Loss given default as a function of the default rate

September 5, 2012

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

Comments on this draft are welcome. Please to not forward, quote, or publish this draft.

The author thanks Jeremy Staum for insightful comments.

Abstract

A recent formula ties a portfolio's loss given default (LGD) rate to its default rate. This study compares the formula to the statistical relationship modeled by linear regression. Data are simulated with a linear model, which gives an advantage to the regression approach. Despite this, the LGD function produces lower mean squared error over an economically meaningful range of parameter values. Unless a data set is extensive enough to show that statistical uncertainty is properly controlled, this suggests that risk managers can benefit by using the LGD formula to model the relationship between default and LGD.

The views expressed are the solely author's and do not necessarily represent the views of the management of the Federal Reserve Bank of Chicago or the Federal Reserve System.

The relationship between the default rate and the loss given default rate (LGD) is part of every portfolio loss model. A model that assumes independence produces less risk than a model that assumes the rates respond to the same conditions. Although every model must express some connection between default and LGD, at present there is no modeling standard. For example, the latest report by Edward Altman shows four different lines fit to data on high-yield bonds: linear, quadratic, logarithmic, and exponential.[1] Any of the lines might be used to interpolate, but the lines extrapolate to different levels of risk.

Banks have many questions about the LGD/default relationship because bank loans are diverse by borrower rating and by loan facility. Numerous regressions might be performed, but the data can reflect only recent years when banks have rated borrowers separately from facilities. Thus, a bank's diversity of questions is not matched by a wealth of data.

A collection of free-form regressions can sometimes be the best way to proceed. An exception is when a simple formula performs better. This study compares the performance of a recently proposed LGD formula to the performance of free-form regression. Performance is measured by root mean squared error (RMSE). The default and LGD data are simulated based on a linear model, which gives an advantage to linear regression. Despite this, the LGD formula outperforms regression over a meaningful range of parameter values.

The reason is that the data sample is short. The impression created by a few years of data might not convey the true connection between default and LGD. This is the case whether LGD is tied directly to default, as in this study, or it is tied to underlying variables and the connection to default is left implicit. Either way, statistical uncertainty is high when data are scarce.

LGD data are naturally scarce, because only loans that default have an LGD. Further, defaults cluster in time. This leaves many years with few defaults and with portfolio LGDs that are consequently noisy. The natural scarcity and noise of LGD data limit confidence in statistical estimates.

By contrast to regression, the LGD formula has no sensitivity parameter that could be over-fit to noisy data. In many cases the formula outperforms linear regression even for a data sample that is decades longer than currently available. A final section attempts to improve the LGD formula by adjusting its slope based on samples of data. Again, decades more data might be required for this to have a beneficial effect.

Unless statistical uncertainty is properly controlled, these experiments suggest that risk managers can benefit by using the LGD formula to model the relationship between default and LGD.

The LGD formula

The LGD formula connects the conditionally expected LGD rate (cLGD) to the conditionally expected default rate (cDR). These rates would be observed in an asymptotic portfolio that

[1] Altman and Kuehne

averages the idiosyncratic effects of an infinite number of identical firms. The rates serve as expectations for finite portfolios or for loans. For example, under given conditions the LGD of a loan is random with mean equal to cLGD.

If the conditionally expected rates of loss and default depend on a single factor, then cLGD can be stated as a function of cDR. Frye and Jacobs derive a particularly simple LGD function.[2] They test the function using Moody's-rated loans and bonds and find that neither steeper nor flatter functions display statistical significance. They conclude that the LGD function adequately describes the relationship exhibited in the Moody's data, and that to add additional parameters entails too much risk of Type 1 Error.

The LGD function is:

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

where $\Phi[\cdot]$ represents the standard normal cumulative distribution function, and k is a positive parameter called the LGD Risk Index.

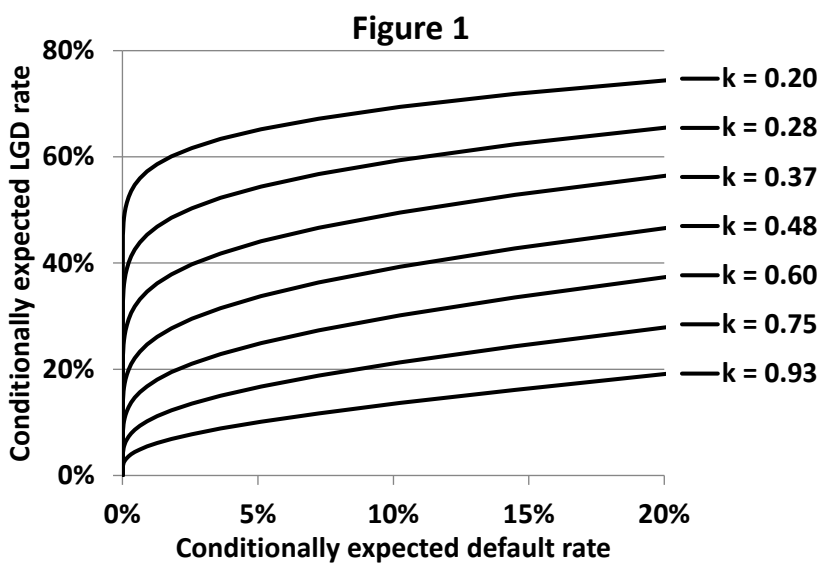


Figure 1 illustrates the LGD formula for seven values of the LGD Risk Index. Each curve is approximately a translation of the next.

Given its value of the LGD Risk Index, a loan moves along a curve as conditions change. If conditions imply a greater default rate, the formula implies a greater LGD rate. For a likely span of cDR, the range of cLGD is moderate. For example, suppose that a loan is apt to produce default rates between 1% and 10%, depending on conditions. Along the top line, the contribution to cLGD would be 11.6% (that is, cLGD would vary from 57.6% to 69.2%), and along the bottom line the contribution would be 7.9% (from cLGD = 5.6% to cLGD = 13.5%). Broadly speaking, there is about a point-for-point connection between default and LGD in this region.

[2] Frye and Jacobs

To apply the formula requires only the value of the LGD Risk Index, and this depends only on parameters already in common use. Specifically, the LGD Risk Index is

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

where ρ denotes correlation and EL denotes the expected loss rate. A loan's EL equals its PD times its expected LGD rate (ELGD). Thus, a loan's basic parameters (PD, EL or ELGD, and ρ) combine in Equation (2) to produce a value of the LGD risk index that establishes the LGD/default relationship of Equation (1).

Each line reflects many different loan types. For example, a loan having parameters PD = 1%, ELGD = 59.3%, ρ = 10% has LGD Risk Index = 0.2, but so does a loan having parameters PD = 15%, ELGD = 73.4%, ρ = 10%. Both loans move along the top line of Figure 1. The first would be found toward the left of the diagram and the second toward the right.

The objective is to understand credit loss. The distribution of credit loss is particularly simple when the LGD function is assumed. This is important because simple models are the standard against which other models must prove themselves. Until a demonstrably better model of credit loss is identified, a simple is to be preferred.

A loan's conditionally expected credit loss rate (cLoss) is the product of cDR and cLGD. It is assumed that cDR obeys the Vasicek Distribution:[3]

$$cDR = \Phi[\,(\Phi^{-1}[\,PD\,] + \sqrt{\rho}\,Z)\,/\,\sqrt{1-\rho}\,], \qquad Z \sim N[0,1] \tag{3}$$

where Z represents economic conditions. Then,

$$\begin{aligned} cLoss = cLGD\ cDR \ &= \ \Phi\left[\Phi^{-1}[cDR] - k\right] \\ &= \Phi\left[\Phi^{-1}\left[\Phi[\frac{\Phi^{-1}[\,PD\,] + \sqrt{\rho}\,Z}{\sqrt{1-\rho}}]\right] - \frac{\Phi^{-1}[\,PD\,] - \Phi^{-1}[\,EL\,]}{\sqrt{1-\rho}}\right] \\ &= \ \Phi[\frac{\Phi^{-1}[\,EL\,] + \sqrt{\rho}\,Z}{\sqrt{1-\rho}}] \end{aligned} \tag{4}$$

Thus, if cDR has a Vasicek Distribution with mean equal to PD, then cLoss has a Vasicek Distribution with mean equal to EL. This credit loss model has only two parameters. Its simplicity is purchased with the precise form of Equation (1).

In general, however, the parameters of an LGD function make their way into the distribution of credit loss. For example, cLGD might be assumed fixed at the value ELGD. The resulting credit loss distribution has three parameters: PD, ELGD, and ρ. More flexibly, cLGD might be assumed to be a linear function of cDR; then, the associated credit loss distribution has four parameters. There is nothing wrong with additional statistical parameters if they are statistically significant.

[3] Vasicek

However, to date no real world data sets have revealed statistical significance for any alternative to the LGD function investigated here.

The position of a line in Figure 1 is determined almost completely by the expectations ELGD and PD. The effect of ρ, by contrast, is muted. The practical range of ρ is between about 5% and about 25%, and this implies that the factor $1/\sqrt{1-\rho}$ in Equation (2) between about 1.02 and 1.16. The resulting 14% effect on the value of the LGD Risk Index is not great enough to shift between the lines in Figure 1. The muted effect of ρ is fortuitous, because the LGD formula is not highly sensitive to uncertainty in the estimate of ρ.

The LGD formula has four intertwined advantages compared to statistical approaches such as linear regression. First, the LGD formula can be applied without statistical research. Estimates of PD and ELGD may exist or may be obtained from vendors, and the value of ρ has a weak effect in any case. Second, the LGD function implies a particularly simple model of credit loss. No alternative to this model has been found statistically significant, and this lends support to the LGD formula. Third, the LGD formula depends primarily on averages, PD and EL. Estimates of averages are generally more reliable than estimates of sensitivities, especially when data are scarce. Fourth, the LGD formula expresses a moderate, positive dependence of LGD on default. This seems to be a more plausible starting place than no dependence whatsoever, which is the null hypothesis under linear regression.

Simulation experiments

Equation (1) predicts LGD as a function of the default rate. This study compares that to the function estimated by linear regression. Simulation begins with a standard normal variable Z that represents economic conditions in a particular year. Z implies cDR using Equation (3). The density of cDR can be found using the "change of variable" technique:

$$f_{Vas}[cDR; PD, \rho] = \frac{\sqrt{1-\rho}}{\sqrt{\rho}} \frac{\phi\,[\,(\Phi^{-1}[PD] - \sqrt{1-\rho}\,\Phi^{-1}[cDR])\,/\,\sqrt{\rho}\,]}{\phi[\,\Phi^{-1}[cDR]\,]} \tag{5}$$

Conditionally expected LGD is simulated with a linear model. This does not reflect a belief that the LGD rate is a linear function of the default rate. Instead, it reflects a belief that a common form of data analysis is linear regression. The intention is to give linear regression an advantage in the comparisons by using a linear data generator:

$$cLGD = a + b\,cDR \tag{6}$$

Simulated portfolios contain 1,000 firms. This is generous compared to the portfolios that generate historical data for a particular combination of PD, ELGD, and ρ. A later section allows the number of firms to differ from 1,000.

The number of defaults has a Binomial Distribution [N = 1,000, p = cDR]. The simulated portfolio default rate (DR) therefore depends on an independent standard normal variable X:

(7) $$DR = CDF^{-1}_{Binomial(N,cDR)}[\Phi[X]] \,/\, N, \qquad X \sim N[0,1]$$

Each default produces an LGD that is assumed to have a Normal Distribution [cLGD, σ^2]. The LGD rate therefore depends on an independent normal variable Y:

(8) $$LGD = cLGD + Y\,\sigma/\sqrt{N\,DR}\,, \qquad Y \sim N[0,1]$$

In the initial simulations, PD = 3% and ρ = 10%, which are not uncommon borrower characteristics. The values a = 0.5 and b = 2.3 are those fit by Altman to high-yield bond data. The value σ = 20% is the estimate provided by Frye and Jacobs. Later sections explore different values for each of these parameters.

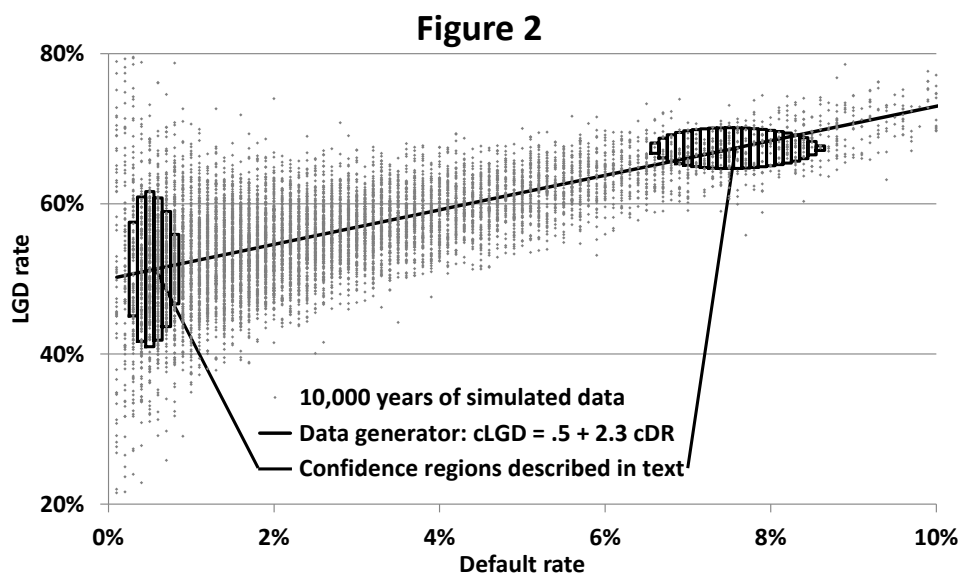


Figure 2 shows the data generator and 10,000 simulated years of data. If linear regression were based on all 10,000 data points, the regression line would match the data generator almost exactly. Without question, linear regression would surely outperform the nonlinear LGD formula if the data sample were many times greater than possible.

The difficulties faced by linear regression are that the data sample is short and that individual years can be far from the data generator. Therefore, a short data sample might lead regression to a material error.

To illustrate, two concentration regions are highlighted in Figure 2. The one at the left is centered at the point cDR = 0.57%, cLGD = 51.3%, reflecting the 5th percentile of the systematic factor Z. The concentration region itself contains 50% of the portfolio-years of data that would be generated under this condition.[4] The low value of cDR insures that portfolio DR is apt to be low. The low default rate implies only a few losses contribute to the portfolio's average LGD, which therefore has high variance. As a result, the concentration region at the left is narrow and tall.

[4] A 50% concentration region is a transformation of a disc of (X, Y) with radius 1.177.

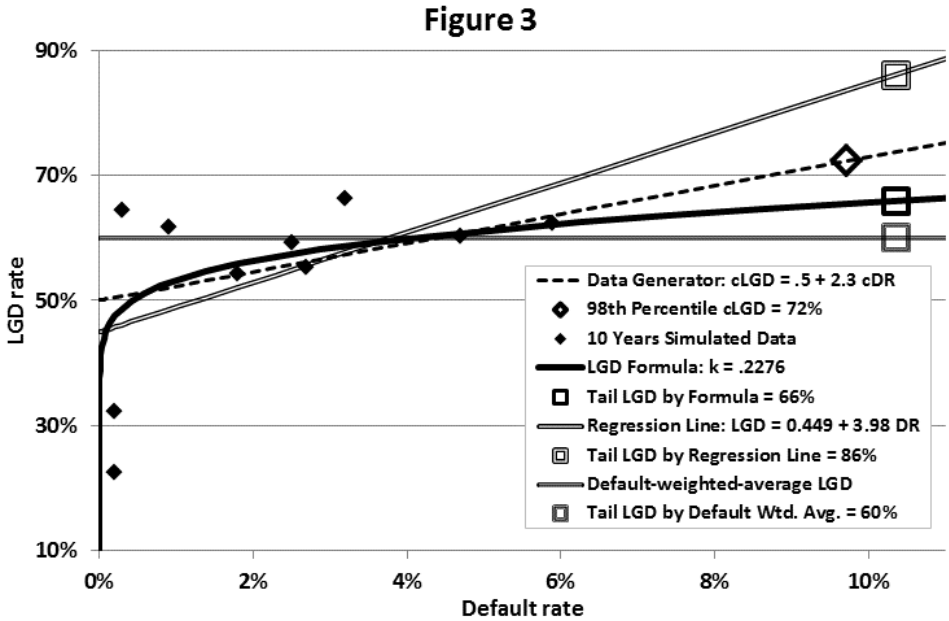
The 50% concentration region at the right reflects the 95th percentile at which cDR equals 7.58% and cLGD equals 67.4%. Compared to the concentration region at the left, defaults are more numerous and the portfolio LGD rate has less variance. However, a wider range of default rates is possible. The resulting concentration region is shorter and wider. Thus, simulated data can depart from the data generator at every level of the systematic factor.

Left and right, half the simulated data is outside the concentration region. A sample of a few years can therefore appear quite different from the data generator. This has a profound effect on linear regression. The effect on the LGD formula is less profound, because its prediction depends on averages rather than the slope of the data swarm.

All data sets are generated by the same procedure. The values of the seven parameters are initially PD = 3%, ρ = 10%, a = 56.9%, b = 2.3, σ = 20%, N = 1,000 firms, and T = 10 years. Each year of data has a cDR chosen at random from the Vasicek distribution. The portfolio default rate is based on a Binomial Distribution with mean equal to cDR. The year's cLGD rate is a linear function of cDR. The portfolio LGD rate is drawn from a normal distribution with mean equal to cLGD and variance reflecting the number of defaults that year. Each data set is analyzed by both regression and the LGD formula, and the results are compared.

Analysis of a single data set

This section describes the analysis of a selected data set. The aim is to predict the 98th percentile of cLGD. Different values of the percentile are considered in later sections.



In Figure 3, the data generator appears as the dashed line. Equations (3) and (6) imply 98th percentile values of cDR and cLGD as follows:

(9) $$cDR = \Phi[\,(\Phi^{-1}[\,3\%\,] + \sqrt{10\%}\,2.054)\,/\,\sqrt{1-10\%}\,] = 9.72\%$$

(10) $$cLGD = 0.5 + 2.3 \times 9.72\% = 72.3\%$$

The point (cDR = 9.72%, cLGD = 72.3%) is indicated by an open diamond on the data generator.

The dots illustrate ten years of simulated data. For both LGD predictors, the estimate of PD is the average annual default rate, 2.24%, and ρ maximizes the following likelihood function:

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

The estimate of ρ for these data is 17.6%. The estimated 98th percentile of cDR is then

(12) $$\widehat{cDR} = \Phi[\,(\Phi^{-1}[\,2.24\%\,] + \sqrt{17.6\%}\,2.054)\,/\,\sqrt{1-17.6\%}\,] = 10.35\%$$

Of the two LGD predictors, the LGD formula is simpler to apply. The estimate of EL is the average annual loss rate, 1.34%. This implies k = 0.2276 and establishes the line for the LGD formula. Along this line, the prediction ($\widehat{cDR}$ = 10.35%, $\widehat{cLGD}$ = 65.9%) is marked with an open square. The LGD formula understates true cLGD by 72.3% - 65.9% = 6.4%.

The regression parameter estimates are $\hat{a}$ = 0.449 and $\hat{b}$ = 3.98. The regression line prediction ($\widehat{cDR}$ = 10.35%, $\widehat{cLGD}$ = 86.1%) is marked with an open square. This overstates cLGD by 13.8%, which is a much greater error than made by the LGD formula.

The regression slope, though steep, does not display significance in a test of 5% size; its p-value is 0.176. Data points do not lie sufficiently close to the line to produce statistical significance.

When a regression slope is not statistically significant, a regression prediction reverts to an average. Average annual LGD equals 53.9%. This has gained the name "time-weighted" average, because it gives equal weight to each year's LGD rate. Time-weighted average LGD is not acceptable under Basel II regulations. Since those regulations drive much of the practical work in LGD modeling, modelers prefer to use the "default-rate-weighted-average" LGD. This is equal to the ratio of estimated EL to estimated PD, 1.34% / 2.24% = 60.0%, and is marked with an open square in Figure 3. This is better than the regression-line-based estimate and better than time-weighted LGD, but makes greater error than the LGD formula estimate.

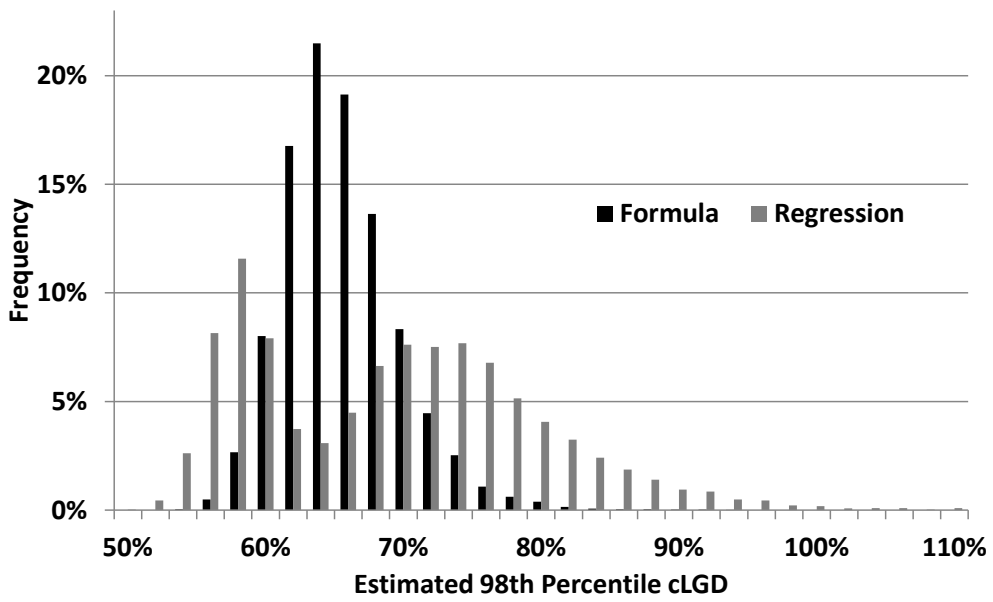
This section highlights features that are far from unique. A year of data can be quite far from the linear data generator, especially if there are only a few defaults. Since only a few years of data are available at present, the entire data sample as a whole may not accurately reflect the data generator. The LGD formula tends to understate tail LGD when the data generator is steep and to overstate tail LGD when the data generator is shallow. A regression might be steep or shallow, and separately, it might be significant or not significant. A significant regression need not resemble the data generator; an insignificant regression concludes, wrongly, that there is no systematic LGD risk.

For these data the 98th percentile of the data generator equals 72.3%. The LGD formula, based on estimates of PD, EL, and correlation, produces an estimate of 65.9%. The regression line would produce an estimate of 86.1%, but its slope is not statistically significant. Therefore, the regression prediction reverts to the estimate of ELGD, which equals 60.0%. This makes an error of 12.3%, which is nearly twice the error made by the LGD formula.

Analysis of 10,000 data sets

This section repeats the foregoing analysis 10,000 times. Figure 4 compares the predictions. Predictions made by the LGD formula average 65.5% and are tightly distributed; over 80% of predictions are within 5% of the average. The root mean squared error (RMSE) of the formula predictor is 7.94%, most of which is downward bias.

Figure 4



The average prediction made by the regression equals 69.2%, but the range of predictions is dispersed from 49.7% to 133%, and 0.56% percent of predictions are greater than 100%. As a result of high dispersion, the RMSE of the regression predictor is 11.04%.

The histogram of regression predictions is bimodal. The greater mode reflects primarily regressions having a statistically significant slope. The lesser mode reflects primarily regressions lacking a significant slope where predictions revert to the default-rate-weighted LGD estimate.

This section compares the two LGD predictors over 10,000 sets of simulated data holding the values of parameters fixed at previous levels. The RMSE of the LGD formula, 7.94%, is substantially less than the RMSE of linear regression, 11.04%.

Values of other parameters

This section extends the comparison by allowing each parameter a range of values, holding fixed the other parameters at the initial levels. Each RMSE reflects 5,000 simulation runs.

The length of the data sample

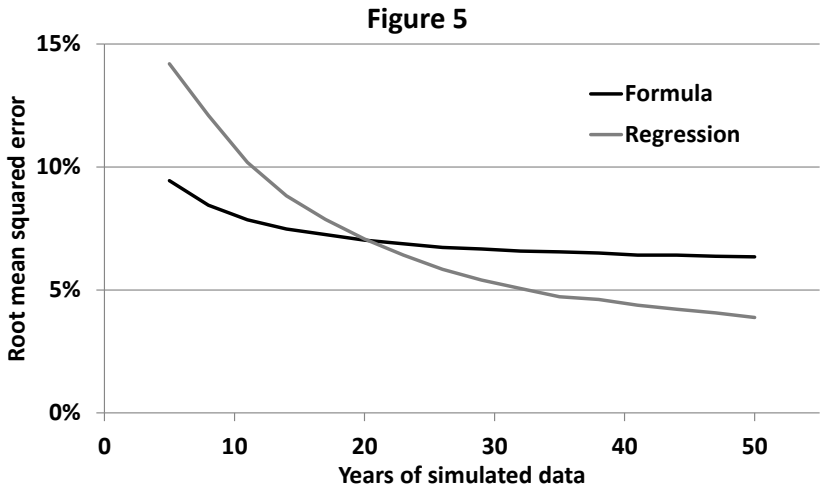


Figure 5 compares the RMSE of the LGD formula to the RMSE of the regression as the sample length ranges from 5 years to 50 years. As the sample increases, the estimates of all parameters become more reliable. For the LGD formula, a lower bound of RMSE is established by asymptotic bias, which is 72.3% - 66.4% = 5.9%. With the given set of parameter values, beyond about 20 years of simulated data the regression begins to produce lower RMSE than the LGD formula.

More than 20 years of data from the real world would probably be required for the regression to outperform the LGD formula. That is because real-world data are auto-correlated: good years tend to be followed by good years, rather than by random years. An additional year of real-world data is less informative than an independent simulation draw. Although it is common to overlook auto-correlation in real-world data, doing so distorts the significance of regression results.

The number of firms in the portfolio

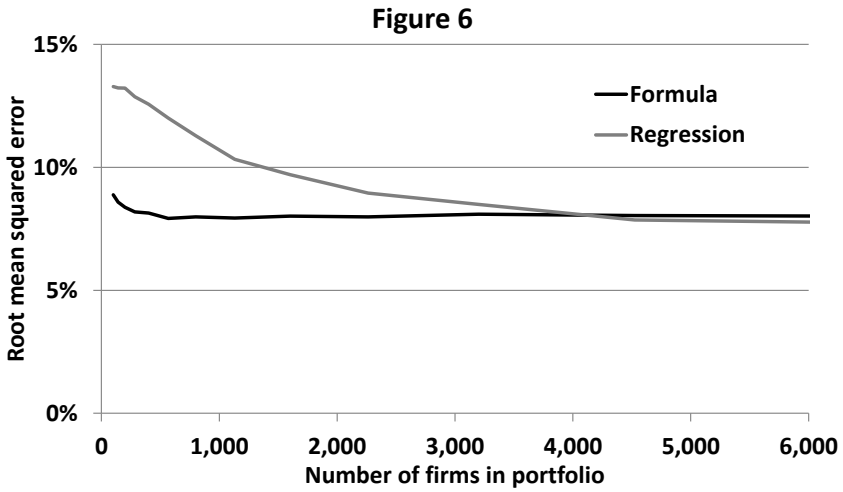


Figure 6 compares the RMSEs for portfolios with firms numbering from 100 to 6,000, holding other variables fixed. But even an asymptotic portfolio does not reveal PD or ρ in a few years of data. Beyond about 5,000 firms the two methods produce about the same RMSE and little further gain appears.

The limit to RMSE is imposed by the short data sample. A large portfolio is no guarantee that a sample of a few years of data accurately reflects the true relation between LGD and default.

Tail percentile

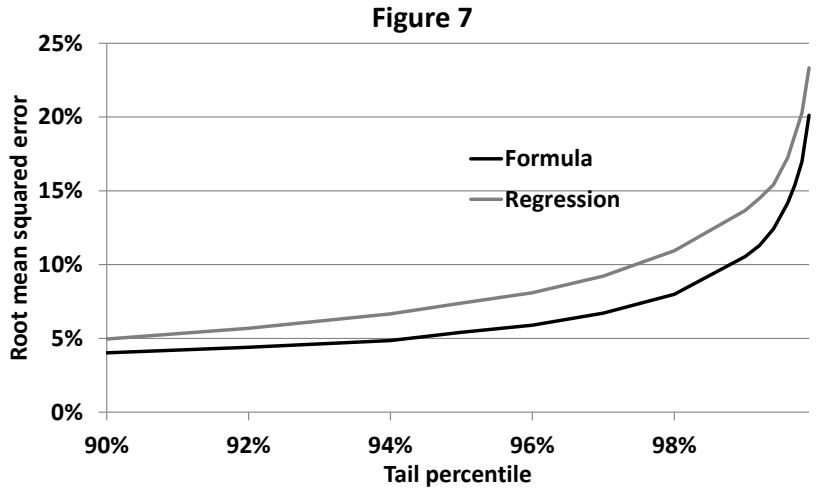


Figure 7 compares RMSEs as the tail percentile ranges from 90% to 99.9%. As would be expected, each predictor becomes less reliable at distant percentiles. Throughout the range the LGD formula outperforms regression.

Correlation

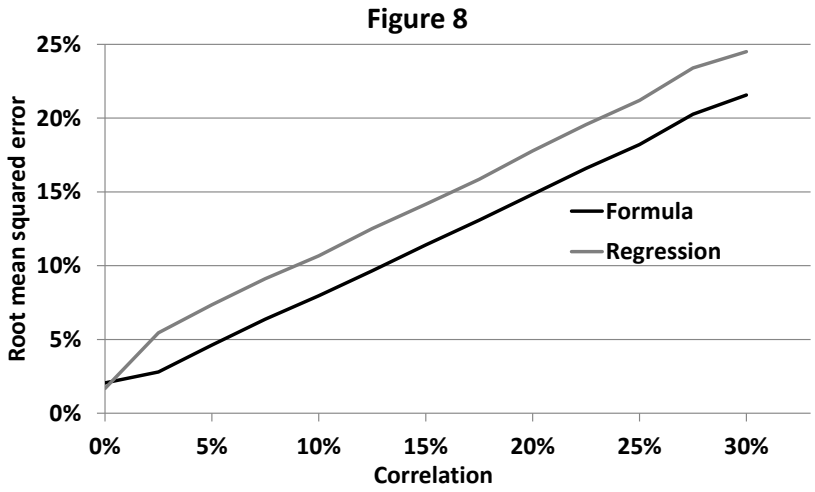


Figure 8 compares RMSE for the LGD formula and the regression as ρ ranges from 0% to 30%. Greater values of ρ imply greater variance of both cDR and cLGD, and this variation affects the RMSE of either method. Throughout the range of ρ, however, the LGD formula produces a lower value of RMSE.

Standard deviation of loan-level LGD

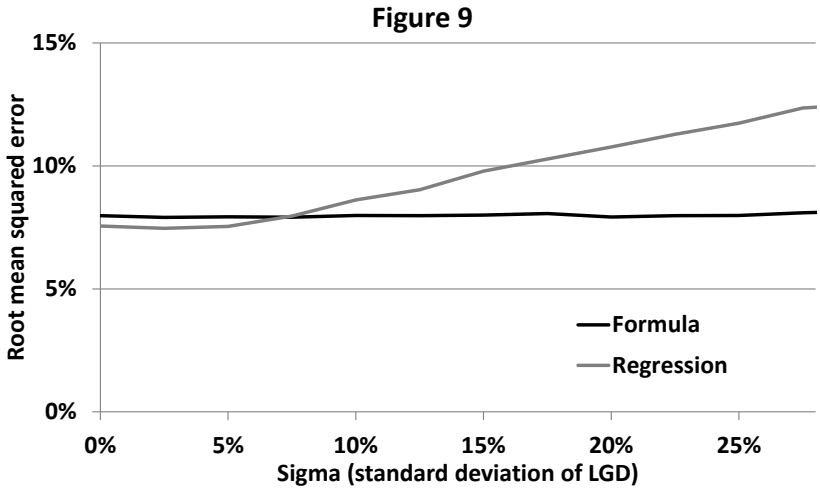


Figure 9 compares RMSEs as σ ranges from 0% to 30%. σ is the standard deviation of a defaulted loan's percentage loss around its conditional expectation. Controlling for rating, seniority, and the default rate, Frye and Jacobs estimate a value of about 20%. This is lower than studies that do not control for systematic LGD variation. For example, Araten et. al. place the standard deviation of LGD, uncorrected for systematic risk, between about 30% and about 40% depending on industry, business line, or collateral.[5]

[5] Araten et al.

At very low values of σ the two predictors have similar RMSEs. At realistic levels, the LGD formula produces a lower value of RMSE than regression.

Slope of data generator

This subsection uses data generators with different values of the slope parameter b. For each value of b, the value of a is set equal to 0.569 – 0.03 b. This makes each data generator pass through (cDR = 3%, cLGD = 56.9%), like the data generator analyzed in previous sections.

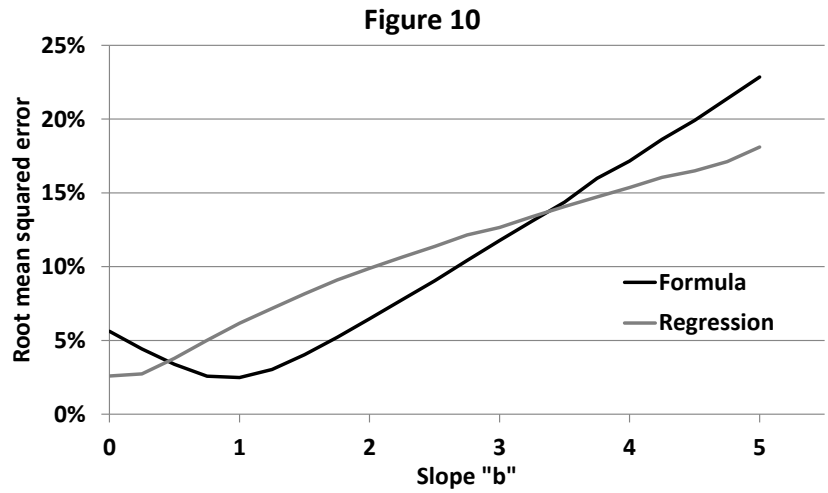


Figure 10 compares the RMSEs as b ranges from 0 to 5. Among all parameters, the slope of the data generator has the most nuanced effect in that it causes the lines to cross twice.

The LGD formula predicts a particular, moderate, response of LGD to the default rate. When the data generator also has moderate sensitivity, the formula produces its minimal values of RMSE. This occurs at about b = 1 given the other parameter values.

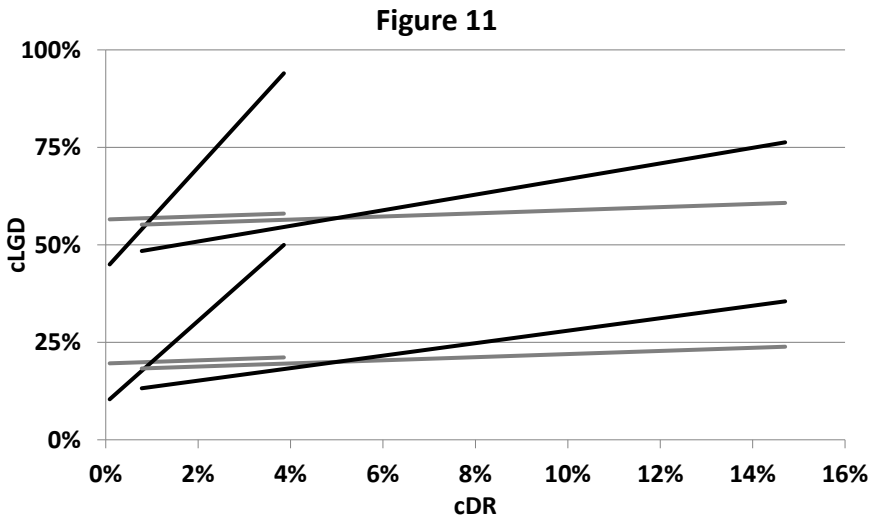
The regression RMSE increases with b. That is partly because a greater value of b increases the variance of the LGD data, which leads to greater forecast errors. It is also partly because a greater value of b makes it more likely that a regression is significant, and significant regressions can predict values of LGD that are quite different from the generator.

Comparing the two predictors, when b = 0 the regression outperforms the formula because most regressions are not significant. When b takes a high value, the regression outperforms because the formula curves downward from the data generator. However, there is a substantial mid-range (0.5 < b < 3.25 in Figure 10) in which the LGD formula outperforms linear regression. In this mid-range, the LGD formula has approximately the right level of sensitivity and less dispersion than linear regression.

Primary loan parameters

This subsection extends the previous analysis to different kinds of loans. The loans have PD equal to either 1% or 5% and produce either high LGDs (a = 0.569 - b PD) or low LGDs (a = .2 – b PD). For each kind of loan, the slope parameter b takes a range of values. The high-LGD

generators pass through (cDR = PD, cLGD = 56.9%) and the low-LGD generators pass through (cDR = PD, cLGD = 20%).



The results are depicted in Figure 11. For each kind of loan there is a pair of crossing lines. The two lines represent the shallowest and steepest data generators for which the LGD formula outperforms regression analysis. At the upper left, the shallowest slope is b = 0.4, and the steepest slope is b = 14. Within this range of slopes, the RMSE of the LGD formula is less than the RMSE of linear regression.

The lengths of the lines correspond to 96% concentration regions of cDR and cLGD. For example, the steep line at the upper left runs from the 2nd percentile (cDR = 0.09%, cLGD = 45.0%) to the 98th percentile (cDR = 3.86%, cLGD = 94.0%). Over the central 96% of the distribution, the systematic contribution to LGD equals 94% - 45% = 49%. This degree of LGD risk is economically meaningful and greater than commonly encountered in real-world data. Even when this degree of risk is present, the LGD function outperforms regression.

Comparing the pair of lines at the upper left to those at the lower left, the range of b is similar. (The range of b for the PD = 1%/low LGD combination is $0.4 \leq b \leq 10.5$.) Similarity is not unexpected. Suppose that a first generator passes through (DR = 1%, LGD = 56.9%) and a second generator passes through (DR = 1%, LGD = 20%), each having the same slope. The first data generator produces LGDs that are 36.9% greater than the second. Regression constants and fitted values also differ by 36.9%. Therefore, the regression produces the same RMSE using either data generator. ELGD, also, is 36.9% greater on the first generator than the second. However, this does not produce an exact parallel shift of the LGD formula. The LGD formula produces similar but slightly different RMSEs for the two generators. The comparison between LGD predictors is therefore slightly different for the two data generators.

Comparing the lines for PD = 1% to those for PD = 5%, greater PD implies a compressed range of b. (For high LGD the range is $0.4 \leq b \leq 2.0$, and for low LGD the range is $0.4 \leq b \leq 1.6$.) The range is compressed for two reasons. The first reason is that a low default rate implies few LGDs and noisy portfolio LGD rates; the noise has a greater effect on the regression than on the LGD

formula and widens the range of b. The second reason is that when PD is greater, the range of likely values of cDR is greater. The greater range of default rates implies a greater range of cLGD and gives greater opportunity for regression to detect systematic LGD risk.

The degree of systematic LGD risk remains great, even where the range of b is most compressed. Along the steeper line at the lower right, the total systematic contribution to LGD risk is 35.5% - 13.2% = 22.3%. The contribution can also be quantified by the difference between ELGD and tail LGD. ELGD is EL divided by PD, and EL equals the probability-weighted product of cLGD and cDR:

(13) $$EL = \int_0^1 (a + b\ DR)\ DR\ \ f_{vas}[DR, PD, \rho]\ \ dDR = 0.012 = PD\ ELGD$$

Thus, ELGD equals 24%. The 98th percentile cLGD, 35.5%, is nearly 50% greater than ELGD. By either measure there is a substantial amount of systematic LGD risk. It might be expected that statistical regression could detect this; however, the LGD formula outperforms regression even when this degree of risk is present.

This subsection shows that the LGD formula has lower RMSE than regression when data is generated using two values of PD, two levels of LGD, and ranges of values of b. Regression outperforms the formula if the regression null hypothesis, b = 0, happens to be true. But formula outperforms regression from a low value of b up to the limits shown in Figure 11. If PD = 1% the limiting slope is quite steep. At PD = 5% the limiting slope is less steep, but the contribution of systematic LGD risk is still substantial. At either level of PD, there is a range of slopes for which the moderate LGD sensitivity of the formula produces a more accurate result than regression analysis.

This section allows each of the eight parameters to take ranges of values. For the entire range of many parameters, the RMSE of the LGD formula is lower than, or nearly as low as, the RMSE of the regression. The slope of the data generator has the nuanced effect described in the previous paragraph. As the sample grows longer the RMSE of the regression declines and eventually becomes less than the RMSE of the LGD formula. However, for samples available at this time, for a range of data generators, the LGD formula outperforms regression analysis.

The LGD formula as a Null Hypothesis

This section adjusts the slope of the LGD formula to improve the fit to a sample of data. This attempt to improve performance in fact degrades it over a range of parameter combinations.

Data are simulated as before. After the effect of the LGD formula is removed, the LGD residuals are regressed on the default rate and a constant. If the supplementary regression is significant, its contribution is added to the tail prediction made by the LGD formula:

(14) $$LGD_{tail} = \frac{\Phi\left[\Phi^{-1}[DR_{tail}] - k\right]}{DR_{tail}} + const + slope\ *\ DR_{tail}$$

where DR_{tail} is found as in Equation (9). Otherwise the prediction reverts to the LGD formula. Only if the supplementary regression is significant does it affect the LGD prediction.

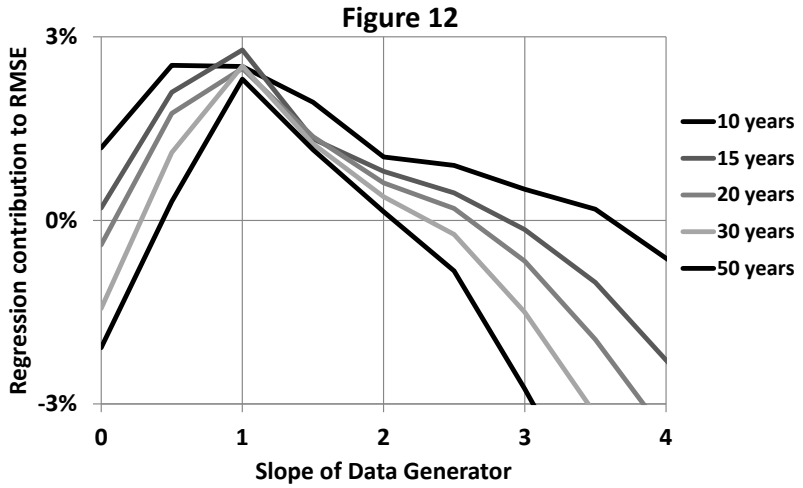


Figure 12 shows the effect of the regression on the RMSE of the LGD formula. The topmost line reflects samples of ten years of data. At the left, the value 1.2% indicates that, far from improving the predictions of the LGD formula, the supplemental regression adds 1.2% to the RMSE of the prediction. The topmost line dips below zero only if the slope of the data generator is greater than about 3.6. If the data generator is this steep or steeper, regression can improve the performance of the LGD formula working with ten years of data. The improvement is of course to make the LGD predictor steeper and the prediction of tail LGD greater.

The other lines of Figure 12 reflect longer data samples. As the sample increases, supplementary regression is better able to make a desirable contribution. With 50 years of data, supplementary regression can improve performance if the data generator has a slope less than 0.5 or greater than 2.0. Within this moderate range, however, supplementary regression degrades the performance of the LGD formula.

The reason that the supplementary regression degrades performance is that the LGD formula, by itself, performs well. In particular, the LGD formula lies close to the data generator at elevated default rates when $b = 1$. If a regression were significant, its data swarm must differ from the LGD formula, and therefore it must differ from the data generator. A significant regression line therefore is likely to point the wrong way. In the moderate range of slopes, approximately $0 < b < 2$, the contribution of regression is to degrade the LGD prediction even when 50 years of data are available.

This section allows linear regression to contribute to the explanation of LGD after accounting for the LGD formula. For moderate slopes of the data generator, regression degrades performance rather than improving it. Only if the data generator is quite steep, and tail values of LGD quite greater than predicted by the LGD formula, can supplementary regression improve performance with the length of data sample banks currently possess.

Conclusion

A risk manager must choose a method to analyze the connection between default and LGD. He might perform ad-hoc regression. He might also use a recently proposed LGD formula that does not require statistical estimation and therefore entails less estimation error.

Each method has its own starting place. Linear regression, unless significant, imagines that there is no connection between default rates and LGD rates. The LGD formula imagines a moderate, positive connection between default and LGD.

This simulation study compares the predictions of the two approaches. Under many conditions, the LGD formula produces lower RMSE than regression. Attempting to adjust the slope of the LGD formula based on supplementary regression under many conditions degrades the performance of the LGD formula by itself.

Three features of the simulations contribute to these results. First, LGD data are noisy. Portfolios of similar loans have limited size, default is a rare event, and the default rate in most years is less than average. As a consequence, portfolio LGD is a noisy realization of the conditionally expected rate of LGD. Second, the data sample is short. This means that the entire sample can deviate from the data generator, and has material effect on regression predictions. Third, the sensitivity of LGD to default is not too different from the moderate, positive connection between default and LGD expressed by the LGD formula.

These factors appear to be in place in the real world. Real-world portfolios have noisy LGDs for the same reasons as the simulated portfolios. Only a few years of carefully observed bank data are available, because separate ratings of borrowers and loan facilities are a recent development. The moderate, positive connection expressed by the LGD formula has survived statistical testing to date.

The distribution of credit loss can be resolved to finer detail when the sample of credit loss data becomes long enough. Until this can be done, risk managers can use the LGD formula to avoid introducing new parameters and unneeded noise into LGD predictions. Unless a data set is extensive enough to show that statistical uncertainty is properly controlled, risk managers can benefit by using the LGD formula to model the relationship between default and LGD.

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.

Araten, M., Jacobs, M., and Varshney, P., The RMA Journal, May 2004, 28-35.

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

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