Loss given default as a function of the default rate

January 27, 2013

Jon Frye
Senior Economist
Federal Reserve Bank of Chicago
230 South LaSalle Street
Chicago, IL 60604
Jon.Frye@chi.frb.org
312-322-5035

The author thanks Greg Gupton, Matt Pritsker, Balvinder Sangha and Jeremy Staum for insightful comments, as well as participants in conferences sponsored by the Federal Reserve Bank of Chicago, Moody's Analytics, and The Financial Engineering Program at Columbia University.

Abstract

A recently derived function ties a portfolio's loss given default rate (LGD) to its default rate. This study compares the predictive performance of the LGD function to that of linear regression using simulated data. The data are simulated using a linear model. Even though this confers an advantage to linear regression, the LGD function produces lower mean squared error over a meaningful range of conditions. This suggests that risk managers can benefit by using the LGD function to model the relationship between default and LGD.

Models of portfolio credit loss contain default rates and loss given default (LGD) rates. If the two rates tend to rise in the same conditions, credit risk is worse than otherwise. Although the connection between default and LGD must be expressed in some way, there has been no standard approach to modeling.

A recently derived LGD function predicts the LGD rate of a portfolio based upon its default rate. If conditions cause the default rate to be elevated, the function predicts that the LGD rate will be elevated by an associated amount. The quantitative connection has survived statistical testing against steeper and flatter alternatives within the Moody's-rated universe of loans and bonds.[1]

This study tests the LGD function with data simulated by linear models. A linear statistical model would seem best to analyze such data, so linear regression is used as a performance benchmark. But when a data set is short, linear regression tends to over fit the data and to make noisy predictions. The LGD function cannot be over fit to the same degree. Over a wide range of conditions, the LGD function therefore outperforms linear regression. This suggests that a risk manager is well served using the LGD function rather than a statistically estimated relationship.

The LGD function

The LGD function connects the conditionally expected LGD rate (cLGD) to the conditionally expected default rate (cDR). These rates would be observed in a homogeneous portfolio with a large number of statistically identical loans. In smaller portfolios, default and LGD rates are random with means equal to cDR and cLGD.

The conditionally expected LGD rate can sometimes be stated as a function of cDR. Frye and Jacobs derive a particularly simple LGD function:

$$cLGD = \Phi\left[\Phi^{-1}[cDR] - k\right]/cDR \tag{1}$$

where $\Phi[\cdot]$ represents the standard normal cumulative distribution function, and k is a positive parameter that characterizes a given loan. The LGD function is strictly monotonic in cDR for all values of k and is bounded on the interval $(0, 1)$.

[1] Frye and Jacobs

Figure 1: LGD Function for seven values of k

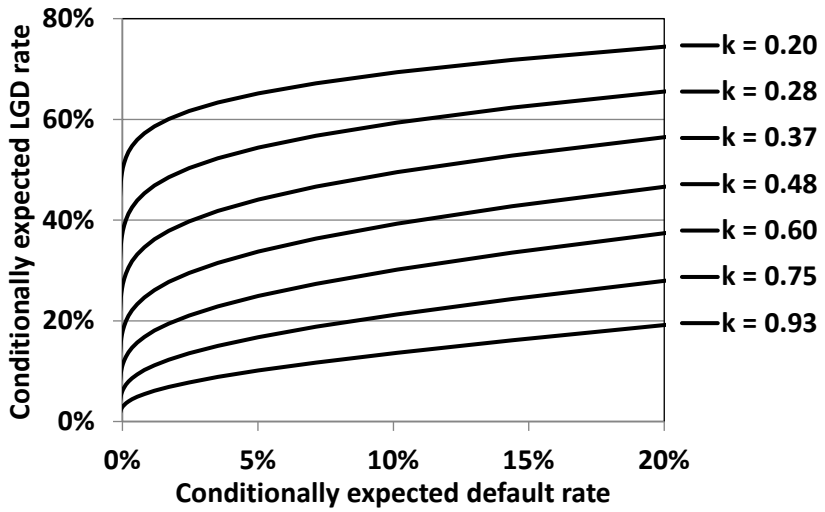


Figure 1 illustrates the moderate, positive LGD function for seven values of *k*. For example, in the range of default rates between 1% and 10%, each of the lines exhibits approximately a nine-point response of cLGD.

Applying the LGD function to a loan is usually straightforward because *k* depends only on parameters that are in common use. Specifically,

$$k = (\Phi^{-1}[PD] - \Phi^{-1}[EL])/\sqrt{1-\rho} \quad (2)$$

where PD denotes a loan's probability of default, EL denotes its expected loss rate, and ρ denotes correlation. EL equals PD times the expected LGD rate (ELGD). A given value of *k* can produced by many combinations of PD, EL, and ρ. Of the three parameters, ρ has least effect.

Some models of the default rate do not condition on observable variables. An example is the widely-used Vasicek Distribution.[2] If cDR has a Vasicek Distribution and the LGD function is valid, then credit loss has a two-parameter distribution:

$$(3) \quad cLoss = cLGD\ cDR\ = \Phi\left[\Phi^{-1}[cDR] - k\right]$$

$$= \Phi\left[\Phi^{-1}\left[\Phi\left[\frac{\Phi^{-1}[\,PD\,] + \sqrt{\rho}\,Z}{\sqrt{1-\rho}}\right]\right] - \frac{\Phi^{-1}[\,PD\,] - \Phi^{-1}[\,EL\,]}{\sqrt{1-\rho}}\right] = \Phi\left[\frac{\Phi^{-1}[\,EL\,] + \sqrt{\rho}\,Z}{\sqrt{1-\rho}}\right];\ Z \sim N[0,1]$$

Specifically, this is a Vasicek Distribution with mean equal to EL. Other specifications of cLGD, even if they appear simpler, tend to produce more complicated distributions of cLoss. For example, if cLGD always equals ELGD, the distribution of credit loss has three parameters:

[2] Vasicek

(4) $$cLoss \;=\; ELGD\;\Phi\left[\frac{\Phi^{-1}[\,PD\,]+\sqrt{\rho}\,Z}{\sqrt{1-\rho}}\right];\; Z\sim N[0,1]$$

In addition to purely random influences, a default rate model might condition on observed variables such as the change in GDP or house prices. The same variables might also affect LGD. Credit loss would then depend on the default rate (with its dependences on underlying variables) and on the LGD rate (with its own dependences).

There is nothing wrong with additional statistical parameters if they are statistically significant. The LGD function makes the testable assertion that available data are not extensive enough to resolve the additional parameters with statistical significance. The simulations performed in this study compare the performance of the simpler LGD function to that of linear regression.

Data Simulation

Data are simulated in two stages. The first stage simulates conditionally expected values, and the second stage adds the randomness inherent in a finite portfolio.

On the default side, the conditionally expected rate is drawn from the Vasicek Distribution. The number of defaults is drawn from the Binomial Distribution with its parameter equal to cDR. On the LGD side, the conditionally expected LGD rate is inferred from a linear function. The LGD rate is drawn from a normal distribution with mean equal to cLGD and variance that depends on the number of defaults. In symbols,

(5) $$Z \sim N[0,1]$$

(6) $$cDR \;=\; \Phi\,[\,(\Phi^{-1}[PD]+\sqrt{\rho}\,Z\,)/\sqrt{1-\rho}\,]$$

(7) $$D \sim Binomial[n, cDR]$$

(8) $$cLGD \;=\; a \;+\; b\,cDR$$

(9) $$LGD \sim N[cLGD, \sigma^2/D]$$

The initial simulations use a particular set of parameter values: PD = 3%, ρ = 10%, and n = 1,000. The values a = 0.5 and b = 2.3 are those fit by Altman and Kuehne to high-yield bond data.[3] The value σ = 20% is provided by Frye and Jacobs. The target for each approach is 98th percentile cLGD, which equals 72.3%:

(10) $$cLGD \;=\; 0.5 + 2.3\;\Phi\,[\,(\Phi^{-1}[0.03]+\sqrt{.1}\,\Phi^{-1}[0.98]\,)/\sqrt{1-.1}\,]$$

$$=\; 0.5 + 2.3 \;*\; 0.0972 \;= 0.723$$

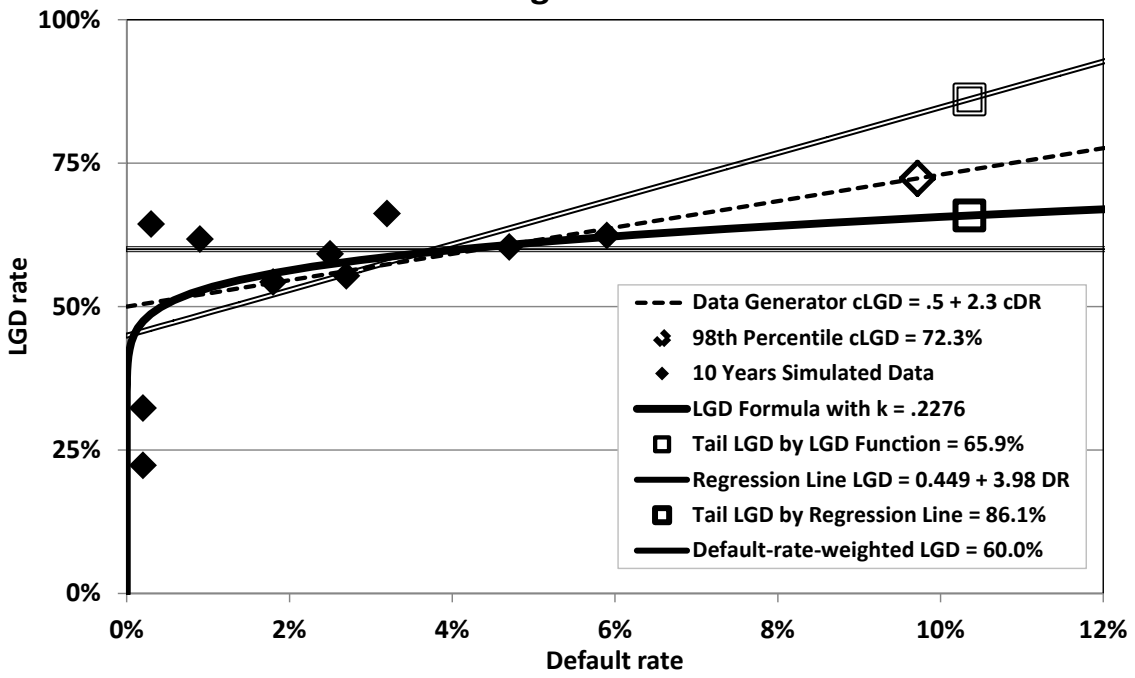
[3] Altman and Kuehne

The length of the data sample is initially set to ten years, because many banks have not had a rigorous definition of default for longer than that. Later experiments explore a range of values for each of the control variables.

Initial simulations

This section details the analysis of one set of data and summarizes the analysis for 10,000 sets. Figure 2 illustrates the data generator, Equation (8), as a dashed line. The 98th percentile is indicated by an open diamond. Ten simulated data points are indicated with solid diamonds.

Figure 2



On the default side, estimated PD is assumed equal to the average annual default rate, 2.24%. Maximizing the following likelihood function produces an estimate of ρ:

$$LnL_{\rho}[\rho] = \sum_{dr_i>0} Log[f_{Vas}[dr_i; \widehat{PD}, \rho] \tag{11}$$

where $f_{Vas}[\,\cdot\,]$ is the PDF of the Vasicek Distribution. For these data the estimate is $\hat{\rho}$ = 17.6%. The estimated 98th percentile of cDR is then

$$\widehat{cDR} = \Phi[\,(\Phi^{-1}[\,0.0224\,] + \sqrt{0.176}\,\Phi^{-1}[0.98]) / \sqrt{1-0.176}\,] = 0.1035 \tag{12}$$

On the LGD side, the LGD function is simple to apply. Estimated EL is the average annual loss rate, 1.34%, which implies k = 0.2276. The LGD function prediction, $\widehat{cLGD}$ = 65.9%, is marked with an open square in Figure 2. It understates true cLGD by 72.3% - 65.9% = 6.4%.

Ordinary least squares (OLS) estimates are $\hat{a}$ = 0.449 and $\hat{b}$ = 3.98. The regression line prediction, $\widehat{cLGD}$ = 86.1%, is marked with an open square and overstates cLGD by 13.8%. However, the regression slope is not statistically significant with a test size of 5%. The regression prediction therefore reverts to an average, and for this we use default-rate-weighted-average LGD, 60.0%. This is an improvement relative to the untested regression, but it understates the target by 12.3%. The error made by OLS is about twice as great as the error made by the LGD function.

Figure 3. 10,000 predictions of cLGD

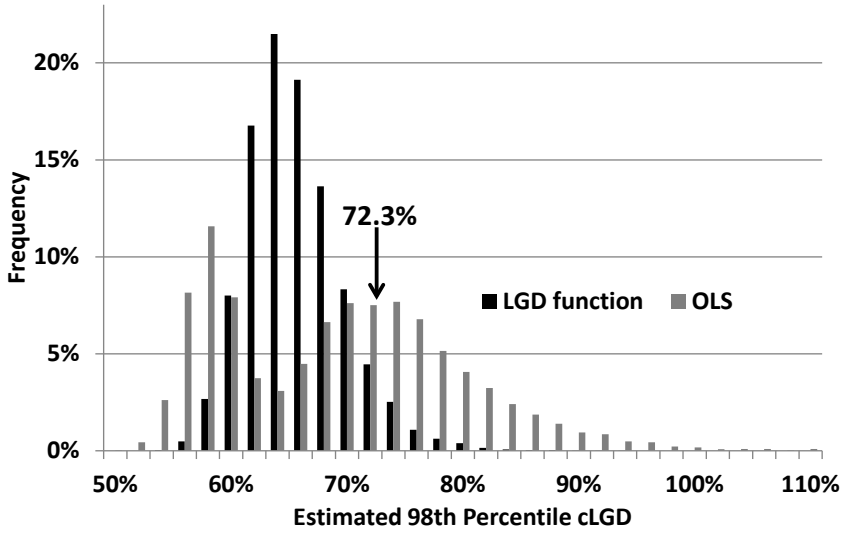


Figure 3 summarizes 10,000 instances of this analysis using randomly generated data sets. Predictions made by the LGD function are tightly distributed and produce root mean squared error (RMSE) equal to 7.9%. Predictions made by OLS range from 49.7% to 133% and produce RMSE equal to 11.0%. The lesser mode reflects primarily non-significant regressions such as the one illustrated in Figure 2.

Thus, the LGD function (RMSE = 7.9%) outperforms OLS (RMSE = 11.0%). This is principally because the LGD function is less affected by the noise that is observed in a short data set.

Robustness

This section allows each control variable to take a range of values. Of the eight control variables, five have little effect on the conclusion that the LGD function outperforms OLS. However, if there are many years of data, or if LGD responds very strongly (or not at all) to conditions, then OLS can sometimes outperform. PD affects the tradeoff between the steepness of the data generator and the relative performance of the two predictive approaches.

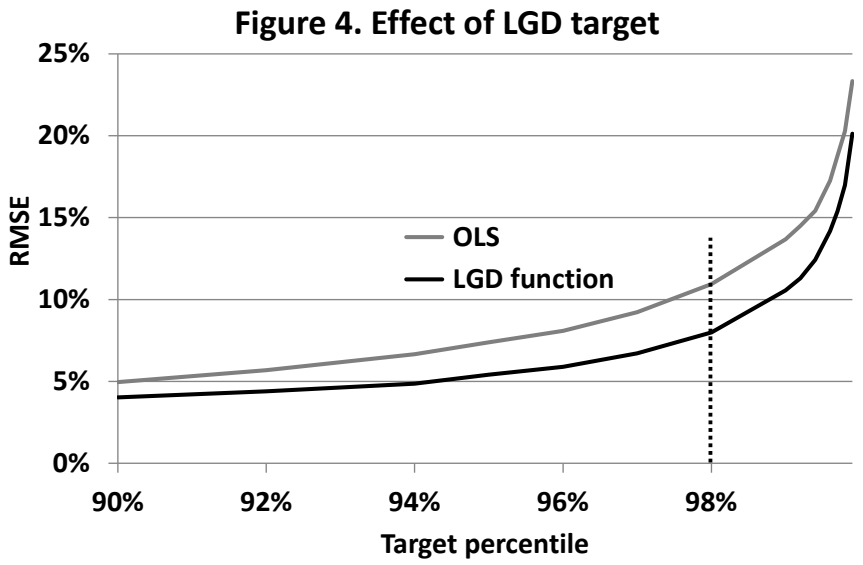
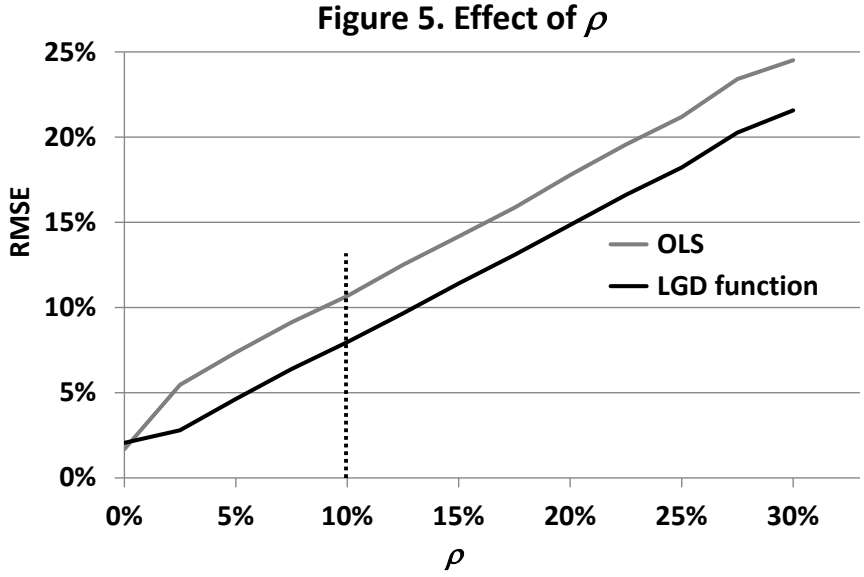
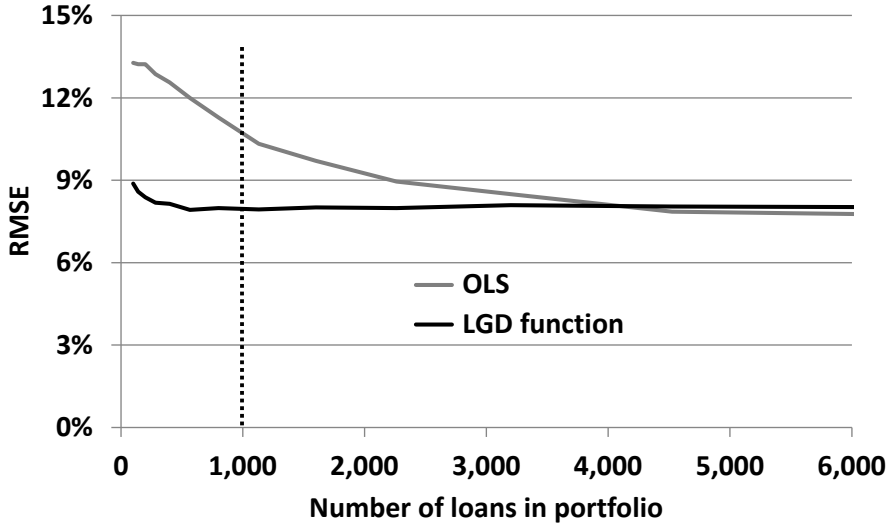


Figure 4 illustrates the effect of the LGD target percentile. The dashed vertical line calls attention to the 98th percentile and to the RMSEs of 7.9% and 11.0%. A greater (lesser) percentile requires greater (lesser) extrapolation from the data and entails greater (lesser) errors by either method. However, the LGD function outperforms OLS at each possible target percentile.



A greater value of ρ implies greater variances of all random variables. Figure 5 shows this leads to greater errors by either method. However, the LGD function outperforms OLS at each possible value of ρ.

Figure 6. Effect of number of loans in portfolio



A greater number of loans reduces the RMSE for OLS until a limit is reached at about 4,000 loans, as shown in Figure 6. Otherwise, the LGD function outperforms OLS.

Figure 7. Effect of intercept of data generator, *a*

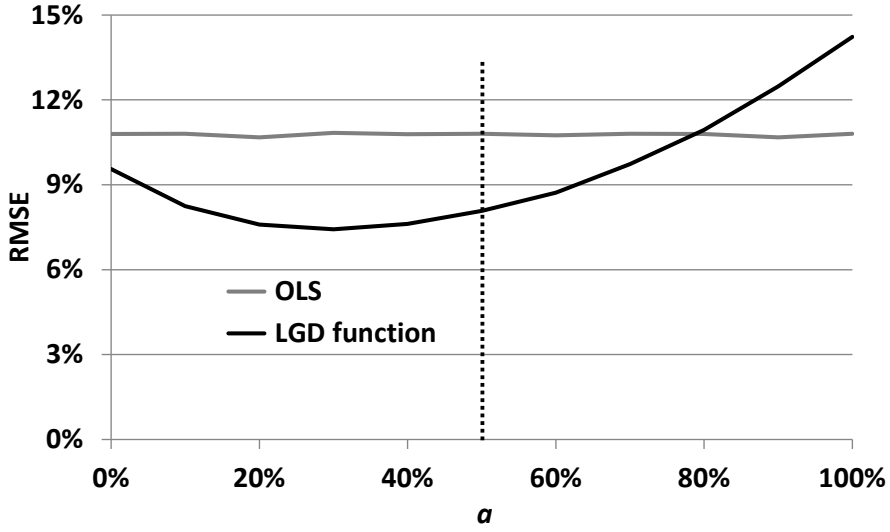
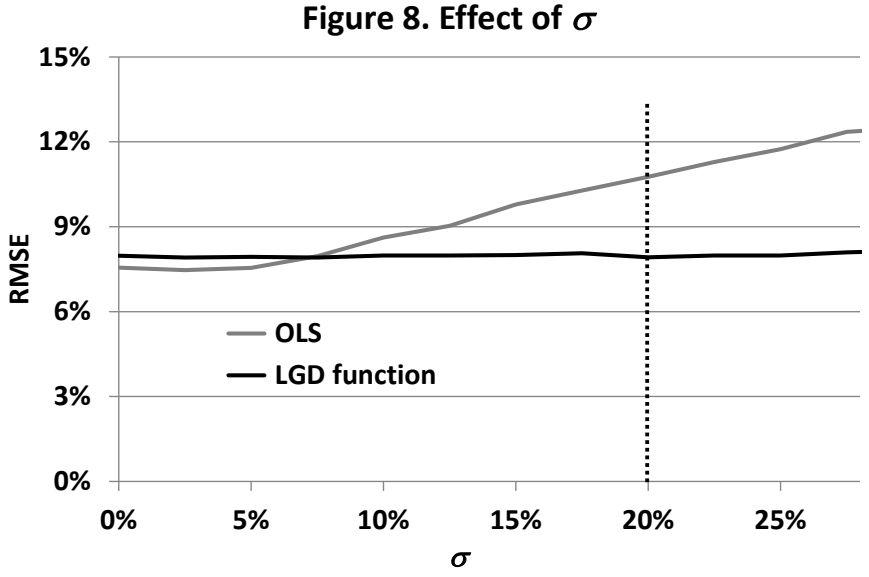
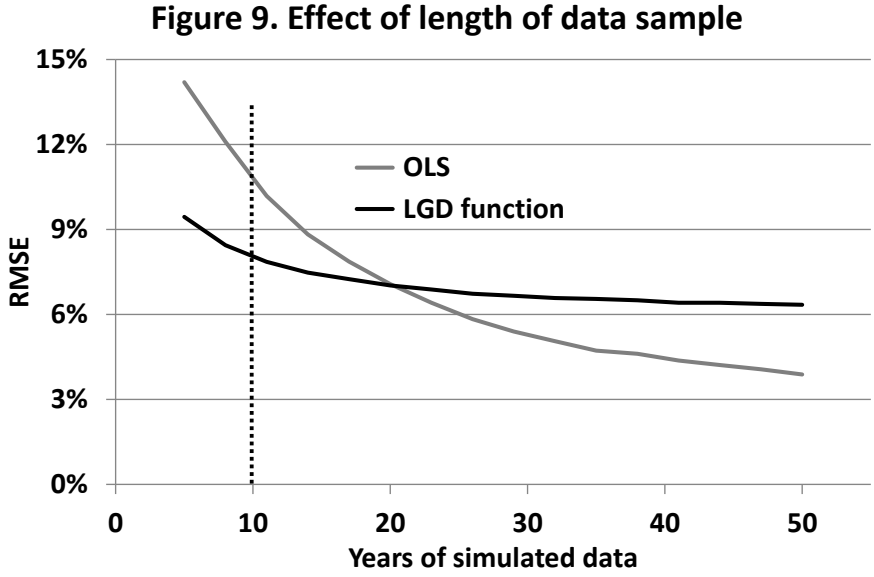


Figure 7 shows that over a broad range of values of intercept *a*, the LGD function outperforms OLS. This range extends from zero up to the value $a = 78.4\%$. Then, 98th percentile cLGD equals $0.784 + 2.3 * 0.097 = 101\%$. This highlights a shortcoming of the linear data generator: it often produces values of cLGD outside the range [0, 1], unlike the LGD function, which is bounded. For more reasonable values of *a*, the LGD function outperforms OLS.



A greater value of σ adds noise and increases the errors of either approach. The initial value, σ = 20%, is less than most LGD studies.[4] Figure 8 shows that at realistic values of σ, the LGD function outperforms OLS.



An increase in the length of the data sample benefits each approach, but regression benefits more. Figure 9 shows that the cross-over, using the initial values of the control variables, is twenty years of data produced by a sequence of random draws. But real-world data is less informative than simulated data because of serial dependence; therefore, more than twenty years of real-world data would be required to produce cross-over.

[4] Jacobs

Figure 10. Effect of slope of data generator

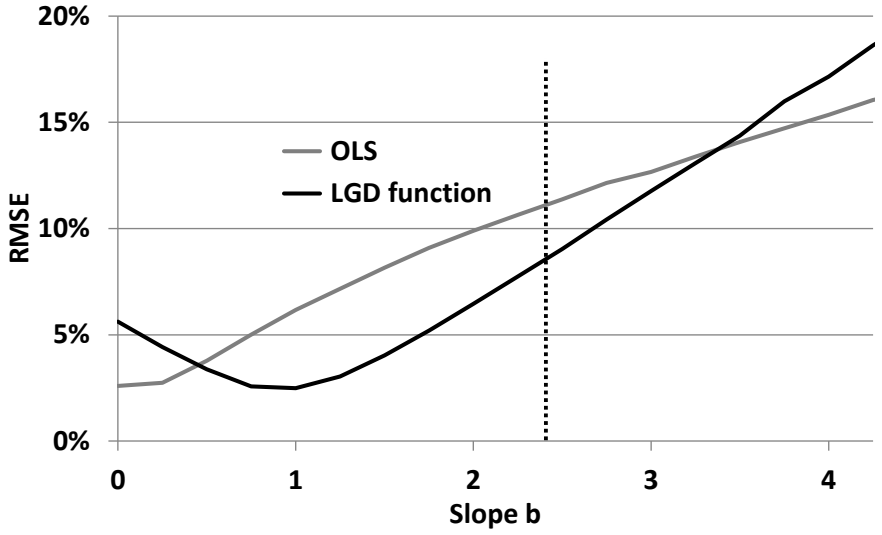
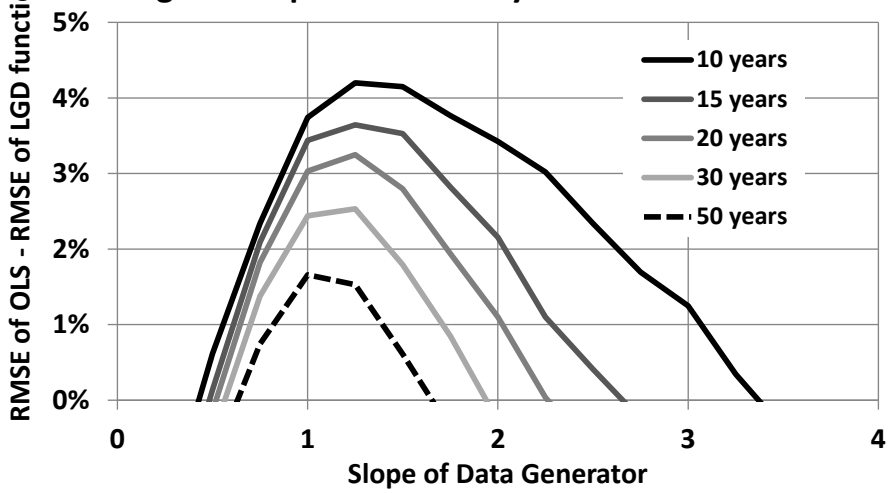


Figure 10 shows that when *b* is nearly zero, the RMSE of OLS is low. This is because few regressions are significant and predictions revert to the average, which is a good predictor under the circumstances. Greater values of *b* entail greater errors by OLS. But greater values of *b* make the data generator a better match for the LGD function. The RMSE of the LGD function declines at first, and this creates a wide range of slopes for which the LGD function outperforms OLS.

Figure 11. Effect of number of years of data on range of outperformance by the LGD function



The top line of Figure 11 shows the difference between the RMSE of OLS and the RMSE of the LGD function when there are ten years of data in the sample. As in Figure 10, the LGD function outperforms OLS for slopes between 0.45 and 3.4. As the number of years increases, the range of outperformance narrows. But even after fifty years of data there remains a range of slopes for which the LGD function outperforms OLS. If the data generator resembles the LGD function itself, outperformance could continue for any number of years.

Figure 12. Effect of PD on the range of *b*

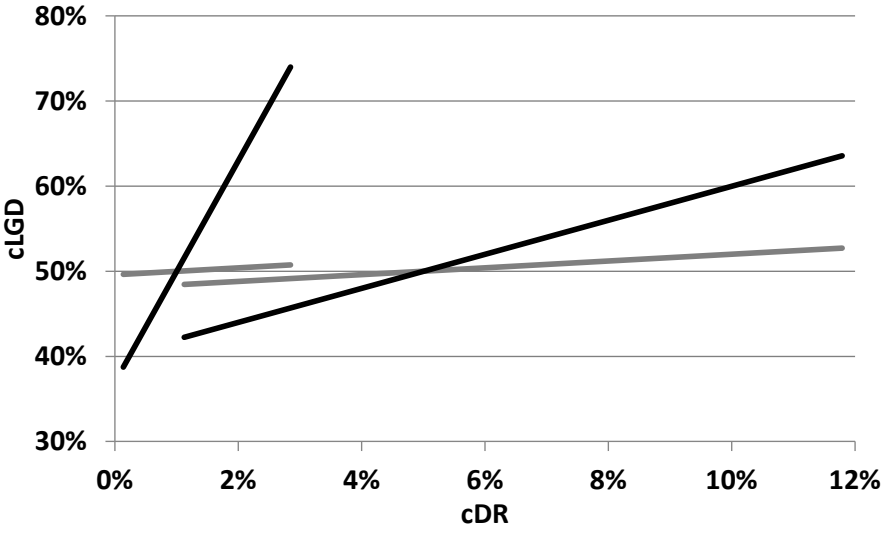


Figure 12 shows the effect of PD on the range of outperformance. At the left, the black line is the steepest data generator passing through (1%, 50%) for which the LGD function outperforms OLS when PD = 1%; its slope is 13.0. The gray line is the shallowest data generator passing through the same point for which the LGD function outperforms. Thus, unless either cLGD has almost no relation to cDR or its sensitivity is extreme, the LGD function produces less error than an estimated relationship. One reason that the range of outperformance is so great is that the portfolio produces few defaults and the LGD data is particularly noisy.

The pair of lines at the right represent the bounds of outperformance when PD = 5%. Although the range of slopes is less, the range of systematic LGD variation is still quite large. The endpoints of all lines in Figure 12 are the 5th and 95th percentiles of the variables. For the steeper bound at the right, the percentiles of cLGD are 42% and 64%. This means that within the central 90% of the distribution of cLGD, a bad year has 150% the LGD of a good year. Thus, a substantial amount of systematic LGD risk would be required for OLS to outperform the LGD function.

This section allows each of the eight control variables to take a range of values. Unless the data sample is longer than currently available to most banks, or unless systematic LGD risk is either near zero or extreme, the LGD function tends to produce lower RMSE than OLS.

Exact regression

The forgoing comparison uses OLS to estimate the relationship between default and LGD. However, default and LGD data unavoidably violate the assumptions under which OLS works best. This section derives the exact distribution and compares the performance of exact regression to the other approaches.

A portfolio's default rate is based on a Binomial Distribution with parameter cDR. The variance of the distribution depends on cDR, so when cDR is high the portfolio default rate is highly random; the observed default rate is therefore a poor guide to cDR. On the other hand, when there are few defaults portfolio LGD is highly random; the observed LGD rate is a poor guide to cLGD. These complications, compared to when OLS assumptions are obeyed, allow random influences to have a greater role shaping data.

The probability density of portfolio average LGD given that the observed number of defaults, D, is greater than zero is symbolized $f_{LGD|D}[LGD]$. It can be derived with two applications of Bayes Rule:

$$(13) \qquad f_{LGD|D}[LGD] = \int_0^1 f_{\{LGD,cDR\}|D}[LGD, cDR]\, dcDR$$

$$= \int_0^1 f_{LGD|cDR,D}[LGD]\, f_{cDR|D}[cDR]\, dcDR$$

$$= \int_0^1 f_{LGD|cDR,D}[LGD]\ f_{D|cDR}[D] f_{cDR}[cDR]\ dcDR \,/\, f_D[D]$$

where $f_{LGD|cDR,D}[LGD]$ is the Normal Distribution of Equation (9), $f_{D|cDR}[D]$ is the Binomial Distribution of Equation (7), $f_{cDR}[cDR]$ is the Vasicek Distribution, and

$$(14) \qquad f_D[D] = \int_0^1 \phi[z] \left(\Phi\left[\frac{\Phi^{-1}[PD] + \sqrt{\rho}\, z}{\sqrt{1-\rho}}\right]\right)^D \left(1 - \Phi\left[\frac{\Phi^{-1}[PD] + \sqrt{\rho}\, z}{\sqrt{1-\rho}}\right]\right)^{n-D} \binom{n}{D} dz$$

The exact distribution contains five parameters to be established by the ten data points. We illustrate with the data points of Figure 2 and take $\widehat{PD} = 2.24\%$ and $\hat{\rho}$ = 17.6% as before. The statistical approach is allowed the additional advantage that it uses the true value of σ rather than an estimate. Maximizing the likelihood in the remaining two parameters produces $\hat{a}$ = 0.543 and $\hat{b}$ = 1.539. These imply that 98th percentile cLGD equals 70.2%. This is a substantial improvement to the estimate produced by OLS before testing, 86.1%.

The tests of the previous sections employed the null hypothesis $b = 0$. This is the null hypothesis most often used in practice because it is part of all statistical toolkits. Yet this hypothesis produces a distribution of credit loss that is more complicated than the one produced by the LGD function itself, as discussed earlier, so this is used in the likelihood ratio test. As with OLS, the exact regression is not statistically significant for the example data at test size 5%. Therefore, the exact-regression estimate of 98th percentile cLGD reverts to the LGD function estimate, 65.9%. Again, this is an improvement compared to the OLS estimate, 60.0%.

Table 1. Exact regression compared to LGD function and OLS			
	All data sets 1,000 cases	Not Significant 582 cases	Significant 418 cases
RMSEs			
LGD function	8.0%	9.3%	5.7%
Exact regression	9.4%	9.3%	9.5%
OLS	10.8%	11.4%	9.9%
Average intermediate estimates			
PD	3.0%	2.6%	3.5%
ρ	9.8%	8.6%	11.5%
ELGD	60.4%	58.5%	63.2%
Tail default rate	9.6%	8.1%	11.7%
Average tail cLGD estimates; target = 72.3%			
LGD function	65.5%	63.4%	68.5%
Exact regression	69.4%	63.4%	77.8%
OLS	69.2%	64.7%	75.5%

Table 1 reports the results of repeating this 1,000 times. Although the exact regression outperforms OLS applied to the same cases, it nonetheless underperforms the LGD function.

The LGD function performs particularly well in cases that exhibit statistical significance for the exact regression. For these cases, RMSE = 5.7%. These cases, compared to others, have on average greater estimates of PD, greater estimates of correlation, and greater estimates of ELGD. Each of these characteristics raises the prediction of the LGD function, and this tends to improve performance. By contrast, exact regression performs no better when it finds significance (9.5%) than when it finds no significance (9.3%).

For these settings of the control variables, the LGD function outperforms exact regression. By continuity, there is a range of slopes of the data generator for which this would occur, and this range extends to slopes at least as low as 1.0, a rough match for the LGD function itself. One might attempt to find a statistical procedure that outperforms the LGD function. Significance tests might be performed at size 1% or 10% or not at all, or other techniques might be tried.

However, statistical tinkering cannot replace facts. There are only a few years of reliable data. Since the data have serial dependence, they are less informative than assumed in many statistical procedures. In a “bad” year, the observed DR is highly dispersed around cDR, which is therefore uncertain. In a “good” year a portfolio does not produce many defaults, so cLGD is uncertain. When a short, serially dependent, noisy data set is subject to statistical modeling, large errors are apt to result.

In many situations large errors must be accepted because there is no alternative. In the case of systematic LGD risk, an LGD function has been derived from simple assumptions. It expresses a moderate, positive relationship between default rates and LGD rates. It can produce lower

errors than statistical analysis even if the statistical analysis uses the distribution that generates the data, and even if it finds statistically significant results.

Conclusion

A portfolio credit loss model must express some connection between default rates and LGD rates. The connection might be established by a free-form statistical technique or by the LGD function that expresses a moderate, positive relationship between default and LGD. Of the two, the LGD function can be applied more readily because it depends only on parameters that are already in common use.

This simulation study compares the predictions of the two approaches using simulated data. Data are simulated with a linear model. Still, the nonlinear LGD function produces lower RMSE than linear regression when two not unlikely conditions hold: the data sample must be short and the sensitivity of LGD to default must be neither zero nor extreme. Until there is evidence that statistical techniques can outperform it, risk managers can use the LGD function to avoid introducing unnecessary parameters into their models and unneeded noise into their predictions.

References

Altman, E. I., and Kuehne, B. J., Defaults and Returns in the High-Yield Bond and Distressed Debt Market: The Year 2011 in Review and Outlook. Report, New York, University Salomon Center, Leonard N. Stern School of Business, 2012.

Frye, J. and Jacobs, M., Credit loss and systematic loss given default, The Journal of Credit Risk (1–32) Volume 8/Number 1, Spring 2012

Jacobs, LGD paper to be determined

Vasicek, O. (2002). Loan portfolio value. Risk 15(12), 160–162.