이 학 석 사 학 위 논 문

Longitudinal Analysis of Barra Multi Factor Model Using R

종단연구를 활용한 Barra의 다중팩터모형

2015년 2월

서울대학교 대학원

통계학과

김 한 나

Longitudinal analysis of Barra Multi Factor Model Using R

종단연구를 활용한 Barra 다중팩터모형

지도교수 Myunghee Cho Paik

이 논문을 이학석사학위논문으로 제출함
2014년 12월

서울대학교 대학원
통계학과
김 한 나

김한나의 석사학위논문을 인준함
2014년 12월

위 원 장	박 병 욱	(인)
부위원장	Myunghee Cho Paik	(인)
위 원	이 상 열	(인)

Abstract

Longitudinal Analysis of Barra Multi Factor Model Using R

Hanna Kim
The Department of Statistics
The Graduate School
Seoul National University

Barra multi factor model has been widely used in this mature capital market. Many investors use this model to obtain the estimates of the market risk and excess returns. In this paper, Barra multi factor model framework is constructed by longitudinal data analysis, linear mixed effects model, in Korea stock market since the idea of each company is correlated and then consideration of random effect is appropriate. This application might be quite useful to fit a data and gives a good explanation.

Keywords : Linear mixed effects model, random effect, Barra multi factor model, MSCI Korea Equity Index

Student Number : 2012-20223

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Markowitz (1952) the beginning of the modern financial theory in investment field, had proposed a method of constructing an effective portfolio. After the Markowitz model, Sharpe (1964) had developed capital asset pricing model (CAPM). The paper of Ross (1976) gives a new way, APT, to do asset pricing. CAPM is one factor model that expected excess returns are proportional to the portfolio's beta. APT is multiple factor model that the linear function of various factors can be modeled to represent the expected return of the financial asset. Fama and French (1992) studied three factors which can explain the returns of stock market, then constructed Fama-French three-factor model. Barra (1976) follows the APT, which uses multi factor method to build model.

There are three types of multi factor models. Frist one is Macro-economic factor model that are observable economic and financial time series factors such as GNP growth and inflation. Second is statistical factor model, factors are unobservable and extracted from asset returns. Last one is fundamental factor model. Under this model, factors are created from observable asset characteristics like industry classification, market capitalization, style classification (value, growth) etc. to determine the common risk factors. Note that BARRA approach is based on fundamental factor model.

Barra Rosenberg had observed that companies possessing similar characteristics

may show returns that are different from the other companies. The market-wide factors are called as common factors, and characteristics that are unique to a particular company are called specific factors. This idea can be summarized with a BARRA multiple factor model.

In this thesis, we try to apply mixed effects model to BARRA model framework. Linear mixed effects model is an extension of a linear regression model for longitudinal data. Due to characteristic of data, each company is correlated and then consideration of random effect is appropriate. The remainder of this paper is organized as follows. Chapter 2 reviews statistical methods used in this paper and the linear mixed effects model. Chapter 3 describes the data which is based on the BARRA Equity model. In chapter 4, we will discuss the data application based on the suggested models from Chapter 2. From this result, we analyze the significant risk factors on portfolio return. Chapter 5 provides conclusion and discussion.

Chapter 2

Statistical Methods

2.1 Linear Mixed Effects Model

Linear mixed effects models, like many types of statistical models, describe a relationship between a response variable and some of the covariates that have been measured along with the response. Linear mixed effects models consists of two parts, fixed and random effects. Fixed effects are the covariate effects that are fixed across subjects. These effects are the ones of our particular interest. For example, the regression coefficients in usual regression model are fixed effects. Random effects are the covariate effects that vary among subjects. So these effects are subject-specific and hence are random.

For a given subject i with n_i repeated measurements, the Laird-Ware model for outcome vector Y_{ij} can be written as

$$Y_{ij} = \underbrace{\beta_1 x_{1ij} + \cdots + \beta_p x_{pij}}_{fixed} + \underbrace{b_{i1} z_{1ij} + \cdots + b_{iq} z_{qij}}_{random} + \epsilon_{ij} \qquad (2.1)$$

- $Y_{ij} =$ response variable of subject i at jth measurement.
- $n_i =$ number of measurement for subject i.
- $m =$ number of subjects.

- $\beta_1,\cdots,\beta_p$ = fixed-effect coefficients, identical for all subjects. $\beta \in R^p$
- $x_{1ij},\cdots,x_{p,ij}$ = fixed-effect regressors for ith subject at jth measurement.
- $b_{i1},\cdots,b_{iq}$ = random-effect coefficients for subject i, assumed to be mutivariately normally distributed. $b_i \in R^q$
- $z_{1ij},\cdots,z_{qij}$ =random-effect regressor for ith subject at jth measurement.

Alternatively but equivalently, in matrix form,

$$Y_i = X_i\beta + Z_i b_i + \epsilon_i, \qquad i = 1,\cdots,m \tag{2.2}$$

$$b_i \sim N_q(0, D)$$
$$\epsilon_i \sim N_{n_i}(0, \sigma^2 \Lambda_i)$$
$$b_i \amalg \epsilon_i$$

Elements along the main diagonal of the D matrix represent the variances of each random effect in b_i, and the off-diagonal elements represent the covariance between two corresponding random effects. In particular, this matrix is used to define the random intercept and random coefficient models. The $\sigma^2\Lambda_i$ matrix, called the residual (or error) covariance structure is a key matrix in marginal models. Unlike standard linear models, it allows errors and therefore observation to be correlated to each other. There are several ways of specifying $\sigma^2\Lambda_i$ to make model better fit the nature of data. Note that the general form of the linear mixed effects model is the same for clustered observations.

2.2 Estimating Parameters in Linear Mixed Effects Model

The most commonly used approaches to parameter estimation in linear mixed effects models are maximum likelihood and restricted maximum likelihood methods. The linear mixed effects model (2.2) can be rewritten as

$$Y = X\beta + Zb + \epsilon \qquad (2.3)$$

where $\begin{pmatrix} Y \\ b \end{pmatrix} \sim N_{n+mq}\left(\begin{pmatrix} X\beta \\ 0 \end{pmatrix}, \begin{pmatrix} R & Z\varsigma \\ \varsigma Z^T & \varsigma \end{pmatrix}\right)$

$$\varsigma = \begin{pmatrix} D & & \\ & \ddots & \\ & & D \end{pmatrix} \in R^{mq \times mq}$$

$$R = \begin{pmatrix} \sigma^2 \Lambda_i & & \\ & \ddots & \\ & & \sigma^2 \Lambda_m \end{pmatrix} = \begin{pmatrix} \Sigma_1 & & \\ & \ddots & \\ & & \Sigma_m \end{pmatrix} \in R^{n \times n}$$

Linear mixed effects model (2.3) can be written as two stage hierarchical model as follows:

$$Y|b \sim N_n(X\beta + Zb, R) \qquad (2.4)$$

$$b \sim N_{mq}(0, \varsigma) \qquad (2.5)$$

Linear mixed effects model (2.3) can be written as marginal model as follows:

$$Y \sim N_n(X\beta, Z\varsigma Z^T + R) \qquad (2.6)$$

If one is only interested in estimating β one can use the ordinary linear model (2.6). However, if one is interested in estimating β and b, one has to use model (2.4) and (2.5).

2.2.1 Estimation of β and b for known ς and R

When estimating β for known ς and R, we have as MLE or WLSE(weighted least squared estimator) using marginal model (2.6).

$$\log L = -\frac{1}{2} ln|V| - \frac{1}{2}(Y - X\beta)^T V^{-1}(Y - X\beta) \qquad (2.7)$$

In order to maximize the log likelihood function (2.7), we solve this equation (2.8):

$$X^T V^{-1}(Y - X\beta) = 0 \qquad (2.8)$$

Therefore, the MLE or WLSE for β is $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} Y$.

From (2.3), the predictor of b is $E(b|Y) = \varsigma Z^T V^{-1}(Y - X\beta)$ with variance covariance $Var(b|Y) = \varsigma - \varsigma Z^T V^{-1} Z\varsigma$ and this predictor is the best linear unbiased predictor of b (BLUP). Therefore, $\tilde{b} := \varsigma Z^T V^{-1}(Y - X\tilde{\beta})$ is the empirical BLUP (EBLUP).

Consider joint log likelihood of $(Y^T, b^T)^T$ with respect to (β, b) in order to estimate parameters.

$$f(Y,b) = f(Y|b) f(b) \qquad (2.9)$$

$$\ln f(Y,b) = -\frac{1}{2}(Y - X\beta - Zb)^T R^{-1}(Y - X\beta - Zb) - \frac{1}{2} b^T \varsigma^{-1} b + constant \qquad (2.10)$$

Therefore, solve below two equations respectively and then gain estimation of β and b.

$$\frac{\partial}{\partial \beta} Q(\beta, b) = -2X^T R^{-1} Y + 2X^T R^{-1} Z b + 2X^T R^{-1} X\beta = 0 \qquad (2.11)$$

$$\frac{\partial}{\partial b} Q(\beta, b) = -2Z^T R^{-1} X\beta - 2Z^T R^{-1} Y + 2Z^T R^{-1} Zb + 2\varsigma^{-1} b = 0 \qquad (2.12)$$

2.2.2 Estimation of β and b for unknown ς and R

We assume the marginal model (2.6). Let $V = Z\varsigma Z^T + R$ that ς and R are only known up to the variance parameter ϑ, i.e. rewrite $V(\vartheta) = Z\varsigma(\vartheta)Z^T + R(\vartheta)$. Based on marginal model, log likelihood for (β, ϑ) as follows:

$$l(\beta, \vartheta) = -\frac{1}{2} ln|V(\vartheta)| + (Y - X\beta)^T V(\vartheta)^{-1} (Y - X\beta) + constant \qquad (2.13)$$

If we maximize (2.13) equation for fixed ϑ with regard to β, we get $\tilde{\beta}(\vartheta) := (X^T V(\vartheta)^{-1} X)^{-1} X^T V(\vartheta)^{-1} Y$. Then the profile log likelihood is

$$\begin{aligned} l_p(\vartheta) &:= l(\tilde{\beta}(\vartheta), \vartheta) \qquad (2.14) \\ &= -\frac{1}{2} ln|V(\vartheta)| + (Y - X\tilde{\beta}(\vartheta))^T V(\vartheta)^{-1} (Y - X\tilde{\beta}(\vartheta)) \end{aligned}$$

Maximizing $l_p(\vartheta)$ with regard to ϑ gives MLE $\hat{\vartheta}^{ML}$. However $\hat{\vartheta}^{ML}$ is biased and this is why one uses often restricted ML estimation (REML).

Therefore the restricted ML of covariance parameter vector ϑ is estimated by $\hat{\vartheta}^{REML}$ which maximizes $l_R(\vartheta) = l_p(\vartheta) - \frac{1}{2} ln|X^T V(\vartheta)^{-1} X|$. And the fixed effects β and random effect b are estimated by $\hat{\beta} = (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} Y$ and $\hat{b} = \hat{\varsigma} Z^T \hat{V}^{-1} (Y - X\hat{\beta})$.

Chapter 3

Data Description

3.1 Data introduction

This paper selects data as the MSCI Korea Equity index, and the time is from January 2014 to August 2014, a total of 8 months. Using index price, calculate rate of returns:

$$R_{i,t} = \frac{\mathrm{Price}_{i,t}}{\mathrm{Price}_{i,t-1}} - 1 \qquad (3.1)$$

Where $R_{i,t}$ is rate of returns of assets i, time t and $\mathrm{Price}_{i,t}$ is the closing price of assets i at time t. In practice, Barra identifies many descriptors and defines style factors applied different weight for each market. However, it is difficult to collect all descriptors which are suggested by barra methodology because of market data constraint. In this paper, choose historical beta, historical alpha, log of market cap and other descriptors as showed in Table 3.1.

Factor	Descriptor	Information
Country	Korea	1 (all)
Industry	Energy	Binary
	Materials	Binary
	Industrials	Binary
	Consumer Discretionary	Binary
	Consumer Staples	Binary
	Health Care	Binary
	Financials	Binary
	Information Technology	Binary
	Telecommunication Service	Binary
	Utilities	Binary
Volatility	HBETA	Historical beta
Momentum	HALPHA	Historical alpha
Size	LNCAP	Log of market cap
Value	PPER	Predicted price to earning ratio
	TPER	Trailing price to earning ratio
	PDR	Price to dividend ratio
	PBR	Price to book ratio
	PCR	Price to cash earning ratio
Growth	GROWTH5	Earning to growth ratio
	PEGR	Price earning to growth ratio

Table 3.1 Descriptors by Factor

In this case, county factor represents the intercept term in the regression. Industries are very important variables in explaining the source of Korea equity return due to applying global industry classification standard (GICS) for the industry factor structure. In this paper, there are top 10 level sectors as industry factors by hierarchical GICS scheme.

Descriptors are combined into style factors. $X_{nk} = \sum w_i d_{n,i}$. Where w_i is the descriptor weight which are determined by optimization algorithm to maximized the explanatory power of model. As a result, there are 5 style factors. Volatility is typically the most important factor since it captures market risk. Momentum differentiates stocks based on recent relative performance. Size captures the

effect of large-cap stocks moving differently from small-cap stocks. Value describes investment style which seeks to identify stocks that are priced low relative to fundamentals. Growth differentiates stocks based on their prospects of sales or earning growth.

3.2 Data Cleansing

Frist of all, since data is structured in the wide form, need to re-arrange data. Owing to reshaping, longitudinal data analysis is appropriate. Second, outcome variable (Rate of return) has missing value. Thus, need to determine missing data mechanism by fitting a logistic regression model to check whether the missingness depends on the measurement at previous time points, covariates and products of them.

3.2.1 Missing Mechanism

The appropriateness of different methods of analysis of incomplete longitudinal data is determined by the missing data mechanism. The missing data mechanism can be thought of as the probability model describing the relation between the missing data R and response data Y processes. A taxonomy of missing data mechanisms, first proposed by Rubin (1976), and further developed in Little and Rubin (2002), is based on the conditional density of the missingness process R given the complete response vector $Y=(Y^o, Y^m)$. The three types of mechanisms are:

Mssing Completely at Random (MCAR)

$$\Pr(R|Y,X,Z)=\Pr(R|X,Z) \qquad (3.2)$$

The probability that responses are missing is unrelated to both the specific values that they would have been obtained and the set of observed responses.

This means that the distribution of the observed data does not differ from the distribution of the complete data. As a result, under MCAR we can obtain valid inference using any valid statistical procedure for the data, while ignoring the missing values.

Missing at Random (MAR)

$$\Pr(R|Y,X,Z) = \Pr(R|Y^o,X,Z) \qquad (3.3)$$

The probability of missingness depends on the set of observed response, but is unrelated to the outcomes that should have been obtained. Missing values can be validly predicted using the observed data under a model for the joint distribution $Y = \{Y^o, Y^m\}$. Under MAR, likelihood-based analyses based on the observed data can provide valid inferences even if we ignore the contribution of r_i, provided that the model for the measurement process y_i is correctly specified.

Missing Not at Random (MNAR)

$$\Pr(R|Y) \quad depends\ on \quad Y^m \qquad (3.4)$$

The probability that responses are missing depends on a subset of the responses we would have observed. Under MAR the observed data do not constitute a random sample from the target population. The model assumed for the missingness process is crucial and must be included in the analysis.

Coefficient	Estimate	Std.Error	P-value
Rtn_Jan	-0.17342	0.15306	0.2572
Rtn_Feb	0.11253	0.0863	0.1922
Rtn_Mar	-0.01023	0.14208	0.9426
Rtn_Apr	0.02594	0.14709	0.86
Rtn_May	-0.10653	0.10816	0.3246

Table 3.2 Fit a logistic regression model to determine missing mechanism

Based on result from Table 3.2 and three types of missing mechanism, p-values are large and hypotheses keep. That is Rtn_Jun's missingness does not depend on previous returns. Therefore missing mechanism is MCAR. Mixed effects model is valid under MCAR when ignoring missing outcome.

In this data, there are also many missing covariates. It is required to compute missing factor. A very simple approach would be to assign zero to all missing values. However, the better approach, regress only non-missing factors against missing factors. The coefficients are used to estimate the missing factors.

3.2.2. Style Factors

After missing issues, descriptors are standardized to a uniform scale. It is the process by which a mean is subtracted from each value. Then each value is divided by a standard deviation. Finally, calculate style factors by equal weighted sum of descriptors. The data given in Table 3.3 shows that the only complete data which is used in this analysis.

Name	Symbol	Weigh	Price	Sector.Nar	Secto	Return	HBETA	HALF	SIZE	PPER	TPER	PDR	PBR	PCR	GROV	PEGR	TIM	Value	Growt
Amorepaci	090430-I	0.53	945.485	Consumer	30	-0.2	-1.25	-0	-0	0.82	0	-1	0.4	0.2	0.32	0.68	1	0.17	0.5
Amorepaci	090430-I	0.6	1,099.77	Consumer	30	16.3	-1.34	0.3	-0	0.78	0.2	-1	0.6	0.4	0.32	0.76	2	0.28	0.54
Amorepaci	090430-I	0.64	1,183.71	Consumer	30	7.63	-1.18	0.3	-0	0.95	0.5	-1	0.7	0.3	0.32	0.87	3	0.36	0.59
Amorepaci	090430-I	0.69	1,291.98	Consumer	30	9.15	-1.11	0.5	-0	1.08	0.7	-1	0.8	0.4	-0	-2.5	4	0.44	-1.3
Amorepaci	090430-I	0.73	1,420.31	Consumer	30	9.93	-1.11	0.8	-0	1.15	0.8	-1	1	0.5	-0	-2.7	5	0.54	-1.4
Amorepaci	090430-I	0.76	1,506.23	Consumer	30	6.05	-1.2	0.9	-0	1.3	0.6	-1	1	0.5	-0	-2.4	6	0.53	-1.2
Amorepaci	090430-I	0.85	1,722.04	Consumer	30	14.3	-1.3	1.3	0.1	1.85	0.9	-1	1.3	0.7	-0	-2.9	7	0.8	-1.4
Amorepaci	090430-I	1.04	2,080.97	Consumer	30	20.8	-1.09	1.7	0.2	2.05	1.4	-1	1.7	1.1	-0	-3.4	8	1.08	-1.7
AmorePaci	002790-I	0.22	443.313	Consumer	30	0.61	-0.81	1.9	-0	0.82	0.1	-1	0	-0	0.67	0.41	1	-0	0.54
AmorePaci	002790-I	0.22	468.384	Consumer	30	5.66	-0.74	2.2	-0	0.6	0.2	-1	0.1	-0	0.67	0.43	2	-0	0.55
AmorePaci	002790-I	0.23	481	Consumer	30	2.69	-0.79	2.1	-0	0.65	-0	-1	0.1	-0	0.67	0.39	3	-0.1	0.53
AmorePaci	002790-I	0.24	516.791	Consumer	30	7.44	-0.69	1.7	-0	0.79	0	-1	0.1	-0	0.61	0.43	4	-0	0.52
AmorePaci	002790-I	0.3	672.417	Consumer	30	30.1	-0.57	2.1	-0	1.05	0.5	-1	0.4	-0	0.61	0.5	5	0.18	0.56
AmorePaci	002790-I	0.32	738.288	Consumer	30	9.8	-0.67	2.3	-0	1.29	0.2	-1	0.5	-0	0.61	0.45	6	0.18	0.53
AmorePaci	002790-I	0.37	861.993	Consumer	30	16.8	-0.77	2.1	-0	1.7	0.5	-1	0.8	-0	0.61	0.5	7	0.4	0.56
AmorePaci	002790-I	0.45	1,043.44	Consumer	30	21.1	-0.17	2.4	-0	2.16	0.9	-1	1.1	0.1	0.61	0.57	8	0.66	0.59
BS Financia	138930-I	0.42	14.473	Financials	40	-2.3	-0.37	0.6	-0	-0.9	-1	0.6	-0	-0	0.25	0.43	1	-0.4	0.34
BS Financia	138930-I	0.41	14.558	Financials	40	0.58	-0.37	0.4	-0	-0.9	-1	0.6	-0	-0	0.25	0.43	2	-0.4	0.34
BS Financia	138930-I	0.39	13.819	Financials	40	-5.1	-0.34	0.3	-0	-0.9	-1	0.6	-0	-1	0.25	0.44	3	-0.4	0.34
BS Financia	138930-I	0.41	14.945	Financials	40	8.15	-0.39	0.8	-0	-1	-1	0.5	-0	-1	0.15	0.69	4	-0.5	0.42
BS Financia	138930-I	0.47	15.242	Financials	40	1.99	-0.23	0.6	-0	-1	-1	0.5	-0	-1	0.15	0.7	5	-0.5	0.43
BS Financia	138930-I	0.44	14.726	Financials	40	-3.4	-0.25	0.5	-0	-1	-1	0.6	-0	-1	0.15	0.67	6	-0.5	0.41
BS Financia	138930-I	0.48	16.102	Financials	40	9.34	-0.22	0.8	-0	-0.9	-1	0.4	-0	-1	0.15	0.73	7	-0.4	0.44
BS Financia	138930-I	0.49	16.569	Financials	40	2.9	-0.15	0.9	-0	-0.9	-1	0.4	-0	-1	0.15	0.74	8	-0.4	0.45
Celltrion, I	068270-I	0.45	39.818	Health Ca	35	15.1	-1.69	0.8	-0	0.44	-0	-1	1.4	-0	1.72	0.25	1	0.07	0.98
Celltrion, I	068270-I	0.47	42.779	Health Ca	35	7.44	-1.56	1.1	-0	-0.3	0.1	-1	1.6	-0	1.72	0.26	2	-0	0.99
Celltrion, I	068270-I	0.43	39.368	Health Ca	35	-8	-1.33	1	-0	0.36	1.1	-1	1.3	2	1.72	0.32	3	0.74	1.02
Celltrion, I	068270-I	0.48	44.333	Health Ca	35	12.6	-1.28	1.5	-0	0.57	1.4	-1	1.5	2.3	0.55	0.72	4	0.93	0.63
Celltrion, I	068270-I	0.47	45.416	Health Ca	35	2.44	-0.48	0.3	-0	0.6	1.4	-1	1.6	2.3	0.55	0.72	5	0.95	0.64
Celltrion, I	068270-I	0.45	43.816	Health Ca	35	-3.5	-0.47	0.1	-0	0.49	2	-1	1.5	4.6	0.55	0.84	6	1.5	0.7
Celltrion, I	068270-I	0.37	36.924	Health Ca	35	-16	-0.41	0.1	-0	0.01	1.5	-1	1.1	3.8	0.55	0.75	7	1.07	0.65
Celltrion, I	068270-I	0.4	39.121	Health Ca	35	5.95	-0.2	-0.2	-0	0.24	1.7	-1	1.2	4	0.55	0.77	8	1.21	0.66

Table 3.3 MSCI Korea equity index from Jan 2014 to Aug 2014.

Data Source: FACTSET

Chapter 4

Application to MSCI Korea Equity Data

Chapter 4 shows that the linear mixed effects model in Chapter 2 are applied to the MSCI Korea Equity Data in Chapter 3 in order to research the effect of company's sectors and style factors on their return and risk. We implemented the model (2.1) to MSCI Korea Equity data.

4.1 Basic Analysis

	Mean
Rtn_Jan	-3.87596
Rtn_Feb	1.472525
Rtn_Mar	0.549798
Rtn_Apr	1.958687
Rtn_May	1.281616
Rtn_Jun	3.169583
Rtn_Jul	2.190208

Table 4.1 Calculate marginal mean

First of all, estimate the marginal mean. The outcomes are shown following Table 4.1. Through Table 4.1, it seems that there is time effect but in order to check this numerically and statistically, test for no change over time. Then p-value converges to zero and then null hypothesis (4.1) is rejected. Therefore,

there is change over time.

$$H_o : E(Y_{i1}) = E(Y_{i2}) = \cdots = E(Y_{i8}) \qquad (4.1)$$

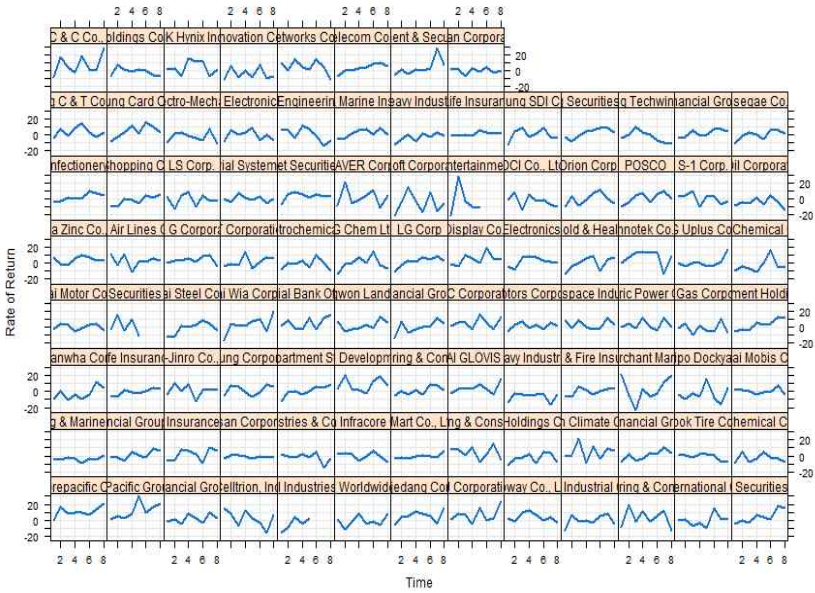

Figure 4.1 Return versus time by subject

Figure 4.1 shows that some individuals have a slowly increasing rate of return, whereas the trend for others is more erratic.

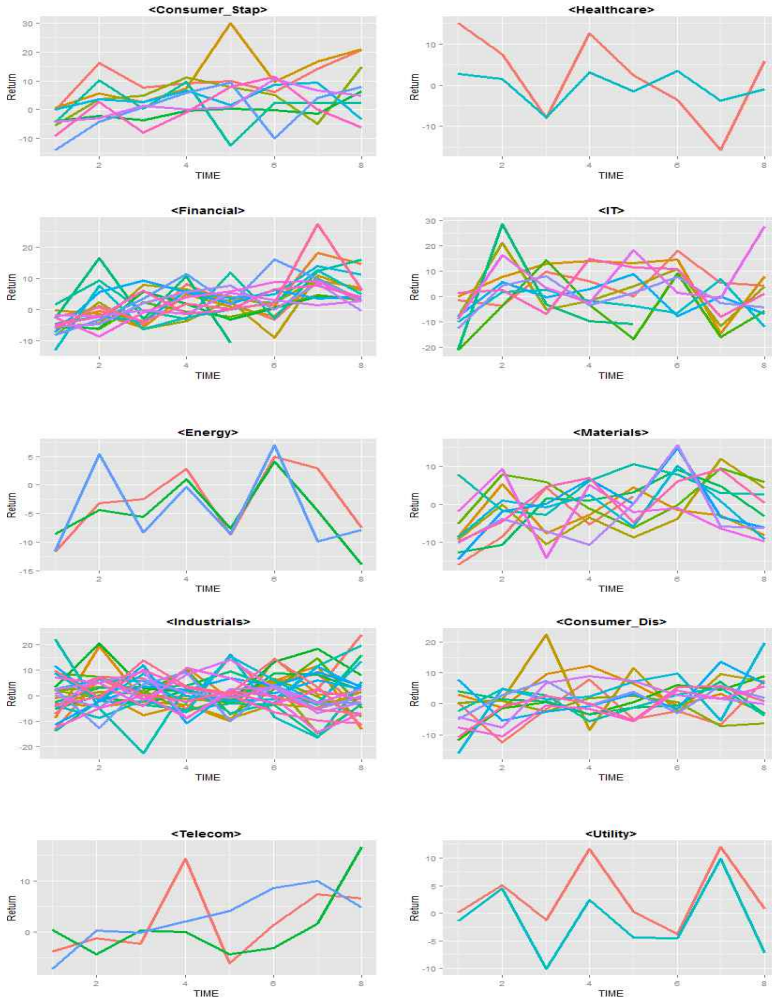

Figure 4.2 Return versus time by industry sectors

Figure 4.2 shows that some industry have a slowly increasing rate of return such as consumer staples and financial, whereas the trend for others is more erratic.

4.2 Linear Mixed Effects Model

The observations from the same subject in a cross-sectional study tend to be more similar to each other than those observations from other subjects. Responses from the same subjects are not independent and are the elements of responses are correlate only because they share common characteristics, called the random effects. Thus, we use mixed effects model approach: model fixed effects and random effects in Chapter2.

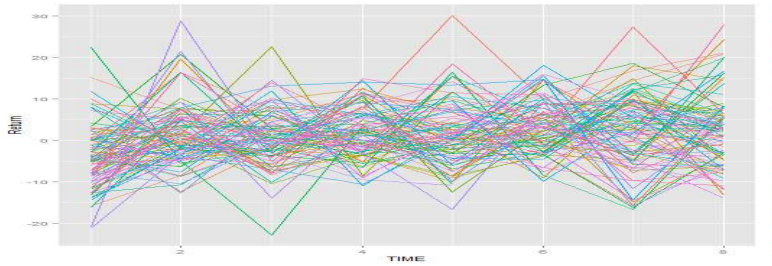

Figure 4.1 Plot rate of return over time

According to Figure 4.1, we could have hypothesis that there is variation of the intercept and slope across each company. Thus fit the linear mixed effects model with a varying intercept and with varying intercept and slope.

	Df	AIC	BIC	Pr(>Chisq)
name_intercept	18	5293.7	5377.6	
name_slope	20	5297.4	5390.7	0.8894

Table 4.2 Result of ANOVA test between LMM with intercept only and LMM with intercept and slope (Name as random effect)

	Df	AIC	BIC	Pr(>Chisq)
sector_intercept	9	5294.3	5336.2	
sector_slope	11	8286.6	5337.9	0.002913

Table 4.3 Result of ANOVA test between LMM with intercept only and LMM with intercept and slope (Sector as random effect)

	Df	AIC	BIC	Pr(>Chisq)
M1	10	5293.4	5340.1	
M2	12	5285.4	5341.4	0.00245
M3	12	5297.4	5353.3	1
M4	14	5289.4	5354.6	0.00245

Table 4.4 Result of ANOVA test 3-level mixed effects model

	Df	AIC	BIC	Chisq	Pr(>Chisq)
sector_slope	11	5286.6	5337.9		
m2	12	5285.4	5341.4	3.1737	0.07483
name_int	18	5293.7	5377.6	3.7218	0.71426

Table 4.5 Result of ANOVA test among three models

Table 4.2 shows that the intercept only model is better than intercept and slope model. That is, there is no difference slope for each company. However, Table 4.3 shows that random intercept and slope model is better when random effect is Sector. In Table 4.4, M1 represents (1|Name)+(1|Sector), M2 model represents (1|Name)+(TIME|Sector), M3 model represents (TIME|Name)+(1|Sector) and M4 model represents (TIME|Name)+(TIME|Sector). Table 4.4 shows that M2 model is the best one based on AIC and BIC. M2 model has random intercept of individual company and random intercept and slope of each sector. Table 4.5

shows that M2 model, 3-level model, is the best one to apply MSCI Korea Equity Index data to linear mixed effects model.

	Estimate	Std.Error	t-value
Intercept	-1.4111	0.5968	-2.365
HBETA	1.0999	0.3546	3.084
HALPHA	2.737	0.3403	8.043
SIZE	-0.1517	0.3107	-0.188
Value	0.6514	0.597	1.091
Growth	-0.2824	0.3933	-0.718
TIME	0.522	0.1646	3.171

Table 4.6 Outcome liner mixed effects model for M2 model (Fixed effect)

	Name	Variance	Std.Dev	Corr
Name	(Intercept)	2.2947	1.5148	
Sector	(Intercept)	0.1371	0.3703	
	TIME	0.1193	0.3453	-0.61
Residual		46.4099	6.8125	

Table 4.7 Outcome liner mixed effects model for M2 model (Random effect)

Table 4.6 shows that HBETA, HALPHA, Value and TIME are significant. Thus if the HBETA, HALPHA, VALUE and TIME increase, rate of return will increase. That is, volatility, momentum, value increase, rate of return will increase. As time goes by, rate of return also increase. In this result, there are no size and growth effects into rate of return. Table 4.7 shows that variability of the intercept across Name is greater than variability of the intercept across Sector. Also correlation between intercept and slope across Sector is negative. That is, when a sector's intercept increase by one unit of standard deviation, that sector's slope would decrease by 0.61 standard deviation. Through Table 4.6 and Table 4.7, we could find out all parameter. $\sigma^2 = 46.4099$, $D_{Name} = 2.2947$, $D_{Sector} = \begin{pmatrix} 0.1371 & -0.3454 \\ -0.3454 & 0.1193 \end{pmatrix}$

Chapter 5

Conclusion and Discussion

Our objectives in this thesis were that we implemented the linear mixed effects model to MSCI KOREA Equity Index data and analyzed how industry factors and style factors could influence on their rate of return.

As a result, this study has indicated that the linear mixed effects model was very suitable in modeling Barra's multi factor model. Because the effect of individual company's unique characteristics, due to correlation, could be applied, the effect of sector's specific characteristics also could be reflected. Futhermore, it is difficult to obtain the covariance matrix of beta in Barra's multi factor model but in this model we do not need to find out covariance matrix and only apply time effect in mixed effects model. However, there is limitation that it could be not appropriate methodology when analyzing long time series data since longitudinal analysis is basically deal with data with length of time is not greater than number of subjects. None the less, we can conclude that the result would be useful and application is very simple.

Bibliography

Agresti, A. (1990), *Categorical Data Analysis*. John Wiley & Sons. New York.

Chi, E.M. and Reinsel, G.C. (1989), Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Association*, **84**, 452-459.

Cox, D.R. and Wermuth, N. (1992) Response models for mixed binary and quantitative variables. *Biometrika*, **79**, 441-461.

Donald, B. Rubin (1976), Inference and Missing Data. *Biometrika*, **63,** 581-592.

Fitzmaurice, G.M and Laird, N.M. (1993) A Likelihood-based method for analysing longitudinal binary responses. *Biometrika*, **80**, 141-151.

Fitzmaurice, G.M and Laird, N.M (1997) Regression models for mixed discrete and continuous responses with potentially missing values. *Biometrics*, **53**, 110-122.

Paik, MC and Wang, C. (2009), Handling missing data by deleting competely observed record. *Journal of Statistical Planning and Inference*, **139,** 2341-2350.

Peter, J. Diggle, Patrick, J. Heagerty, Kung-Yee Liang and Scott L. Zeger. (2002), *Analysis of Longitudinal Data*: Second Edition, Oxford University Press, New York.

Verbege, G. and Lesaffre, E. (1996) A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, **91**, 217-221.

Verbeke, G. and Molenberghs, G. (2000), *Linear Mixed Models for Longitudinal Data*, Springer-Verlag, New York.

Zhang, H and Paik, MC. (2009), Handling missing responses in generalized linear mixed model without specifying missing mechanism. *Journal of Biopharmaceutical Statistics*, **19,** 1001-1017.

국 문 초 록

종단연구를 활용한 Barra 다중팩터모형

Longitudinal Analysis of
Barra Multi Factor Model Using R

Barra 다중팩터모형은 자본시장에서 굉장히 폭넓게 사용되고 있다. 많은 투자자들이 이 모형을 활용하여 시장리스크와 초과수익률을 추정하고 있다. 본 논문에서는 Barra 다중팩터모형의 기본적인 개념을 바탕으로 MSCI 대한민국 주식 지수 데이터를 활용하여 종적자료연구 중 하나인 선형혼합모형에 적용해 보고자 한다. 선형혼합모형을 적용하게 된 가장 큰 이유는 각 기업이 가지는 특수한 특성이 존재하여 그들 사이에 상관관계가 있다는 아이디어가 선형혼합모형에서의 랜덤효과에 접목시킬 수 있기 때문이다. 이 적용법은 자료를 적합 하는 데 유용하고 좋은 결과와 설명력을 제공해주고 있다.

주요어 : 선형혼합모형, 랜덤효과, Barra 다중팩터모형, MSCI Korea 주식 지수

Student Number : 2012-20223