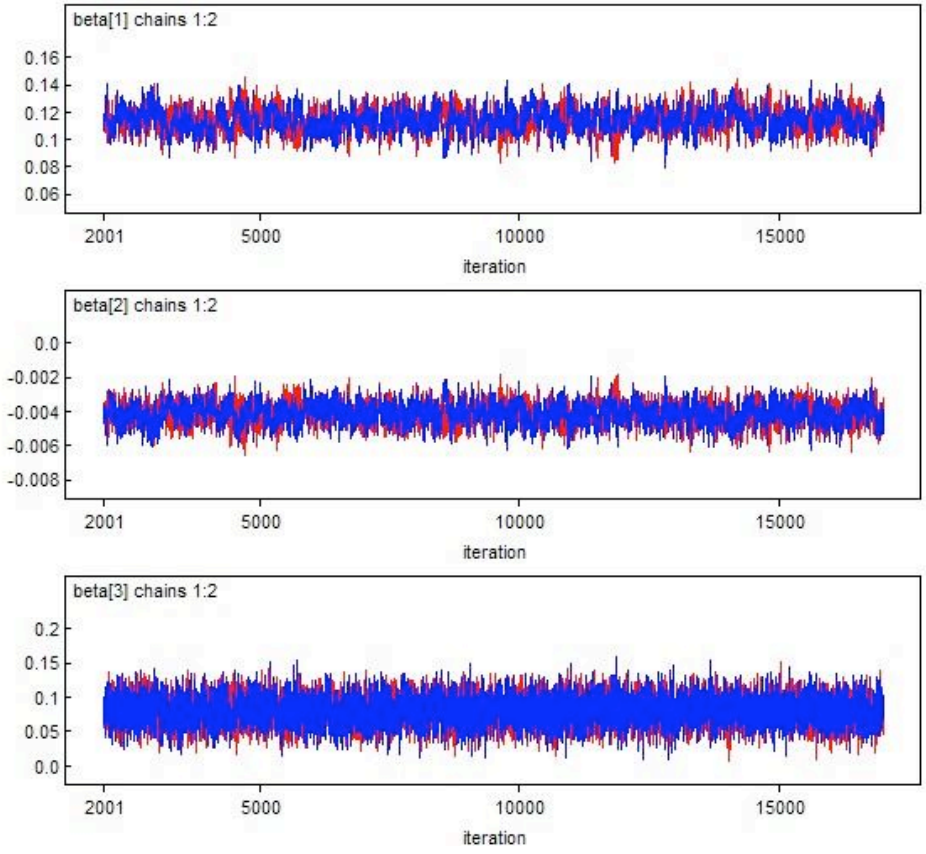

Figure E.39: Convergence chains of β_k, $k = 2, 3, 4$

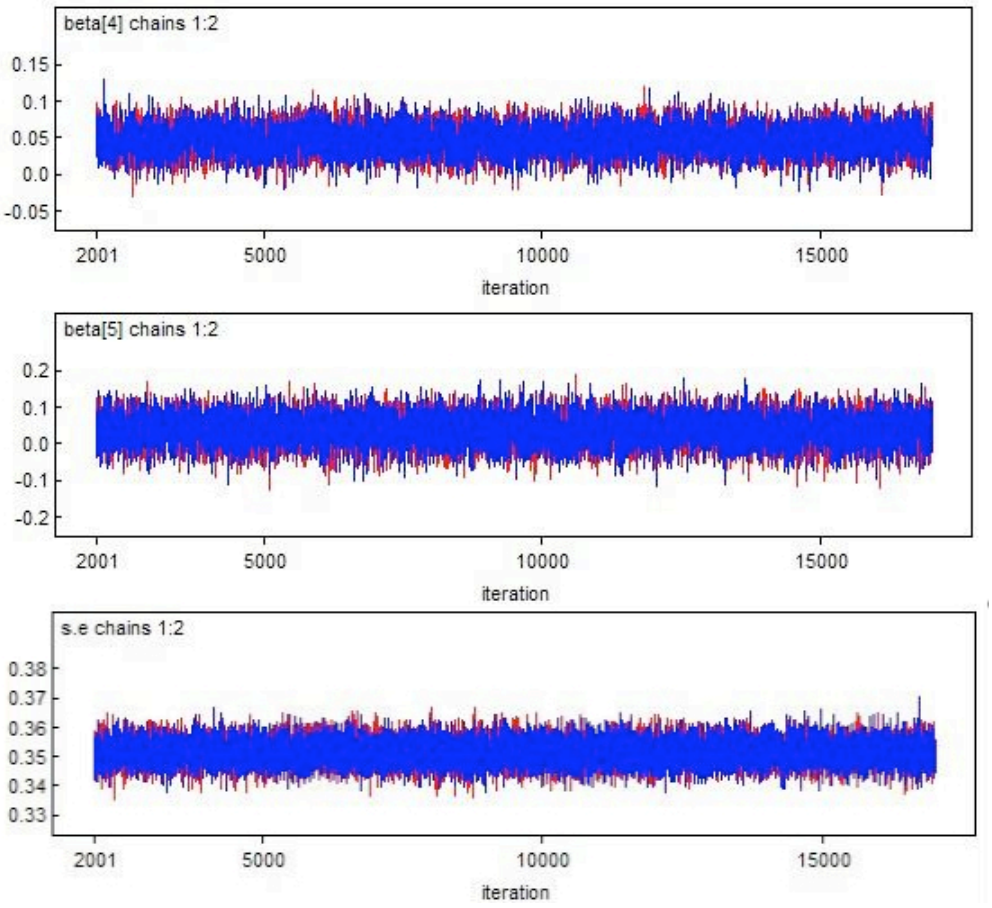

Figure E.40: Convergence chains of β_k, $k = 5, 8$ and σ_ε

The posterior distribution for each estimate are

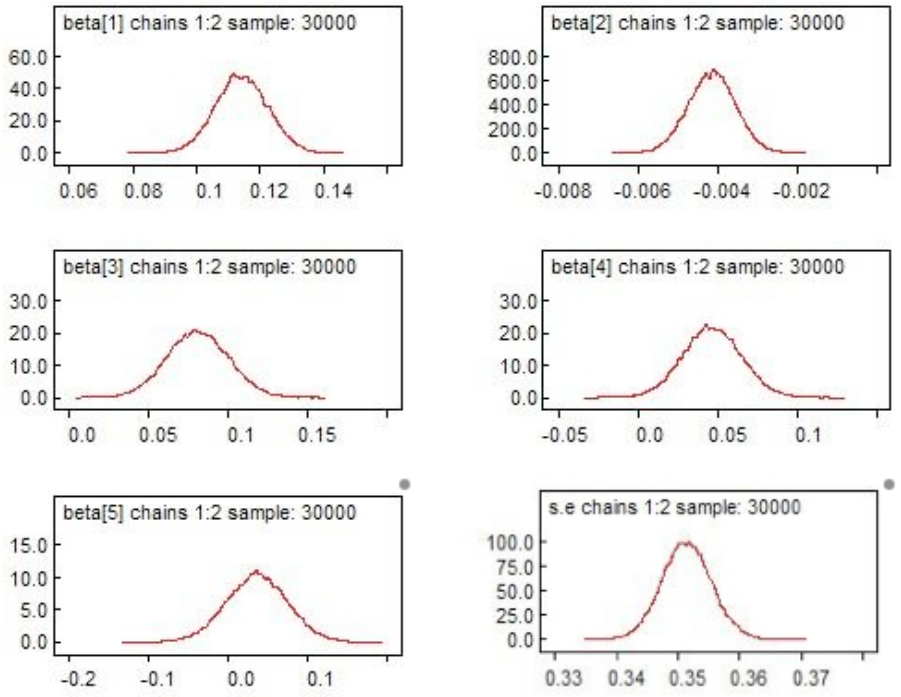

Figure E.41: Posterior distribution of β_k, $k = 2, 3, 4, 5, 8$ and σ_ε

E.3.3 PL Model

The Gelman-Rubin diagnostics for each estimate are

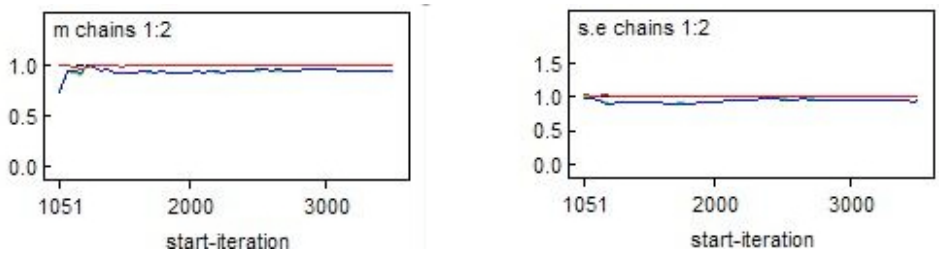

Figure E.42: Gelman-Rubin diagnostics of β_0 and σ_ε

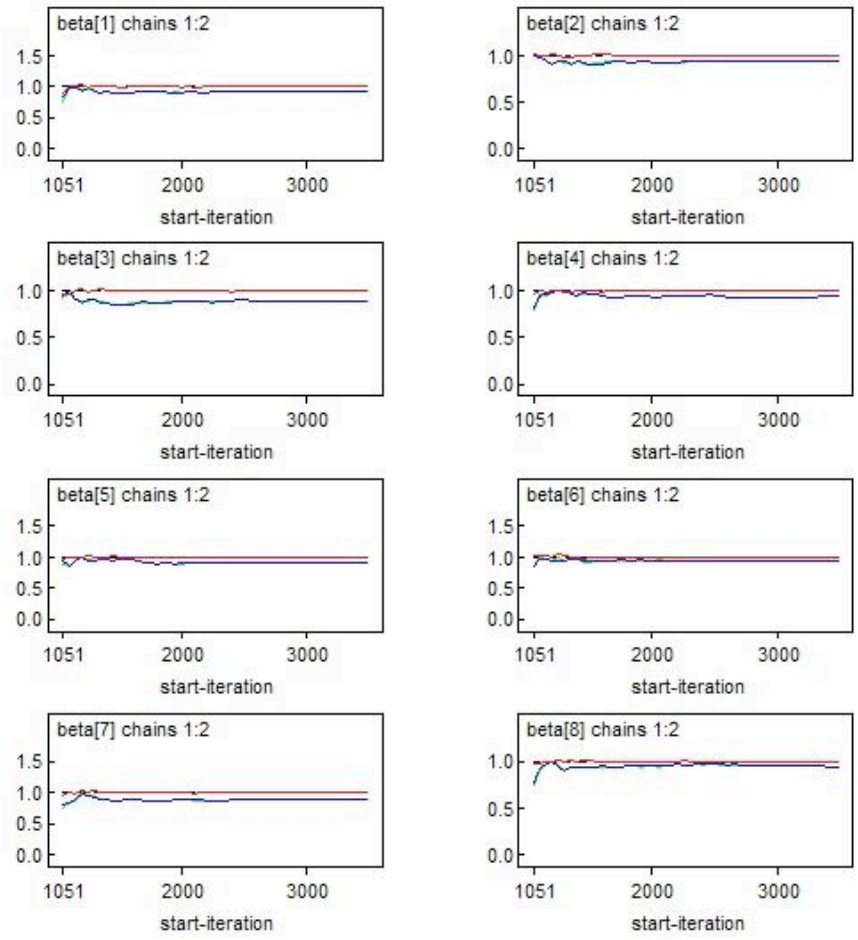

Figure E.43: Gelman-Rubin diagnostics of β_k, $k = 1, \cdots, 8$

The convergence chains for each estimate are

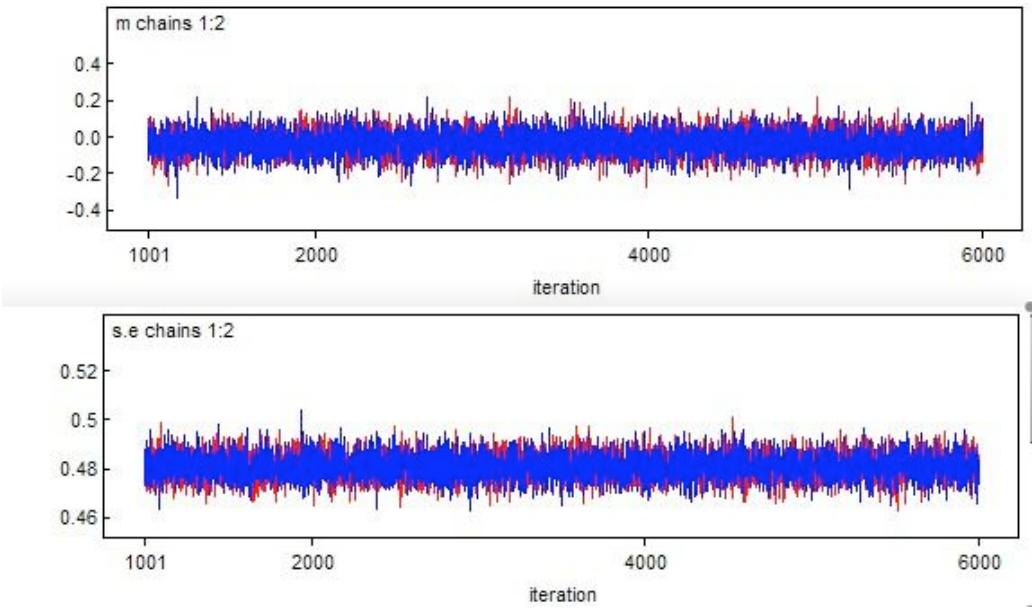

Figure E.44: Convergence chains of β_0 and σ_ε

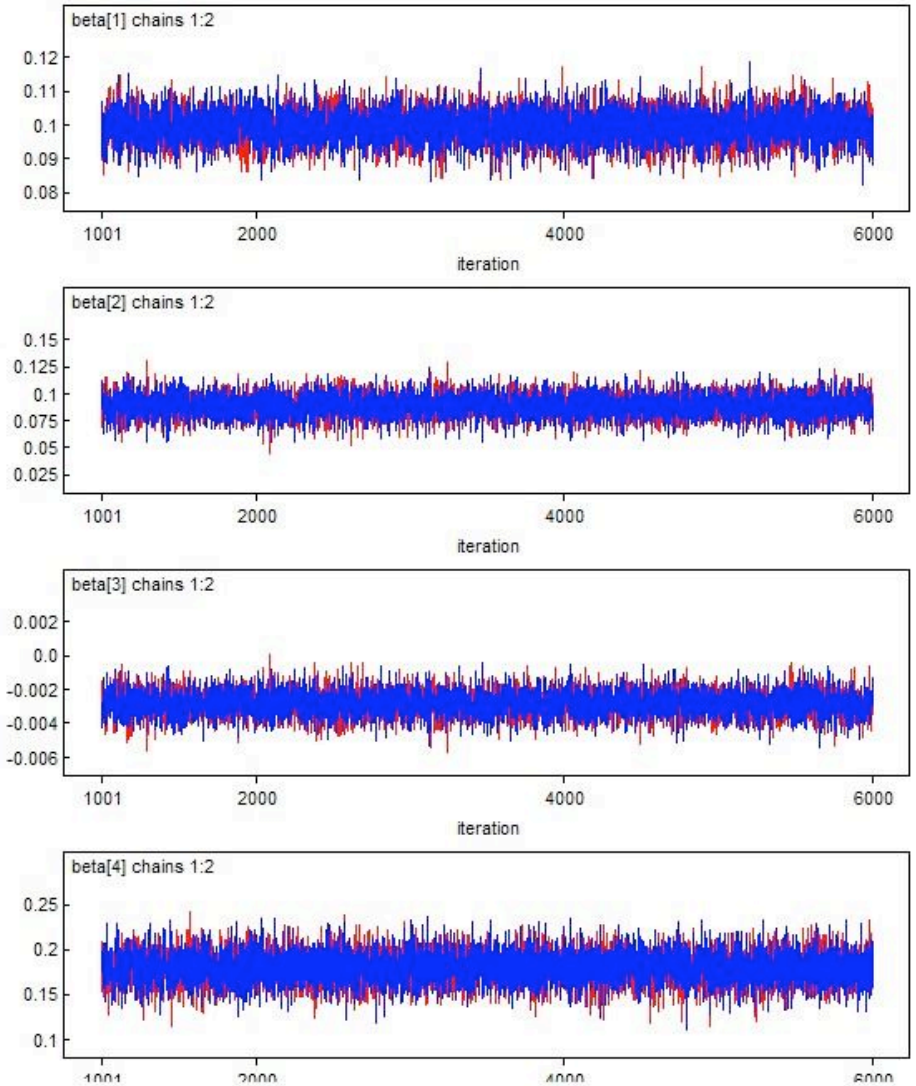

Figure E.45: Convergence chains of β_k, $k = 1, \cdots, 4$

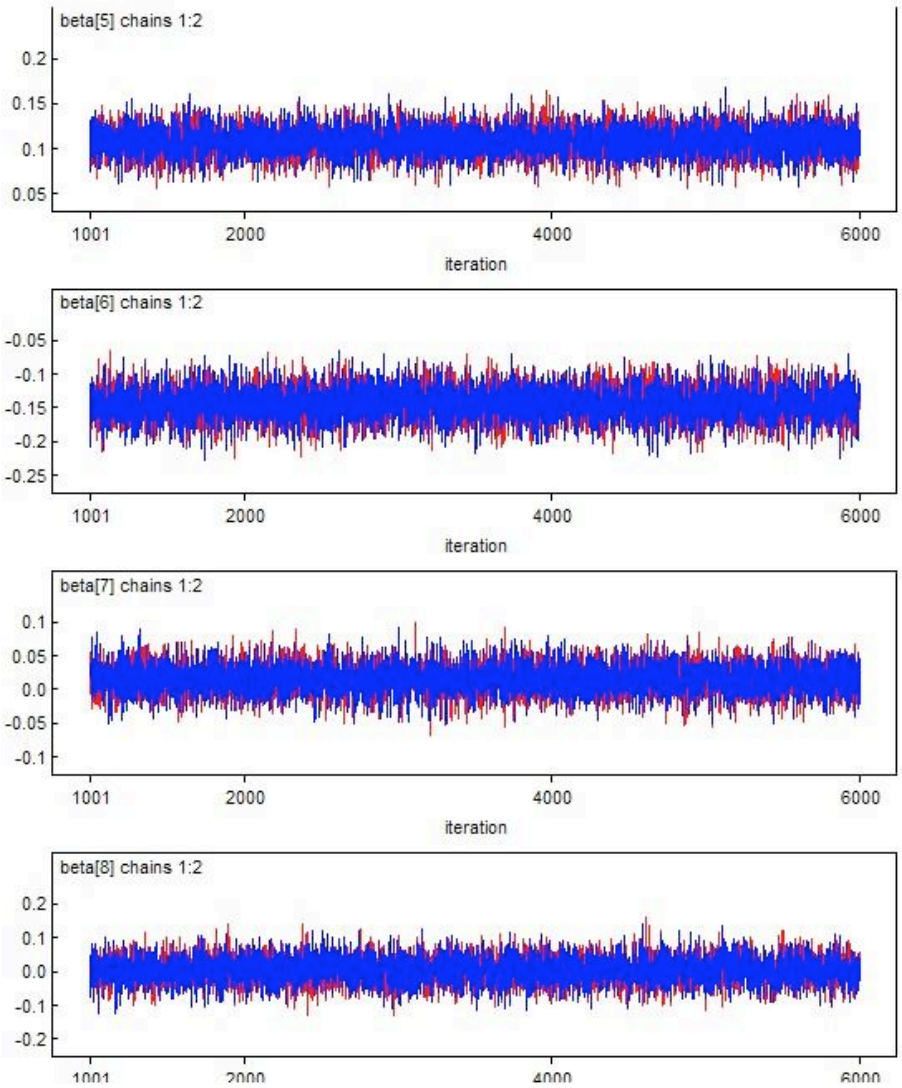

Figure E.46: Convergence chains of β_k, $k = 5, \cdots, 8$

The posterior distribution for each estimate are

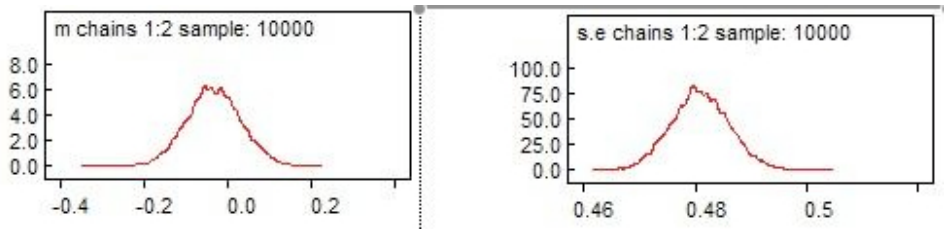

Figure E.47: Posterior distribution of β_0 and σ_ε

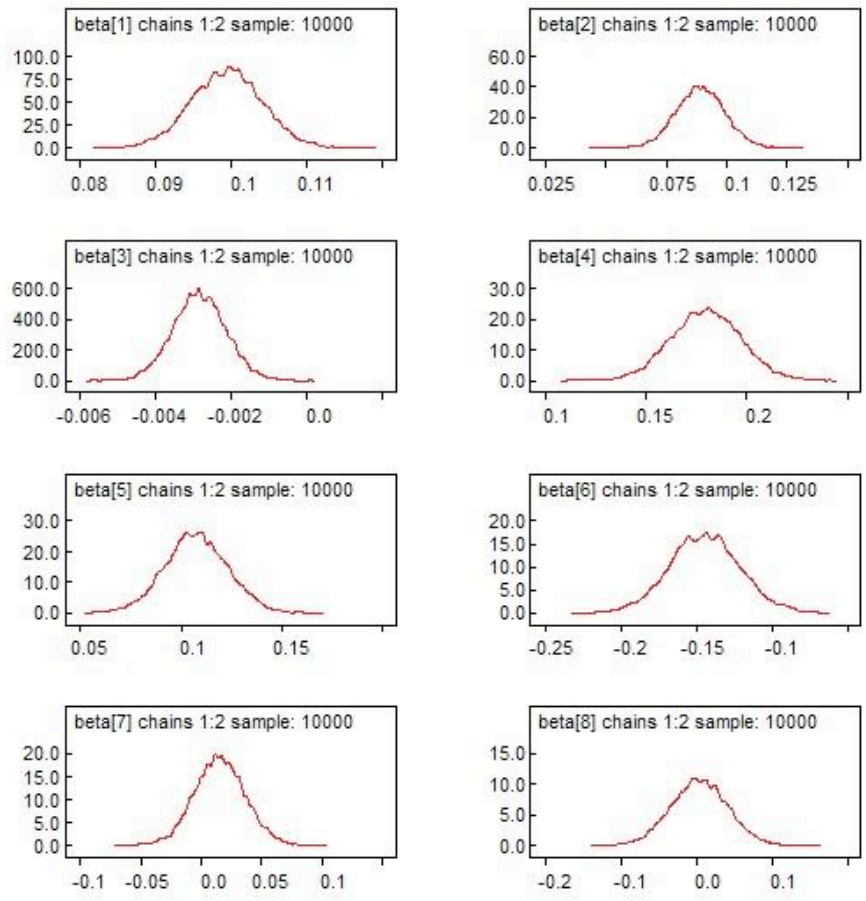

Figure E.48: Posterior distribution of β_k, $k = 1, \cdots, 8$

E.3.4 MF Model

The Gelman-Rubin diagnostics for each estimate are

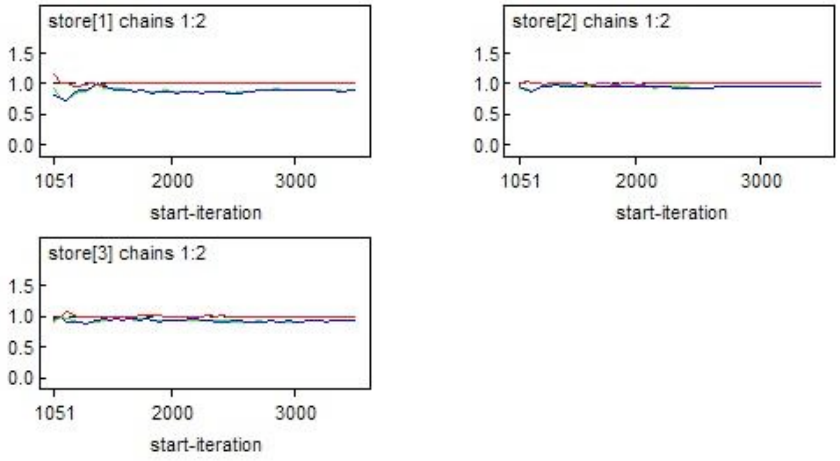

Figure E.49: Gelman-Rubin diagnostics of β_0, σ_ε and σ_w

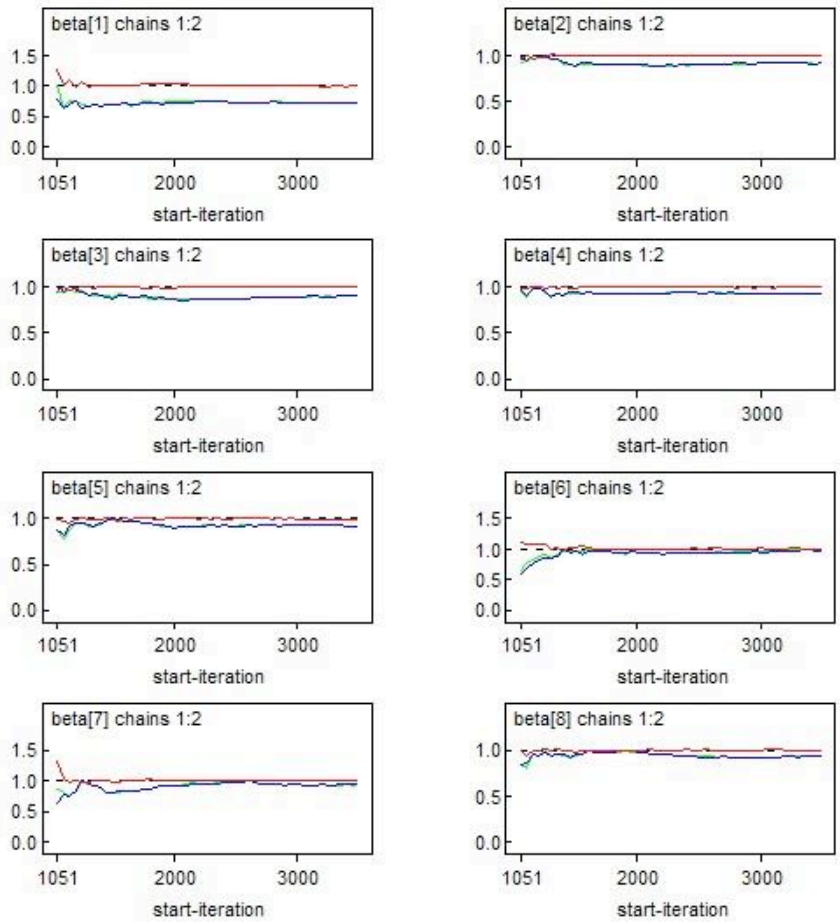

Figure E.50: Gelman-Rubin diagnostics of β_k, $k = 1, \cdots, 8$

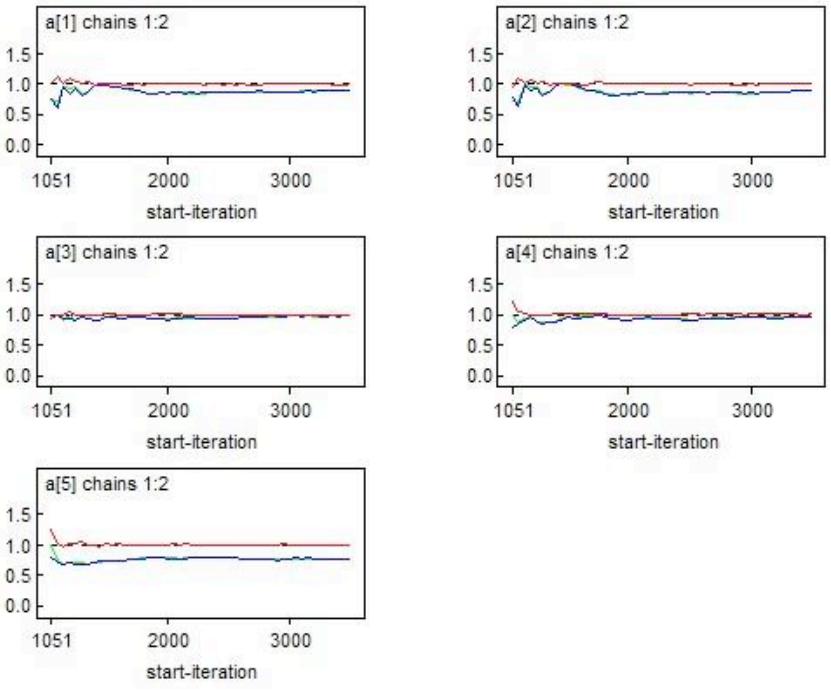

Figure E.51: Gelman-Rubin diagnostics of ρ_p, $p = 1, \cdots, 5$

The convergence chains for each estimate are

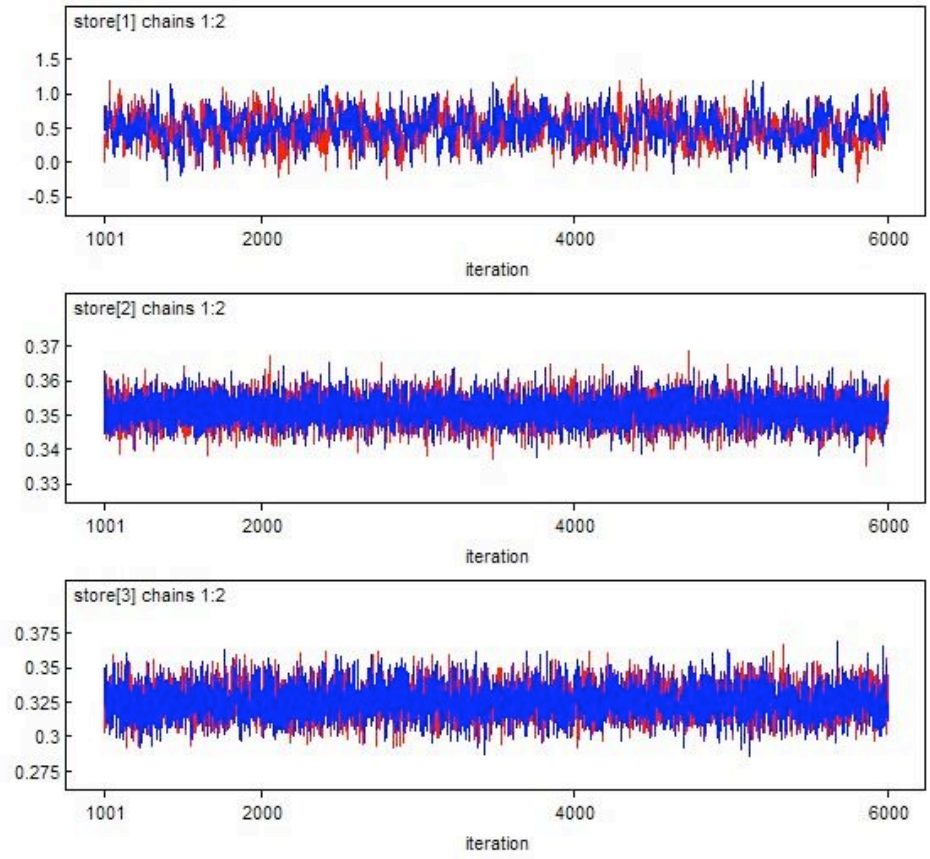

Figure E.52: Convergence chains of β_0, σ_ε and σ_w

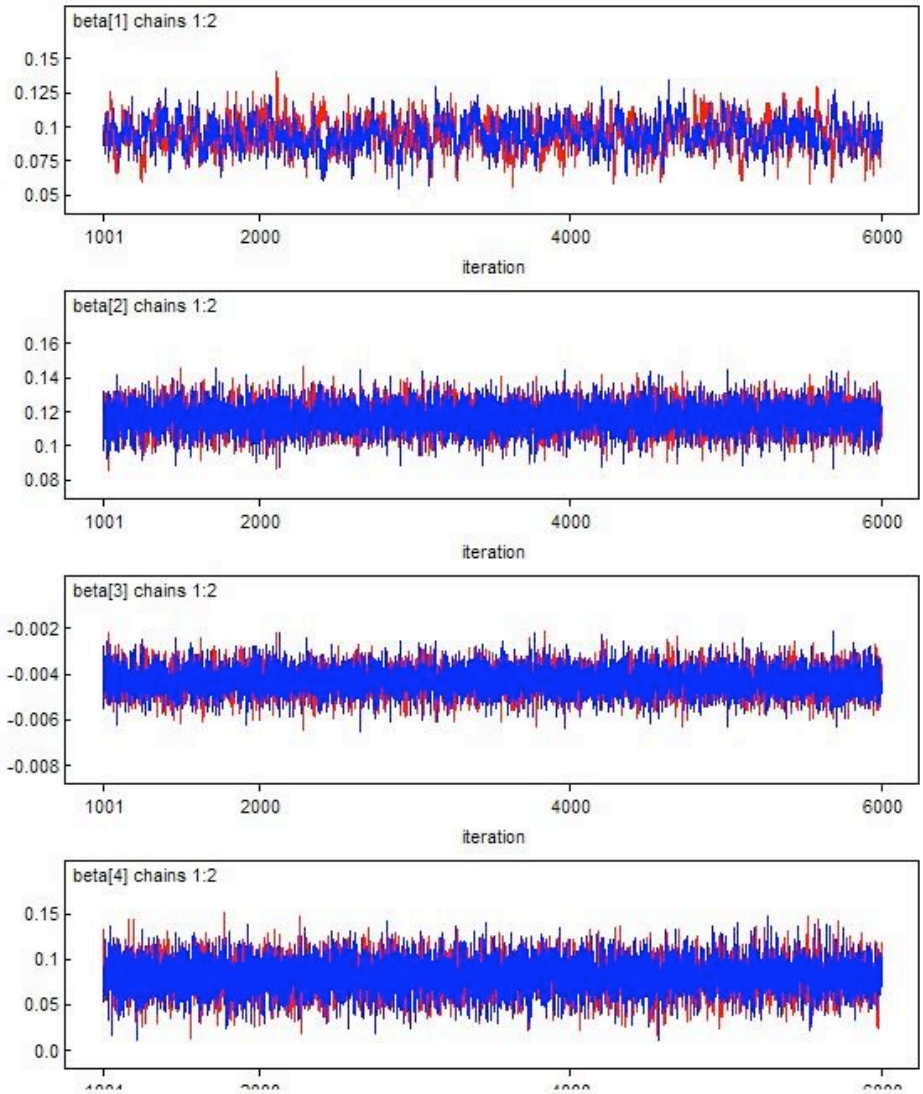

Figure E.53: Convergence chains of β_k, $k = 1, \cdots, 4$

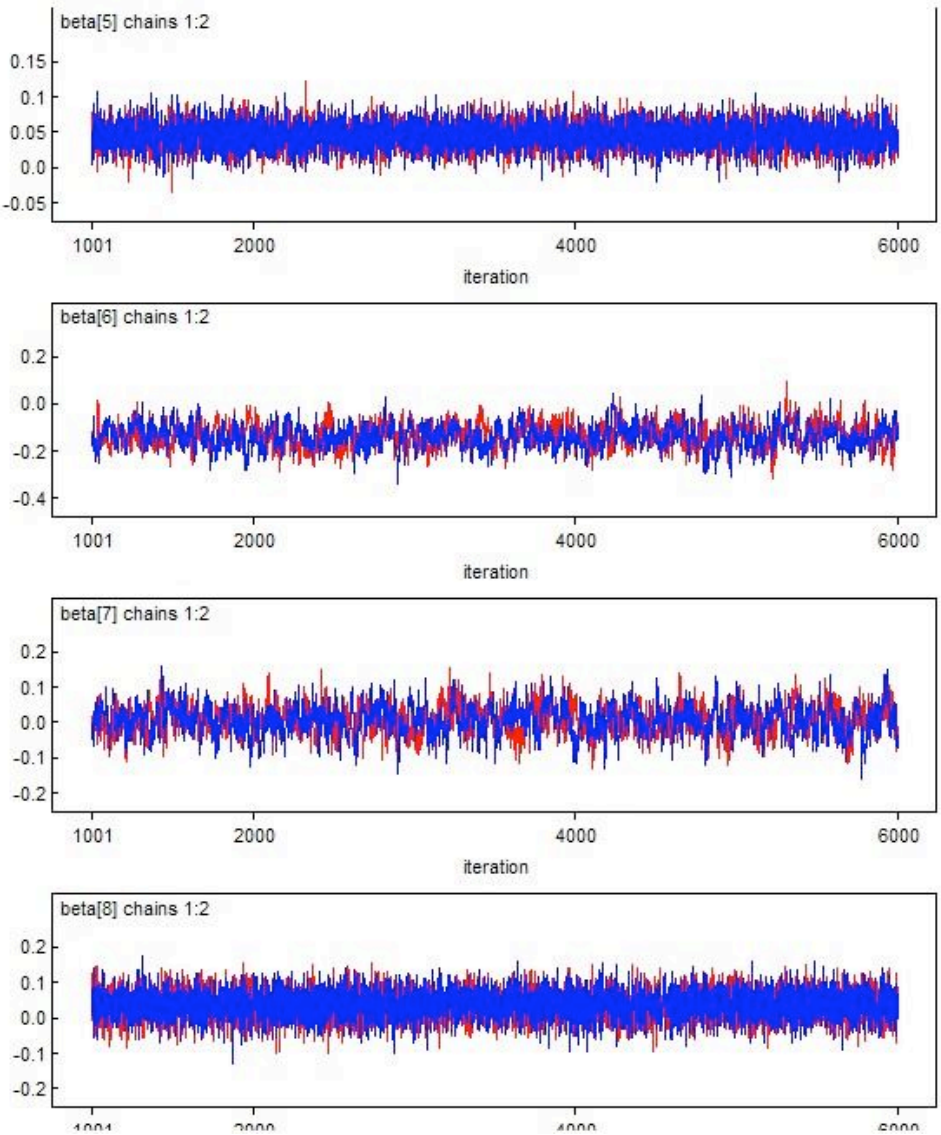

Figure E.54: Convergence chains of β_k, $k = 5, \cdots, 8$

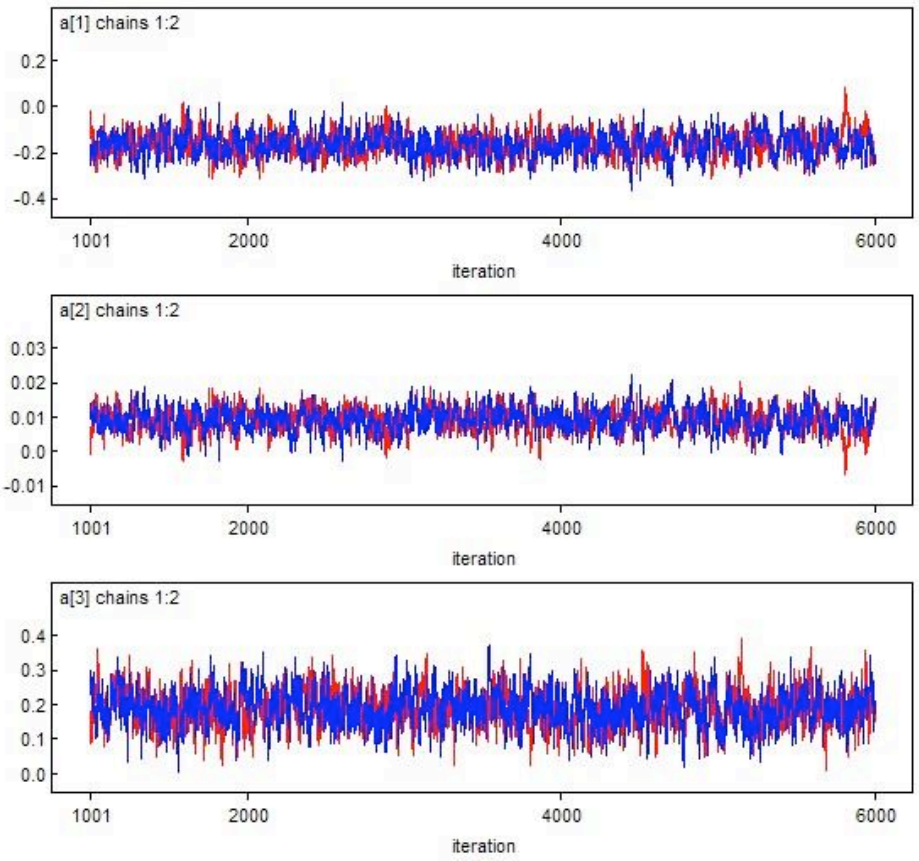

Figure E.55: Convergence chains of ρ_p, $p = 1, \cdots, 3$

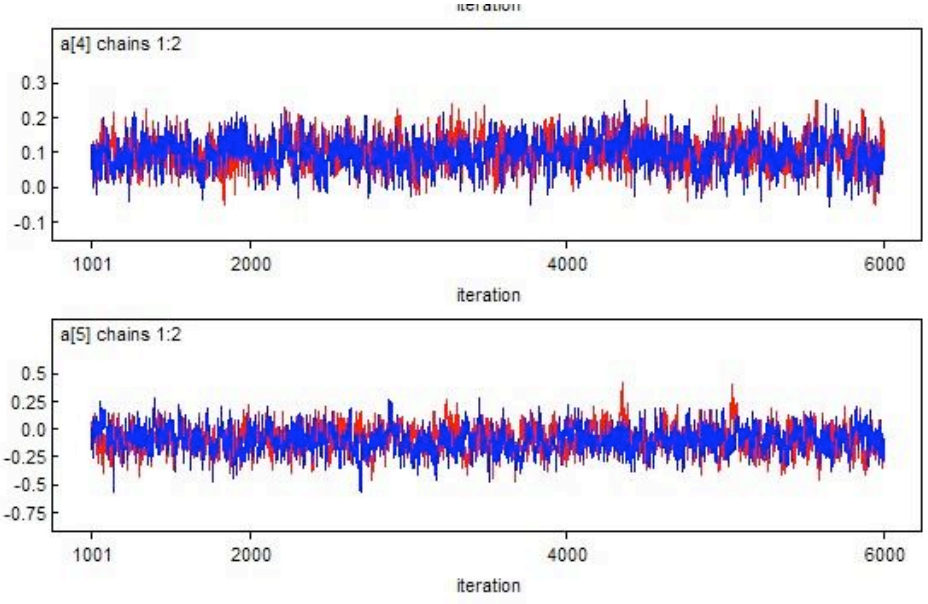

Figure E.56: Convergence chains of ρ_p, $p = 4, 5$

The posterior distribution for each estimate are

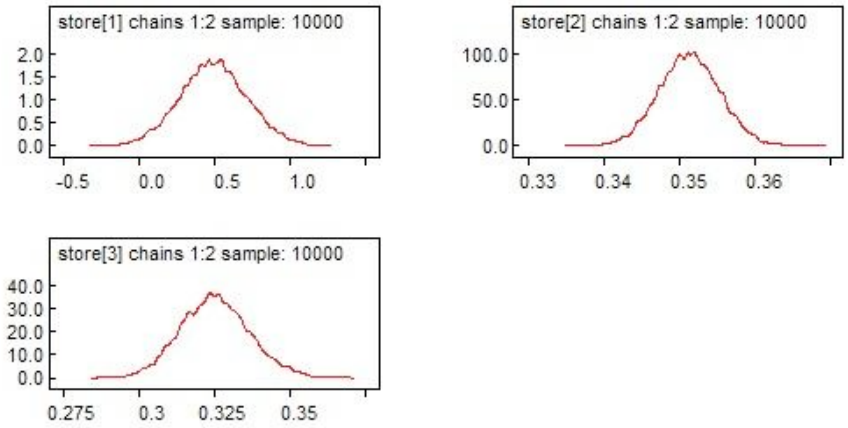

Figure E.57: Posterior distribution of β_0, σ_ε and σ_w

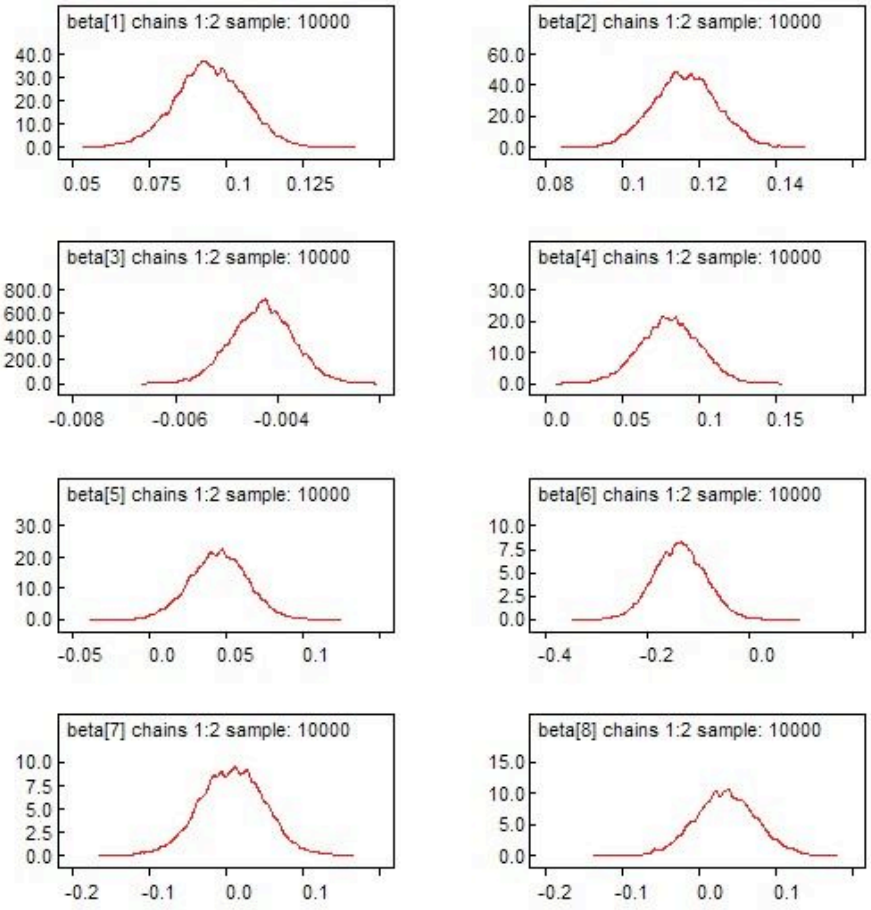

Figure E.58: Posterior distribution of β_k, $k = 1, \cdots, 8$

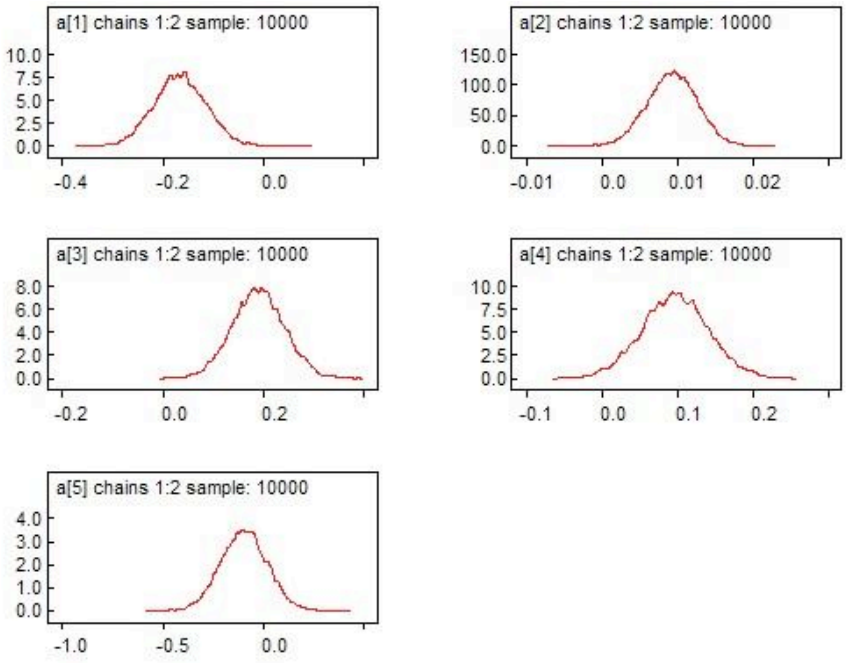

Figure E.59: Posterior distribution of ρ_p, $p = 1, \cdots, 5$

Bibliography

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19:716–723, 1974.

R. Arnold and I. M. Liu. Sas longitudinal data analysis course notes. Technical report, Victoria University of Wellington, New Zealand, 2004.

A. C. Atkinson. *Plots, Transformations and Regression*. Oxford University Press, 1985.

S. P. Brooks and A. Gelman. General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7:434–455, 1997.

K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference*. Springer, 2002.

A. C. Cameron and P. K. Trivedi. *Microeconometrics: Methods and Applications*. Cambridge University Press, 2005.

E. Demidenko. *Mixed Models: Theory and Applications*. Wiley, 2004.

P. J. Diggle, P. Heagerty, K. Y. Liang, and S. L. Zeger. *Analysis of Longitudinal Data*. Oxford University Press, 2002.

A. J. Dobson. *An Introduction to Generalized Linear Models*. Chapman & Hall, London, 2002.

E. W. Frees. *Longitudinal and Panel Data-analysis and Applications in the Social Sciences*. Cambridge University Press, New York, 2004.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7:457–511, 1992.

J. Geweke. *Contemporary Bayesian Econometrics and Statistics*. Wiley, 2005.

W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57 (1):97–109, 1970.

F. Hayashi. *Econometrics*. Princeton University Press, 2000.

C. Hsiao. *Analysis of Panel Data*. Cambridge University Press, 2003.

J. Johnson and J. DiNardo. *Econometric Methods*. The McGraw-Hill Companies, Inc., Singapore, 2007.

J.M. Marin and C.P. Robert. *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer, 2007.

A. M. McCarthy. *Bayesian Methods for Ecology*. Cambridge University Press, 2007.

Y. Mundlak. On the pooling of time series and cross section data. *Econometrica*, 46(1):69–85, 1978.

H. D. Patterson and R. Thompson. Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58 (3):545–554, 1971.

M. R. Rosenzweig and T. P. Schultz. Consumer demand and household production: The relationship between fertility and child mortality. *The American Economic Review (AER)*, 73(2):38 – 42, 1983.

D. J. Spiegelhalter, N. G. Best, and A. V. Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society*, 64:583–639, 2002.

A. Thomas, D. J. Lunn, N. Best, and D. Spiegelhalter. Winbugs - a bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337, 2000.

F. Vella and M. Verbeek. Whose wages do unions raise? A dynamic model of unionism and wage rate determination for young men. *Applied Econometrics*, 13:163–183, 1998.

M. Verbeek. *A Guide to Modern Econometrics*. Wiley, 2004.

J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, 2002.

J. M. Wooldridge. *Introductory Econometrics: a Modern Approach*. South-Western Cengage Learning, 2009.