Loss given default as a function of the default rate

10 September 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, Jeremy Staum, and Dirk Tasche 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. Two anonymous referees contributed important suggestions.

Risk managers have used complex models or ad-hoc curve fitting to incorporate LGD risk into their models. Here, Jon Frye provides a function that is simpler to use and which works better.

Credit loss models contain default rates and loss given default (LGD) rates. If the two rates respond to the same conditions, credit risk is greater than otherwise. The risk affects loan pricing, portfolio optimization and capital planning.

A study by Frye and Jacobs predicts LGD as a function of the default rate. Their function does not require a user to calibrate new parameters. Models that require such calibration do not significantly improve the description of instrument-level data.

This study compares the LGD function to earlier LGD models and tests it with thousands of sets of simulated data. The comparison shows that the earlier models resemble a version of the LGD function that was not found to be statistically significant. The simulations show that the predictions of the LGD function are more accurate than those of regression and may remain more accurate for decades. Risk managers appear better served by the LGD function than by statistical models calibrated to available data.

The LGD function

The LGD function connects the conditionally expected LGD rate (cLGD) to the conditionally expected default rate (cDR). These are the rates that would be observed in an asymptotic portfolio. The asymptotic portfolio is an abstraction, like the perfect vacuum or absolute zero. It contains an infinite number of loans of which each has the same probability of default (PD) and each has the same expected loss (EL).

To derive the LGD function, suppose that cDR has a Vasicek Distribution. The associated cumulative distribution function (CDF) provides the quantile, q:

(1) $$q = F_{cDR}[cDR] = \Phi\left[\frac{\sqrt{1-\rho}\,\Phi^{-1}[cDR] - \Phi^{-1}[PD]}{\sqrt{\rho}}\right]$$

where $\Phi[\cdot]$ is the CDF of the Normal Distribution and $\Phi^{-1}[\cdot]$ is the inverse CDF. Suppose that the conditionally expected loss rate (cLoss) obeys a comonotonic Vasicek Distribution with the same value of ρ. Then cLoss can be stated as a function of cDR:

(2) $$cLoss = F_{cLoss}^{-1}[q] = \Phi\left[\frac{\Phi^{-1}[EL] + \sqrt{\rho}\,\Phi^{-1}[q]}{\sqrt{1-\rho}}\right] = \Phi\left[\Phi^{-1}[cDR] - \frac{\Phi^{-1}[PD] - \Phi^{-1}[EL]}{\sqrt{1-\rho}}\right]$$

Dividing Equation (2) by cDR produces the LGD function:

(3) $$cLGD = \Phi\left[\Phi^{-1}[cDR] - k\right]/cDR; \quad \text{where } k = (\Phi^{-1}[PD] - \Phi^{-1}[EL])/\sqrt{1-\rho}$$

Thus, a loan's PD, ρ, and EL imply the value of its LGD Risk Index, k, which fully determines its LGD function.

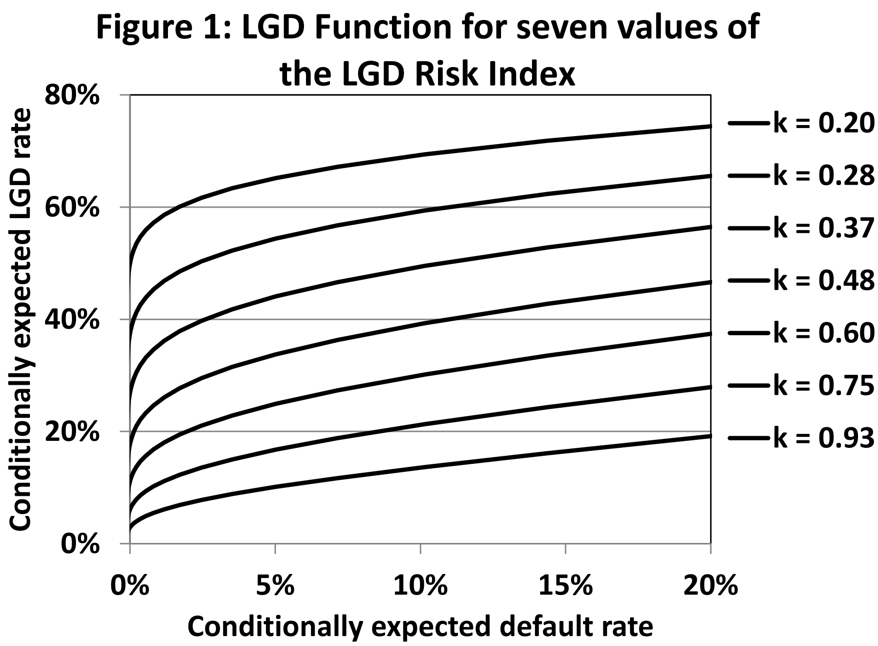


Figure 1 illustrates the LGD function for seven values of the LGD Risk Index. In each instance, cLGD has approximately the same moderate, positive sensitivity to cDR.

The LGD function says that if conditions produce an elevated value of cDR, they also produce an elevated value of cLGD. This fills a gap because LGD modeling is subject to significant difficulties that trace back to data scarcity. We restrict attention to the connection between cDR and cLGD without denying that other variables might be discovered to make a contribution.

Many banks have estimates of EL, ρ, and PD. EL should be part of the spread charged on any loan. Correlation, ρ, is probably the most common measure of dispersion. EL and ρ may be enough to describe the distribution of loss in the asymptotic portfolio, according to Frye (2010). To decompose the distribution of loss into variables default and LGD, EL must be decomposed into expectations PD and ELGD. The values of PD, ρ, and EL are so important that a minor industry now supplies estimates.

Earlier LGD models

Several earlier models involve the rates of LGD and default. This section compares the LGD functions of five of them to the present one. Doing so reveals a strong similarity. (The LGD functions are derived in a mathematical appendix that is available here: http://www.chicagofed.org/webpages/people/frye_jon.cfm#.)

Table 1. Frye-Jacobs and five earlier LGD models.

Model	Implied LGD Function	Parameter values illustrated in Figure 2
Frye-Jacobs	$\Phi[\Phi^{-1}[cDR]-k]/cDR$ $k=$ LGD risk index $=(\Phi^{-1}[PD]-\Phi^{-1}[EL])/\sqrt{1-\rho}$	$k=0.470$
Frye (2000)	$1-(\mu+\sigma q(\sqrt{1-\rho}\,\Phi^{-1}[cDR]-\Phi^{-1}[PD])/\sqrt{\rho})$ $\mu=$ recovery mean, $\sigma=$ recovery SD, $q=$ recovery sensitivity	$\mu=0.696$ $\sigma q=0.0447$
Pykhtin	$\Phi\left[\frac{-\mu/\sigma-\beta Y}{\sqrt{1-\beta^2}}\right]-Exp\left[\mu+\sigma\beta Y+\frac{\sigma^2}{2}(1-\beta^2)\right]\Phi\left[\frac{-\mu/\sigma-\beta Y}{\sqrt{1-\beta^2}}-\sigma\sqrt{1-\beta^2}\right];$ $Y=(\Phi^{-1}[PD]-\sqrt{1-\rho}\,\Phi^{-1}[cDR])/\sqrt{\rho}$ $\mu=$ log recovery mean, $\sigma=$ log recovery SD, $\beta=$ recovery correlation	$\mu=$ -0.384 $\beta=0.251$ $\sigma=0.3$
Tasche	$\int_{-\Phi^{-1}[cDR]}^{\infty}\phi[z]\,BetaCDF^{-1}[\frac{\Phi[\sqrt{1-\rho}\,\Phi^{-1}[cDR]-\Phi^{-1}[PD]+\sqrt{1-\rho}\,z]-1+PD}{PD},$ $a=\frac{ELGD(1-v)}{v}, b=\frac{(1-ELGD)(1-v)}{v}]\,dz\,/\,cDR$ ELGD = expected LGD; $v=$ fraction of maximum variance of Beta distribution	ELGD = 0.333 $v\rightarrow 1$
Giese	$1-a_0(1-PD^{a_1})^{a_2}$ a_1, a_2, a_3 = values to be determined	$a_0=0.872$ $a_1=0.278$ $a_2=0.5$
Hillebrand	$\int_{-\infty}^{\infty}\Phi[a-\frac{bdc}{e}+\frac{bd}{e}\,\Phi^{-1}[cDR]-b\sqrt{1-d^2}\,x]\,\phi[x]\,dx$ a, $b=$ parameters of cLGD in second factor; $d=$ correlation of latent factors; $c=\Phi^{-1}[PD]/\sqrt{1-\rho};\ e=\sqrt{\rho}/\sqrt{1-\rho}$	$a-\frac{bdc}{e}=0.253$ $\frac{bd}{e}=0.422$ $b\sqrt{1-d^2}=0.5$

Table 1 details the LGD functions. They arise from diverse premises. Frye (2000) assumes that recovery is a linear function of the normal risk factor associated to the Vasicek Distribution. Pykhtin parameterizes the amount, volatility, and systematic risk of a loan's collateral and infers the loan's LGD. Tasche assumes a connection between LGD and the systematic risk factor at the loan level; the idiosyncratic influence is integrated out. Giese makes a direct specification of the functional form linking cLGD to cDR. Hillebrand introduces a second systematic factor that is integrated out to produce cLGD given cDR.

Figure 2.
Six LGD functions with chosen parameter values

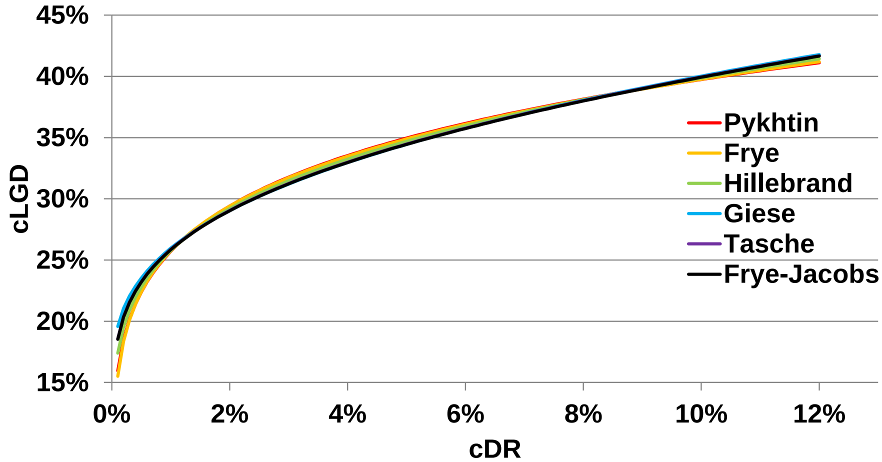


Figure 2 illustrates the six LGD functions. Each function reflects a loan with PD = 3% and ρ = 10%. Setting EL = 1% fully determines the Frye-Jacobs LGD function. The other functions have the chosen parameter values shown in Table 1. Clearly, each of the earlier models can closely approximate Frye-Jacobs for a loan having PD = 3%, ρ = 10%, and EL = 1%. Experimentation suggests that any of the earlier models can closely agree with Frye-Jacobs for a wide range of PD, ρ, and EL.

Thus, compared to Frye-Jacobs, each of the earlier models asserts that something else matters. Instead of ELGD alone, earlier models say that two or three LGD parameters are needed. The extra parameter(s) make the earlier models more flexible than Frye-Jacobs, and this flexibility makes them more attractive to some workers.

Careful workers, however, require a model that displays statistical significance. A model lacking significance is likely to make Type 1 Error; it has inputs that are not relevant. This causes managers to make the wrong decisions, because their decisions are based on the wrong factors. Irrelevant factors are worse than nothing. They actively throw off the results by calibrating to the noise of a data set, rather than to the signal.

To investigate the significance of an earlier model, all the parameters can be freely fit to historical data. Separately, the parameters can be restricted to values that make the model close to Frye-Jacobs. A careful risk manager would use the simpler model of Frye-Jacobs unless the difference in in explanatory power were shown to be significant.

Such tests have been performed using specially created alternatives. Frye and Jacobs' Alternative A has the following form:

$$(4) \quad LGD_A[\mathrm{DR}] = (\mathrm{EL/PD})^a\, \Phi\left[\Phi^{-1}[cDR] - \frac{\Phi^{-1}[PD] - \Phi^{-1}[EL/(\mathrm{EL/PD})^a]}{\sqrt{1-\rho}}\right] / cDR$$

When parameter a is set to zero, Alternative A equals the Frye-Jacobs function. Other values of a produce an LGD function that has the same EL but is steeper or flatter. Calibrating to fourteen years of senior secured loans in five rating grades, a equals 0.01. This is very close to the Frye-Jacobs LGD function and very far from statistical significance. It seems doubtful that any of the earlier models would display statistical significance if calibrated to the same data, though the detailed tests are left for later research.

Data Simulation

Real world data are the standard against which any scientific hypothesis must be judged. However, real world credit loss data, such as used in the tests described above, have a number of shortcomings. Credit model researchers do not publish data as in other sciences. The effects of the assumptions made while handling the data are therefore hard to judge. Critics can claim that results are driven by data imperfections, but these claims can neither be established nor refuted. Many such shortcomings are overcome by using data simulated from fully specified structures.

Here the simulated data are used to make competing predictions of tail LGD. The predictors are the LGD function and linear regression. Simply to give linear regression an advantage in this contest, we generate the data with a linear model. Thus, the nonlinear LGD function competes against a linear model when a linear model has generated the data. Despite this uneven start, the LGD function performs better over a wide range of conditions. These conditions include samples of data longer than those that will be available this decade.

A year's cDR is drawn from the Vasicek Distribution. The number of defaults, D, has the Binomial Distribution with probability equal to cDR. The year's cLGD is inferred from a linear function of cDR. Portfolio average LGD, denoted simply as LGD, is drawn from a distribution with mean equal to cLGD and variance that depends inversely on the number of defaults.

Portfolio average LGD has a normal distribution when there are many defaults, according to the Central Limit Theorem. Researchers sometime restrict individual LGDs to the interval [0, 1]. However, some historical LGDs lie outside the interval, and Frye and Jacobs report that some annual average LGDs also lie outside it. Therefore, we do not restrict LGD to [0, 1], and we use the normal distribution to simulate it.

Stating this in symbols, the simulation of a single year of data proceeds as follows:

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

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

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

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

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

A complete data set consists of T years of (D, LGD). From these data the LGD function and linear regression make their respective predictions of tail cLGD. Since we know the true value of tail cLGD, it is easy to determine the winner.

Altogether there are eight control variables. Each of the control variables is allowed a range of values in a later section, but the initial simulations use the common 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 their heterogeneous set of high-yield bonds. The value σ = 20% is provided by Frye and Jacobs. Analysis is initially conducted at the 98th. At that percentile, cDR is 9.72%:

(10) $$cDR \;=\; \Phi\,[\,(\Phi^{-1}[0.03] + \sqrt{.1}\;\Phi^{-1}[0.98]\,)/\sqrt{1-.1}\,] \;=\; 0.0972$$

The target of the comparison is then 98th percentile cLGD, which equals 72.3%:

(11) $$cLGD = \;\; 0.5 + 2.3\, *\, 0.0972 \;= 0.723$$

The eighth control variable, T, is set to ten years. This is because many banks established rigorous definitions of default, and began to measure the LGDs of loans, less than ten years ago.

Initial simulations

Using the forgoing set of values of the control variables, this section details one simulation run and summarizes the analysis of 10,000 runs.

Figure 3. A selected simulation run

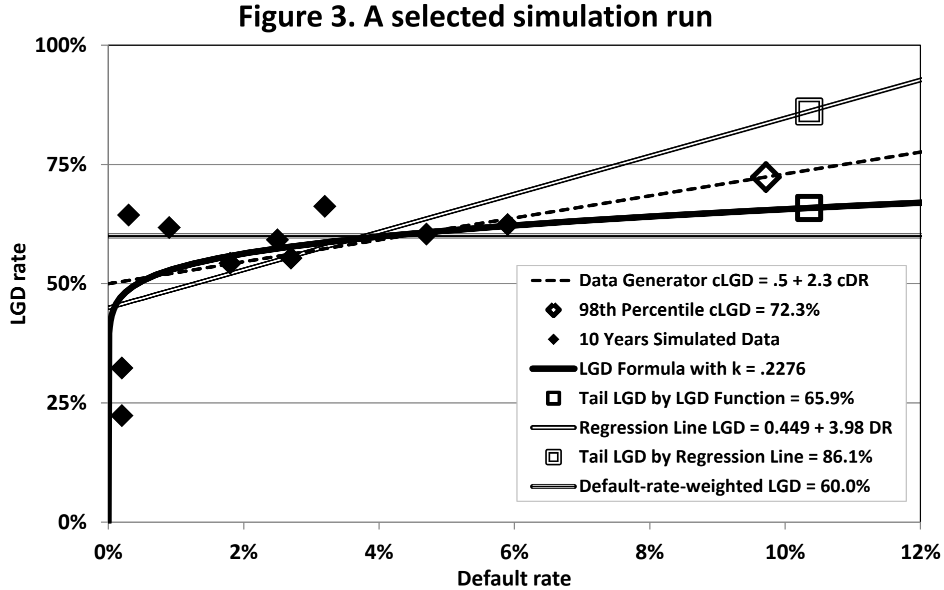


Figure 3 illustrates the data generator, Equation (8), as a dashed line. The 98th percentile is indicated by an open diamond. Ten simulated data points are indicated by solid diamonds.

Analyzing the simulated data, we estimate PD is the average annual default rate, 2.24%. Maximizing the following likelihood function produces the estimate $\hat{\rho}$ = 17.6%:

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

where f_{Vas} [·] is the probability density of the Vasicek Distribution. The estimated 98th percentile of cDR is then

(13) $$c\widehat{DR} \ = \ \Phi[\ (\Phi^{-1}[\ 0.0224\ ] + \sqrt{0.176}\ \Phi^{-1}[0.98])\ /\ \sqrt{1-0.176}\ ] \ = \ 0.1035$$

The LGD function is easy to apply. Estimated EL is the average annual loss rate, 1.34%. This implies k = 0.2276. The LGD function prediction is then $c\widehat{LGD}$ = 65.9%. 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, $c\widehat{LGD}$ = 86.1%, is marked with an open square. It overstates cLGD by 13.8%. However, the slope of the regression is not statistically significant with a test size of 5%. The regression prediction therefore reverts to an average. For this we use default-rate-weighted-average LGD, 60.0%. This is an improvement relative to the untested regression, but in the end OLS understates the target by 12.3%. This error is about twice as great as the error made by the LGD function.

This example illustrates that even in the best of circumstances, LGD data are far from ideal. Ten data points are not much. Most of the ten will tend to be "good" years in which the default rate is low, there are few defaults, and portfolio LGD might be high or low depending on the luck of a few draws. More rarely, cDR is elevated. Then, the variance of D is elevated, and the observed default rate can be a poor reflection of conditions.

Figure 4. Summary of 10,000 simulation runs

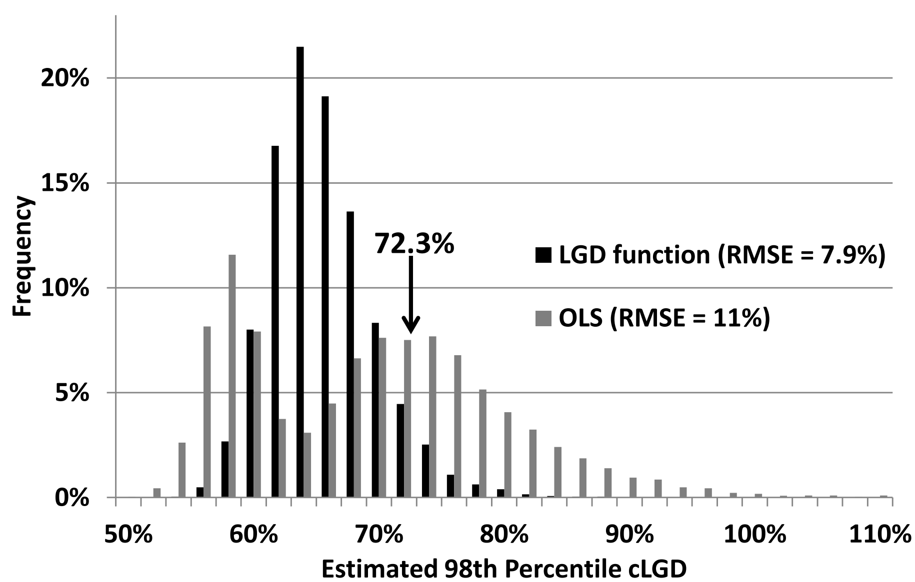


Figure 4 summarizes 10,000 simulation runs using the initial values of the control variables. Predictions made by the LGD function are tightly distributed, while those made by OLS range from 50% to well over 100%. Of the OLS predictions, the lesser mode reflects mostly regressions lacking significance; although regression itself is unbiased, after testing it is biased downward. Overall, the LGD function (root mean squared error (RMSE) = 7.9%) is more accurate than OLS (RMSE = 11.0%) for the initial set of values of the control variables.

Robustness

This section allows each control variable to take a range of values. Throughout the ranges shown in Table 2, the LGD function produces more accurate predictions than OLS.

Of the eight variables, five have little effect on the conclusion. Two variables can reverse it: if there are many years of data or if LGD responds very strongly to default, OLS can sometimes outperform the LGD function. The final variable, PD, affects the tradeoff between these variables and the relative performance of the two predictive approaches. More detail is available in a working paper: http://www.chicagofed.org/webpages/people/frye_jon.cfm#.

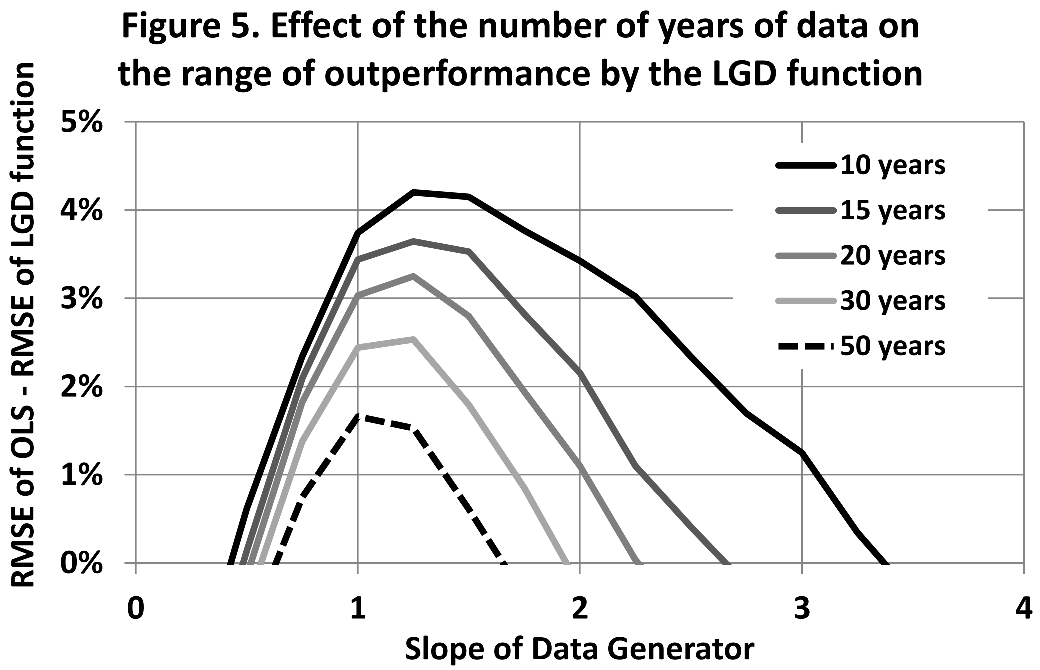
Table 2. Range of parameter values in robustness checks

Variable	Description	Initial Value	Range of Values
q	LGD target quantile	98%	90% – 99.9%
ρ	Correlation	10%	0% – 50%
n	Number of loans in portfolio	1,000	0 – 10,000
σ	Standard deviation of individual LGD	20%	0% – 30%
a	Intercept of data generator	50%	0% – 78%
T	Number of years of data	10 years	0 – 20 years
b	Slope of data generator	2.3	0.45 – 3.4
PD	Probability of default	3%	0% – 3%

The ranges shown in Table 2 include practical situations. Risk models are rarely developed for outcomes less extreme than the 90^{th} percentile. In a literature review Chernih and co-authors find no estimates of correlation greater than 21%. The data of Frye and Jacobs contain no year with as many as 1,000 bonds or 1,000 loans, and Altman's high-yield universe has had more than 1,000 bonds for less than ten years. If the intercept of the data generator is 78%, then target cLGD is 0.78 + 2.3 * 0.0972 = 101%, an exceptional, even unrealistic, situation. Few banks have had definitions of default for more than 20 years, let alone long histories of loss given default. Frye and Jacobs estimate greater, but not significantly greater, systematic LGD risk in rated bonds than in rated loans.

Beyond the ranges shown in Table 2, T and *b* can reverse the usual conclusion. With a long enough data set, OLS eventually outperforms the LGD function. Table 2 shows that this occurs when the number of independently simulated data points exceeds 20, given the values of the other control variables.

The slope of the data generator, *b*, also affects relative accuracy. The LGD function performs best when the slope of the data generator is approximately 1.0. If the slope of the data generator is much steeper, as in the initial set of simulations, the LGD function tends to under predict. If the data generator is steep enough, OLS can outperform the LGD function. OLS can also outperform if the data generator is very shallow. That is because few of the regressions display statistical significance, and the regression reverts to the forecast that systematic LGD risk does not exist.



Variables T and *b* interact as illustrated in Figure 5. The top line shows that with ten years of data the LGD function is more accurate than OLS for a range of slopes from 0.45 to 3.4. As the number of years increases, the length of this range declines. But even with 50 years of simulated data, the LGD function continues to produce more accurate predictions than OLS if the data generator has a moderate, positive slope similar to the LGD function.

Figure 5 understates the real-world data requirement. A year of real-world data is less informative than an independent draw, because each year tends to resemble the previous one. The lines in Figure 5 would be higher, and the ranges of outperformance wider, if it were based on simulations containing serial dependence like real-world data.

The value of the last control variable, PD, affects the tradeoff between T and *b*. If PD takes lower values, there are fewer defaults and fewer LGDs, and regression can discover less about their connection. If PD is greater than 3%, regression can outperfrom the LGD function sooner or with a shallower data generator.

In these simulations, data from one linear model are analyzed by another linear model. The resulting predictions are outperformed by the curved LGD function for broad ranges of control variables. To reverse this conclusion appears to require decades more data than currently available. But even decades of data might not be sufficient. If the data generator itself were curved, a linear statistical procedure might never outperform the LGD function.

Exact regression

The forgoing experiments use OLS to estimate the relationship between default and LGD. Although OLS is the most common method of linear regression, the simulated data violate the assumptions under which OLS works best. This section derives the exact regression and compares its performance to the other approaches.

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

$$(14) \qquad f_{LGD|D}[LGD] = \int_0^1 f_{\{LGD,cDR\}|D}[LGD, cDR]\ dcDR$$

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

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

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 PDF of the Vasicek Distribution, and

$$(15) \qquad f_D[D] = \int_0^1 \phi[z] \left(\Phi\left[\frac{\Phi^{-1}[PD] + \sqrt{\rho}\, z}{\sqrt{1-\rho}}\right]\right)^D \left(1 - \Phi\left[\frac{\Phi^{-1}[PD] + \sqrt{\rho}\, z}{\sqrt{1-\rho}}\right]\right)^{n-D} \binom{n}{D}\ dz$$

Equation (14) contains five parameters. We illustrate with the data points of Figure 3 and take $\widehat{PD} = 2.24\%$ and $\hat{\rho} = 17.6\%$ as before. We give the statistical approach the true value of σ, which is 20%. Maximizing the likelihood produces $\hat{a} = 0.543$ and $\hat{b} = 1.539$. These imply that 98th percentile cLGD equals 70.2%. This is an improvement to the OLS prediction of 86.1%. However, regression is not a significant improvement to the simpler LGD function. The exact-regression prediction therefore reverts to the LGD function prediction, 65.9%.

Table 3. Exact regression compared to LGD function and OLS

	Root mean squared error		
	1,000 regressions all cases	582 regressions not significant	418 regressions with significance
LGD function	8.0%	9.3%	5.7%
Exact regression	9.4%	9.3%	9.5%
OLS	10.8%	11.4%	9.9%

Table 3 reports the results of 1,000 independent runs. Over all, the LGD function produces more accurate predictions than exact regression.

When the exact regression is not significant, the two approaches are identical by definition. When the exact regression is significant it performs worse than otherwise, but the LGD function performs particularly well. This is because this collection of cases has greater estimates of PD, correlation, and ELGD. These elevate the predictions of the LGD function and partly offset its tendency to under predict for a steep data generator like this one.

Overall, exact regression is outperformed by the LGD function. Although exact regression uses data more efficiently than OLS, improvements in statistical technique cannot substitute for data.

Data shortcomings are therefore the impediment to modeling LGD risk. There are only a few years of real-world data at present. These data have serial dependence, so they are less informative than independent draws. In most years, little is learned about LGD because there are few defaults. When such a short, serially dependent, noisy data set is subject to statistical modeling, large errors and low significance are the result. The simpler LGD function, which uses the data less intensely, performs better.

Conclusion

Every model containing a default rate and an LGD rate must connect them in some way. The connection can be expressed by a recently introduced LGD function. The function introduces no new parameters; therefore, it can be applied readily.

This study compares the accuracy of the LGD function to linear regression. To give regression an advantage, the data are simulated with a linear model. Despite this, the LGD function produces more accurate predictions when two conditions hold: the data sample is less than a few decades and the sensitivity of LGD to default is not extreme. Both conditions hold in practice. Now and perhaps for several more decades, risk managers can use the LGD function to avoid unnecessary parameters in their models and to avoid unnecessary noise in their forecasts.

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.

Chernih, A., Vanduffel, S., and Henard, L. (2006). “Asset correlations: a literature review and analysis of the impact of dependent loss given defaults,” http://www.econ.kuleuven.be/insurance/pdfs/CVH-AssetCorrelations_v12.pdf

Frye, Jon. “Depressing Recoveries,” *Risk*, November 2000, pages 108-111.

Frye, Jon. “Modest Means,” *Risk*, January 2010, pages 94-98.

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

Giese, Guido, “The impact of PD/LGD correlations on credit risk capital,” *Risk*, April 2005, pages 79-84.

Gordy, Michael B. (2003). "A Risk-Factor Model Foundation for Ratings-Based Bank Capital Rules," *Journal of Financial Intermediation*, vol. 12, no. 3, pp. 199-232.

Hillebrand, Martin. “Modeling and estimating dependent loss given default”, *Risk*, September 2006, pages 120-125.

Pykhtin, Michael. “Unexpected recovery risk,” *Risk*, August 2003, pages 74-78.

Tasche, Dirk. “The single risk factor approach to capital charges in case of correlated loss given default rates,” Working Paper, February 17, 2004, http://papers.ssrn.com/sol3/papers.cfm?abstract_id=510982

Vasicek, O. (2002). “Loan portfolio value,” *Risk*, December 2002, pages 160-162.