경제학박사학위논문

Forecasting Bankruptcy More Frequently:

Information Update via High Frequency Data

회귀모형에서 혼합주기 자료를 이용한

정보 업데이트 방법에 관한 이론 및 실증연구

2016 년 2 월

서울대학교 대학원

경제학부

김 명 원

Forecasting Bankruptcy More Frequently:

Information Update via High Frequency Data

지도교수 류 근 관

이 논문을 경제학박사학위논문으로 제출함

2015 년 10 월

서울대학교 대학원

경제학부 경제학 전공

김 명 원

김명원의 박사학위논문을 인준함

2015 년 12 월

위 원 장 김 세 직 (인)

부 위 원 장 류 근 관 (인)

위 원 이 석 배 (인)

위 원 임 형 석 (인)

위 원 박 용 린 (인)

Abstract

Forecasting Bankruptcy More Frequently:

Information Update via High Frequency Data

Myungwon Kim

Department of Economics

The Graduate School

Seoul National University

This paper considers the econometric problems arising from using outdated data in a regression model in which the independent variable is observed less frequently than the dependent variable. Specifically, OLS estimates may suffer from a form of omitted variable bias if outdated data is correlated with information during the time no observation takes place. We claim that using data correlated with the independent variable but with a shorter observation period to update the independent variable can eliminate the bias, as well as reducing uncertainty in estimating the dependent variable. We test the theory with an empirical model of bankruptcy forecast for medium sized firms. We present a more accurate default forecast model that updates the average change in firms' financial standing with monthly business cycle information. Financial institutions may use the monthly estimates to monitor losses on their loan portfolios more accurately and more frequently.

Keywords: mixed frequency data, information update, outdated data, omitted variable bias, default forecasting, credit risk monitoring

Student Number: 2013 - 30058

Contents

List of Tables

List of Figures

1. Introduction

1.1 Motivation

Any recorded data is observed at discrete points in time. This means that the time that one makes use of some data may not coincide with the time of its observation. The difference can be negligible if the data is observed frequently enough, but this is not always the case. Even data containing sufficient information for decision-making may lead to inferior decisions if it is not observed in time.

Consider a game of baseball, for example. You do not have time to watch the game yourself, so a friend texts you the runs scored at each inning. It is now the top of the 9^{th} inning. If the messages you received until the 8^{th} inning are as shown in Table 1, which team would you believe is in the lead? There are many factors such as closers, lineups, or the coachs' strategies. But assuming the scores of period are values that have been independently sampled from identical probability distributions, you would probably believe that Team B is winning.

[Table 1] Team scores at each inning

Team	1	2	3	4	5	6	7	8	9	Total
A	0	0	0	0	0	0	0	0		0
B	0	0	0	3	0	0	0	0		3

Now assume that a different friend of yours texts you every time there is a hit, so you have access to the current number of hits for each team, as well as scores after each inning.

[Table 2] Team hits at each inning

Team	1	2	3	4	5	6	7	8	9	계
A	0	0	0	0	0	0	1	0	**4**	5
B	0	1	0	2	0	0	0	2		5

If your information regarding the number of hits for each team is as described in Table 2, which team would you now believe in winning? As Team A makes additional hits in the top of the 9th, you are likely to become less certain of B's lead. After you hear that Team A has made 4 hits, you might as well believe that Team A is on the verge of turning the game around, if not already in the lead. [1] While the information on scores was yet to be updated, the information on hits changed your prediction.

The current score of each team is sufficient to determine which team is on the lead. Hence if you always have access to the current scores, the number of hits does not provide additional information in determining the winning team. But at moments when you do not have the current scores, updated data on the number of hits may become useful in guessing the winner. Even though the current scores provide sufficient information to determing the current winning team, when this information is not available, we can make use of variables that are related to scores but have shorter observation periods (number of hits, or on-base data) to obtain better estimates of the game.

The above example can be formalized as below.

[1] Numbers in Table 1 and Table 2 are based on the WBSC Premier 12 semifinal match between South Korea and Japan. In the actual game, Team A (Korea) took the lead after the fourth hit in the top of the 9th and won the game.

$$f(y_t|x_t) = f(y_t|x_t, z_t) \quad (1)$$

$$f(y_t|x_{t-1}) \neq f(y_t|x_{t-1}, z_t) \quad (2)$$

$$\text{where } y_t = \begin{cases} 1 & if\ x_t > 0 \\ 0 & if\ x_t = 0 \\ -1 & if\ x_t < 0 \end{cases}$$

x_t is defined as Team A's score minus Team B's score at time t, and z_t is a vector representing the number of hits from t-1 to t for each team. y_t is a variable taking the value of 1 if Team A is winning, -1 if Team A is losing, and 0 if the game is tied, at time t. Equation (1) implies that if the current score difference (x_t) is given, the number of hits (z_t) does not provide additional information on identifying the winning team. Since the number of hits (z_t) affects the game result (y_t) only through score difference (x_t), the score difference provides sufficient information on the game result. But if the current score difference is unobservable and all we have is the last observed value (x_{t-1}), the number of hits can provide additional information about the game result because it holds information about the current score differential.

This paper focuses on phenomena that resemble the relationship between score difference and the number of hits. x_t and z_t are closely related, but once x_t is given z_t does not have additional information on y_t. But because z_t is observed more frequently than is x_t, z_t can be used to update information on x_t at times the latter is not observable. If we use outdated data in regression models, and the outdated data is correlated with information during the time no observation takes place, the estimated coefficient suffers from omitted variable bias. Using an auxiliary variable z_t that is correlated with x_t but has a shorter observation period can eliminate or mitigate bias.

We first characterize econometric problems and phenomena that arise in such cases with the linear regression model, then perform an empirical analysis to test our theory. The empirical analysis applies our information updating method to a bankruptcy forcast model for firms. The

critical factor that determines whether a firm goes bankrupt is its financial standing.[2] But while a firm's credit risk changes continuously, its financial statement is disclosed only once a year. If a firm's financial standing is related to the business cycle, we may update the information on firms' average financial status by using monthly business cycle data, thereby measuring firms' credit risk more accurately and frequently. Such methods can be utilized by financial intermediaries, such as banks, for whom loss management is at the core of risk management.

[2] Various factors including the business cycle, executive competence, or outlet numbers can affect a firm's bankruptcy. However, we may claim that all of these factors cause bankruptcy through the channel of deterioration in financial standing.

1.2 Contributions and outline of the paper

This paper contributes to the literature in both theoretical and empirical sides.

Theoretically, we show that in a linear regression model with mixed frequency data, where the independent variable has a longer observation period than the dependent variable, it is possible to obtain timely estimates of the dependent variable by using an auxiliary variable with shorter periods of observation that is correlated with the independent variable. The values of the independent variable that are not observed during the intervals between each observation can be seen as missing data. The existing literature, such as Little(1992), provide imputations for such missing data in various environments by focusing on the randomness of the missing data's pattern or generating process. Moreover, using the latest financial data instead of the current financial standing of a firm which is unobservable may also be considered a imputation as in Shumway(2001), Chava and Jarror(2004).

We establish that using outdated data leads to omitted variable bias on the estimated coefficient if the omitted information is correlated with the independent variable. We then show that by using an auxiliary variable with shorter periods of observation that is correlated with the independent variable, we may update the omitted information and thereby mitigate or eliminate the bias, obtaining timely estimates of the dependent variable. We also describe the information updating channel by analyzing, via the relationship between the independent and the auxiliary variables, how the length of the time the independent variable is omitted affects the uncertainty of estimation.

Empirically, we apply our model of information update via auxiliary variable to a bankruptcy forecast model for firms. Our analysis contributes to the literature in the following aspects.

First, we present a method to shorten the period of forecasting firm bankruptcy to a month. A firm goes bankrupt generally due to bad financial standing. But financial standing of a firm is not observed at every moment; financial statements are usually disclosed annually. This means that bankruptcy forecasts based on firms' financial statements are also updated only once every year. This paper uses monthly business cycle data to update information on firms' unobserved financial standing, thus enabling monthly bankruptcy forecasts. Papers including Chava and Jarrow(2004) use macroeconomic variables and the discrete-time hazard model to estimate monthly bankruptcy forecast models, but this paper takes a different approach in using data. The current literature uses the latest financial statements and macroeconomic information of the corresponding month, without taking into consideration the fact that the point in time we use financial statements differs from the point they are disclosed. Since macroeconomic information preceding the settlement is already reflected in the statements, we accumulate business cycle information from the settlement to the point of evaluation to shorten the length of time that information is omitted. Intuitively, this method is more efficient because it reduces information loss. The actual estimations on the model support our approach as well.

Second, our model can be used by financial institutions such as banks that specialize in loans to monitor credit risk. Banks, which are required to hold capital over a specified level by the financial regulator, can estimate bankruptcy rate and expected loss more accurately and more frequently. It is also possible to compare losses under hyperthetical economic situations.

Third, our information updating method can be applied to diverse fields. If firms can be classified into groups that share common characteriscs such as business type or firm size, and an index is available for each group, we may be able to customize our prediction based on group-specific information source. Predicting winners in baseball games or monitoring health status may also benefit from timely information updates.

The remainder of the paper is as follows. In Chapter 2, we describe the econometric problems that arise when one cannot use data in a timely manner because the independent variable has a long observation period. Then we show that by using data that is correlated with the independent variable and has a shorter observation period, one can update information on the independent variable and eliminate or mitigate the problems. Chapter 3 uses the information update method to estimate a bankruptcy forecast model for firms with a hazard model. We also provide specific ways for financial institutions to apply our method. Chapter 4 conlcudes.

2. Is it problematic to use outdated data in regression models?

We first introduce the characteristics of mixed frequency data, and then discuss the problems that arise when a regression model uses outdated data instead of current but unobservable data, as well as how to overcome these obstacles. Specifically, we show that in a linear regression model, using mixed frequency data with observation period of the independent variable longer than that of the dependent variable leads to bias in estimated coefficient from Ordinary Least Squares (OLS). This bias can be interpreted as omitted variable bias that arises because of information that is correlated with the latest observation but has not yet been updated. Using an auxiliary variable, correlated with the independent variable but with a shorther observation period, can not only remove the bias but also reduce the uncertainty of estimation.

2.1 Introduction

Many forecast models are used by research centers or financial institutions in order to cope effectively with future economic situations, from GDP to firms' bankruptcy. A forecast model, by definition, must provide an ex ante prediction of the probability of some event; as such, accurate forecasting requires timely acquisition of the necessary data. Consider a model that precisely predicts suicides a week before by using data on individuals' emotional state. The model would be useless in predicting, let alone preventing, suicides if change in the data is not updated in 7 days. It is very important in using forecast models to use timely information, and this is made easier if the observation period of the data is short.

Not all data has short observation periods, however, and it is often the case that different

types of data have different observation periods. For instance, GDP is observed every quarter while price indices and unemployment rate are reported monthly. In such cases we say that the data are unsynchronized, and the group of data is called mixed frequency data. This paper shows that when dealing with mixed frequency data where the observation period is longer for the independent variable than the dependent variable, we can use an auxiliary variable that has a shorter observation period than the independent variable and affects the dependent variable only through the independent variable, to update the change in the independent variable taking place while it is unobservable.

Research on bankruptcy forecast models for firms such as Shumway(2001) or Chava and Jarrow(2006) use financial statements from the most recent year. This is done under the assumption that the unobserved data follows a martingale process under the given information set. But since other types of information related to financial standing are being updated even while the financial statement is not, the information set changes constantly and there seems to be room for reflecting new information. This paper shows that using outdated data in the place of current data for the independent variable in a linear regression model leads to omitted varable bias in the estimated coefficient. However, using an auxiliary variable can update the information on the independent variable to reduce or eliminate the bias. Furthermore, the regression variance can be shown to decrease as we use more information.

Using mixed frequency data is also related to imputation for missing data. That data have different observation periods means that not all data are observed at every given moment. At a given moment some data are unavailable; thus we can view the data as missing and use imputation. Resorting to outdated data can also be considered as a type of imputation where the outdated data is used as a proxy for the missing data. How to process missing data depends on how one assumes the data goes missing. Little(1992) suggests ways to estimate the linear

regression model with only the given independent and dependent variables when the independent variable is missing for different reasons.

As development of information technology has improved the way we store and process data, there is growing attention on nowcasting. Most of the studies in nowcasting focus on the situation, contrary to our model, where the dependent variable has a longer observation period than the independent variable. Giannone et al(2008) develop a GDP forecast model that updates the GDP estimate every month by using monthly data. They extract common factors from a large volume of data useful for predicting GDP, and bridge these factors to the data on GDP. Ghysels et al. (2003) and Andreou et al. (2010) introduce an estimation of mixed frequency model using MIDAS(Mixed Data Sampling) regression model. MIDAS regression model sets weights on data with high frequency data so that both high and low frequency data can be used. Weights for Intra-period data are usually determined from Beta Lag, which uses the beta distribution, or Exponential Almon Lag, using the exponential distribution(Foroni, 2013). Parameters of the distributions are estimated along with the regression coefficients by NLS(Non-linear Least Squares) method.

This chapter proceeds as follows. Section 2 shows that outdated data leads to omitted variable bias and that making use of auxiliary variable can remove the bias. Section 3 shows that using an auxiliary variable with a short observation period can reduce uncertainty of estimation. Section 4 concludes.

2.2 Omitted variable bias from using outdated data and information update with auxiliary variable

Let us start from a simple linear regression model with a single independent variable. At time t,

x and y have the following relationship in the population.

True model

$$y_t = \alpha + \beta x_t + u_t \quad (3)$$

where $t = 1, 2, \dots, T$.

y_t consists of a deterministic part and a random part. It is assumed that random disturbance term u_t has expectation 0 and variance σ_u^2 for all t. In addition, exogeneity condition $\mathrm{E}(u_t|x_t) = 0$ is satisfied for all t. This implies that in predicting the disturbance term at time t, the independent variable at time t provides no information.

We further make the following simplifying assumptions.

Assumptions

A1. The observation period of y is 1 and the observation period of x is an integer p(> 1).

A2. z has the same observation period (=1) as y and has the following relationship with x.

$x_t = x_{t-1} + \gamma \Delta z_t + \varepsilon_t$, where ε_t is a random disturbance term with expectation 0 and variance σ_ε^2, and $\Delta z_t \equiv z_t - z_{t-1}$.

A3. For any t, $\mathrm{t}' \leq T$, $\mathrm{E}(\varepsilon_{t'}|\Delta z_t) = 0$.

For any $\mathrm{t}' > \mathrm{t}$, $\mathrm{E}(\varepsilon_{t'}|x_t) = 0$.

A4. For any $\mathrm{t}' > \mathrm{t}$, $\mathrm{E}(u_{t'}, \varepsilon_t) = 0$.

Let us interpret the above assumptions with the example of bankruptcy forecast model for firms. Assumption A1 means that the observation period of x is longer than other variables. A firm's bankruptcy(y) is observable daily or monthly but its financial statements are observed only once a year. A2 implies that the change in x_t can be divided into a systematic part from change in z_t and unsystematic part from the disturbance term ε_t. If z_t is a common factor to all firms, a firm's financial standing can be explained by the systematic factor of business cycles and factors unique to the firm. This is similar to the CAPM(Capital Asset Pricing Model), where

the return on an asset is deteremined by the market return and idiosyncratic factors. A2 means that during the time a firm's fianancial standing is unobservable, business cycle information can be used as a source to update its current financial status. Assumption A3 ensures that past financial standing and systematic factors provide no information in determining the unsystematic factor affecting the current financial standing. In other words, business cycle information and past firm-specific (financial) factors do not affect current firm-specific (financial) factors. This assumption may seem strong since firm-specific factors may cause serial correlation. However, the problem could be mitigated if repeated observations are available by using fixed-effects model to reflect firm-specific unobserved heterogeneity. Assumption A4 means that the non-financial disturbance term u_t [3] and the financial disturbance term ε_t are not correlated.

The above assumptions lead to the following equations. As we have seen, z_t provides no additional information on y_t when current x_t is given, but does provide information when x_t is not known.

$$\mathrm{f}(y_t|x_t) = \mathrm{f}(y_t|x_t, z_t) \quad (4)$$

$$\mathrm{f}(y_t|x_{t-p}) \neq \mathrm{f}(y_t|x_{t-p}, z_{t-p+1}, \dots, z_t) \quad (5)$$

Writing the population equation for each observation leads to

$$y_{it} = \alpha + \beta x_{it} + u_{it} \quad (6)$$

$$\Delta x_{it} = \gamma \Delta z_{it} + \varepsilon_{it} \quad (7)$$

where i=1,2,…, N and t = 1, 2, …, T.

It may be practically meaningful to assume that at a given point in time z_{it} is constant in

[3] The non-financial factors are those that do not affect a firm's financial standing but do affect its bankruptcy. Defining bankruptcy as the state of net assets being lesser than 0 rules out the existence of such factors; however, in the real world, firms with the same financial standing may differ in bankruptcy due to the entrepreneur's credibility or institutional devices for bond preservation.

i.[4] This is because only average, and not individual, data may be available. For example, if Δx_{it} represents firm *i*'s change in financial standing, $\gamma\Delta z_t$ would be the average financial change for all firms, and likely to be closely related with macroeconomic variables such as the business cycle. Of course, assuming z_t has different values for each observation does not change our analysis. Below we suppress the subscript *i* unless there is ambiguity or need for emphasis.

According to the above data generating process, the observation period of x is p, so that it is updated only between intervals the length of p. On the other hand, z has observation period 1 so that it is observable at every period.

[Table 3] Data generation and observated data

	t=0	t=1	…	t=p-1	t=p	t=p+1	…	t=2p-1	t=2p	…
True x_t	x_0	x_1	…	x_{p-1}	x_p	x_{p+1}	…	x_{2p-1}	x_{2p}	…
Observed x_t	x_0	x_0	…	x_0	x_p	x_p	…	x_p	x_{2p}	…
Observed z_t	z_0	z_1	…	z_{p-1}	z_p	z_{p+1}	…	z_{2p-1}	z_{2p}	…
Observed y_t	y_0	y_1	…	y_{p-1}	y_p	y_{p+1}	…	y_{2p-1}	y_{2p}	…

If we only use fully observed data on x and y, the amount of data decreases as the observation period of x becomes longer and the number of observations on y that we discard increases. Also, it is difficult to get real-time estimates of y because we lack information about current x. If we use all observed data, we may replace unobserved values of x with the most

[4] One must be careful when z_{it} constant for all i. If one uses data only from the same period and all data sets have the same length of omission, z_{it} is no differnet from a constant term. The regression equation is identifiable only when data differs in period or length of omission.

recent observations, as do Shumway(2001) and Chava and Jarrow(2004). But such usage of outdated x may suffer from omitted variable bias.

$$y_t = \alpha + \beta x_t + u_t \quad (8)$$

$$= \alpha + \beta x_{t-\tau} + \beta(x_t - x_{t-\tau}) + u_t \quad (9)$$

$$= \alpha + \beta x_{t-\tau} + u'_t \quad (10)$$

Here, $\tau = \mathrm{t} - \left[\frac{t}{p}\right]p$, $[x]$ is the larest integer not greater than x, and $u'_t = \beta(x_t - x_{t-\tau}) + u_t$. τ is the length of time x has been omitted since its last observation.

According to the above equations, when $x_{t-\tau}$ and $(x_t - x_{t-\tau})$ are correlated, there arises an endogeneity problem in estimating the parameters because of the omitted values of x. In other words, the omitted information $(x_t - x_{t-\tau})$ cause omitted variable bias. To analyze this in detail, equation (8) can be written as below by using A2 (refer to Appendix A.1).

$$y_t = \alpha + \beta x_{t-\tau} + \beta\,\gamma\Delta_\tau z_t + \beta\sum_{j=1}^{\tau}\epsilon_{t-\tau+j} + u_t \quad (11)$$

where $\Delta_\tau z_t \equiv z_t - z_{t-\tau}$.

Now y_t consists of a constant and 4 other terms. The first term implies that the last observed x is reflected in y by a ratio of β, while the second term says that change in x that can be explained by change in z is also reflected by a ratio of β. The third term represents that part of cumulative change in x, taking place while the variable is unobservable, that is not explainable by z. The last term is the disturbance term from the original equation.

In equation (10), let $\hat{\beta}_s$ be the estimated coefficient when we perform OLS estimation only with the constant term and observed data $x_{t-\tau}$, ignoring the updated information ("short

regression"). The omitted variables bias formula gives the following relationship between $\hat{\beta}_s$ and β.

$$\text{plim}\,\hat{\beta}_s = \frac{cov(x_{t-\tau}, y_t)}{var(x_{t-\tau})} = \beta + \beta\gamma\delta_{zx} \neq \beta \quad (12)$$

Here, β and γ are population regression coefficients in equation (11) and δ_{zx} is the population regression coefficient when we regress $(z_t - z_{t-\tau})$ on $x_{t-\tau}$. According to equation (12), β_s and β are not equal unless γ or δ_{zx} is 0. That is, if the change in x that has not been updated is correlated with the latest observation, there is an omitted variable bias. The same result holds in a multiple linear regression model with more than one independent variable (refer to Appendix A.2).

Equation (11) is a from of Berkson measurement error model [5] introduced by Berkson(1950). Unlike the classical measurement error model, the true value is equal to the observed value plus error, so that if observed value and measurement error are independent OLS estimation does not suffer from attenuation bias. GLS estimation is more efficient, however, if data differ in the length of time observations are missing. In the above equation, $\gamma = 0$ means that there exists no data z that can meaningfully explain x. By A2, this implies x follows a martingale process. Therefore, under a given information set the best information on current x_t is the last observation $x_{t-\tau}$. In this case we can estimate β with only $x_{t-\tau}$ and not suffer from bias.

On the other hand, δ_{zx} can be expressed as below. $\delta_{zx} = 0$ means the observed x and

[5] In the classical measurement error model, the observed value(w) is equal to the true value(x) added by the error term(u), so that w=x+u holds. On the other hand, in the Berkson error model the true value is equal to the observed value plus the error term(x=w+u). This model can be used to measure uncertainty when we replace individual data with group average or conduct an experiment that controls variables such as the temperature or wind speed.

information updated through z are uncorrelated.

$$\delta_{zx} = \frac{cov(z_t - z_{t-\tau}, x_{t-\tau})}{var(x_{t-\tau})} \quad (13)$$

If x is white noise, for instance, the observed and the updated information are uncorrelated. If the observation period of x is 1, at each period information added by z is 0 so that β can be estimated without bias. δ_{zx} cannot be 0 if z is an autocorrelated time series.[6]

Although we can get an unbiased estimate of the coefficient if $\delta_{zx} = 0$, there is still a need to control the information updated through z if it is meaningful to get accurate estimates of y_{it}, such as in forecast models. Even if the correlation between the information updated through z and $x_{i,t-\tau}$ is 0, the value of the dependent variable estimated without the update ($\hat{y}_{it}^{s}$) differs systematically from that with the update ($\hat{y}_{it}^{l}$).

When x_{it} is observable, we can compare the estimators $\hat{y}_{it}$, $\hat{y}_{it}^{s}$, and $\hat{y}_{it}^{l}$ can be compared in large sample as follows. $\hat{y}_{it}$ is estimated value of y_{it} when x_{it} is available.

$$\text{plim}\,\hat{y}_{it} - \text{plim}\,\hat{y}_{it}^{s}$$

$$= \beta\,\gamma\left(\Delta_\tau z_{it} - \mu_{\Delta_\tau z_t}\right) - \beta\gamma\delta_{zx}\left(x_{i,t-\tau} - \mu_{x_{t-\tau}}\right) + \beta\sum_{j=1}^{\tau}\epsilon_{i,t-\tau+j} \quad (14)$$

$$\text{plim}\,\hat{y}_{it} - \text{plim}\,\hat{y}_{it}^{l} = \beta\sum_{j=1}^{\tau}\epsilon_{i,t-\tau+j} \quad (15)$$

where $\mu_{\Delta_\tau z_t} = \text{E}(\Delta_\tau z_{it})$, $\mu_{x_{t-\tau}} = \text{E}\left(x_{i,t-\tau}\right)$.

The difference between estimates of y_{it} when using $x_{i,t-\tau}$ and x_{it} can be divided into three parts – the part from information not updated through z for τ periods, the part from

[6] Since $cov(z_t - z_{t-\tau}, x_{t-\tau}) = cov\left(z_t - z_{t-\tau}, x_0 + \gamma(z_{t-\tau} - z_0)\right)$, δ_{zx} is not 0 when Δz_t has autocorrelation.

omitted variable bias of coefficient estimate, and the part from information in x_{it} that is not explained by $x_{i,t-\tau}$ or $\Delta_\tau z_{it}$. If we update information through z, only the difference due to the last, unsystematic part remains.

If we wish to obtain estimates of group average instead of individual observations (e.g. expected loss for a portfolio), the above difference in estimates decreases by the law of large numbers. In this case, adding $\Delta_\tau z_{it}$ to the equation may make little difference. However, average accuracy of individual estimates increases as we use more information, so uncertainty can still be reduced. Hence our method of information updating can be useful in obtaining more accurate group or portfolio estimates through individual estimates.

In addition, we analyze what happens when the auxiliary variable z_{it}, our source of information updates, has measurement error. The result is that OLS coefficient estimate shows attenuation bias as usual. From the period information was last updated, the size of the bias may be increase or decrease depending on the relative size of variances of updated information and measurement error. Refer to Appendix A.3 for precise results.

2.3 Information update and regression variance

We now discuss how regression variance differs as the regression model uses different information sets. Regression variance is constant in time if x_t is observed at every period because there is no omitted information which depends on time. When the observation period of x_t is p, uncertainty from using outdated information increases as information update is delayed. Specifically, regression variance increases as τ increases, is eliminated once x is updated ($\tau = 0$), and then increases again. When we partially update x_t through z_t in every period, on

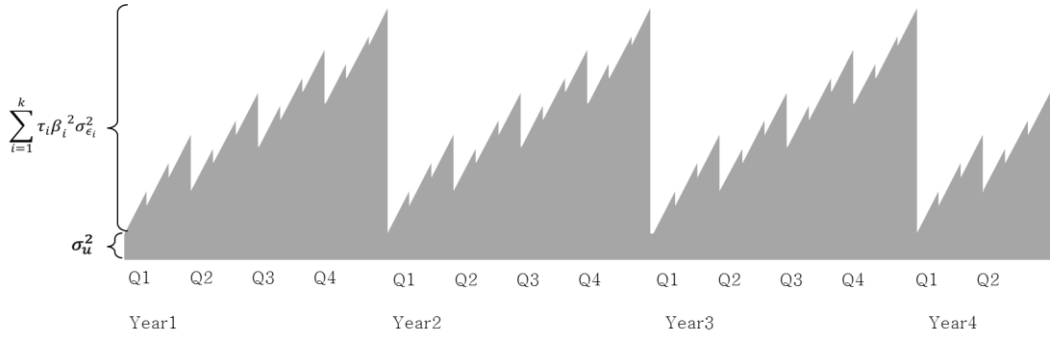
the other hand, uncertainty increases less from delayed update as z_t contains more information about x_t. Regression variance in each information set is as follows.

$$var\big(y_t - E(y_t|x_t)\big) = var(u_t) = \sigma_u^2 \equiv A$$
$$var\big(y_t - E(y_t|x_{t-\tau}, \Delta_\tau z)\big) = \tau\beta^2\sigma_\epsilon^2 + \sigma_u^2 \equiv \mathrm{B}$$
$$var\big(y_t - E(y_t|x_{t-\tau})\big) = \tau\beta^2\sigma_\epsilon^2 + \sigma_u^2 + \beta^2\gamma^2\sigma_{\Delta_\tau z}^2 \equiv \mathrm{C}$$

where $\sigma_{\Delta_\tau z}^2 = var(\Delta_\tau z)$.

From the above equations it is easy to check that $\tau > 0$ implies $A < \mathrm{B} < \mathrm{C}$. Using more accurate information about x_t reduces uncertainty in y_t. Using outdated data causes uncertainty, which is the largest right before information is updated and the smallest at the point of update. When updating through z, the maximum value of the regression variance is $Max\left(var\big(y_t - E(y_t|x_{t-\tau}, \Delta_\tau z)\big)\right) = (p-1)\beta^2\sigma_\epsilon^2 + \sigma_u^2$, so uncertainty increases as the observation period p increases. For example, if the independent variable is mixture of sets of data that are updated monthly, quarterly, or annually, the regression variance at each period takes the form in Figure 1.

[Figure 1] Information update and regression variance



Defining recovery rate of information as the portion of uncertainty from outdated data that is recovered by using auxiliary variable, it can be written as below.

$$\text{Recovery rate of information } = \frac{var(y_t - E(y_t|x_{t-\tau})) - var(y_t - E(y_t|x_{t-\tau}, \Delta_\tau z))}{var(y_t - E(y_t|x_{t-\tau})) - var(y_t - E(y_t|x_t))} = \frac{\gamma^2 \sigma^2_{\Delta_\tau z}}{\tau\sigma^2_\epsilon + \gamma^2 \sigma^2_{\Delta_\tau z}}$$

Recovery rate of information becomes close to 0 as time delayed(τ) is long, and becomes close to 1 as the change in z is large and z recovers more information. For example, if the economy has suffered a big shock since the last disclosure of financial statements, the current financial standing of firms can be explained more by the shock than past statements. If there has been little business cycle variability since the disclosure, using the past information implies a relatively small loss in information.

2.4 Conclusion

This paper suggests a method to update the independent variable, while it is unobservable, with an auxiliary variable that has a shorter observation period. In the linear regression model, using outdated data leads to biased OLS estimates. This can be seen as an omitted variable bias, arising because the unobserved data is correlated with the last observed data. Using the auxiliary variable with a shorter observation period not only removes the bias but reduces uncertainty in estimation. When the auxiliary variable contains measurement error, OLS estimate exhibits attenuation bias as usual.

The information update via mixed frequency data we suggest has three advantages. First, it eliminates the bias inherent in using outdated data. Second, it provides timely forecasts of the

dependent variable. Finally, it can be easily applied to different areas. In Chapter 2, we apply our information updating model to an actual bankruptcy forecast model for firms.

3. Forecasting Bankruptcy More Frequently: Information Update via High Frequency Data

In this chapter, we apply our information updating method to a bankruptcy[7] forecast model for firms. Though a firm's default is closely related with its financial standing, financial statements are generally issued only once a year. This means that changes occurring after financial disclosure are often neglected in measuring default risk. We present a more accurate default forecast model that updates the average change in firms' financial standing with monthly business cycle information. Financial institutions may use the monthly estimates to not only monitor losses on their loan portfolios more accurately and frequently, but also predict losses under various macroeconomic risks via stress tests.

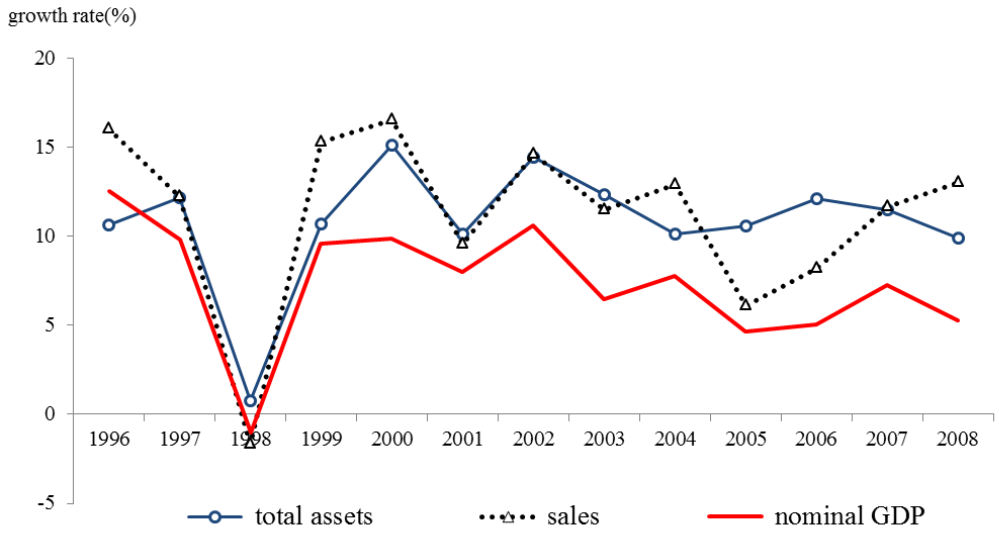
3.1 Introduction

In general, a firm's financial statement is issued every year and the statement is available in audit report if the firm is subject to external audit. But once a statement has been issued, financial changes remain a black hole until the new disclosure. On the other hand, business cycle data holds information about firms' production and is closely related to their average financial standing. For example, drop in production, a symptom of economic downturn, often implies reduced profits and weaker financial standing. Figure 2 below compares the GDP growth rate to the growth rate of total assets and sales of around 56,000 medium sized businesses, which implies that firms average financial standing is closely related to business

[7] Actually our model is a default forecast model.

cycle.

[Figure 2] GDP growth, total asset growth, and sales growth



: Growth rate of total assets(sales) is computed for firms with at least 2 years of consecutive financial disclosure, by weighting individual firm's growth rate of total assets(sales) with the ratio of its total assets(sales) to the sum of total assets(sales) of all firms. We use 184,982 financial statements from 55,856 medium sized businesses.

If we could observe a firm's financial standing at every moment, business cycle data may provide little additional information on its default possibility. As we have noted, however, the observation term of financial statements is a year, while business cycle data is available on a monthly or quarterly basis. We may thus use business cycle data after the last financial disclosure as a proxy variable for financial changes of the firm during the same period.

Controlling credit risk is at the core of risk management for financial institutions such as banks that specialize in loans. Therefore it is essential to monitor borrowers' default risk and there is a need to reduce the observation term to prepare for unexpected losses. Credit assessment based on financial statements implies that changes in credit risk are not observed for

a year after each evaluation.[8] It is common to use the most recent and available financial statement measure credit risk, as do Shumway(2001), Chava and Jarrow(2004). But such practice leads to omitted variable bias when the unobserved data is correlated to past data. Estimates from only observed data are less reliable when there are enormous changes in the unobserved data –after an economic shock such as a financial crisis, for example, last year's financial statements are not very reliable. As a result, business cycle data can be useful in estimating average change in financial standing.

Shumway(2001) and Chava and Jarrow(2004) note that for listed companies, share prices are also relevant to default. If business cycle data is duly reflected in share price, it may be more accurate to use the latter information as stock prices are updated instantaneously. Agarwal et al.(2007) also claim that for a business of very small size, credit information of the company's representative is important in predicting bankruptcy. Information concentration institutions such as the Korea Federation of Banks provide credit information of representatives, including financial transaction data, to financial institutions on an almost real-time basis. Therefore, this paper focuses on medium sized firms[9], for which there is little information on individual financial standing except financial statements.

We adopt the hazard model as our bankruptcy forecast model. This entails a few advantages. Shumway(2001) shows that single-period models are subject to selection bias and that using the hazard model removes the bias. In addition, a discrete-time hazard model can easily incorporate time-varying variables so that we can identify the changes in financial standing as a firm falls into bankruptcy. In this case, the likelihood function is identical to those

[8] Banks with their own monitoring systems may reevaluate firms that show signs of insolvency such as short term overdue loans. But such reevaluation is usually performed qualitatively by credit analysts, and evaluation based on financial statements will only result in identical results unless a new statement is obtained.

[9] Defined as firms with sales below 60 billion KRW and total assets not less than 2 billion KRW. We further elaborate in section 3.1.

of regression models with dichotomous dependent variables, such as the multiperiod logit model, so that we can estimate the model with ease (Cox, 1970 and Brown, 1975 and Allison, 1982). Chava and Jarrow(2004) empirically confirm these advantages of Shumway(2001)'s model by extending the bankruptcy data. Unlike other studies, they construct monthly instead of yearly data to show that the hazard model with industry effects provides better predictions. This paper also constructs monthly data based on the hazard model of Shumway(2001). Whereas Chava and Jarrow(2004) only take into account the macroeconomic information of the corresponding month, we accumulate all business cycle information from the latest financial disclosure to the time of prediction to update firms' financial changes.

The chapter proceeds as follows. Section 2 shows that using outdated data in a hazard model lead to inconsistent estimators with a simple example. Section 3 provides a method that financial institutions which specialize in loans are able to shorten the measurement term of credit risks (expected losses) by using monthly business cycle data. Section 4 briefly introduces the data we use, and in Section 5 we show that business cycle data after financial disclosure is very significant in predicting default and expected loss, proving empirically the importance of timely usage of data. Section 6 conducts robustness checks under various environments, such as estimation with validation set and estimation with financial variables used in previous studies(Altman(1968), Zmijewski(1984)) augmented by information update. We conclude that business cycle data after disclosure remains very significant. Section 7 explores how our forecast model may be utilized by financial institutions to manage credit risk. Section 8 concludes.

3.2 Hazard model and outdated data

Using the hazard model to forecast firm's bankruptcy has a few advantages. First, single-period model, used by Altman(1968) and Zmijewski(1984), only consider a single financial statement at a certain point in time. That is, firms' survival time and changes over time in financial standing are neglected. On the other hand, the hazard model as used in Shumway(2001) directly considers the survival period of firms and can easily incorporate time-varying variables. It is thus possible to obtain a consistent estimator without selection bias. Even with the hazard model, however, estimation can be inconsistent if the independent variable is not observed all the time and we use the most recent observation instead. We explain this with a simple modification of the example in Shumway(2001), used to show the selection bias of single-period models.

Let there be two periods (t=1, 2) and assume that firms can go bankrupt at any period. y_{it} equals 1 if firm *i* goes bankrupt at time t, and 0 if otherwise. x_{it} also takes the values 0 or 1 and has the following relationship to firm's default probability.

$$\text{Prob}(y_{it} = 1|x_{it}) = \theta x_{it} \tag{16}$$

We start with N firms and both x_{it} and y_{it} are observable. At t=2, only those firms that have not gone bankrupt at t=1 are observable. Observations of each firm are independently and identically distributed. The log likelihood function for estimating θ with the given data is as follows.

$$L = \ln\left\{\prod_{i=1}^{N}(\theta x_{i1})^{y_{i1}}[(1-\theta x_{i1})(\theta x_{i2})^{y_{i2}}(1-\theta x_{i2})^{1-y_{i2}}]^{1-y_{i1}}\right\} \tag{17}$$

Differentiating with respect to θ to obtain the first order condition, and using the fact that x_{it} and y_{it} are either 0 or 1, the estimator for θ can be derived as below (Shumway, 2001, p.107).

$$\hat{\theta} = \frac{\sum_{i=1}^{N}(y_{i1} + y_{i2})}{\sum_{i=1}^{N}(x_{i1} + x_{i2})} \tag{18}$$

This represents the ratio of bankrupt firms among those that were exposed to bankruptcy risk($x_{it} = 1$)[10] and can be understood intuitively as bankruptcy probability. $\hat{\theta}$ is an unbiased estimator of θ because $\mathrm{E}(y_{it}|x_{it}) = \theta x_{it}$, and the law of large numbers it is also consistent. But if at t=2 x_{i2} is unobservable so that we use x_{i1} in its place, $\hat{\theta}$ no longer enjoys the above properties. Replacing x_{i2} with x_{i1} and taking expectation,

$$E(\hat{\theta}|x_{i1}) = E\left(\frac{\sum_{i=1}^{N}(y_{i1} + y_{i2})}{\sum_{i=1}^{N}(x_{i1} + x_{i1})}|x_{i1}\right) = \frac{\theta}{2}\left(1 + \frac{\sum_{i=1}^{N}E(x_{i2}|x_{i1})}{\sum_{i=1}^{N}x_{i1}}\right) \tag{19}$$

The estimate is biased unless $\sum_{i=1}^{N}x_{i1}$ and $\sum_{i=1}^{N}E(x_{i2}|x_{i1})$ have the same values. In other words, we do not have an unbiased estimator when under the given information set, x_{i1} is not the optimal predictor of x_{i2} ($x_{i2}{}^{OP} = \mathrm{E}(x_{i2}|x_{i1}) \neq x_{i1}$). If we have additional information besides x_{i1}, however, so that we can obtain the optimal predictor $x_{i2}{}^{OP}$ under the given information set, we can get an unbiased estimator of θ without having to observe x_{i2}.

[10] Firms that are exposed to risk in both periods are counted twice.

3.3 Estimating expected portfolio loss using monthly business cycle data

This section briefly describes how a financial institution calculates the expected loss (EL) of its loan portfolio. We then show that using monthly business cycle data to update a firm's financial changes after financial disclosure can lead to more accurate and frequent estimates.

The expected loss of firm *i* at time t can be written as

$$EL_{it} = PD_{it} \times LGD_{it} \times EAD_{it} \tag{20}$$

PD_{it} is the probability of default for firm *i* at time t, and LGD_{it} and EAD_{it} represent the loss at default and exposure at default, respectively. The expected loss from firm *i* equals the probability that the firm will not pay back, multiplied by the amount it borrowed but is not covered by collateral. Basel II requires a financial institution to hold at least as much capital as the difference between the expected loss and the unexpected loss[11] as a buffer. As a result, accurate and more frequent measurement of credit risk enables financial institutions to prepare more adroit response to shocks.

Probability of default is a key to computing expected loss. Probability of default can be estimated by a bankruptcy forecast model, which in turn depends on financial statements issued every year. This is where we incorporate monthly business cycle data.

Assume that the credit risk of firm *i* at time t is determined by a K-dimensional financial ratio vector $f_{it} = (f_{it}^1, f_{it}^2, \cdots, f_{it}^K)'$. Weighting by $\beta = (\beta_1, \beta_2, \cdots, \beta_K)'$, we take the sum $f_{it}'\beta (\equiv x_{it})$ to be the financial standing of firm *i* at time t. Letting $\Lambda(\ \ )$ denote the function that maps financial standing to default probability, the expected default rate equals $\Lambda\,(f_{it}'\beta)$,

[11] Unexpected loss is calculated from hyperthetically extreme level of default probability (at the top 0.001 percentile).

where $i = 1, 2, \ldots, \mathrm{I}$. We assume without loss of generality that LGD(Loss Given Default) and EAD(Exposure at Default) are equal to 1. The expected loss from I firms can then be expressed as

$$\sum_{i=1}^{I} EL_i = \sum_{i=1}^{I} E[D_{it} = 1 \mid f_{it}] = \sum_{i=1}^{I} \Pr(D_{it} = 1 \mid f_{it}) \equiv \sum_{i=1}^{I} \Lambda\,(x_{it}) \tag{21}$$

In this case, the sum of probability of default of all firms is equal to expected loss.

Assume further that a firm's financial standing (x_{it}) is determined by systematic risk and firm-specific risk. Systematic risk can be understood as business cycle risk that all firms face. Letting z_t be the economic environment at time t, its change from time t to t+1 equals $\Delta z_{t+1}(=z_{t+1} - z_t)$. Assume that all firms are equally sensitive to the business cycle by a factor of γ, the financial standing of firm *i* at time t+1 is

$$x_{it+1} = x_{it} + \gamma \Delta z_{t+1} + \varepsilon_{it+1} \tag{22}$$

where $i = 1,\ 2,\ \ldots,\ I$ and $\varepsilon_{it} \sim Normal(0, \sigma_{it}^2)$.

If the observation term of financial statements is an integer p and the last observation occurred at period t, financial standing after $\tau(< p)$ periods $x_{it+\tau}$ is[12]

$$x_{it+\tau} = x_{it} + \gamma(z_{t+\tau} - z_t) + \textstyle\sum_{j=1}^{\tau} \varepsilon_{it+j} \tag{23}$$

---

[12] $x_{it+1} = x_{it} + \gamma \Delta z_{t+1} + \varepsilon_{it+1}$
$x_{it+2} = x_{it+1} + \gamma \Delta z_{t+2} + \varepsilon_{it+2}$
$\vdots \qquad\qquad \vdots$
$x_{it+\tau} = x_{it+\tau-1} + \gamma \Delta z_{t+\tau} + \varepsilon_{it+\tau}$
Taking summation of both sides gives the equation.

Let us estimate the model by replacing the unobserved $x_{it+\tau}$ with x_{it} and $(z_{t+\tau} - z_t)$, and denote the estimators of β and γ by $\hat{\beta}$ and $\hat{\gamma}$, respectively. If the model is a linear probability model, as we have seen in Chapter 2, then $\hat{\beta}$ and $\hat{\gamma}$ are consistent estimators of β and γ, respectively, and using only the past financial statement (x_{it}) causes omitted variable bias if this data is correlated with business cycle data $(z_{t+\tau} - z_t)$.

Using the estimated model, the expected loss of firm *i* at time t+τ is written as

$$\Lambda\,(\,\hat{x}_{it+\tau}) = \Lambda\left(f'_{it}\hat{\beta} + \hat{\gamma}(z_{t+\tau} - z_t)\right) \tag{24}$$

Since the default rate function $\Lambda\,(\quad)$ may not be linear, as is the case with logit functions, approximating to a linear function at $\hat{x}_{it+\tau}$ gives

$$\tilde{\Lambda}\,(x) = \Lambda\,(\,\hat{x}_{it+\tau}) + \Lambda'(\,\hat{x}_{it+\tau})(x - \,\hat{x}_{it+\tau}) \tag{25}$$

Where $\tilde{\Lambda}\,(x)$ is the linear approximation of $\Lambda\,(x)$ at $\hat{x}_{it+\tau}$. Taking linear approximation of the default rate function at the unobserved true value $x_{it+\tau}$, and using the relationship between $x_{it+\tau}$ and the observed values, we obtain

$$\tilde{\Lambda}\,(f'_{it+\tau}\beta) = \Lambda\,(\,\hat{x}_{it+\tau}) + \Lambda'(\,\hat{x}_{it+\tau})(f'_{it+\tau}\beta - \,\hat{x}_{it+\tau}) \tag{26}$$

$$= \Lambda\,(\,\hat{x}_{it+\tau}) + \Lambda'(\,\hat{x}_{it+\tau})\left\{f'_{it}\beta + \gamma(z_{t+\tau} - z_t) + \sum_{j=1}^{\tau}\varepsilon_{it+j} - f'_{it}\hat{\beta} - \hat{\gamma}(z_{t+\tau} - z_t)\right\} \tag{27}$$

$$= \Lambda\,(\,\hat{x}_{it+\tau}) + \Lambda'(\,\hat{x}_{it+\tau})\left\{f'_{\ it}\left(\beta - \hat{\beta}\right) + (z_{t+\tau} - z_t)(\gamma - \hat{\gamma}) + \sum_{j=1}^{\tau}\varepsilon'_{it+j}\right\} \tag{28}$$

If $\hat{\beta}$ and $\hat{\gamma}$ are consistent estimators of β and γ, respectively, and if the number of firms

I is large enough, by the law of large numbers the expected loss of the actual portfolio and the estimated expected loss are approximately equal.

$$\sum_{i=1}^{I} \tilde{\Lambda}(f'_{it+\tau}\beta) \quad \simeq \quad \sum_{i=1}^{I} \Lambda\left(f'_{it}\hat{\beta} + \hat{\gamma}(z_{t+\tau} - z_t)\right) \tag{29}$$

In other words, even if financial standing is not observed every month, we can use business cycle data that is updated monthly to estimate expected loss on a monthly basis. This method allows for a more timely response to credit risk; using outdated data may lead to underestimation of expected loss in the case of economic shocks, something a sound risk management must avoid.

3.4 Data construction and basic description

We estimate the bankruptcy forecast model by using financial statements and bad credit information of medium sized firms in Korea from 1996 to 2009, obtained from NICE Information Service, the largest credit information company in Korea.[13] Since financial standings of large and small firms can often be inferred by share prices or credit scores of company representatives,[14] we limit our analysis to medium sized firms[15] with annual sales below 60 billion KRW and total assets not less than 2 billion KRW. To estimate the hazarad

[13] Data is only used for research purposes, and firm ID numbers have been generated via an encryption process.

[14] With adequate assumptions, we may add changes after settling month in stock price or credit scores of representatives as risk factors.

[15] Under the credit risk measurement standard for small and medium sized businesses of the new BIS treaty(Basel II), the Financial Supervisory Service in Korea classifies as medium sized firms those with annual sales under 60 billion KRW.

model, we flow sample firms that have been established after January 1st, 1996.

To control for differences in firm size, we group firms that hold less than 7 billion in total assets as "Medium sized firms 1," and those with more than 7 billion as "Medium sized firms 2."[16] We also classify the firms into light industry, heavy industry, construction, wholesale and resale, and services, according to the standard industrial classification; we remove financial statements of some business types, including financial institutions and nonprofit organizations.

We construct the data based on monthly observed data as done by Chava and Jarrow(2004). Of course, financial standing remains constant for a year. For monthly business cycle data we use the cyclical component of coincident index[17] to consider only the cyclical factors. We also consider availability of data at the corresponding point in time, however, and construct the data to imitate real-time analysis. Because cyclical component of coincident index of month n is released on the last day of month n+1, we use indices of up to month n to forecast defaults in month n+2. We differ from the existing literature in that we accumulate all the business cycle data after the last announcement of financial statement, instead of using only the data of the current month.

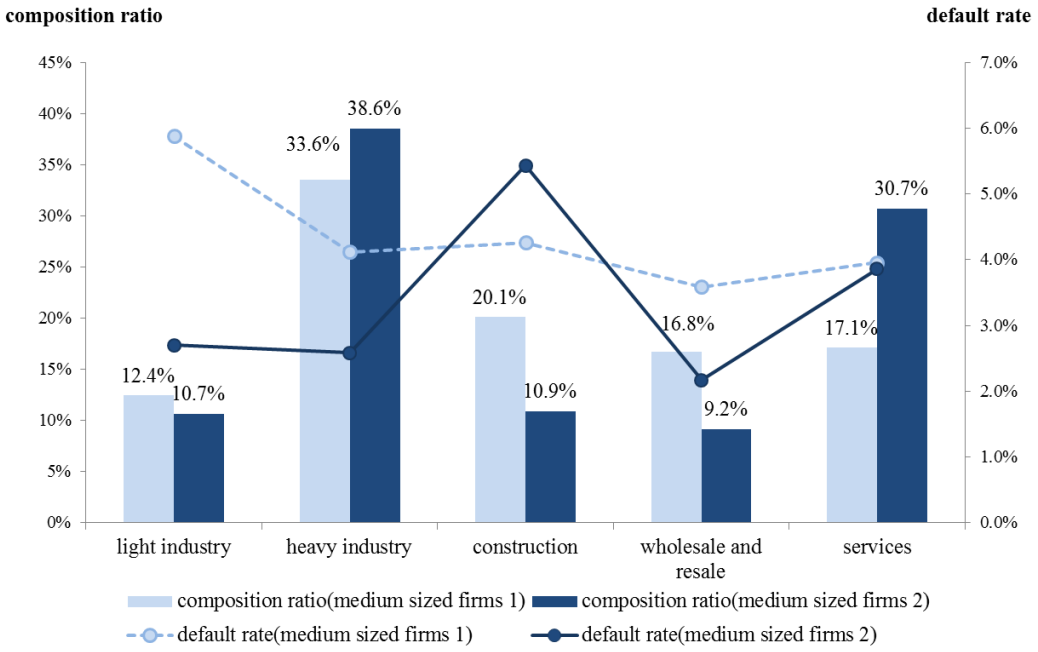
To check for robustness, we divide the sample into two sets of almost equal size, training set and validation set. We use stratified sampling to ensure that the default rate and the ratio of each business type are the same in the two samples. Yearly summaries of the sample are shown in Appendix A.4 and the composition of the sample according to business type, as well as default rates, are shown in Figure 3. "Medium sized firms 1" have a relatively large ratio of

[16] The critierion is based on the fact that for the most of the period in our sample, a firm is subject to external audit if its total assets exceed 7 billion. A firm is classified as "Medium sized firms 2" if its total assets exceeds 7 billion at least once during the whole period.

[17] Cyclical component of coincident index is the coincident index that has been detrended, and represents business cycle movement not affected by the trend of economic growth. It is announced by the National Statistical Office every month.

construction and wholesale and resales and show the highest default rate in light industries. “Medium sized firms 2” have a high ratio of heavy industries and services and show the highest default rate in construction.

[Figure 3] Composition and default rate according to business type



Bankruptcy is defined as the bad credit information corresponding to default.[18] Firms that default on principal or interest, that are registered on the credit management information system according to the credit information management rule of the Korea Federation of Banks, and firms that are undergoing bankruptcy and regenerative process are included.

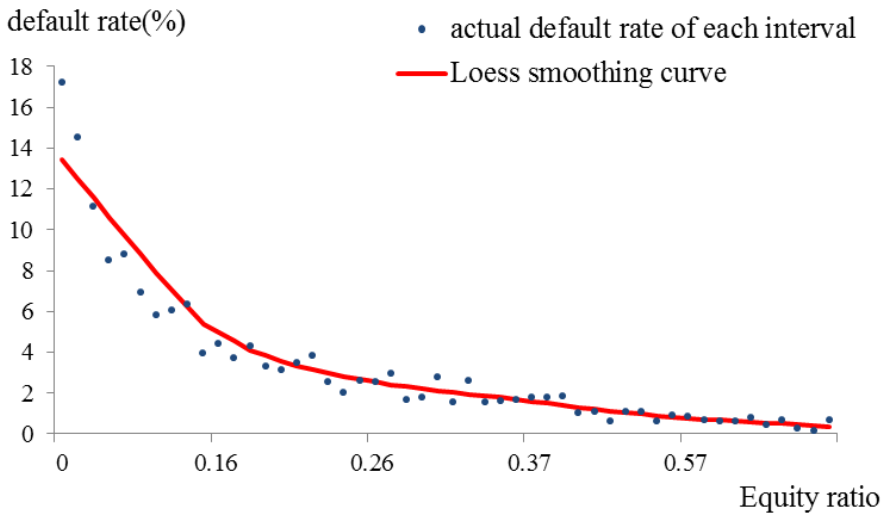
Before we estimate the model, we investigate the relationship between individual financial ratios and default rate. We use the financial ratios used by Altman(1968) and Zmijewski(1984),

[18] The definition of default often differs from one study to another. Banks use different definitions in practice, and ours makes use of the common factors among the definitions of different banks.

as well as cash and cash equivalent to total assets (CA/TA) and accounts receivables to sales (AR/SA). The list of ratios is in Appendix A.5. CA/TA represents firms' liquidity risk and AR/SA is a measure of default risk due to credit transactions between firms. As in previous studies (Shumway, 2001 and Chava and Jarrow, 2004), we replace those values whose percentiles are below 1 or above 99 with the value of 1st and 99th percentiles, respectively, so that the analysis is not unduly affected by outliers.

Figure 4 shows the empirical relationship between equity ratio and actual default rate. Equity ratio is ordered from low to high and is divided into 50 buckets, each representing 2 percentiles. The horizontal axis represents the average equity ratio in each bucket. The vertical axis represents actual default rates at each bucket. We note that the two variables have a strong negative correlation.

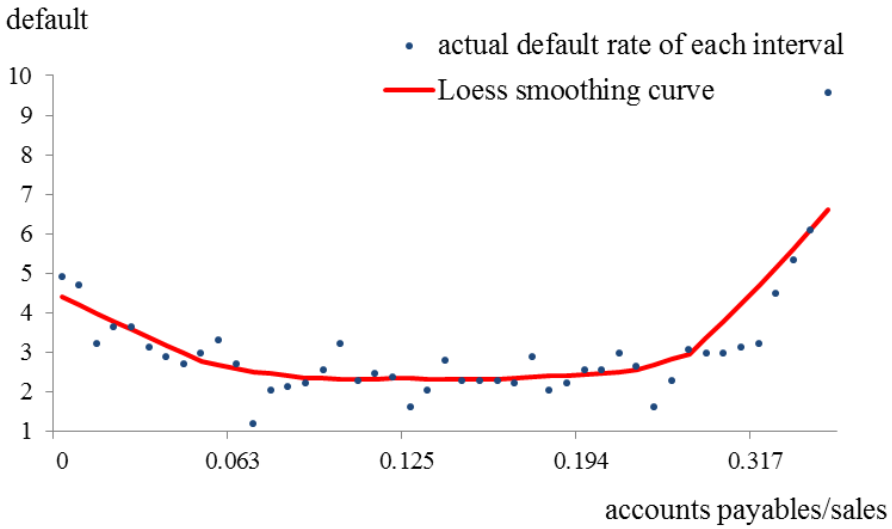
[Figure 4] The relationship between equity ratio and actual default rate



: Equity ratio is ordered from low to high and is divided into 50 buckets, each representing 2 percentiles. The horizontal axis represents the average equity ratio in each bucket. The vertical axis represents actual default rates at each bucket. The dots are actual default rates while the solid line is the Loess smoothing curve (bandwidth=0.4).

Figure 5 plots AR/SA in the same way. Accounts receivables are retrievable assets, so to a certain level increase in AR/SA reduces default rate. But because there is risk that other parties may default, after the critical level AR/SA can act as a risk factor.

[Figure 5] The relationship between AR/SA and actual default rate



: AR/SA is ordered from low to high and is divided into 50 buckets, each representing 2 percentiles. The horizontal axis represents the average AR/SA in each bucket. The vertical axis represents actual default rates at each bucket. The dots are actual default rates while the solid line is the Loess smoothing curve (bandwidth=0.4).

AR/SA differs from other financial ratios in that a firm may go bankrupt regardless of its business results due to risk transferred from its transacting parties. According to recursive moral hazard model of Kim and Shin(2012), delayed payments which can be considered implicit inter-firm credit relationships has a function of mitigating incentive problems in production chains by building up interlocking claims and obligations. In fact, many firms hold large amounts of both accounts receivables and accounts payables. AR/SA represents risk from other parties that is not captured by other ratios related to firm's profitability or stability. We define as excess accounts

receivables ratio the amount of AR/SA exceeding 20%[19], and add this as an independent variable.

3.5 Estimation results

Estimation results of the bankruptcy forecast model show that compared with using only financial statements or using business cycle data without any adjustment, accumulating business cycle data after the settling month leads to increased goodness-of-fit and accuracy ratio.

Table 4 shows the estimation results with "Medium sized firms 1". Among dozens of financial ratios, we finally use "net income to total assets(NI/TA)", "total debt to total assets(TD/TA)", "cash and cash equivalent to total assets(CA/TA)", and "excess accounts receivables ratio($|AR/SA-0.2|^{+}$)", which best represent the characteristics of medium sized firms in Korea.[20] Since we wish to determine the significance of business cycle data after the settling month in predicting defaults, we take as benchmarks estimation only with financial ratios (Model 1) and estimation with cyclical component of coincident index (referred to from now on as cyclical component) as well as financial ratios (Model 2). Model 3 uses financial ratios along with the change in cyclical component since the settling month, which we will refer to as "accumulated cyclical component".

All three models show high accuracy ratios and goodness-of-fits. However, Model 2 is better than Model 1 in both aspects. Model 3 shows the highest accuracy ratio and goodness-of-

[19] Default rate starts to increase as AR/SA rises above 20%.

[20] We first used t-test and KS test to select out of around 100 financial ratios dozens of ratios that were closely related to default rate. Then, in view of statistical significance and the meaning of the variables, we chose the above 4 ratios that represent profitability, stability, and liquidity.

fit among three models. The estimated coefficient of cyclical component is -0.0337 in Model 2, while the estimated coefficient of accumulated cyclical component is -0.1303 in Model 3. Also, the respective t-values are 2.7 and 8.9, which means that the accumulated cyclical component is statistically more significant than cyclical component. This implies that business cycle data from settling month to evaluation month includes relevant information to forecast defaults of firms. On the other hand, current business cycle data shows relatively less significant level since the information betwwen the settling month and evaluation mont has been omitted.

[Table 4] Estimation of the hazard model (Medium sized firms 1)

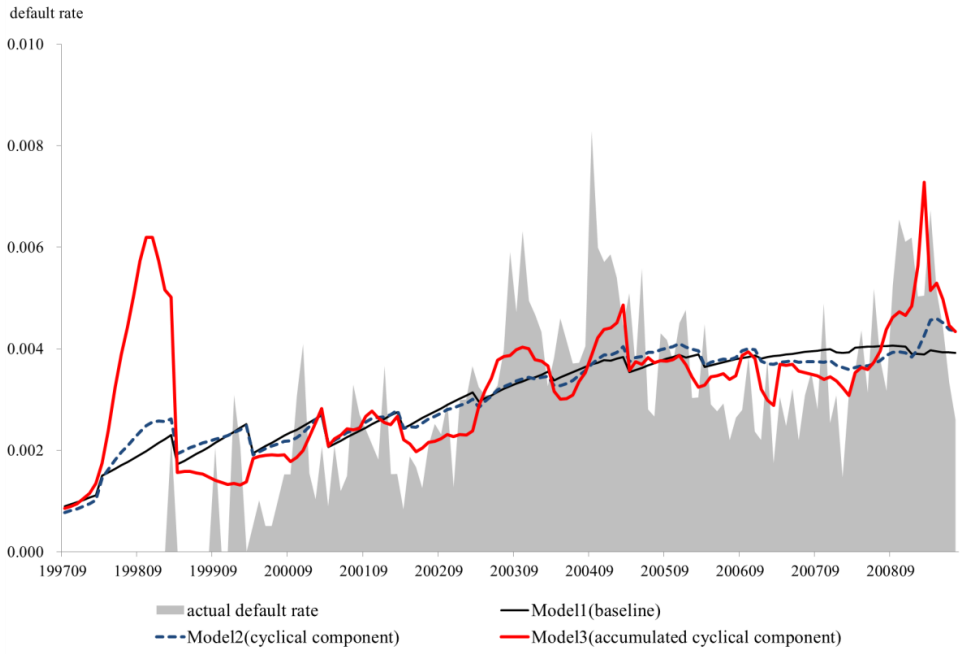
	Model1		Model2		Model3	
	(baseline)		(cyclical component)		(accumulated cyclical component)	
Variable	Coefficient	t	Coefficient	t	Coefficient	t
const.	-8.2136 ***	54.7	-8.2219 ***	54.7	-8.2824 ***	54.9
dummy(construction)	0.5006 ***	8.6	0.4997 ***	8.6	0.4970 ***	8.5
NI/TA	-0.0152 ***	10.5	-0.0152 ***	10.5	-0.0153 ***	10.6
TD/TA	0.0107 ***	11.2	0.0107 ***	11.2	0.0104 ***	11.0
CA/TA	-0.0733 ***	12.2	-0.0732 ***	12.2	-0.0728 ***	12.1
|AR/SA-0.2|+	0.0126 ***	9.2	0.0126 ***	9.2	0.0126 ***	9.2
cyclical component(CC)			-0.0337 ***	2.7		
accumulated CC(after settling month)					-0.1303 ***	8.9
Age	0.0563 ***	14.4	0.0566 ***	14.5	0.0581 ***	14.9
Age2	-0.00032 ***	11.6	-0.00033 ***	11.7	-0.00034 ***	12.2
Model Fit	1390.5 ***		1397.2 ***		1482.7 ***	
#firms	9,751		9,751		9,751	
#firm-month obs.	622,821		622,821		622,821	
#defaults	2,163		2,163		2,163	
-2LogL	27,425		27,418		27,333	
AUROC	0.714		0.716		0.720	

: Model 1 does not use business cycle information, while Model 2 uses the data without adjustment. Model 3 accumulates business cycle data from the last settling month. By comparing Model 3 to Model 1 and 2, we can discern the effect of updating business cycle data information. Cluster-robust standard errors are used to inference. *, **, *** represent significance at levels 10%, 5%, 1%.

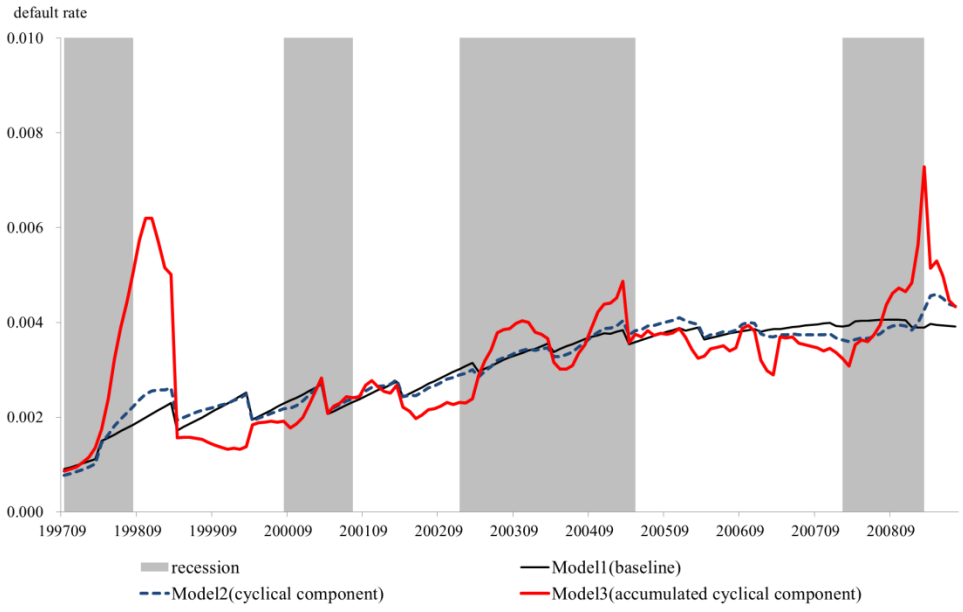
"net income to total assets(NI/TA)" and "cash and cash equivalent to total assets(CA/TA)" are typical indicators of firm profitability and liquidity, so the coefficients are negative as one would expect. On the other hand, the coefficients of "total debt to total assets(TD/TA)" and "excess accounts receivables to sales($|AR/SA-0.2|^{+}$)" are positive. TD/TA is the ratio of debt to total assets and equals equity ratio subtracted from 1. Higher debt ratio means lower financial stability and thus higher default rate. AR/SA is also referred to as accounts receivable collection period because it represents the speed at which receivables occurring at a constant rate are turned into cash. That the estimated coefficient of excess accounts receivables to sales ($|AR/SA-0.2|^{+}$) is positive and very significant implies that AR/SA levels over 0.2 increase default rates. Unlike other financial ratios, account receivables can be a medium through which one firm's default is transmitted to other transacting firms, causing a contagion throughout the system. If we have access to firms' transaction data, we may analyze how an industry is impacted by bankruptcy of a single firm via account receivables.

We now look at the difference between actual and estimated default rates. Define the actual default rate as the percentage among firms surviving at the end of each month that go bankrupt in less than a month. Let average expected default rate be the average of expected default rates of all individual firms. Fixing LGD (loss given default) and EAD (exposure at default) at 1, the expected loss is equal to the sum of expected default rates. Hence expected loss at time t is computed by (average expected default rate at time t) × (number of firms at time t). Actual default rates and average expected default rates are plotted for "Medium sized firms 1" in Figure 6 below. Since estimation results of "Medium sized firms 2" are quite similar to that of "Medium sized firms 1", we don't present here.

[Figure 6] Actual and average expected default rates (Medium sized firms 1)



[Figure 7] Recessions and average expected default rate (Medium sized firms 1)



Default rates are very high in the late 1990s and 2008 due to Asian financial crisis and the global financial crisis. Model 3 with accumulated cyclical component is the most sensitive to

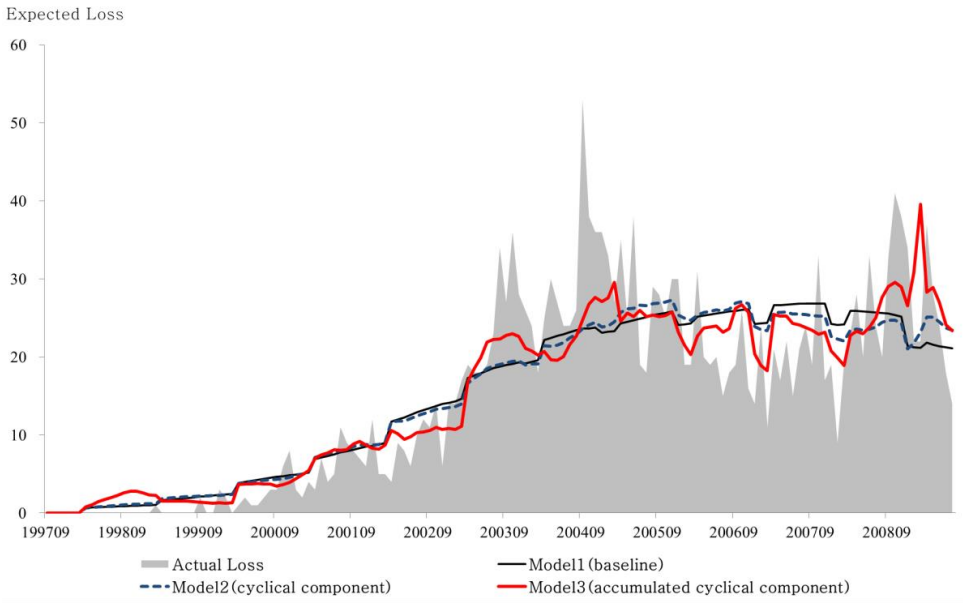
economic volatility during such periods. The disaccord in the late 90s between actual rates, mostly 0, and expected rates, sometimes above 0.006, occurs because few firms before the year of 2000 are in the sample so that many firms that went bankrupt are not included. After 2000, when data is ample, the actual and estimated rates are quite similar.

Figure 7 removes the actual default rates from Figure 6 and instead shades the periods deemed as recession according to the reference date released by the National Statistical Office.

Expected default rates tend to increase during recessions and decrease during booms. Model 3 with accumulated cyclical component responds instantaneously to changes in the business cycle.

Next, we compare expected loss to actual loss in terms of portfolio. As we have noted above, the sum of expected default rates equals expected loss if both LGD and EAD are 1. This is the expected loss of a financial institution that lends an amount of 1 to each firm, with no collateral. Figure 8 shows the expected loss for "Medium sized firms 1".

[Figure 8] Expected and actual loss (Medium sized firms 1)



Expected loss shows a rapid rise after 2000 because in our sample, the number of financial statements increases sharply at that time. A larger number of firms lead to higher expected loss, though of course changes in the business cycle affect expected loss by changing the probability of default. Using RMSE(Root-Mean-Square Error) to compare the actual loss and expected loss, we see that expected loss becomes more accurate as we move from Model 1 to Model 3 as in Table 5.

[Table 5] Root-Mean-Square Error(RMSE) of actual and expected loss

	Model1 (baseline)	Model2 (cyclical component)	Model3 (accumulated cyclical component)
Medium sized firms 1	6.43	6.28	5.68
Medium sized firms 2	3.44	3.43	2.97

Cyclical component of coincident index may be revised after release. This paper aims to construct the data to make the anlaysis as close to real-time as possible, but non-revised data is available only after January, 2007, and the changes in data due to revision are usually negligible, so that we use revised data. But to discern the effect of revision, we addionally calculated RMSPE(Root-Mean-Square Prediction Error) using real-time data(available at that time, not revised) in validation set for the period of 2008. 1 ~ 2009. 7. To accumulate information after the settling month, estimation is feasible after January, 2008. The result, presented below, shows that Model 3 exhibits the best predictive power regardless of revision.

[Table 6] Root-Mean-Square Prediction Error(RMSPE) of actual and expected loss

	Model1 (baseline)	Model2 (cyclical component)	Model3 (accumulated cyclical component)
Medium sized firms 1	8.92	8.59	8.07
Medium sized firms 2	5.70	5.46	5.04

3.6 Robustness checks

We now undergo robustness checks to determine if the estimation results of section 5 remain consistent when different data or financial ratios are used. We consider 4 types of robustness checks. First, we check if cyclical component of coincident index can be replaced by industrial production index or market interest rates in updating business cycle data. Second, we test if the bankruptcy forecast model with business cycle data is still valid in out-of-sample. We apply the estimated values from the training set to the validation set and test the predictive power on loss with out-of-sample data. We use RMSPE(Root-Mean-Square Prediction Error) to compare the actual and expected loss for each model. Third, to rule out arbitrary selection of financial ratios, we use the ratios used in previous studies and reestimate the model. We use the ratios used in Altman(1968) and Zmijewski(1984). Finally, to check that accumulated cyclical component is more useful than business cycle information for an arbitrary period, we consider models that use as independent variable business cycle data of 3, 6, and 12 months leading to the time of prediction. These four analyses show that business cycle data after settling month is consistently significant in predicting defaults, regardless of which data or financial ratios are used.

3.6.1 Estimation with industrial production index and market interest rate

We replace cyclical component of coincident index with industrial production index and use the same updating method as before. Industrial production index is detrended by Hodrick-Prescott filter so that we only use the cyclical component of the index. The result is shown below.

[Table 7] Estimation with industrial production index

	Medium sized firms 1			Medium sized firms 2		
Variable	Coefficient		t	Coefficient		t
const.	-8.2229	***	54.9	-8.7800	***	39.1
dummy(construction)	0.4967	***	8.5	0.6831	***	6.7
NI/TA	-0.0153	***	10.6	-0.0201	***	10.2
TD/TA	0.0105	***	11.1	0.0194	***	13.7
CA/TA	-0.0730	***	12.1	-0.0689	***	6.0
|AR/SA-0.2|+	0.0126	***	9.2	0.0115	***	5.4
accumulated industrial production index	-0.0468	***	10.8	-0.0373	***	5.1
Age	0.0567	***	14.6	0.0332	***	6.3
Age2	-0.0003	***	11.8	-0.00015	***	4.5
Model Fit	1484.6	***		1047.2	***	
#firms	9,751			4,435		
#firm-month obs.	622,821			288,354		
#defaults	2,163			837		
-2LogL	27,331			10,404		
AUROC	0.72			0.796		

: The Hodrick-Prescott filter is used for detrending Industrial production index. Accumulated industrial production index is the change in cyclical component of the Industrial production index since the settling month. Cluster-robust standard errors are used to inference. *, **, *** represent significance at levels 10%, 5%, 1%.

Industrial product index leads to better estimation as does cyclical component of coincident index. However, industrial product index has a high composition of mining and manufacturing industries so that the significance of coefficient estimates falls below that of the coincident index.

We also estimate the model with interest rate data. Interest rate provides information about the current economic environment, as well as debt repayment of firms. We must be careful of reverse causality, however. According to Kim(2008), if the central bank preemptively changes the base rate in anticipation of future economic changes, production is not only affected by but also affects interest rates through forward-looking monetary policy of the central bank. This paper removes the impact of changes in the base rate on the market interest rate by extracting the market's information of credit risk from the variable (market rate – base rate). As we have done with coincident index, we take the accumulated change in (market rate – base rate) since the settling month.

The first column in Table 8 uses the market interest rate per se. The interest rate on 3-year corporate bonds rated BBB- is used as the market interest rate. We note that counterintuitively, the coefficient of interest rate is negative. This may be caused by the reverse causality we mention above. When we use (market rate – base rate), however, the coefficient is positive and statistically significant, as shown in the second column.

[Table 8] Estimation with interest rate

	(1)			(2)		
Variable	Coefficient		t	Coefficient		t
const.	-7.6664	***	35.0	-7.9764	***	51.9
dummy(construction)	0.5109	***	8.6	0.5129	***	8.7
NI/TA	-0.0155	***	10.6	-0.0158	***	10.8
TD/TA	0.0107	***	11.0	0.0104	***	10.8
CA/TA	-0.0734	***	12.1	-0.0731	***	12.1
|AR/SA-0.2|+	0.0127	***	9.2	0.0126	***	9.2
Δ_τ market rate	-0.0508	**	2.1			
Δ_τ (market rate - base rate)				0.0909	***	5.3
Age	0.0492	***	12.3	0.0518	***	13.1
Age2	-0.00028	***	10.0	-0.00031	***	11.0
Model Fit	1304.5	***		1336.5	***	
#firms	9,654			9,654		
#firm-month obs.	581,912			581,912		
#defaults	2,104			2,104		
-2LogL	26,555			26,523		
AUROC	0.712			0.717		

: Corporate bond rates are available after October, 2000 in Economic Statistics System operated by the Bank of Korea. Previous data is not included in estimation. Cluster-robust standard errors are used to inference. *, **, *** represent significance at levels 1%, 5%, 10%.

3.6.2 Out-of-sample test

We use validation data to measure via RMSPE(Root-Mean-Square Prediction Error) the difference between actual and expected losses for Model 1(baseline), Model 2(cyclical component), Model 3-1(accumulated cyclical component), Model 3-2(industrial product index), and Model 3-3(interest rate). As shown in Table 9, Model 3-1 exhibits lower RMSPE than Model 1 or Model 2 for both groups of firms. This implies that accumulated cyclical component holds important information for estimating defaults even outside the sample. Model

3-2 which updates information through the Industrial Production Index falls short of Model 1 and Model 2, while Model 3-3 which updates information through interest rates outperforms both.

[Table 9] Root-Mean-Square Prediction Error(RMSPE)

	Model1	Model2	Model3-1	Model3-2	Model3-3
Information source	-	-	Coincident index	Industrial production index	Coporate bond yield
Medium sized firms 1	7.79	7.73	6.80	8.66	7.22
Medium sized firms 2	3.48	3.44	3.14	3.53	3.20

: RMSPE(Root-Mean-Square Prediction Error) between actual and expected losses during the period 1997.6 ~ 2009.7 is employed to evaluate different models.

3.6.3 Fiancial ratios in Altman(1968) and Zmijewski(1984)

We estimate our model by using the financial ratios of the models in Altman(1968) and Zmijewski(1984), which laid the foundations of bankruptcy forecast models and are still widely used. As shown below, the results remain consistent; Model 3 with accumulated cyclical component shows the highest fit and discriminative power. Detailed results are given in Appendix A6.

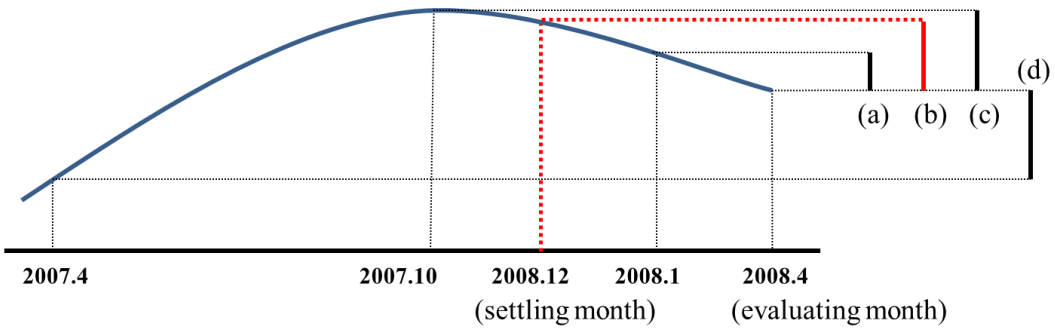
[Table 10] Estimation with financial ratios of Altman(1968) and Zmijewski(1984)

		Model Fit	AUROC
Altman(1968)	Model1(A)	811	0.657
	Model2(B)	819	0.659
	Model3(C)	912	0.668
	(B) - (A)	**7.6**	**0.002**
	(C) - (A)	**100.7**	**0.011**
Zmijewski(1984)	Model1(A)	842	0.662
	Model2(B)	849	0.664
	Model3(C)	941	0.673
	(B) - (A)	**6.6**	**0.002**
	(C) - (A)	**98.9**	**0.011**
Kim	Model1(A)	1391	0.714
	Model2(B)	1397	0.716
	Model3(C)	1483	0.720
	(B) - (A)	**6.7**	**0.002**
	(C) - (A)	**92.2**	**0.006**

3.6.4 Accumulating business cycle data for a fixed period(3/6/9 months)

We compare our result of accumulating business cycle data after the latest disclosure with estimations that use business cycle data for a fixed period of 3, 6, and 12 months. Figure 9 shows how to accumulate business cycle data for a given period.

[Figure 9] Accumulating data for a given period



(a) business cycle data for 3 months
(b) business cycle data after settling month
(c) business cycle data for 6 months
(d) business cycle data for 12 months

Estimation results are shown in Table 11. Even though the same data source is used, using data after the settling month has a much better fit and discriminatory power compared to using data for a fixed period of time. The coefficient of accumulated cyclical component (Model3) is also more statistically significant than values in other models.

[Table 11] Accumulating business cycle data for a fixed period (Medium sized firms 1)

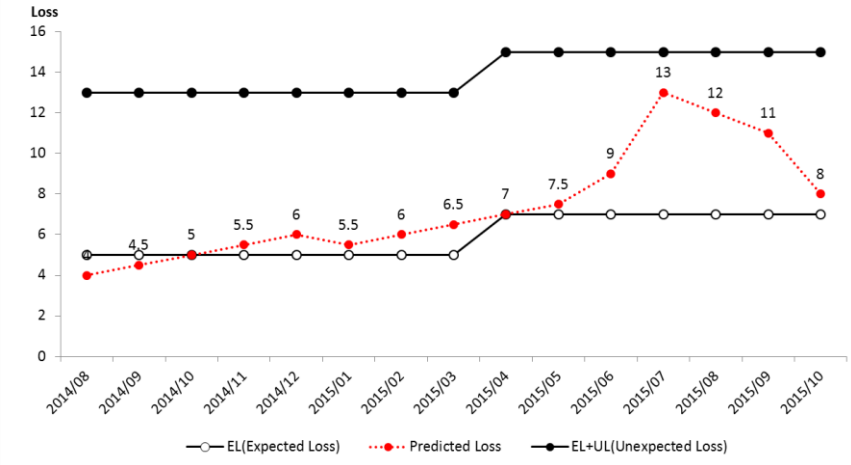
	3 months		6 months		12 months		After settling month	
Variable	Coefficient		Coefficient		Coefficient		Coefficient	
const.	-8.245	***	-8.259	***	-8.245	***	-8.282	***
dummy(construction)	0.502	***	0.503	***	0.502	***	0.497	***
NI/TA	-0.015	***	-0.015	***	-0.015	***	-0.015	***
TD/TA	0.011	***	0.011	***	0.011	***	0.010	***
CA/TA	-0.073	***	-0.073	***	-0.073	***	-0.073	***
|AR/SA-0.2|+	0.013	***	0.013	***	0.013	***	0.013	***
accumulated CC (3 months)	-0.119	***						
accumulated CC (6 months)			-0.084	***				
accumulated CC (12 months)					-0.047	***		
accumulated CC (after settling month)							-0.130	***
Age	0.057	***	0.058	***	0.057	***	0.058	***
Age2	0.000	***	0.000	***	0.000	***	0.000	***
Model Fit	1432.7	***	1438.7	***	1421.0	***	1482.7	***
#firms	9,751		9,751		9,751		9,751	
#firm-month obs.	622,821		622,821		622,821		622,821	
#defaults	2,163		2,163		2,163		2,163	
-2LogL	27,383		27,377		27,395		27,333	
AUROC	0.719		0.721		0.717		0.72	

: The first three columns show results for estimations that use business cycle data for 3, 6, and 12 months. Cluster-robust standard errors are used to inference. *, **, *** represent significance at levels 10%, 5%, 1%.

3.7 Application in financial institutions

Managing credit risk and loss is very important for financial institutions. Especially, banks that have introduced Internal Ratings-Based Approach under Basel II use probability of default computed from their own credit rating systems to calculate expected loss and regulatory capital. Accuracy of probability of default and monitoring of expected loss is therefore essential. The information updating method we present can help to make the predictions more accurate and timely. Furthermore, by setting the business cycle variable at 0, "*business-cycle-neutral*" expected loss is also easily obtained. Based on this value, we may estimate average default rates in the long run and use this as the baseline rate or the rate for calculating regulatory capital. We can also compare the rate to expected loss under actual business cycle conditions, thereby monitoring the adequacy of capital levels every month. Figure 10 shows a simple example which can be applied to monitor the adequacy of capital on a monthly basis.

[Figure 10] Example of monitoring capital adequacy on a monthly basis



: EL(expected loss) and EL+UL(expected and unexpected loss) correspond to business-cycle-neutral expected loss and expected loss under extreme circumstances, respectively. The difference between the two values equals UL(unexpected loss). Predicted loss reflects monthly business cycle data and makes it possible to monitor expected loss, as well as capital adequacy, on a monthly basis.

These practical analyses do not require a complex calculation; all that is necessary is to add as independent variable business cycle data after the latest observation of financial statement. This simplicity is a big advantage for financial institutions as it is costly to monitor numerous small business firm by firm.

3.8 Conclusion

This chapter presents a method to update financial information of firms with monthly business cycle data while financial statements are yet to be updated. Even if financial statements hold sufficient information about default, if business cycle data is observed more frequently than financial statements, the former can provide additional information during the period the latter is not observed. We use data on medium sized firms in Korea to test empirically if using business cycle data from the settling month to the point of evaluation has predictive power in forecasting defaults. The answer is positive; using business cycle data enhances prediction power, and using the data from the latest financial disclosure is shown to be better than using the data for a fixed number of months.

The result is also shown to be robust to out-of-sample test and using financial ratios from previous studies. This method can be useful for financial institutions in monitoring credit risk in an accurate and timely manner.

Finally, we present “excess accounts receivables ratio” as a new, significant risk factor that affects firm bankruptcy. Further study is called for to ascertain how financial risk is transmitted from a defaulting firm through the production chains.

4. Concluding remarks

This paper shows that in a regression model with the observation term of independent variable longer than that of dependent variable, using outdated data can lead to omitted variable bias. We also show that this bias can be removed, and the uncertainty of estimation reduced, if we use an auxiliary variable that is correlated with the independent variable but observed more frequently.

We apply our method to bankruptcy forecast model for firms. Our empirical analysis shows that using monthly business cycle data to augment financial statements leads to better default forecasts. Cyclical component of coincident index, industrial product index, market interest rate, or different indices that fit the characteristics of the firms under consideration may be used as the source of information update for business cycle. Financial institutions can use this information updating mechanism to more precisely monitor expected losses from loan portfolios, as well as estimating losses under various economic environments.

References

[1] Agarwal, S., Chomsisengphet, S., and C. Liu, 2007, Determinants of small business default, *The Analytics of Risk Model Validation*, Palgrave-Macmillan Publishing, 1-12.

[2] Allison, P. D., 1982, Discrete-time methods for the analysis of event histories, *Sociological methodology* 13(1), 61-98.

[3] Altman, E.I., 1968, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *Journal of Finance* 23(4), 589-609.

[4] Andreou, E., Ghysels, E., and Kourtellos, A., 2010, Regression models with mixed sampling frequencies, *Journal of Econometrics* 158(2), 246-261.

[5] Berkson, J., 1950, Are there two regressions?, *Journal of the American Statistical Association* 45(250), 164-180.

[6] Brown, C. C., 1975, On the use of indicator variables for studying the time-dependence of parameters in a response-time model, *Biometrics* 31, 863-872.

[7] Chava, S., and Jarrow, R. A., 2004. Bankruptcy prediction with industry effects, *Review of Finance* 8(4), 537–569.

[8] Cox, D. R., 1970, *The Analysis of Binary Data*, London: Methuen.

[9] Foroni, C. and Marcellino, M. G., 2013, A survey of econometric methods for mixed-frequency data, *Norges Bank Working Paper* 2013-6.

[10] Giannone, D., Reichlin, L., and Small, D., 2008. Nowcasting: The real-time Informational content of macroeconomic data, *Journal of Monetary Economics* 55(4), 665–676.

[11] Kim, M. W., 2008, 통화정책에 따른 장단기 금리스프레드의 경기 예측력에 관한 연구[Do forward-looking monetary policies weaken the predictive power of the term spread?] *(Unpublished master's thesis), Seoul National University*, Seoul, South Korea.

[12] Kim, S. J. and Shin, H. S., 2012, Sustaining production chains through financial linkages, *The American Economic Review* 102(3), 402-406.

[13] Little, R. J. A., 1992, Regression with missing X's: A review, *Journal of the American Statistical Association* 87(420), 1227-1237.

[14] Shumway, T., 2001, Forecasting bankruptcy more accurately: A simple hazard model, *The Journal of Business* 74(1), 101-124.

[15] Zmijewski, M. E., 1984, Methodological issues related to the estimation of financial distress prediction models, *Journal of Accounting Research* 22, 59-82.

Appendix

A.1 Proof of the equation (11)

At each time, data x with observation period p has the following relationship with data z with observation period 1.

$$x_1 = x_0 + \gamma \Delta z_1 + \varepsilon_1$$

$$x_2 = x_1 + \gamma \Delta z_2 + \varepsilon_2$$

$$\vdots$$

$$x_{p-1} = x_{p-2} + \gamma \Delta z_{p-1} + \varepsilon_{p-1}$$

$$x_p = x_p$$

$$x_{p+1} = x_p + \gamma \Delta z_{p+1} + \varepsilon_{p+1}$$

$$\vdots$$

$$x_t = x_{t-1} + \gamma \Delta z_t + \varepsilon_t$$

Taking summation, x_t can be written as

$$x_t = \begin{cases} x_{t-\tau} + \gamma(z_t - z_{t-\tau}) + \sum_{j=1}^{\tau} \epsilon_{t-\tau+j} & \tau > 0 \\ x_{t-\tau} & \tau = 0 \end{cases}$$

where $\tau = \mathrm{t} - \left[\frac{t}{p}\right] p$.

Substituting the above in $y_t = \alpha + \beta x_t + u_t$ gives,

$$y_t = \alpha + \beta \left(x_{t-\tau} + \gamma(z_t - z_{t-\tau}) + \sum_{j=1}^{\tau} \epsilon_{t-\tau+j} \right) + u_t$$

$$= \alpha + \beta x_{t-\tau} + \beta\, \gamma(z_t - z_{t-\tau}) + \beta \sum_{j=1}^{\tau} \epsilon_{t-\tau+j} + u_t$$

A.2 Multiple outdated variables

At time t, let x and y have the following relationship in the population.

True model

$$y_t = \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_K x_{Kt} + u_t, \tag{30}$$

where $t = 1, 2, \ldots, T$.

y_t consists of a deterministic part and a random part. The random disturbance term u_t has expectation 0 and variance σ_u^2 for all t. Assume also the exogeneity condition $\mathrm{E}(u_t | x_{1t}, x_{2t}, \ldots x_{Kt}) = 0$ for all t. This ensures that in predicting the disturbance term at time t, the independent variables at the same time are not informative at all.

Assumptions

A1. The observation period is 1 for y and p_k for x_k, where each p_k is an integer greater than 1 and satisfies $p_1 < p_2 < \cdots < p_K$.

A2. z has the same observation period(=1) as y and has the following relationship with x_k.

$x_{k,t} = x_{k,t-1} + \gamma_k \Delta z_t + \varepsilon_{k,t}$ where ε_{kt} is a disturbance term with expectation 0 and variance $\sigma_{\varepsilon_k}^2$ and $\Delta z_t \equiv z_t - z_{t-1}$.

A3. $\mathrm{E}(\varepsilon_{kt+1} | x_{10}, x_{11}, \ldots, x_{1t}, \ldots, x_{K0}, x_{K1}, \ldots, x_{Kt}, z_0, z_1, \ldots) = 0$

A3. For all t, $\mathrm{t}' \leq T$ and k, $\mathrm{E}(\varepsilon_{k,t'} | \Delta z_t) = 0$.

For all $\mathrm{t}' > \mathrm{t}$, and for all k, k′, $\mathrm{E}(\varepsilon_{k',t'} | x_{k,t}) = 0$.

A4. For all $\mathrm{t}' > \mathrm{t}$ and for all k, $\mathrm{E}(u_t, \varepsilon_{kt}) = 0$.

With A2, x_t can be written as

$$x_{kt} = x_{k,t-\tau_k} + \gamma_k (z_t - z_{t-\tau_k}) + \sum_{j=1}^{\tau_k} \epsilon_{k,t-\tau_k+j} \tag{a1}$$

where $\tau_k = \mathrm{t} - \left[\frac{t}{p_k}\right] p_k$.

y_t can then be expressed as [21]

$$y_t = \alpha + \sum_{k=1}^{K} \beta_k x_{kt-\tau_k} + \sum_{k=1}^{K} \beta_k \gamma_k \left(z_t - z_{t-\tau_k}\right) + \sum_{k=1}^{K} \sum_{j=1}^{\tau_k} \beta_k \epsilon_{k,t-\tau_k+j} + u_t \tag{31}$$

Estimating the above equation with only outdated data, omitted variable bias formula gives

$$\beta_k{}^{s} = \frac{cov\left(y_t, \tilde{x}_{k,t-\tau_k}\right)}{var\left(\tilde{x}_{k,t-\tau_k}\right)} = \beta_k + \sum_{j=1}^{K} \beta_j \gamma_j \, \delta_{zx_j} \neq \beta_k \tag{32}$$

Here, $\tilde{x}_{k,t-\tau_k}$ is the residual when the equation is regressed on all $x_{j,t-\tau_j}$ with j ≠ k. δ_{zx_j} is the population regression coefficient when $\left(z_t - z_{t-\tau_j}\right)$ is regressed on $\tilde{x}_{k,t-\tau_k}$. Thus it follows that if neither γ_j nor δ_{zx_j} is 0, $\beta_k{}^{s}$ is not equal to β_k.

Next, the regression variance when updating through z is as follows.

$$\begin{aligned} & var\left(y_t - E\left(y_t | x_{1,t-\tau_1}, \dots, x_{K,t-\tau_K}, \Delta_{\tau_1} z, \dots, \Delta_{\tau_K} z\right)\right) \\ & = var\left(\sum_{k=1}^{K} \sum_{j=1}^{\tau_k} \beta_k \epsilon_{k,t-\tau_k+j} + u_t\right) = \sum_{k=1}^{K} \tau_k \beta_k{}^{2} \sigma_{\epsilon_k}^2 + \sigma_u^2 \end{aligned} \tag{33}$$

[21] The calculation is as below.

$$\begin{aligned} y_t &= \alpha + \beta_1 x_{1t} + \beta_2 x_{2t} + \cdots + \beta_K x_{Kt} + u_t \\ &= \alpha \\ &\quad + \dots + \beta_K \left(x_{K,t-\tau_K} + \gamma_K \left(z_t - z_{t-\tau_K}\right) + \sum_{j=1}^{\tau_K} \epsilon_{K,t-\tau_K+j}\right) + u_t \\ &= \alpha + \sum_{k=1}^{K} \beta_k x_{kt-\tau_k} + \sum_{k=1}^{K} \beta_k \gamma_k \left(z_t - z_{t-\tau_k}\right) + \sum_{k=1}^{K} \sum_{j=1}^{\tau_k} \beta_k \epsilon_{k,t-\tau_k+j} + u_t \end{aligned}$$

A.3 Measurement error in the auxiliary variable z

We consider the case when the auxiliary variable z_t suffers from measurement error. For example, an estimated financial statement of a firm, obtained from a research center some months after the firm's disclosure, is a data with measurement error. Also, in a survey on health status conducted some time after periodic checkup, the subjective health status is equal to the actual health status plus measurement error.

To incorporate measurement error, let x_t and z_t satisfy the following equations.

$$x_t = x_{t-1} + \gamma \Delta z_t^* \tag{34}$$

$$\Delta z_t = \Delta z_t^* + v_t \tag{35}$$

Assume that v_t is the measurement error with expectation 0 and variance σ_v^2.

The difference is that instead of Δz_t^*, which directly affects x_t, Δz_t with measurement error is observed. y_t can then be expressed as below. We assume the dependent and independent variables are all differences from the mean and suppress the constant term.[22]

$$y_t = \beta x_{t-\tau} + \beta\,\gamma(z_t - z_{t-\tau}) - \beta\gamma \sum_{j=1}^{\tau} v_{t-\tau+j} + u_t \tag{36}$$

---

[22] The derivation is as follows. In the equation below, Δz_t^* is unobservable.

$$x_t = x_{t-1} + \gamma \Delta z_t^*$$

Instead, we observe Δz_t that contains measurement error.

$$\Delta z_t = \Delta z_t^* + v_t$$

x_t can be written as

$$x_t = \begin{cases} x_{t-\tau} + \gamma(z_t - z_{t-\tau}) - \gamma \sum_{j=1}^{\tau} v_{t-\tau+j} & \tau > 0 \\ x_{t-\tau} & \tau = 0 \end{cases}.$$

Substituting this into $y_t = \alpha + \beta x_t + u_t$ gives

$$y_t = \alpha + \beta x_{t-\tau} + \beta\,\gamma(z_t - z_{t-\tau}) - \beta\gamma \sum_{j=1}^{\tau} v_{t-\tau+j} + u_t.$$

where $\tau = \mathrm{t} - \left[\frac{t}{p}\right]p$. It is assumed that measurement error v_t is not correlated with any other variable except Δz_t. Defining $\theta \equiv \beta\gamma$ and suppressing subscript t gives

$$\mathrm{y} = \beta x + \theta\Delta_\tau z - \theta v^\tau + u \quad \text{or} \quad \mathrm{y} = \beta x + \theta\Delta_\tau z^* + u \tag{37}$$

Here, $\Delta_\tau z \equiv z_t - z_{t-\tau}$ and $v^\tau \equiv \sum_{j=1}^{\tau} v_{t-\tau+j}$. Note that v_τ and $\Delta_\tau z$ are functions of τ. While x is observed accurately, Δz_t contains measurement error. If x is correlated with $\Delta_\tau z^*$ but not with v_τ ($\sigma_{xz^*} = \sigma_{xz}$), the estimator $\hat{\theta}$ for θ can be expressed as

$$\operatorname{plim}\hat{\theta}(\tau) = \frac{cov(\tilde{z}, y)}{var(\tilde{z})} \tag{38}$$

$$= \frac{var(x)cov(y, \Delta_\tau z) - cov(x, \Delta_\tau z)cov(y, x)}{var(\Delta_\tau z)var(x) - cov(x, \Delta_\tau z)^2} \tag{39}$$

where $\tilde{z}$ is the residual when $\Delta_\tau z$ is regressed on x.

Defining $\sigma^2_{\Delta z^*} \equiv var(\Delta_\tau z^*)$, $\sigma_{x\Delta z^*} \equiv cov(x, \Delta_\tau z^*)$, $\sigma_{x\Delta z} \equiv cov(x, \Delta_\tau z)$, $\sigma_x^2 \equiv var(x)$, and $\sigma^2_{v_\tau} \equiv var(v^\tau)$ leads to

$$\operatorname{plim}\hat{\theta}(\tau) = \frac{\sigma_x^2\left(\theta\sigma^2_{\Delta_\tau z^*} + \beta\sigma_{x\Delta_\tau z^*}\right) - \sigma_{x\Delta_\tau z}\left(\beta\sigma_x^2 + \theta\sigma_{x\Delta_\tau z^*}\right)}{\sigma_x^2\left(\sigma^2_{\Delta_\tau z^*} + \sigma^2_{v_\tau}\right) - (\sigma_{x\Delta_\tau z})^2} \tag{40}$$

$$= \frac{\theta\left(\sigma_x^2\sigma^2_{\Delta_\tau z^*} - \sigma_{x\Delta_\tau z}\sigma_{x\Delta_\tau z^*}\right) + \beta\sigma_x^2\left(\sigma_{x\Delta_\tau z^*} - \sigma_{x\Delta_\tau z}\right)}{\sigma_x^2\left(\sigma^2_{\Delta_\tau z^*} + \sigma^2_{v_\tau}\right) - (\sigma_{x\Delta_\tau z})^2} \tag{41}$$

$$= \frac{\theta\left(\sigma_x^2\sigma^2_{\Delta_\tau z^*} - \left(\sigma_{x\Delta_\tau z^*}\right)^2\right)}{\sigma_x^2\left(\sigma^2_{\Delta_\tau z^*} + \sigma^2_{v_\tau}\right) - \left(\sigma_{x\Delta_\tau z^*}\right)^2} \tag{42}$$

$$= \theta\lambda_\tau < \theta \tag{43}$$

Here $\lambda_\tau = \frac{\sigma_x^2\sigma_{\Delta_\tau z^*}^2 - \left(\sigma_{x\Delta_\tau z^*}\right)^2}{\sigma_x^2\left(\sigma_{\Delta_\tau z^*}^2 + \sigma_{v_\tau}^2\right) - \left(\sigma_{x\Delta_\tau z^*}\right)^2} = \frac{\lambda_\tau^0 - \rho_{x\Delta_\tau z}^2}{1 - \rho_{x\Delta_\tau z}^2}$, $\lambda_\tau^0 = \frac{\sigma_{\Delta_\tau z^*}^2}{\sigma_{\Delta_\tau z^*}^2 + \sigma_{v_\tau}^2}$, and $\rho_{x\Delta_\tau z}$ is the correlation coefficient of x and $\Delta_\tau z$.

The first equality follows from equation (37), and the third from $\sigma_{x\Delta_\tau z^*} = \sigma_{x\Delta_\tau z}$. The inequality is due $\sigma_x^2\sigma_{v_\tau}^2 > 0$, which implies to $\lambda_\tau < 1$. Thus we see that attenuation bias occurs as in usual measurement error models. When there is no measurement error, there is no bias in coefficient estimation with using only the outdated information if the outdated information x and the updated information $\Delta_\tau z^*$ are uncorrelated. But if updated information has measurement error, even with updating we still have attenuation bias.

Intuitively, the regression coefficient measures the marginal effect on the dependent variable of a change in the independent variable by 1 unit. When there is more than one independent variable, the coefficient of a given independent variable is determined by the relationship between the part not explained by other independent variables and the dependent variable. If there is measurement error, a larger part is left unexplained by other variables but the part related to the dependent variable remains, causing an attenuation bias. Also, as the correlation between independent variables increase, the unexplained part decreases while the variance of measurement error remains the same, so that attenuation bias increases. If the correlation is strong enough, even the sign of coefficient estimates may change.

If x and $\Delta_\tau z$ have 0 correlation, λ_τ is equal to reliability or signal-to-total variance ratio in the simple linear regression model with measurement error, i.e. $\frac{\sigma_{\Delta_\tau z^*}^2}{\sigma_{\Delta_\tau z^*}^2 + \sigma_{v_\tau}^2}$, and the bias from measurement error decreases as the ratio of variance of noise to variance gets smaller for $\Delta_\tau z^*$. We should note here that the bias is a function of τ, the time during which information is not updated. As τ increases, if the updated information $\sigma_{\Delta_\tau z^*}^2$ dominates $\sigma_{v_\tau}^2$, the variance due to measurement error, λ_τ converges to 1; otherwise, attenuation bias becomes greater. A real-life

example would be the fluctuation in the stock market after firms' financial statements are disclosed, called "earning surprise" or "earning shock". Such fluctuations may arise because estimates by research centers, produced while the statements were not being updated, had been excessively reflected compared to actual changes in firms' fundamental. This may have caused estimates to differ considerably from actual values at the time of new disclosure.

In summary, θ measures how well y is explained by that part of $\Delta_\tau z^*$ that remaning unexplained by x. But if we use $\Delta_\tau z$ with measurement error, variability is greater than $\Delta_\tau z^*$ but the part explained by x is the same, so that only inaccurate information has been added to the unexplained part. That is, measurement error increases variability unrelated to y and thus causes attenuation bias.

Let us check if $\hat{\beta}$, the estimate of a coefficient without measurement error, is also biased.

$$\begin{aligned} \operatorname{plim}\hat{\beta}(\tau) &= \frac{cov(\tilde{x}, y)}{var(\tilde{x})} \\ &= \frac{var(\Delta_\tau z)cov(x,y) - cov(x, \Delta_\tau z)cov(\Delta_\tau z, y)}{var(x)var(\Delta_\tau z) - cov(x, \Delta_\tau z)^2} \end{aligned} \tag{44}$$

Rearranging the above as we have done above gives [23]

$$\operatorname{plim}\hat{\beta}(\tau) = \beta + \frac{\theta\sigma_{v_\tau}^2\rho_{x\Delta_\tau z}^2}{\sigma_{x\Delta_\tau z}\left(1-\rho_{x\Delta_\tau z}^2\right)} \tag{45}$$

[23] This is derived as below.

$$\begin{aligned} \operatorname{plim}\beta(\tau) &= \frac{\left(\sigma_{\Delta_\tau z^*}^2 + \sigma_{v_\tau}^2\right)\left(\beta\sigma_x^2 + \theta\sigma_{x\Delta_\tau z}\right) - \sigma_{x\Delta_\tau z}\left(\beta\sigma_{x\Delta_\tau z^*} + \theta\sigma_{\Delta_\tau z^*}^2\right)}{\sigma_x^2\left(\sigma_{\Delta_\tau z^*}^2 + \sigma_{v_\tau}^2\right) - \sigma_{x\Delta_\tau z}^2} \\ &= \frac{\beta\left(\sigma_x^2\left(\sigma_{\Delta_\tau z^*}^2 + \sigma_{v_\tau}^2\right) - \sigma_{x\Delta_\tau z}^2\right) + \theta\sigma_{x\Delta_\tau z}\sigma_{v_\tau}^2}{\sigma_x^2\left(\sigma_{\Delta_\tau z^*}^2 + \sigma_{v_\tau}^2\right) - \sigma_{x\Delta_\tau z}^2} \\ &= \beta + \frac{\theta\sigma_{x\Delta_\tau z^*}\sigma_{v_\tau}^2}{\sigma_x^2\left(\sigma_{\Delta_\tau z^*}^2 + \sigma_{v_\tau}^2\right) - \sigma_{x\Delta_\tau z}^2} = \beta + \frac{\theta\sigma_{v_\tau}^2\rho_{x\Delta_\tau z}^2}{\sigma_{x\Delta_\tau z}\left(1-\rho_{x\Delta_\tau z}^2\right)} \end{aligned}$$

We can see that even the coefficient of a variable with no measurement error is biased. However, the sign of the bias is determined by $\sigma_{x\Delta_\tau z^*}$, and the bias increases with measurement error.

A.4 Yearly financial statements data summary

• Number of financial statements by year (Medium sized firms 1)

year	Training set			Validation set		
	nondefault	default	total	nondefault	default	total
1997	449	1	450	482	3	485
1998	960	14	974	1,022	9	1,031
1999	1,909	44	1,953	1,960	39	1,999
2000	3,231	88	3,319	3,367	72	3,439
2001	4,578	158	4,736	4,606	153	4,759
2002	5,408	325	5,733	5,463	325	5,788
2003	5,918	419	6,337	5,947	395	6,342
2004	6,134	283	6,417	6,137	297	6,434
2005	6,222	221	6,443	6,180	211	6,391
2006	5,995	276	6,271	5,944	284	6,228
2007	5,330	299	5,629	5,262	332	5,594
2008	4,823	35	4,858	4,713	42	4,755
total	50,965	2,163	53,128	51,088	2,162	53,250

*. Default means that the firm went default in less than a year after the financial disclosure.

• Number of financial statements by year (Medium sized firms 2)

year	Training set			Validation set		
	nondefault	default	total	nondefault	default	total
1997	273	0	273	220	3	223
1998	530	4	534	471	5	476
1999	843	9	852	804	6	810
2000	1,353	18	1,371	1,324	22	1,346
2001	1,886	34	1,920	1,863	37	1,900
2002	2,233	68	2,301	2,225	77	2,302
2003	2,534	112	2,646	2,526	119	2,645
2004	2,843	90	2,933	2,775	100	2,875
2005	3,010	98	3,108	2,958	97	3,055
2006	3,182	130	3,312	3,148	124	3,272
2007	3,135	219	3,354	3,115	180	3,295
2008	2,990	55	3,045	2,979	66	3,045
total	24,820	837	25,657	24,408	836	25,244

*. Default means that the firm went default in less than a year after the financial disclosure.

A.5 Financial ratios

Financial ratio	Altman (1968)	Zmijewski (1984)	Kim (Model3)	computation
Net income/total assets		O	O	(net income/total assets)*100
EBIT/total assets	O			{(earnings before interest rate+financial expenses)/total assets}*100
Total debt/total assets		O	O	(total debt/total assets)*100
Working capital/total assets	O			{(current assets-current debt)/total assets}*100
Retained earnings/total assets	O			(retained earnings/total assets)*100
Current assets/current liabilities		O		(current assets/current liabilities)*100
Cash and cash equivalent/total assets			O	(cash and cash equivalent/total assets)*100
Sales/total assets	O			(sales/total assets)*100
Accounts receivables/sales			O	(accounts receivables/sales)*100

*. “O” means that the ratio was used in the corresponding model.

A.6 Estimation results with financial ratios in Altman(1968) and Zmijewski (1984)

• Estimation results(Medium sized firms 1, financial ratios in Altman(1968))

	Model1		Model2		Model3	
	(baseline)		(cyclical component)		(accumulated cyclical component)	
Variable	Coefficient	t	Coefficient	t	Coefficient	t
const.	-7.1858 ***	55.4	-7.1912 ***	55.6	-7.2820 ***	55.6
dummy(construction)	0.1820 ***	3.0	0.1803 ***	3.0	0.1839 ***	3.0
EBIT/TA	-0.0101 ***	5.6	-0.0103 ***	5.7	-0.0105 ***	5.8
WC/TA	-0.0031 ***	4.6	-0.0031 ***	4.6	-0.0030 ***	4.4
RE/TA	-0.0027 ***	3.7	-0.0026 ***	3.6	-0.0025 ***	3.5
SA/TA	-0.0027 ***	11.2	-0.0027 ***	11.2	-0.0027 ***	11.1
cyclical component(CC)			-0.0359 ***	2.9		
accumulated CC(after settling month)					-0.1363 ***	9.4
Age	0.0576 ***	14.7	0.0580 ***	14.8	0.0595 ***	15.2
Age2	-0.0003 ***	12.1	-0.0003 ***	12.2	-0.0004 ***	12.7
Model Fit	811.1 ***		818.7 ***		911.8 ***	
#firms	9,751		9,751		9,751	
#firm-month obs.	622,821		622,821		622,821	
#defaults	2,163		2,163		2,163	
-2LogL	28,004		27,997		27,904	
AUROC	0.657		0.659		0.668	

: Model 1 does not use business cycle information, while Model 2 uses the data without adjustment. Model 3 accumulates business cycle data from the last settling month. By comparing Model 3 to Model 1 and 2, we can discern the effect of updating business cycle data information. Cluster-robust standard errors are used to inference. *, **, *** represent significance at levels 10%, 5%, 1%.

• Estimation results(Medium sized firms 1, financial ratios in Zmijewski (1984))

	Model1		Model2		Model3	
	(baseline)		(cyclical component)		(accumulated cyclical component)	
Variable	Coefficient	t	Coefficient	t	Coefficient	t
const.	-8.8215 ***	60.1	-8.8289 ***	60.1	-8.8966 ***	60.1
dummy(construction)	0.4217 ***	7.0	0.4210 ***	7.0	0.4192 ***	7.0
NI/TA	-0.0150 ***	11.1	-0.0150 ***	11.1	-0.0151 ***	11.2
TD/TA	0.0153 ***	16.1	0.0153 ***	16.1	0.0151 ***	15.9
CA/CL	-0.0001 **	2.2	-0.0001 **	2.2	-0.0001 **	2.3
cyclical component(CC)			-0.0336 ***	2.6		
accumulated CC(after settling month)					-0.1348 ***	9.3
Age	0.0575 ***	14.6	0.0579 ***	14.7	0.0595 ***	15.1
Age2	-0.0003 ***	11.8	-0.0003 ***	11.9	-0.0004 ***	12.4
Model Fit	842.1 ***		848.7 ***		941.0 ***	
#firms	9,751		9,751		9,751	
#firm-month obs.	622,821		622,821		622,821	
#defaults	2,163		2,163		2,163	
-2LogL	27,974		27,967		27,875	
AUROC	0.662		0.664		0.673	

: Model 1 does not use business cycle information, while Model 2 uses the data without adjustment. Model 3 accumulates business cycle data from the last settling month. By comparing Model 3 to Model 1 and 2, we can discern the effect of updating business cycle data information. Cluster-robust standard errors are used to inference. *, **, *** represent significance at levels 10%, 5%, 1%.

• Estimation results(Medium sized firms 2, Kim(Model3))

	Model1		Model2		Model3	
	(baseline)		(cyclical component)		(accumulated cyclical component)	
Variable	Coefficient	t	Coefficient	t	Coefficient	t
const.	-8.7865 ***	39.1	-8.7881 ***	39.1	-8.8317 ***	39.5
dummy(construction)	0.6786 ***	6.7	0.6787 ***	6.7	0.6800 ***	6.7
NI/TA	-0.0201 ***	10.2	-0.0201 ***	10.2	-0.0203 ***	10.3
TD/TA	0.0195 ***	13.8	0.0195 ***	13.8	0.0191 ***	13.6
CA/TA	-0.0687 ***	6.0	-0.0687 ***	6.0	-0.0685 ***	6.0
|AR/SA-0.2|+	0.0114 ***	5.3	0.0114 ***	5.3	0.0116 ***	5.5
cyclical component(CC)			-0.0078	0.4		
accumulated CC(after settling month)					-0.1239 ***	6.2
Age	0.0333 ***	6.4	0.0334 ***	6.4	0.0345 ***	6.6
Age2	-0.00016 ***	4.5	-0.00016 ***	4.5	-0.00017 ***	4.8
Model Fit	1018.1 ***		1018.2 ***		1054.0 ***	
#firms	4,435		4,435		4,435	
#firm-month obs.	288,354		288,354		288,354	
#defaults	837		837		837	
-2LogL	10,433		10,433		10,397	
AUROC	0.79		0.79		0.797	

: Model 1 does not use business cycle information, while Model 2 uses the data without adjustment. Model 3 accumulates business cycle data from the last settling month. By comparing Model 3 to Model 1 and 2, we can discern the effect of updating business cycle data information. Cluster-robust standard errors are used to inference. *, **, *** represent significance at levels 10%, 5%, 1%.

• Estimation results(Medium sized firms 2, financial ratios in Altman(1968))

	Model1		Model2		Model3	
	(baseline)		(cyclical component)		(accumulated cyclical component)	
Variable	Coefficient	t	Coefficient	t	Coefficient	t
const.	-7.1742 ***	39.2	-7.1748 ***	39.2	-7.2712 ***	39.4
dummy(construction)	0.8767 ***	9.1	0.8765 ***	9.1	0.8705 ***	9.0
EBIT/TA	-0.0203 ***	7.4	-0.0204 ***	7.4	-0.0205 ***	7.4
WC/TA	-0.0071 ***	7.6	-0.0071 ***	7.6	-0.0069 ***	7.4
RE/TA	-0.0034 **	2.7	-0.0034 **	2.7	-0.0033 **	2.6
SA/TA	-0.0025 ***	5.4	-0.0025 ***	5.4	-0.0025 ***	5.5
cyclical component(CC)			-0.0040	0.2		
accumulated CC(after settling month)					-0.1354 ***	6.8
Age	0.0363 ***	7.1	0.0364 ***	7.1	0.0380 ***	7.4
Age2	-0.0002 ***	5.2	-0.0002 ***	5.2	-0.0002 ***	5.7
Model Fit	619.4 ***		619.5 ***		661.3 ***	
#firms	4,435		4,435		4,435	
#firm-month obs.	288,354		288,354		288,354	
#defaults	837		837		837	
-2LogL	10,832		10,832		10,790	
AUROC	0.716		0.717		0.722	

: Model 1 does not use business cycle information, while Model 2 uses the data without adjustment. Model 3 accumulates business cycle data from the last settling month. By comparing Model 3 to Model 1 and 2, we can discern the effect of updating business cycle data information. Cluster-robust standard errors are used to inference. *, **, *** represent significance at levels 10%, 5%, 1%.

• Estimation results(Medium sized firms 2, financial ratios in Zmijewski (1984))

	Model1		Model2		Model3	
	(baseline)		(cyclical component)		(accumulated cyclical component)	
Variable	Coefficient	t	Coefficient	t	Coefficient	t
const.	-9.2887 ***	43.6	-9.2897 ***	43.6	-9.3275 ***	44.0
dummy(construction)	0.6932 ***	7.1	0.6933 ***	7.1	0.6889 ***	7.0
NI/TA	-0.0181 ***	9.6	-0.0181 ***	9.5	-0.0182 ***	9.7
TD/TA	0.0227 ***	16.5	0.0227 ***	16.5	0.0223 ***	16.4
CA/CL	-0.0001 *	1.8	-0.0001 *	1.8	-0.0001 *	2.0
cyclical component(CC)			-0.0069	0.4		
accumulated CC(after settling month)					-0.1236 ***	6.3
Age	0.0351 ***	6.7	0.0352 ***	6.7	0.0362 ***	6.9
Age2	-0.0002 ***	4.7	-0.0002 ***	4.7	-0.0002 ***	5.0
Model Fit	861.7 ***		861.8 ***		897.5 ***	
#firms	4,435		4,435		4,435	
#firm-month obs.	288,354		288,354		288,354	
#defaults	837		837		837	
-2LogL	10,590		10,589		10,554	
AUROC	0.761		0.762		0.766	

: Model 1 does not use business cycle information, while Model 2 uses the data without adjustment. Model 3 accumulates business cycle data from the last settling month. By comparing Model 3 to Model 1 and 2, we can discern the effect of updating business cycle data information. Cluster-robust standard errors are used to inference. *, **, *** represent significance at levels 10%, 5%, 1%.

국문초록

김명원

경제학부

서울대학교 대학원

본 논문에서는 선형회귀모형에서 설명변수의 관측주기가 종속변수의 관측주기보다 긴 경우에 설명변수에 적시성이 떨어지는 자료를 사용하면 최소자승법을 통한 추정계수에 편의가 발생할 수 있음을 이론적으로 확인한다. 이러한 편의는 설명변수가 관측되지 않는 기간 동안 업데이트 되지 않은 정보와 최근 관측치가 서로 상관되어 있는 경우에 업데이트되지 않은 정보로 인해 발생하는 누락변수편의로 해석할 수 있다. 설명변수보다 관측주기가 짧으면서 설명변수와 상관있는 보조변수를 통해 설명변수에 대한 정보를 업데이트하면 누락변수편의가 제거될 뿐만 아니라 추정의 불확실성도 함께 축소된다.

실증분석에서는 본 논문에서 제시하는 정보업데이트 방식을 기업의 부도예측모형에 적용하여 그 유용성을 확인한다. 결산재무제표는 일반적으로 1 년에 한번 관측되며 부도위험을 측정할 때는 이미 수개월이 지난 재무정보를 이용하게 된다. 따라서 결산월 이후부터 자료를 이용하는 시점 사이에 발생하는 재무상태의 변동을 반영하지 못한다. 본 논문에서는 재무정보가 업데이트되지 않는 기간 동안 기업들의 평균적인 재무상태의 변동정보를 월별 경기변동정보를 이용하여 업데이트함으로써 부도예측의 정확성을 높일 수 있음을 확인한다. 나아가 금융기관은 이러한 방식의 정보업데이트를 이용하여 대출자산에 대한 신용위험을 보다 정확하게 모니터링할 수 있다.

주요어: 혼합주기자료, 정보업데이트, 때늦은 자료, 누락변수편의, 부도예측모형, 신용위험 모니터링

학 번: 2013 - 30058