International Journal of Forecasting 33 (2017) 513–522

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Forecasting loss given default of bank loans with multi-stage model

Yuta Tanoue [a,b,*], Akihiro Kawada [c,d], Satoshi Yamashita [d]

[a] The Graduate University for Advanced Studies, Tachikawa, Tokyo, Japan
[b] Japan Society for the Promotion of Science, Tokyo, Japan
[c] PricewaterhouseCoopers Aarata - Governance, Risk and Compliance, Tokyo, Japan
[d] The Institute of Statistical Mathematics, Tachikawa, Tokyo, Japan

ARTICLE INFO

Keywords:
Credit risk modeling
Loss given default
Multi-stage model
Probability of default
Expected loss

ABSTRACT

Probability of default (PD) and loss given default (LGD) are key risk parameters in credit risk management. The majority of LGD research is based on the corporate bond market and few studies focus on the LGD of bank loans even in Japan because of the lack of available public data on bank loan losses. Consequently, knowledge concerning Japanese bank loan LGD is scarce. This study uses Japanese bank loan data to analyze the influencing factors of LGD and to develop a (multi-stage) model to predict LGD and expected loss (EL). We found that collateral, guarantees, and loan size impact LGD. Further, we confirmed that our multi-stage LGD model has superior predictive accuracy than the corresponding OLS model, Tobit model and Inflated beta regression model.

1. Introduction

The Basel II/III Accord allows banks to estimate their credit risk capital requirements using an internal ratings-based (IRB) approach. The probability of default (PD) and loss given default (LGD) are the most important credit risk parameters in an IRB approach. If banks select the foundations internal ratings-based (FIRB) approach, no proprietary LGD predictive model is required, but if they select the advanced internal ratings-based (AIRB) approach, they will have to build a proprietary predictive model for LGD.

We analyze LGD using data provided by three Japanese banks. The aims of this study are as follows. First, we analyze the factors that influence the Japanese LGD. Second, we develop an expected loss (EL) predictive model, consisting of the PD predictive and LGD predictive models. Third, we compare the performances of LGD models and other models. This paper builds on the work of Kawada and Yamashita (2013). We then investigate the factors that influence LGD and improve the LGD predictive model proposed by Kawada and Yamashita (2013).

The remainder of this paper is structured as follows. Section 2 presents a literature review of the influencing factors and LGD modeling methods. Section 3 defines the terms default and LGD, as used in this study. Section 4 describes the dataset of bank loans used in this study. Section 5 discusses the factors that influence LGD. Section 6 proposes the EL predictive model, consisting of the PD and LGD predictive models. Section 7 evaluates the predictive accuracy of the LGD predictive model. Section 8 presents our conclusions.

2. Literature review

2.1. The factors that influence LGD

Here, we summarize the relationship between LGD and the factors that influence it, as reported in previous studies.

* Corresponding author at: The Graduate University for Advanced Studies, Tahikawa, Tokyo, Japan.
E-mail address: tanoue.yuta@ism.ac.jp (Y. Tanoue).

http://dx.doi.org/10.1016/j.ijforecast.2016.11.005
0169-2070/

Several previous studies (e.g., Araten, Jacobs, & Varshney, 2004; Dermine & De Carvalho, 2006; Grunert & Weber, 2009; Miura, Yamashita, & Eguchi, 2010) have confirmed that collateral has a reduction effect on LGD.

The relationship between LGD and the loan size has been the subject of considerable research. Dermine and De Carvalho (2006) and Felsovalyi and Hurt (1998) reported that the loan size has a negative impact on recovery rates. In contrast, Grunert and Weber (2009) confirmed that a greater loan size led to a lower LGD, while Miura et al. (2010) reported no relationship between the loan size and LGD. According to Gürtler and Hibbeln (2011), the longer the duration of the workout process, the greater is LGD.

However, the results of previous analyses of the relationships between LGD and borrower company size (Asarnow & Edwards, 1995; Felsovalyi & Hurt, 1998; Grunert & Weber, 2009), LGD and borrower creditworthiness (Grunert & Weber, 2009), and LGD and business cycles (Bellotti & Crook, 2012; Caselli, Gatti, & Querci, 2008; Dermine & De Carvalho, 2006; Felsovalyi & Hurt, 1998; Grunert & Weber, 2009) have varied, and the effectiveness of each influencing factor has depended on the method of analysis. Furthermore, previous studies have been based on certain bank data that merely capture the idiosyncratic characteristics of banks.

2.2. LGD models

LGD typically lies in the interval [0, 1], and is concentrated at 0 and 1. These LGD features imply that a simple linear regression model may not have a high level of predictive power. Thus, some previous studies have proposed various different models that considered these features of LGD (Bastos, 2010; Loterman, Brown, Martens, Mues, & Baesens, 2012; Matuszyk, Mues, & Thomas, 2010). The study by Miura et al. (2010) is based on Japanese bank loans, and is closely related to this study. In view of the extended duration of the workout process in Japan, the authors proposed that the time since a default be incorporated into the model.

The multi-stage model proposed in this study consists of binary decision models and a regression model. Multi-stage models have been proposed for LGD modeling by various studies (e.g., Bellotti & Crook, 2012; Gürtler & Hibbeln, 2011; Lucas, 2006), and we introduce some of these here. Lucas (2006) suggested a two-stage model for modeling the LGDs associated with mortgages. The author first divided the workout process according to whether the property is repossessed or not, then calculated the loss in the case of repossession. Gürtler and Hibbeln (2011) found significant differences between the characteristics of recovered and written-off loans. They accounted for these differences by dividing the defaults into two types, recoveries and write-offs, through a logistic regression, then conducting a separate regression for each case. Bellotti and Crook (2012) also proposed a multi-stage model for predicting LGD. Because LGD is concentrated at 0 and 1, the authors present three cases of LGDs ($LGD = 0$, $\mathrm{LGD} = 1$, and $0 < \mathrm{LGD} < 1$) using logistic regression models. They then use an ordinary least squares (OLS) regression model for the case of $0 < \mathrm{LGD} < 1$.

3. Definition

This section defines the terms default and LGD, as used in this study.

Japanese banks typically have internal rating systems as follows.

- Non-default ratings
  1. Normal, performing borrowers.
  2. Performing borrowers with some future concerns.
- Default ratings
  3. Performing borrowers that require monitoring.
  4. Non-performing and probably irrecoverable borrowers.
  5. Practically uncollectible borrowers.
  6. Uncollectible borrowers.

In the actual internal rating system of bank borrowings, each rating is further divided into parts. We define default as a transition from a rating of "Normal, performing borrowers" or "Performing borrowers with some future concerns" to a rating of "others".

LGD can be defined in various different ways (Basel Committee on Banking Supervision, 2005). The LGD calculation formula used in this study is as follows:

$$\mathrm{LGD} = \frac{\text{Total write-off amount}}{\text{EAD}}, \tag{1}$$

where the total write-off amount is the sum of the amount written-off from the default until the end of the default, and EAD represents the exposure at default.

4. Data

The dataset used in this study consists of loan data from three Japanese banks (Bank A, Bank B, and Bank C) for the period from 2004 to 2011. Banks A, B, and C are located in western Japan, eastern Japan, and western Japan, respectively, meaning that our dataset is not region-specific. The observation frequency is once every six months. Our dataset consists of 679,607 records of 81,931 borrowers, with 8,732 default records, and contains the following fields:

(1) Exposure (total loan amount of each borrower).
(2) Bank's internal borrower rating (rating of each borrower).
(3) Write-off amount (amount written-off from an account).
(4) Creditworthiness score (a variable that is synthesized exogenously based on the borrower's financial information, and used for estimating the probability of default).
(5) Collateral quota (the quota for each type of collateral, namely real estate, commercial bills, deposits, and marketable securities).
(6) Credit guarantee quota (the quota of credit guarantee from the National Federation of Credit Guarantee Corporation[1]).

[1] The National Federation of Credit Guarantee Corporations is unique to Japan and is established to smooth the financing of small and medium enterprises (SMEs). SMEs typically find it difficult to obtain finance from banks because SME loans are considered highly risky. The National Federation of Credit Guarantee Corporations guarantees SME loans, thus smoothing SME financing.

Table 1
Median, mean, and standard deviation of variables (dataset A).

Variable	Median	Mean	SD
Creditworthiness score	49.000	48.372	16.860
Collateral quota (real estate)	0.000	0.240	0.485
Collateral quota (commercial bills)	0.000	0.050	0.174
Collateral quota (deposits)	0.000	0.019	0.127
Collateral quota (marketable securities)	0.000	0.004	0.066
Credit guarantee quota	0.513	0.501	0.439
Exposure in hundred million yen	0.203	1.103	5.502

(7) Duration of the workout process (the length of time between the default and the end of the workout process).

4.1. Description of whole dataset

Table 1 describes the whole dataset, consisting of both non-default and default records.

The average creditworthiness score is 48.372. This score, which indicates the creditworthiness of borrowers, is adjusted to make the average value 50 for all borrowers. This result is natural because the dataset consists of both defaulted and non-defaulted borrowers. Real estate makes up the majority of the total collateral, with the average collateral quota (real estate) being 0.240. The average credit guarantee quota is 0.501, which includes only the guarantee from the National Federation of Credit Guarantee Corporations. Guarantees other than those provided by the National Federation of Credit Guarantee Corporations are not included. The average exposure is 1.103 hundred million yen.

4.2. Description of default records

4.2.1. Treatment of censored data

The treatment of censored data (workout proceeding data) presents some problems when analyzing LGD. Since the typical length of a workout process is months or years, the dataset contains a considerable amount of censored data. However, only the analysis based on defaulted records with completed workout processes may be biased.

Previous studies have proposed several techniques for avoiding this bias (Gürtler & Hibbeln, 2011; Zhang & Thomas, 2012). However, we do not consider it here, as this bias is only an issue if the observation period is not sufficiently long, while the data observation period in this study is long enough to allow for the duration of the workout process. Thus, our analysis of LGD is based only on defaults with completed workout processes.

4.2.2. LGD fundamental statistics

Of the 8,732 defaults with completed workout processes that occur during the observation period, we analyzed 5,664. While 1,334 borrowers could be recovered, 4,330 had to be written off, of which 3,214 caused no loss and 1,116 caused loss.

The fundamental LGD statistics for each bank are given in Table 2. The mean LGD value of all banks is 0.089. The average LGD in Japan is significantly lower than the levels suggested by FIRB. We find some differences in regard to individual banks: while the medians of all banks are the same, there are some differences in their means and standard deviations. Bank A's mean LGD is 0.060, Bank B's is 0.095, and Bank C's is 0.155. Thus, Bank C's mean LGD is more than double that of Bank A.

Table 2
Median, mean, and standard deviation of LGD.

	Median	Mean	SD
All	0.000	0.089	0.228
Bank A	0.000	0.060	0.183
Bank B	0.000	0.095	0.237
Bank C	0.000	0.155	0.289

Table 3
Duration of the workout process (in years) for each bank.

	Median	Mean	SD
Bank A	1.000	1.370	1.149
Bank B	1.000	1.453	1.166
Bank C	0.500	0.926	0.797

The average duration of the workout process for the three banks is 1.339 years, with the mean duration of the workout process varying between banks. Table 3 shows the duration of workout process for each bank. The average durations are 1.370 for Bank A, 1.453 for Bank B, and 0.926 for Bank C. There is a difference of approximately six months between the durations of Banks B and C, meaning that the sufficient observation period for estimating LGD may differ between banks.

5. Linear regression analysis

This section conducts a linear regression analysis to determine the factors that influence LGD. LGD lies in the range [0, 1], and we therefore use the logit-transformed linear regression model to find the factors that influence LGD. The values predicted using the logit-transformed linear regression are guaranteed to lie in the range [0, 1]. For the logit-transformed linear regression, we convert LGD $= 1$ to 0.99 and 0 to 0.01.

Table 4 gives the logic-transformed linear regression results. The creditworthiness score is not significant, meaning that it has a weak relationship with LGD, although Grunert and Weber (2009) confirmed that the LGD is high for borrowers with low creditworthiness. All types of collateral are statistically significant and reduce LGD. Credit guarantee quotas are likewise significant and reduce LGD. EAD is statistically insignificant at the 5% level, meaning that LGD and EAD are related only weakly. This is not

Table 4
Linear regression analysis results of the factors that influence LGD.

	Estimate	Std. error	t-value	Pr(> |t|)
(Intercept)	1.858	0.091	20.476	0.000
Creditworthiness score	−0.001	0.002	−0.680	0.497
Collateral quota (real estate)	1.160	0.070	16.488	0.000
Collateral quota (commercial bills)	1.660	0.187	8.878	0.000
Collateral quota (deposits)	1.701	0.244	6.984	0.000
Collateral quota (marketable securities)	3.490	0.735	4.747	0.000
Credit guarantee quota	2.651	0.068	38.930	0.000
ln (EAD)	0.029	0.016	1.731	0.084
Duration of the workout process (in years)	−0.055	0.022	−2.451	0.014
Year of default				
2004	0.416	0.083	5.039	0.000
2005	0.439	0.062	7.069	0.000
2006	0.109	0.060	1.819	0.069
2007	0.113	0.054	2.083	0.037
2008	−0.042	0.054	−0.790	0.430
2009	−0.124	0.061	−2.033	0.042
2010	−0.191	0.078	−2.433	0.015
2011	−0.720	0.148	−4.859	0.000
Adjusted R^2	0.268			
Number of observations	5664			

consistent with the results of Felsovalyi and Hurt (1998) or Bastos (2010). The duration of the workout process is significant: an extended workout process leads to a high LGD, although Querci (2005) reported no relationship between the length of the workout process and LGD.

LGD is considered to be affected by the business cycle, with Altman, Resti, and Sironi (2001) confirming that LGD is affected by economic fluctuations. As Table 4 shows, some of the default dummy years[2] are significant. However, we cannot address the relationship between LGD and the business cycle because it is difficult to define peaks and troughs in business cycles.

6. The EL forecasting model

6.1. Overview of the EL forecasting model

EL is typically calculated as PD × LGD. This section builds both the PD and LGD predictive models for estimating EL, and they are explained below.

As Fig. 1 shows, the EL forecasting model consists of (i) the PD model and the multi-stage LGD model. The multi-stage LGD model consists of (ii) the Pr(Recovery) model, (iii) the Pr(LGD > 0) model, and (iv) the LGD regression model. While the Pr(Recovery) model predicts the probability of recovery for a default borrower, the Pr(LGD > 0) model predicts the probability of a loss being caused when a default borrower is written-off, and the LGD regression model predicts the LGD when a loss occurs ($\mathrm{LGD}_{\mathrm{LGD}>0}$). Each model will be explained in detail in Section 6.2. The LGD predicted from the multi-stage LGD model is expressed as follows:

$$\mathrm{LGD} = (1 - \Pr(\mathrm{Recovery})) \times \Pr(\mathrm{LGD} > 0) \times \mathrm{LGD}_{\mathrm{LGD}>0}. \quad (2)$$

[2] We impose sum-to-zero constraints on the default dummy year variables when estimating parameters.

The output of the multi-stage model is the expectation of each LGD case. Since recovered borrowers generally do not cause loss, we regard them as causing no loss.

The data used for model building vary between models, and are given in Table 5. We use all of the data to build the PD model.

6.2. Model coefficients from regressions

We present the model coefficients from regressions, using Akaike's information criterion (AIC) (Akaike, 1973) as the variable selection method.

6.2.1. Estimation results for the PD model

We use the logistic regression model for the PD model. The logistic regression model is expressed as follows:

$$\mathrm{PD} = \frac{1}{1 + \exp(Z^I)}$$
$$Z^I = \alpha^I \sum_k \beta_k^I x_k^I, \quad (3)$$

where x^I represents the explanatory variables, and α^I and β^I are regression parameters. The regression results are shown in Table 6. All of the explanatory variables are statistically significant. Loan and borrower characteristics have explanatory power to predict PD, with borrowers with high creditworthiness scores being found to be unlikely to default. This is natural because the creditworthiness score synthesizes the default probability estimation. All types of collaterals are statistically significant, indicating that the PD of borrowers with high collateral quotas is low. However, the PD of borrowers with high credit guarantee quotas is high. Thus, banks require the loans of borrowers with low creditworthiness scores to be secured by credit guarantees. We find a negative relationship between exposure and PD. Although borrowers with substantial exposure are

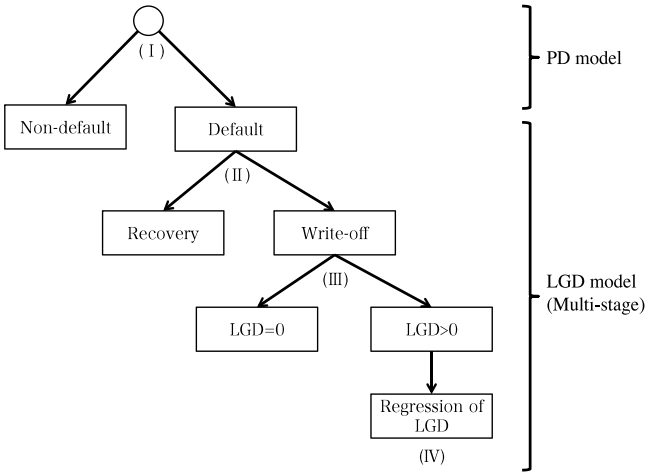

Fig. 1. EL forecasting model.

Table 5
Data used to build each model.

model	Non-default	Default			
		Recoveries	Write-offs (LGD=0)	Write-offs (LGD>0)	Censored
(I) PD model	○	○	○	○	○
(II) Pr(Recovery) model		○	○	○	
(III) Pr(LGD>0) model			○	○	
(IV) Regression model				○	

Table 6
Estimation results for the PD model.

	Estimate	Std. error	z-value	Pr($>\|z\|$)
(Intercept)	1.180	0.036	32.900	0.000
Creditworthiness score	0.072	0.001	96.546	0.000
Collateral quota (real estate)	0.637	0.035	18.012	0.000
Collateral quota (commercial bills)	0.714	0.094	7.557	0.000
Collateral quota (deposits)	0.450	0.122	3.681	0.000
Collateral quota (marketable securities)	0.886	0.258	3.427	0.001
Credit guarantee quota	−0.101	0.033	−3.060	0.002
ln (Exposure)	−0.203	0.008	−26.171	0.000
AUC	0.815			
Number of observations	679607			
Number of defaults	8732			

generally considered unlikely to default, we obtain the opposite result in this study.

6.2.2. *Estimation results for the* Pr(*Recovery*) *model*

We use a logistic regression model for the Pr(Recovery) model, with the regression results being shown in Table 7. The creditworthiness score, some types of collaterals, the credit guarantee, and the EAD are all statistically significant. As the creditworthiness score increases, the probability of recovery increases. Borrowers with high collateral quotas (real estate) are easily recoverable. Borrowers who are secured by collateral (commercial bills) and credit guarantees are difficult to recover. This result indicates that the impacts on the probability of recovery varies by the type of security. EAD also affects the probability of recovery, with a substantial EAD leading to a high probability of recovery.

6.2.3. *Estimation results for the* Pr(*LGD* > 0) *model*

We use the logistic regression model for the Pr(LGD > 0) model, with the regression results being shown in Table 8. All of the variables, except for the creditworthiness score, are statistically significant. A loss is less likely when the collateral and credit guarantee quotas are high. Borrowers with a large EAD are likely to cause loss.

6.2.4. *Estimation results for the LGD regression model*

We use the logit-transformed OLS model for the LGD regression model:

$$\log\left(\frac{\mathrm{LGD}_{\mathrm{LGD}>0}}{1-\mathrm{LGD}_{\mathrm{LGD}>0}}\right) = -\left(\alpha^{\mathrm{IV}} + \sum_n \beta_n^{\mathrm{IV}} x_n^{\mathrm{IV}}\right), \tag{4}$$

where x^{IV} represents the explanatory variables, and α^{IV} and β^{IV} are regression parameters. The regression results

Table 7
Estimation results for the Pr(Recovery) model.

	Estimate	Std. error	z-value	Pr(> |z|)
(Intercept)	0.959	0.115	8.315	0.000
Creditworthiness score	−0.013	0.002	−5.167	0.000
Collateral quota (real estate)	−0.323	0.095	−3.418	0.001
Collateral quota (commercial bills)	1.702	0.339	5.015	0.000
Collateral quota (deposits)				
Collateral quota (marketable securities)				
Credit guarantee quota	0.353	0.095	3.730	0.000
ln (EAD)	−0.382	0.025	−15.568	0.000
AUC	0.701			
Number of observations	5664			
Number of recoveries	1334			

Table 8
Estimation results for the Pr(LGD > 0) model.

	Estimate	Std. error	z-value	Pr(> |z|)
(Intercept)	−2.205	0.107	−20.701	0.000
Creditworthiness score				
Collateral quota (real estate)	2.590	0.208	12.455	0.000
Collateral quota (commercial bills)	2.487	0.295	8.440	0.000
Collateral quota (deposits)	3.221	0.589	5.466	0.000
Collateral quota (marketable securities)	5.384	2.237	2.407	0.016
Credit guarantee quota	3.421	0.122	28.052	0.000
ln (EAD)	−0.654	0.037	−17.621	0.000
AUC	0.904			
Number of observations	4330			
Number of LGD > 0s	1116			

Table 9
Estimation results for the LGD regression model.

	Estimate	Std. error	t-value	Pr(> |t|)
(Intercept)	−1.523	0.072	−21.256	0.000
Creditworthiness score				
Collateral quota (real estate)	2.142	0.207	10.334	0.000
Collateral quota (commercial bills)	2.786	0.312	8.935	0.000
Collateral quota (deposits)	4.414	0.640	6.897	0.000
Collateral quota (marketable securities)	6.690	2.465	2.715	0.007
Credit guarantee quota	4.431	0.133	33.198	0.000
ln (EAD)	0.213	0.031	6.788	0.000
Adjusted R^2	0.539			
Number of observations	1116			

are shown in Table 9. All of the variables, except for the creditworthiness score, are statistically significant. As the collateral and credit guarantee quotas increase, $\mathrm{LGD}_{\mathrm{LGD}>0}$ decreases. A substantial EAD leads to a low $\mathrm{LGD}_{\mathrm{LGD}>0}$. This result differs from that in Section 6.2.3 because the dataset employed is different. The analysis was conducted on a dataset that included the write-off categories (LGD $>$ 0) and (LGD $=$ 0) in Section 6.2.3; however, the dataset we use contains only the category of write-offs (LGD $>$ 0). Although borrowers with a large EAD are likely to cause loss, the LGD is relatively small when a loss occurs.

6.3. Significance of factors in the multi-stage LGD model

Here, we discuss the significance of factors from the perspective of the whole multi-stage LGD model. As was explained in Sections 6.2.2–6.2.4, the creditworthiness score is statistically significant only in the Pr(Recovery) model. As the creditworthiness score increases, so does the probability of recovery. Since recovered borrowers generally do not cause loss, a high credit score makes the LGD value estimated using the multi-stage model low.

Collateral quota (real estate) is statistically significant in the Pr(Recovery), Pr(LGD $>$ 0), and LGD regression models. As was confirmed in Sections 6.2.2–6.2.4, borrowers with high collateral quotas (real estate) are easy to recover and are less likely to cause loss, with $\mathrm{LGD}_{\mathrm{LGD}>0}$ decreasing as the collateral quota (real estate) increases. Therefore, a high collateral quota (real estate) makes the LGD value estimated using the multi-stage model low.

Although borrowers with high collateral quotas (commercial bills) and credit guarantee quotas are less likely to be recovered easily, they are unlikely to cause a loss and their $\mathrm{LGD}_{\mathrm{LGD}>0}$ is small. Therefore, a high collateral quota (commercial bills) and credit guarantee quota make the LGD value estimated by the multi-stage model low.

The collateral quota (deposits) and collateral quota (marketable securities) are statistically insignificant in the Pr(Recovery) model, but significant in both the Pr(LGD $>$

Table 10
Ten-fold cross-validation predictive performance.

	R^2	Spearman's rho	MAE	RMSE	RAE
Multi-stage	0.3190	0.4802	0.1049	0.1901	0.7251
Tobit	0.2955	0.4714	0.1079	0.1920	0.7448
OLS	0.2946	0.4506	0.1134	0.1925	0.7832
Inflated beta	0.3016	0.4709	0.1059	0.1918	0.7311

Table 11
The estimation results for OLS.

	Estimate	Std. error	t-value	Pr($>\|t\|$)
(Intercept)	0.309	0.006	52.376	0.000
Creditworthiness score				
Collateral quota (real estate)	−0.144	0.008	−17.458	0.000
Collateral quota (commercial bills)	−0.218	0.022	−9.820	0.000
Collateral quota (deposits)	−0.213	0.029	−7.454	0.000
Collateral quota (marketable securities)	−0.400	0.084	−4.773	0.000
Credit guarantee quota	−0.317	0.008	−40.419	0.000
ln (EAD)	−0.010	0.002	−4.988	0.000

0) model and the LGD regression model. Sections 6.2.2–6.2.4 confirmed that borrowers with high collateral quotas (deposits) and collateral quotas (marketable securities) are unlikely to cause loss, have a small $\mathrm{LGD}_{\mathrm{LGD}>0}$. Therefore, a high collateral quota (deposits) and collateral quota (marketable securities) make the LGD value estimated by the multi-stage model low.

Although a substantial EAD leads to a high probability of recovery and makes $\mathrm{LGD}_{\mathrm{LGD}>0}$ low, borrowers with a large EAD are likely to cause loss. Therefore, as was explained in Section 6.2.4, although borrowers with a large EAD are likely to cause loss, the LGD is relatively small when a loss occurs.

7. Validation of the multi-stage LGD model

We assess the predictive performance of the multi-stage LGD model developed in Section 6 using several performance measures, namely R^2, Spearman's rho, the mean absolute error (MAE), the root mean squared error (RMSE), and the relative absolute error (RAE). Models with high values of R^2 and Spearman's rho and low values of the MAE, RMSE, and RAE have superior predictive accuracy.

Further, we compare the predictive performance of our multi-stage LGD model with those of the OLS model, the Tobit model (McDonald & Moffitt, 1980; Sigrist & Stahel, 2010; Yashkir & Yashkir, 2013), and the inflated beta regression model (Ospina & Ferrari, 2010; Swearingen, Castro, & Bursac, 2012; Yashkir & Yashkir, 2013).

7.1. Performance assessment with 10-fold cross-validation

We conduct a 10-fold cross-validation in order to avoid overfitting the data, following Bastos (2010), who also evaluated the predictive performance of the LGD model using 10-fold cross-validation. Here, the entire sample is divided randomly into 10 equal-sized subsets, then nine of these are used to build the model and the remaining one is used to assess the model. This procedure is repeated 10 times, with each subset being used once as the assessment subset.

The OLS, Tobit, and inflated beta model regression results are shown in Tables 11–16. These give the results of the cross-validation of the single subsets. Table 10 shows each model's 10-fold cross-validation predictive performance. Our multi-stage model is superior to all of the other models for all parameters. The R^2 and Spearman's rho values of our multi-stage model are higher than those of the other models, while the MAE, RMSE, and RAE of our multi-stage model are lower than those of the other models.

7.2. Performance assessment with out-of-time data

We also assess the predictive power of our multi-stage LGD model using out-of-time data. For this, we conduct five sets of out-of-time predictive performance assessments. In the first set, the models are fitted with the data of borrowers whose defaults are resolved before 2006 and the performance is assessed using the data of borrowers whose defaults are resolved in 2007. The second to fifth assessment sets are carried out in the same manner, fitting the models with the data of borrowers whose defaults are resolved before 2007, 2008, 2009, and 2010 and assessing the performance using the data of borrowers whose defaults are resolved in 2008, 2009, 2010, and 2011, respectively.

Tables 17–21 report the results of each model's predictive performance in the out-of-time samples, respectively. Our multi-stage model's R^2 and Spearman's rho are higher than those of other models in out-of-time predictive performance assessment sets 1–3 (Tables 17–19). Although the R^2 of the multi-stage model is lower than that of other models in our out-of-time predictive performance assessment sets 4–5 (Tables 20–21), the difference is relatively small. In addition, the MAE, RMSE, and RAE of the multi-stage model are lower than those of the other models in all of the out-of-time predictive performance assessment sets.

Table 12
The estimation results for Tobit model.

	Estimate	Std. error	t-value	Pr(> |t|)
(Intercept)	0.281	0.022	13.052	0.000
Creditworthiness score				
Collateral quota (real estate)	−0.769	0.055	−13.880	0.000
Collateral quota (commercial bills)	−0.536	0.086	−6.272	0.000
Collateral quota (deposits)	−0.866	0.164	−5.288	0.000
Collateral quota (marketable securities)	−2.493	0.648	−3.846	0.000
Credit guarantee quota	−1.069	0.037	−29.055	0.000
ln (EAD)	0.046	0.008	5.567	0.000
ln Sigma	−0.567	0.026	−21.573	0.000

Table 13
The estimation results for the inflated beta model (μ).

	Estimate	Std. error	t-value	Pr(> |t|)
(Intercept)	0.967	0.057	16.850	0.000
Creditworthiness score				
Collateral quota (real estate)	−1.689	0.159	−10.652	0.000
Collateral quota (commercial bills)	−2.205	0.250	−8.830	0.000
Collateral quota (deposits)	−2.982	0.372	−8.016	0.000
Collateral quota (marketable securities)	−4.964	1.106	−4.489	0.000
Credit guarantee quota	−3.312	0.092	−36.144	0.000
ln (EAD)	−0.127	0.023	−5.604	0.000

Table 14
The estimation results for the inflated beta model (σ).

	Estimate	Std. error	t-value	Pr(> |t|)
(Intercept)	0.156	0.045	3.499	0.000
Creditworthiness score				
Collateral quota (real estate)	−0.310	0.153	−2.019	0.044
Collateral quota (commercial bills)				
Collateral quota (deposits)	−1.375	0.469	−2.929	0.003
Collateral quota (marketable securities)	−3.893	1.556	−2.502	0.012
Credit guarantee quota	−1.357	0.081	−16.756	0.000
ln (EAD)				

Table 15
The estimation results for the inflated beta model (τ).

	Estimate	Std. error	t-value	Pr(> |t|)
(Intercept)	−4.141	0.571	−7.255	0.000
Creditworthiness score	0.031	0.012	2.534	0.011
Collateral quota (real estate)				
Collateral quota (commercial bills)				
Collateral quota (deposits)				
Collateral quota (marketable securities)				
Credit guarantee quota	−4.924	1.765	−2.790	0.005
ln (EAD)	−0.246	0.104	−2.371	0.018

Table 16
The estimation results for the inflated beta model (ν).

	Estimate	Std. error	t-value	Pr(> |t|)
(Intercept)	−0.665	0.139	−4.803	0.000
Creditworthiness score	0.007	0.003	2.185	0.029
Collateral quota (real estate)	2.166	0.197	10.999	0.000
Collateral quota (commercial bills)	1.043	0.267	3.913	0.000
Collateral quota (deposits)	1.817	0.494	3.675	0.000
Collateral quota (marketable securities)	7.056	2.190	3.223	0.001
Credit guarantee quota	2.485	0.113	22.067	0.000
ln (EAD)	−0.241	0.028	−8.704	0.000

Table 17
Out-of-time predictive performance of set 1 (2006).

	R^2	Spearman's rho	MAE	RMSE	RAE
Multi-stage	0.3203	0.4811	0.0864	0.1783	0.6694
Tobit	0.2561	0.4709	0.0899	0.1886	0.6966
OLS	0.2809	0.4737	0.0922	0.1866	0.7150
Inflated beta	0.2047	0.4659	0.0903	0.1890	0.7003

Table 18
Out-of-time predictive performance of set 2 (2007).

	R^2	Spearman's rho	MAE	RMSE	RAE
Multi-stage	0.3205	0.5013	0.0925	0.1804	0.6804
Tobit	0.2325	0.4818	0.0975	0.1924	0.7171
OLS	0.2931	0.4860	0.0990	0.1871	0.7280
Inflated beta	0.2537	0.4853	0.0960	0.1883	0.7060

Table 19
Out-of-time predictive performance of set 3 (2008).

	R^2	Spearman's rho	MAE	RMSE	RAE
Multi-stage	0.3486	0.5009	0.1115	0.1991	0.6677
Tobit	0.2990	0.4831	0.1158	0.2100	0.6930
OLS	0.3438	0.4964	0.1169	0.2062	0.6995
Inflated beta	0.3234	0.4929	0.1124	0.2035	0.6729

Table 20
Out-of-time predictive performance of set 4 (2009).

	R^2	Spearman's rho	MAE	RMSE	RAE
Multi-stage	0.3304	0.5153	0.1157	0.2065	0.6740
Tobit	0.3424	0.5145	0.1179	0.2109	0.6867
OLS	0.3388	0.4806	0.1219	0.2105	0.7096
Inflated beta	0.3366	0.5078	0.1158	0.2071	0.6746

Table 21
Out-of-time predictive performance of set 5 (2010).

	R^2	Spearman's rho	MAE	RMSE	RAE
Multi-stage	0.3841	0.5464	0.1098	0.2063	0.6377
Tobit	0.3897	0.5371	0.1138	0.2147	0.6610
OLS	0.3670	0.5178	0.1183	0.2149	0.6870
Inflated beta	0.3839	0.5384	0.1104	0.2089	0.6415

8. Conclusions

This study has calculated fundamental data and analyzed the relationship between the influencing factors and LGD. Using the data for Japanese bank loans, we built LGD and EL forecasting models and proposed methods for their estimation, taking into account Japanese banking practices.

We obtained the following results. The LGD levels of Japanese banks are lower than those suggested by FIRB. The duration of the workout process varies between banks, meaning that the length of observation period for the estimation of LGD that is sufficient differs between banks. Collateral, credit guarantees, and EAD are important factors that influence LGD. We confirmed that the multi-stage LGD model has a superior predictive accuracy relative to the OLS, Tobit, and inflated beta regression models.

We then incorporated a recovery model (Pr(Recovery)) into the LGD prediction model to examine the significant differences between the characteristics of recovered and written-off loans. This study considered only Japanese bank data. In view of the significant differences between the characteristics of recovered and written-off loans in Japan, we feel that the LGD prediction model could perform better using bank data from other countries.

The tasks for future research on bank loan credit risks are as follows. This study used only a limited amount of censored data. This could have led to estimation biases when the duration of the workout process impacted LGD significantly. Studies need to model the workout process so as to include censored data in the analysis. By increasing the number of banks that provide data, we can further investigate the differences between banks, determine whether the business sectors of borrowers affect LGD, and consider the downturn LGD estimates required by the Basel Accord. This paper has shown that the LGD level differs from year to year; however, we could not address the relationship between the business cycle and LGD. Considering more data would allow us to survey the impact of the business cycle on LGD.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory* (pp. 267–281). Akademinai Kiado.

Altman, E. I., Resti, A., & Sironi, A. (2001). *Analyzing and explaining default recovery rates. ISDA report*.

Araten, M., Jacobs, M., & Varshney, P. (2004). Measuring LGD on commercial loans: an 18-year internal study. *The RMA Journal*, *86*(8), 96–103.

Asarnow, E., & Edwards, D. (1995). Measuring loss on defaulted bank loans: A 24-year study. *The Journal of Commercial Lending*, *77*(7), 11–23.

Basel Committee on Banking Supervision (2005). Guidance on paragraph 468 of the framework document.

Bastos, J. A. (2010). Forecasting bank loans loss-given-default. *Journal of Banking & Finance*, *34*(10), 2510–2517.

Bellotti, T., & Crook, J. (2012). Loss given default models incorporating macroeconomic variables for credit cards. *International Journal of Forecasting*, *28*(1), 171–182.

Caselli, S., Gatti, S., & Querci, F. (2008). The sensitivity of the loss given default rate to systematic risk: new empirical evidence on bank loans. *Journal of Financial Services Research*, *34*(1), 1–34.

Dermine, J., & De Carvalho, C. N. (2006). Bank loan losses-given-default: A case study. *Journal of Banking & Finance*, *30*(4), 1219–1243.

Felsovalyi, A., & Hurt, L. (1998). Measuring loss on Latin American defaulted bank loans: a 27-year study of 27 countries.

Grunert, J., & Weber, M. (2009). Recovery rates of commercial lending: Empirical evidence for German companies. *Journal of Banking & Finance*, *33*(3), 505–513.

Gürtler, M., & Hibbeln, M. (2011). *Pitfalls in modeling loss given default of bank loans. Technical report, Working Paper*. Technische Universität Braunschweig.

Kawada, A., & Yamashita, S. (2013). The multi stage model for LGD and EL: Empirical research for Japanese banks. *FSA Research Review*, *7*, 1–42. in Japanese.

Loterman, G., Brown, I., Martens, D., Mues, C., & Baesens, B. (2012). Benchmarking regression algorithms for loss given default modeling. *International Journal of Forecasting*, *28*(1), 161–170.

Lucas, A. (2006). Basel ii problem solving. In *QFRMC workshop and conference on Basel II & credit risk modelling in consumer lending*.

Matuszyk, A., Mues, C., & Thomas, L. C. (2010). Modelling LGD for unsecured personal loans: Decision tree approach. *Journal of the Operational Research Society*, *61*(3), 393–398.

McDonald, J. F., & Moffitt, R. A. (1980). The uses of tobit analysis. *The Review of Economics and Statistics*, 318–321.

Miura, K., Yamashita, S., & Eguchi, S. (2010). The statistical model for recovery rates and expected losses in the internal ratings based approach. *FSA Research Review*, (6), 174–205. in Japanese.

Ospina, R., & Ferrari, S. L. (2010). Inflated beta distributions. *Statistical Papers*, *51*(1), 111–126.

Querci, F. (2005). *Loss given default on a medium-sized Italian bank's loans: an empirical exercise*. Milan, Italy: European Financial Management Association.

Sigrist, F., & Stahel, W.A. (2010). Using the censored gamma distribution for modeling fractional response variables with an application to loss given default. arXiv preprint arXiv:1011.1796.

Swearingen, C. J., Castro, M. S. M., & Bursac, Z. (2012). Inflated beta regression: Zero, one and everything in between. In *SAS global forum* (pp. 325–2012). Citeseer.

Yashkir, O., & Yashkir, Y. (2013). Loss given default modeling: a comparative analysis. *The Journal of Risk Model Validation*, *7*(1), 25.

Zhang, J., & Thomas, L. C. (2012). Comparisons of linear regression and survival analysis using single and mixture distributions approaches in modelling LGD. *International Journal of Forecasting*, *28*(1), 204–215.

Yuta Tanoue was received the B.A. degree in Commerce from Waseda University, Tokyo Japan, in 2012. He is now a doctoral course student of the Graduate University for Advanced Studies, Tokyo Japan. His research interests include statistical finance and statistical theory.

Akihiro Kawada received the Ph.D. degree in Engineering from Tokyo University of Science, Tokyo, Japan, in 2015. He works for PricewaterhouseCoopers Aarata, Tokyo, Japan. His research interests include governance, risk and compliance.

Satoshi Yamashita received the Ph.D. degree in Engineering from Kyoto University, Kyoto, Japan, in 1997. He is now a professor of the Institute of Statistical Mathematics, Tokyo, Japan. His research interests include statistical finance and statistical theory.