Basle II Capital Requirement Sensitivity to the Definition of Default

Jiří Witzany[1]
University of Economics, Prague

Abstract

The paper is motivated by a disturbing observation according to which the outcome of the regulatory formula significantly depends on the definition of default used to measure the probability of default (PD) and the loss given default (LGD) parameters. Basel II regulatory capital should estimate with certain probability level unexpected credit losses on banking portfolios and so it should not depend on a particular definition of default that does not change real historical and expected losses. We provide an explanation of the phenomenon based on the Merton default model and test it using a Monte Carlo simulation. Moreover we shall develop an analytical method to model LGD unexpected risk and to combine it with the PD unexpected risk. The developed formula and in particular its simplified version could be used to improve the current regulatory formula. The analysis at the same time provides a different insight into the issue of regulatory capital sensitivity on the definition of default. Finally we perform a structural model based simulation to test the hypothesis according to which scoring functions developed with a soft definition of default provide weaker predictive power than the ones developed with a hard definition of default.

1. Introduction

The Basle II capital requirement formula (see Basle, 2006, e.g. paragraph 328 for mortgages) has in principle the form

(1) $$C = \text{Unexpected Loss} - \text{Expected Loss} = UDR \times LGD - PD \times LGD,$$

[1] The research has been supported by the Czech Science Foundation grant no. 402/06/0890 "Credit Risk and Credit Derivatives"
First version December 2007, this version April 2008

Keywords: credit risk, correlation, recovery rate, regulatory capital
JEL: G21, G28, C14

where *UDR* is a regulatory estimate of unexpected default rate, *PD* estimated average default rate, and *LGD* the loss given default. The resulting percentage capital requirement (*C*) is multiplied by *EAD*, the exposure of the receivable at default (in most cases equal to the actual outstanding amount), to get the account level capital requirement in absolute terms. The calculation is done on account level but the result should reflect only the account contribution to unexpected risk within a well diversified portfolio. The *UDR* is given by a regulatory formula that depends only on *PD* (other key parameters of the formula, i.e. correlation and probability level are set by the regulator as given constants with a possible adjustment depending on *PD* in the case of correlation). In other words the capital requirement *C* is a function of *PD* and *LGD*

(2) $$C = f(PD,LGD) = [UDR(PD) - PD] \times LGD$$

The logic of the formula (2) is to give en estimation of unexpected systematic loss relative to the expected loss on the 99.9% probability level. This goal is achieved mainly through the first part of the formula *UDR(PD)-PD* modeling the difference between the unexpected default rate value and the expected mean default rate. On the other hand the *LGD* parameter is defined in principle as a conservative estimate of long-term average loss rates on defaulted loans. The regulator in addition requires the banks rather vaguely to reflect economic downturn conditions, dependence on *PD* or other factors to capture the relevant risks. Nevertheless under normal circumstances the long term average is considered as satisfactory. In that case the possibility of unexpectedly high loss rate is not captured by the formula at all.

In practice the parameters *PD* and *LGD* are produced from historical data on homogenous (in terms of product, segment, and rating) sets of receivables. The values strongly depend on the definition of default which determines the number of loans marked as defaulted. The definition of default must satisfy certain regulatory conditions but the banks have a significant discretion in its implementation. Assume that a bank suffered a total loss L on its historical portfolio P and recorded a number of defaults N_D. Some of the defaulted receivables might have been recovered fully while others contributed to the total loss L. If $\bar{E} = V / N$ (where V is the total portfolio P volume and N the total number of receivables) is the average exposure of the observed receivables and $\bar{L} = L / N_D$ is the average loss on defaulted receivables then the loss given default parameter can be actuarially estimated as $LGD = \bar{L} / \bar{E}$. Similarly *PD* can be estimated as N_D/N. Consequently we get that the product $PD \times LGD = L/V$ is the average percentage observed loss rate $LR = L/V$ on the portfolio, which does not depend on the definition of default. The definition just influences the decomposition of LR into the product $PD \times LGD$, i.e. given PD (based on a definition of default) we calculate LGD as LR/PD and vice versa. This relationship holds as long as LGD estimate is not stressed. However the product $PD \times LGD$ serves as the bank's estimate of future losses which should be used for loan pricing, budgeting etc. Hence the stressing of $LR=PD \times LGD$ should be quite consistent with experienced losses which might be moderately adjusted to certain negative macroeconomic or other developments. Hence the historical loss rate LR should be replaced with a more conservative expected loss rate EL looking into the future. Then we still have the identity $PD \times LGD=EL$. Thus the capital requirement formula then takes the form

$$C = f(PD, EL/PD) \tag{3}$$

where the expected loss EL is a fixed input parameter. One may ask why PD should be variable if observed from historical data. The point is as already mentioned that there is the notion of default that influences strongly the counted number of defaults. According to the Basle II requirement receivables that are more than 90 days past due, where the debtor is legally proclaimed bankrupt, or that are considered probable not to be repaid in full must be flagged as defaulted. The regulation leaves significant space to banks to use a softer or harder definition of default in particular due to the condition “considered probable not to be repaid in full.” A bank may stick to the simple definition of 90 days past day plus legal events like bankruptcy etc. Or it may opt to use a soft definition of default where many other clients are marked as defaulted earlier based on a number of additional conditions just making sure that the overall probability of absorbing default (staying in default until maturity) remains above 50% to be consistent with the common sense interpretation of the regulatory requirement. Another important parameter in the definition of default is the materiality threshold – only amounts above the threshold may cause a default. If the materiality threshold is too low then many clients that by a mistake left very small amounts unpaid may be marked as defaulted. There is also the issue of cross-cross default of retail clients which might be dealt with differently by different banks – does a default on credit card imply a default on a mortgage which is being paid etc. In the hard default approach the bank may count on a historical data-set just 2,5% default rate while in terms of the soft default definition the rate may come out as high as 5%. Of course the different definitions of default should not and according to our assumptions do not change the observed and expected percentage loss. Since the underlying fundamental financial data remain the same the capital requirement modeling the difference between the 99,9% quantile of future loss and the mean (expected) loss should remain identical under different definitions of default. Nevertheless Exhibit 1 shows that the risk weight (+ using the regulatory formula for retail loans) significantly declines from 115% to 62% when *PD* goes from 2,5% to 5%.

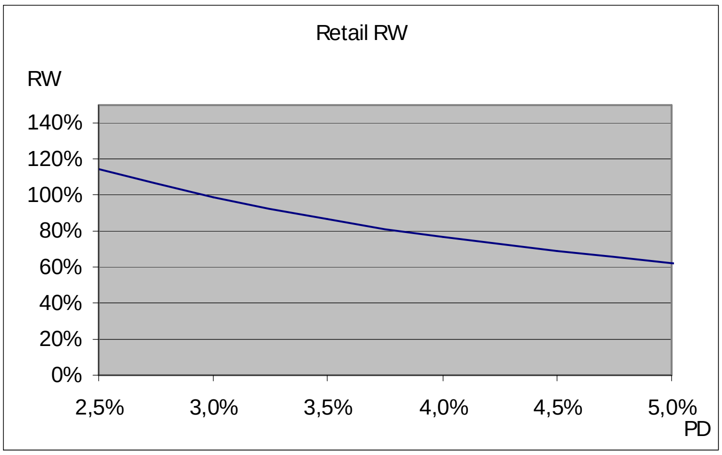


Exhibit 1 (Dependence of regulatory $RW=12{,}5\times C$ on *PD*)

This appears as a significant problem of the regulatory formula. On one hand side regulators might wish to motivate banks to use a conservative (soft) definition of default in order to start collection and workout process soon. On the other hand the sensitivity of the capital to the definition of default as shown in Exhibit 1 is surprisingly dramatic. Soft definition of default may be desirable in terms of conservative classification of assets, but it could pollute the quality if default statistics if the definition is consistently used for development of scoring functions (the Basle II regulation does require banks to use the definition of default consistently throughout the credit process). Many clients marked as defaulted using a soft definition may return to ordinary repayment schedule later. The reason of the default in payment may be not in real financial difficulties but just disorganization on client side when repayments are put into order after one or two reminders. Prediction of this type of default should not be mixed with prediction of real or absorbing default, i.e. default where the client does not return to ordinary repayments and the bank suffers a real loss. This notion of absorbing default (rather than soft default) should be captured by the hard definition of default.

In Section 2 we provide an explanation of the observed sensitivity to the definition of default and propose alternative approaches to *LGD* unexpected risk modeling. In Section 3 we will try to analyze effects of the definition of default on quality of scoring functions developed using a standard technique and based on a structural credit default model simulation.

2. Unexpected Credit Losses

2.1. Economic versus Regulatory Capital

Regulatory capital requirements for credit risk exposures aim to approximate the economic capital that captures unexpected credit losses in a given time horizon and on a probability level. The two parameters in the context of Basle II are 1 year and 99,9%. Thus in principle we need to model the probability distribution of a random variable $V(1)$, the future value of the bank's credit portfolio 1 year from the calculation time. The expected value of $V(1)$ should take into account not only accrued interest but also expected credit losses, so the 99,9% - quantile of $V(1) - E(V(1))$ is the value we need to estimate. The credit economic capital or Credit Value at Risk (VaR) is in practice estimated by a number of techniques. We can mention for example wellknown JP Morgan CreditMetrics (JP Morgan) based on rating migration probabilities, reduced CreditMetrics+ model (CS First Boston a Credit Suisse), and PortfolioManager (KMV Moody's) based on the structural Merton model.

The regulatory model must certainly accept a number of simplifications. Although the regulatory capital requirement is calculated on account level it needs to take into account only the portfolio systematic risk not the specific risk. Currently the regulatory model does not reflect the level of diversification and specific correlation structure of the bank portfolio. It uses overall average correlations differentiated for basic product classes and segments with certain dependence on the probability of default. The key simplification lies in the fact that the formula models principally unexpected default rate and not the unexpected loss rate. The unexpected loss rate ULR is according to Basle II in estimated as

$$ULR = UDR \times LGD \qquad (4)$$

disregarding the maturity adjustment applied only for non-retail clients. Thus the regulatory UEL formula does not take into account the unexpected risk of losses after default. It is rather obvious and has been confirmed by a number of studies that the additional unexpected risk is quite significant (see e.g. Altman et al. (2004)). This fact already partially explains the empirical fact that the Basle II capital requirement depends on the definition of default. We will develop several alternative models extending the unexpected loss calculation beyond the UDR to capture the (systematic or portfolio) risk of unexpected losses and to explain the sensitivity shown in Exhibit 1.

2.2. Basle II Unexpected Default Rate

The Basle II formula can be expressed as follows

(5) $$UDR = N\left(\frac{N^{-1}(PD) + \sqrt{r} \cdot N^{-1}(0{,}999)}{\sqrt{1-r}}\right)$$

where N is the cumulative standardized normal distribution function, N^{-1} its inverse, r is the correlation set up by the regulator (15% for mortgage loans, 4% for revolving loans, and somewhere between depending on *PD* for other retail loans.) It will be useful to recall the principle of the formula (5) that was firstly discovered by Vasicek (1987).

For a client j let T_j be the time to default on a client's debts. It is assumed that everyone will default once and as the time of the future event is unknown at present the time T_j < ¥ is a random variable. If Q_j is the cumulative probability distribution of T_j then it can be easily verified that the transformed variable $X_j=N^{-1}(Q_j(T_j))$ is standardized normal (mean 0, standard deviation 1). The advantage is that after the transformation we can take the assumption that the variables are multivariate normal and given their mutual correlation r properties of normal variables can be used to obtain an analytic result. This approach is called the Gaussian copula model. The following one-factor model is used

(6) $$X_j = \sqrt{r} \cdot M + \sqrt{1-r} \cdot Z_j$$

where M captures the systematic factor and Z_j the client specific. All Z_j's and M have independent standard normal distributions.

The one-year probability of default *PD* of the client j can be expressed as

(7) $$\begin{aligned}
&\Pr\left[T_j \le 1\right] = \Pr\left[X_j \le N^{-1}(Q(1))\right] = \\
&= \Pr\left[\sqrt{r} \cdot M + \sqrt{1-r} \cdot Z_j \le N^{-1}(Q(1))\right] = \Pr\left[Z_j \le \frac{N^{-1}(Q(1)) - \sqrt{r}M}{\sqrt{1-r}}\right] = \\
&= N\left(\frac{N^{-1}(Q(1)) - \sqrt{r}M}{\sqrt{1-r}}\right)
\end{aligned}$$

The next step is to consider *M* as the systematic driver of portfolio default rates. The model can be used for a simulation as follows: first generate randomly the value of an *M* from a standardized normal distribution and then independently all Z_j. If the portfolio is large enough then the simulated default rate on the portfolio will be given by the formula above. If *M* is large the simulated default rate will be low if *M* is smaller then the portfolio default rate will be higher. For a given probability level x the critical point of *M* is given by the quantile $N^{-1}(x)$. When *M* is replaced by $N^{-1}(x)$ and *Q(1)* by the given average *PD* we get exactly the regulatory formula (5).

2.3. Unexpected Loss Given Default - The Merton Model

The Merton model has been developed (Merton, 1974) in particular for corporate debtors. It is used for example as a theoretical foundation by the CreditMetrics Methodology - JP Morgan (1997). The idea is that a company is able to pay back its loans as long as the value of its assets is above the amount of the external liabilities. At the time of granting a new loan the bank normally verifies that the assets are well above the external liabilities (i.e. that capital cushion is positive and sufficient). Nevertheless later the value of the assets stochastically goes up or down depending on a number of external and internal factors. To model the value of the assets we may use the geometrical Brownian motion equation $dA = \mu A dt + \sigma A dz$ or in a simplified approach the Brownian motion stochastic equation $dA = \mu dt + \sigma dz$ allowing negative values of *A*. The default at time t takes place iff $A(t) < L$ where *L* is the actual value of debtor's liabilities. The time of default can be defined as $T_{Def} = \inf\{t \geq 0; A(t) < L\}$. Modeling unexpected default rate UDR we simply count the number of defaults, i.e. the probability *PD* that $T_{Def} \leq 1$ in a stressed scenario. The final loss in case of default, or its estimation at time 1, clearly depends on the terminal value *A(1)* that have evolved randomly from the time T_{Def} to the time 1. The model should also take into account the effect of a distressed sale or bankruptcy proceeding when assets are sold with a discount. Hence setting $ULR = PD \times LGD$ with an average (i.e. deterministic) *LGD* we loose the variability of the development of the assets value *A* from the time T_{Def} to 1. This approach is equivalent replacing of the value A(t) with *L*(1-LGD)* for all $t \geq T_{Def}$.

On the other hand if at the time 1 we have new information on the value of *A* then the best estimate of expected losses (BEEL) generally differs from the initial estimate at the time of default. If A went up again then the new estimate would be better than the average LGD and if the value of assets A has deteriorated since the default then the estimate would be lower. A possible estimation for BEEL could be given by the expression $\max(0, L - H(A(1)))$ where $H(\cdot)$ is a haircut function reflecting the distress liquidation. In practice the recovery process starts at the time of default (or sooner) and the first experiences show soon what the realistic recovery rate might be.

The structural model could be improved (in particular for retail loans and mortgages) as follows: let $Q(t)$ model the liquid assets and let a partially or fully independent variable $A(t)$ model the value the debtor's illiquid assets. The default in payment occurs when Q is less than the amount to be paid P. It is then and usually with a delay that the bank starts to investigate the real value of the assets A. Once the assets are properly assessed it is usually clear whether the value would be sufficient to cover the amount of the loan or not (though the process of liquidation may be time consuming). Consequently BEEL at time 1 is normally a good account level estimate of the final loss; substantially better then the estimate of LGD set at time of default when the illiquid assets have not been inspected yet. Consequently we conclude that the losses recognized at time 1, normally a few months after default, are almost as precise as the losses calculated when the recovery process is completed (which may take a few years).

The analysis above shows that the regulatory formula can be interpreted as a model capturing the stochastic process A up to the time of default T_{Def} then replacing the values of A by the average LGD fully disregarding the element of unexpected LGD losses that can be almost fully recognized at the end of the year. Moreover the sooner defaults take place the more randomness of the after-default process is lost (see Exhibit 2).

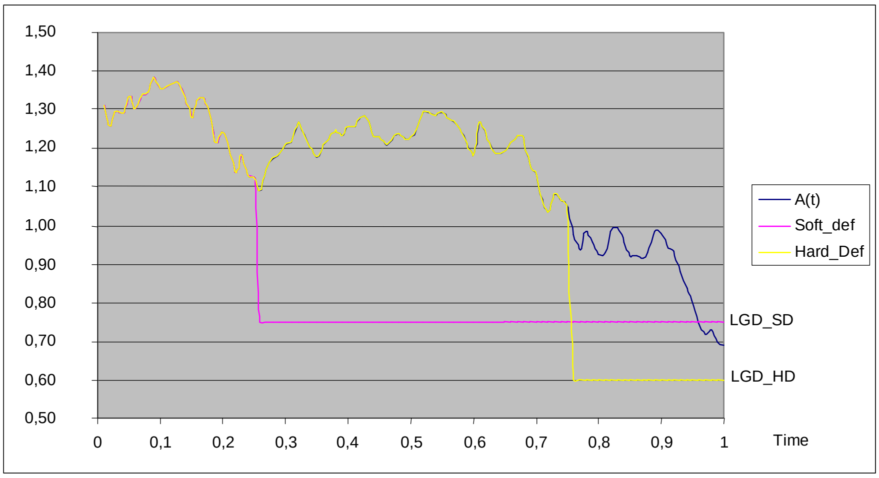


Exhibit 2 (Stochastic modeling of assets up to the time of default)

In this context we can interpret soft default as a larger threshold $L_1 > L$ arguing that some debtor's assets may not be liquid enough to cover insufficient cash for loan repayment on time. If T_1 is defined using the threshold L_1 then generally $T_1 < T$. The two models yield $PD_1 > PD$. The loss given defaults then must be set to keep the expected loss $EL = PD \cdot LGD$ constant, i.e. $LGD_1 < LGD$. Since in the case of soft definition of default the model is less stochastic and more deterministic on should not be surprised by the effect observed in Exhibit 1.

2.4 A Monte Carlo Simulation

We will use the simple structural model $dA = m dt + s\, dz$ to test the phenomenon of the dependence of PD unexpected risk measures on the definition of default. The simulation will be done on a large portfolio level with a correlation ρ involved (we shall use the regulatory level of ρ=10%). Analogously to Section 2.2 we decompose each random element dz_j into its systematic and debtor-specific part $dz_j = \sqrt{r}\, dw + \sqrt{1-r}\, dx_j$. We start with a large portfolio of old and new non-defaulted loans, for the sake of simplicity all with outstanding amount 1. We will use a Monte Carlo simulation to plot the distribution of the loss rate on the portfolio level:

a) directly simulating the values of the loans at time 1 as $\min(1, A_j(1))$ with a random haircut involved if $A_j(1) < 1$,

b) simulating the values through the hard definition of default as 1-LGD_H (where LGD_H is the average loss given hard default) if there is a default during the first year and 1 otherwise,

c) simulating similarly the values through a soft definition of default with a threshold S>1.

The steps of the simulation may be summarized as follows:

1. To simulate end-of-month $A_j(t)$ for each debtor $j=1,\ldots,N$ (N is the number of debtors or receivables in the portfolio) we need to set up initial characteristics $A_j(0)>1$, m_j, s_j. The values are randomly sampled with normal distributions such that the resulting mean portfolio loss and default rate correspond to normal values, e.g. 2% to 3%.

2. For $k=1,\ldots,M$ (M is the number of Monte Carlo portfolio simulations) simulate the vector of monthly systematic factors $Dw(k,0), \ldots, Dw(k,11)$ with the normal $N(0, Dt)$, distribution ($Dt=1/12$).

3. For each k simulate the vectors of account specific monthly systematic changes $Dx(k,j,0), \ldots, Dx(k,j,11)$ for $j=1,\ldots, N$ again with the normal $N(0, Dt)$ distribution. All generated random numbers are sampled independently. Set

$$\Delta z(k, j, m) = \sqrt{r}\,\Delta w(k,m) + \sqrt{1-r}\,\Delta x(k, j, m) \text{ for } m=0,\ldots,11.$$

4. Calculate inductively starting with $A(k,j,0)=A_j(0)$ the simulated asset values

$$A(k,j,m/12)=A(k,j,(m-1)/12) + m_j Dt + s_j Dz(k,j,m-1) \text{ for } m=1,\ldots,12.$$

5. For each k and j calculate the account level loss rate as LR(k,j)=0 if A(k,j,1)≥1 and as $LR(k, j) = (1 - A(k, j,1))(1 - haircut)$ if A(k,j,1)<1 where randomly generated haircut reflects partially the fact that assets must be sold in a distress to cover at least partially the loan. The portfolio loss rate is then

$$LR^k = \frac{1}{N}\sum_{j=1}^{N} LR(k, j) \qquad (8)$$

and the mean loss rate, i.e. expected loss as $EL = \frac{1}{M}\sum_{k=1}^{M} LR^k$. The values of LR^k can be plotted into a histogram (Exhibit 3); the 99,9% quantile $q_{99,9\%}(LR^k)$ (unexpected loss rate) and the loss rate VaR $LR_{VaR} = q_{99,9\%}(LR^k) - EL$ are then estimated from the obtained distribution (Exhibit 4).

6. For each k count the number N_H^k of hard defaults, i.e. of debtors j for which $A(k,j,m)<1$ for some m. The observed (simulated) portfolio default rate is then $PD_H^k = N_H^k / N$. Calculate the mean default rate as $PD_H = \frac{1}{M}\sum_{k=1}^{M} PD_H{}^k$ and $LGD_H=EL/PD_H$. The regulatory unexpected loss formula is based on an estimation of the unexpected default rate. We estimate $q_{99,9\%}(PD_H^k)$ from the distribution of PD_H^k and set $LR_{Var,H} = (q_{99,9\%}(PD_H^k) - PD_H)\cdot LGD_H$. This calculation is equivalent to resetting $A(k,j,m)=LGD_H$ for all defaulted accounts from the month of default on. The estimated unexpected loss would then be the same if we proceeded as in step 5. The distribution of the losses can be also depicted by the histogram of $PD_H^k \cdot LGD_H$.

7. Repeat the calculations for a soft default threshold S > 1.

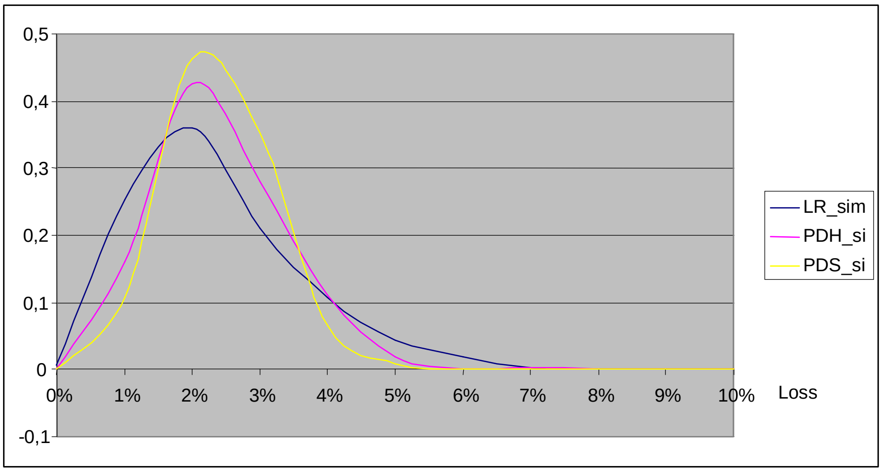


Exhibit 3 Histograms of simulated losses (N=200, M=1000)

	EL (mean)	PD	LGD	LR std	LR_Var(99,9%)
Loss Rate simulation	1,90%	N/A	N/A	1,25%	6,96%
PD Hard Def. Simulation	1,90%	7,43%	25,63%	0,95%	4,50%
PD Soft Def. Simulation	1,90%	13,70%	13,90%	0,75%	3,03%

Exhibit 4 Characteristics of the simulated loss rate distributions

Exhibits 3 and 4 clearly demonstrate that unexpected losses based on PD simulation significantly underestimate unexpected losses based on full loss rate simulation. Moreover the softer definition of default reduces variability of the simulated loss distribution measured e.g. by the standard deviation or by the 99,9% quantile (4,5% in case of hard default simulation while 3,03% in case of soft default simulation). The difference would be even larger if we simulated separately the liquid and illiquid assets (with a correlation factor). The value of illiquid assets recognized short after default and which may be significantly below 1, then mostly determines the final loss on the account. In the simulation above the account level loss rate depends on the simulated value of A(k,j,1) which would not be to far from the level where the assets have been a few months before.

2.5. Gaussian Copula and Beta Distribution Based Model

In this section we are going to propose, in a spirit similar to the Basle II approach, an analytic formula for the unexpected loss on a homogenous portfolio of defaulted receivables due to lower than expected recoveries. The resulting LGD unexpected risk is an interesting measure per se which can be used for example for a consistent definition LGD economic capital etc. Moreover the resulting distribution of LGD losses will be in the next section combined with the (regulatory) distribution of probabilities of default to get an integral measure of unexpected total credit losses.

Let LR_j denote the percentage loss rate (i.e 1 – the recovery rate) on a defaulted receivable j $(j=1,\ldots,n)$ observed from the perspective of a completed recovery process. The completion may take up to 3-5 years, however as we argued in the previous section the loss given default recognition takes much shorter time (e.g. revaluation of collateral, negotiations with the client etc.) while implementation of the chosen recovery strategy takes a long time. Hence by modeling the distribution of LR_j (on account or portfolio level) in full recovery time horizon we get a realistic estimation of the *LGD* unexpected risk in a one year (or time of default to one year) horizon.

Since the portfolio is *LGD* homogenous we can assume that the distribution of all LR_j is the same with certain cumulative probability distribution function Q. LR_j can be transformed as in Section 2.2 to a standardized normal variable $Y_j=N^{-1}(Q(LR_j))$. Standardized loss (recovery) rates are by the way used for example by the KMV Loss Calc methodology (see Gupton (2005) or Kim (2006)). Let us use again use the one factor model for Y_j

(9) $$Y_j = \sqrt{r} \cdot V + \sqrt{1-r} \cdot W_j$$

with independent standardized normal V and W_j. A number of studies (see Altman (2004)) have confirmed not only that there is a correlation between the rates of default and the recovery rates but moreover that the two variables are driven by a common economic factor. This is in particular intuitive in the case of mortgages when a poor state of economy drives not only more clients to default but also reduces the value of the collaterals. The correlation could be estimated from historical data. However if there are only limited data (which is mostly the case) it makes sense to use the same regulatory correlation coefficients (with an average taken for the Basle II class of "other receivables" where the correlation coefficient depends on PD). For a given probability level x (e.g. 99,9%) similarly as above the losses on a portfolio level are generated just by the value of V as the independent values W_j diversify away for a large n. The unexpected portfolio loss rate takes place if V is at the high level expressed by the quantile $N^{-1}(x)$. The unexpected loss rate is

(10) $$ULR = E\left[Q^{-1}(N(\sqrt{r}N^{-1}(x) + \sqrt{1-r} \cdot W)\right]$$

where the expectation is taken over all values of a standardized normal variable W. It might be tempting to use instead the median

$$ULR \cong Q^{-1}(N(\sqrt{r}N^{-1}(x)),$$

but since the function Q^{-1} is generally strongly nonlinear the error (difference between the median and mean) is too significant as our empirical testing confirmed. Consequently we strongly recommend using the expression (10) where the expected value is calculated numerically using the standardized normal density of W, i.e.

(11) $$ULR = \int_{-\infty}^{+\infty} Q^{-1}(N(\sqrt{r}N^{-1}(x) + \sqrt{1-r} \cdot w)\frac{1}{2p}e^{-\frac{w^2}{2}}dw.$$

Compared to the Vasicek formula we have not eliminated the probability distribution function Q. Note that there is one difference: while in the case of default rates we model on an account level a variable taking only two values (0 – no default and 1 – default) in the case of LGD we model a variable taking in general any value in the range [0,1] and so the distribution Q does matter even at the portfolio level. Given the correlation and probability level x we still need to model the LGD distribution Q.

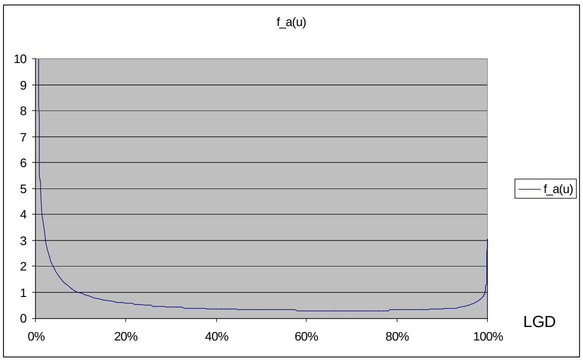
Empirical studies (Gupta (2005), Schuermann (2004)) show that the beta distribution is relatively appropriate for LGD modeling. The beta distribution probability density function with minimum 0, maximum 1, and parameters α, β is

$$(12) \qquad Beta(x,a,b)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$$

where $\Gamma(a)=\int_0^\infty x^{a-1}e^{-x}dx$ is the standard gamma function. The parameters α, β can be calculated from the mean μ and standard deviation σ of the modeled variable

$$(13) \qquad \begin{aligned} a &= m\left(\frac{m(1-m)}{s^2}-1\right), \\ b &= (1-m)\left(\frac{m(1-m)}{s^2}-1\right) \end{aligned}$$

To calibrate the distribution function we can use either the mean and the standard deviation from historical LGD data on the product or, if not available, public data which are unfortunately as far as we know available only for public corporate bonds and bank loans (see e.g. Altman et al. (2004)). Exhibit 5 shows how a bimodal account level beta distribution is transformed into a portfolio beta distribution.



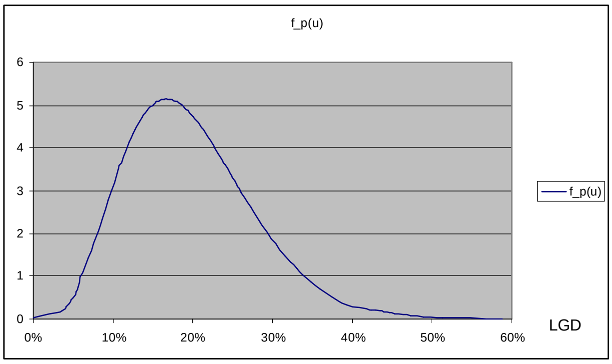


Exhibit 5 Account level beta distribution (μ=20% and σ=30%) and its transformation into a portfolio distribution (ρ=10%)

2.6 Lognormal distribution Based Model – PD and LGD convolution

Portfolio level total loss rate expressed as *LR=PD·LGD* can be modeled through a distribution of the two variables *PD* and *LGD* and their (log) convolution. *LGD* here is not taken as an average but as a random variable defined as the (in advance unknown) loss rate achieved on the defaulted part of the loan portfolio. The regulatory (Vasicek) model provides a reasonable model for *PD* and the technique above a model for the *LGD* variable.

We propose a practical approach based on lognormal distributions to combine the two factors. If we assume that *PD* and *LGD* are jointly lognormal, then *LR* is lognormal as well as ln *LR* = ln *(PD*×*LGD)* = ln *PD* + ln *LGD*. Given the means and standard deviations of *PD* and *LGD* (or rather of ln *PD* and ln *LGD*) and assuming that the logs of the two variables have a correlation r we can calculate the mean and standard deviation of *LR* and all quantiles needed. It can be empirically verified that it is feasible to approximate the distribution of *PD* with the lognormal one. The approximation is more questionable for *LGD* as it takes values in [*0,1*] with mean somewhere in the middle. Nevertheless as *LR* has again a distribution empirically close to a lognormal one the approach may be considered as acceptable. Recall that if X is a lognormal random variable such that ln(X/X_0) is $N(m,s^2)$, i.e. normal with mean m and standard deviation s, then

(14) $$E(X) = X_0 e^{m+\frac{s^2}{2}}$$
$$s(X) = X_0 e^m \sqrt{e^{s2}-1}$$

If we are given $E(X)=X_0$ and $s(X)= s_0 < X_0$ then the equations (14) are solved setting

(15) $s = \sqrt{-\ln(1-(s_0/X_0)^2)}$ and m = -s^2/2.

So given the key parameters of the distributions of *PD* and *LGD*, $E(PD)=p_0$, $s(PD)=s_{PD}$, $E(LGD)=l_0$, $s(LGD)= s_{LGD}$, and a correlation r of ln *PD* and ln *LGD* we calculate the standard deviations s_{PD}, s_{LGD} and the means m_{PD}, m_{LGD} of ln *PD* and ln *LGD* according to (15). The variable ln $LR/(p_0l_0)$ = ln PD/p_0 + ln LGD/l_0 then has the mean $m = m_{PD} + m_{LGD}$ and the standard deviation

(16) $$s = \sqrt{s_{PD}^2 + 2r s_{PD} s_{LGD} + s_{LGD}^2}\ .$$

It is then easy to verify that

(17) $E(LR) = p_0 l_0 e^{r s_{PD} s_{LGD}}$ and

(18) $s(LR) = E(LR)\sqrt{1-Ae^{-2r s_{PD} s_{LGD}}}$ where $A = \left(1-(s_{PD}/p_0)^2\right)\left(1-(s_{LGD}/l_0)^2\right) < 1$.

The Exhibit 6 shows the result for different input parameters of *PD* and *LGD*. The column “PDLR std” and “VaR_PDLR” shows the standard deviation and 99,9% Var of the variable $LR=PD\times l_0$ with the given parameters, i.e. of the loss rate modeled just based on the *PD* distribution in the regulatory manner. On the other hand the columns “LR std” and “VaR_LR” show the measures based on the convolution *PD*×*LGD* using the formulas (17) and

(18). It again shows how significantly the regulatory *PD* based modeling approach underestimates the total unexpected loss if *LGD* variability is taken into account.

PD mean	PD std	LGD mean	LGD std	Rho	**LR mean**	**PDLR std**	**LR std**	**VaR_PDLR**	**VaR_LR**
2,0%	1,50%	60,00%	15,0%	15,0%	**1,24%**	**0,90%**	**0,98%**	**12,0%**	**14,6%**
4,0%	2,50%	30,00%	10,0%	10,0%	**1,23%**	**0,75%**	**0,86%**	**7,0%**	**9,7%**
4,0%	2,50%	30,00%	10,0%	25,0%	**1,27%**	**0,75%**	**0,92%**	**7,0%**	**11,2%**
6,0%	4,00%	60,00%	15,0%	15,0%	**3,71%**	**2,40%**	**2,64%**	**25,1%**	**31,5%**
10,0%	5,00%	50,00%	15,0%	15,0%	**5,13%**	**2,50%**	**3,03%**	**17,7%**	**26,3%**

Exhibit 6 The mean, standard deviation, and 99,9% VaR of $LR=PD \times LGD$

We need to keep in mind that these are portfolio level distribution parameters. Hence we do need to use the copula approach explained in Section 2.5 to transform an account specific beta distribution with observed parameters into a portfolio distribution with a mean and standard deviation used in the lognormal model.

The model also provides another explanation of the regulatory capital sensitivity on the definition of default. The idea is to fix parameters of the loss distribution *LR* and recalculate the "regulatory" *PD* based economic capital. I.e. for a given *PD* mean p_0 find the standard deviation s_{PD} and estimate the regulatory-like *PD* based capital based on the 99,9% quantile of the distribution $PD^* \, l_0$.

To do that we must to solve two equations for three unknowns l_0, s_{LGD}, and s_{PD}. To get a unique solution we need to add a relationship between l_0, s_{LGD} that could be based for example on empirical observations shown in Exhibit 7.

Loan/bond seniority	**Median (%)**	**Mean (%)**	**Standard dev. (%)**	**Portfolio st.dev. (%)**
Senior-secured loans	27,00	31,50	24,40	8,52
Senior-unsecured loans	49,50	45,00	28,40	9,57
Senior-secured bonds	45,51	47,16	23,05	8,20
Senior-unsecured bonds	57,73	65,11	26,62	9,08
Senior-subordinated bonds	67,65	69,83	24,97	8,65
Subordinated bonds	68,04	70,97	22,53	8,07
Discount bonds	81,75	79,07	17,64	6,96
Total sample bonds	**59,95**	**65,69**	**24,87**	**8,65**

Exhibit 7 Loss Given Default on Public Corporate Bonds (1974–2003) and Bank Loans (1989-Q3–2003)[2]

Let us apply simple linear regression to the data to get the approximate relationship for the portfolio LGD standard deviation

[2] Based on prices just after default on bonds and 30 days after default on loans. Source: K. Emery (2003) - Bank Loans - and Altman and Fanjul (2004) - Bonds. The last column shows transformed portfolio loss distribution standard deviations with 10% correlation based on (11).

(19) $s_{LGD} = 9{,}84\% - 2{,}4\% l_0$.

Intuitively speaking since s_{LGD} remains almost constant and the standard deviation of the total losses remains constant as well the *PD* standard deviation s_{PD} remains almost constant as well. Consequently the ratio s_{PD} / p_0 is decreasing going from the hard to the soft definition of default. This means that in the PD based approach where unexpected losses are modeled through the variable $PD \cdot l_0$ the standard deviation $s(PD_S \cdot l_S) = s_{PD,S} \cdot l_S \approx \frac{s_{PD,S}}{p_S} EL$ for the soft definition default is lower than $s(PD_H \cdot l_H) \approx \frac{s_{PD,H}}{p_H} EL$ in the hard default case.

To perform a numerical simulation we may proceed as follows. Set *LR* mean $EL=E(LR)$ and standard deviation $s(LR)$ at a level corresponding to observed data. Fixing the two parameters of the *LR* distribution let p_0 range from an initial (hard default value) PD_H up to PD_H*2 due to gradual softening of the default. Next we need to solve the equations (17), (18), and (19) for the unknown l_0, s_{LGD}, and s_{PD}. Initially set $s_{PD} = s_{LGD} = 0$, then calculate l_0 from (17), then s_{LGD} and s_{LGD} from (19) and (15), and finally s_{PD} from (16). The calculation is repeated until s_{PD} converges to a value with a given precision.

Our numerical simulation (Exhibit 8) indeed shows a strong dependence of the $PD \cdot l_0$ 99,9% VaR (i.e. a proxy of the regulatory capital) based on the described model where we start with the hard default probability 3% and change the definition of default so that the PD goes up to 6% (the other parameters are: E(LR)=1,25%, σ(LR)=1%, and ρ=15%.) The LGD standard deviation follows the equation (19) for changing definition of defaults while the PD standard deviation and the LGD mean are recalculated according to (17) and (18) in order to keep the total loss rate mean and standard deviation σ(LR) constant. It is obvious that the dependence based on the convolution of lognormal distributions is quite similar to the initially observed dependence (Exhibit 1).

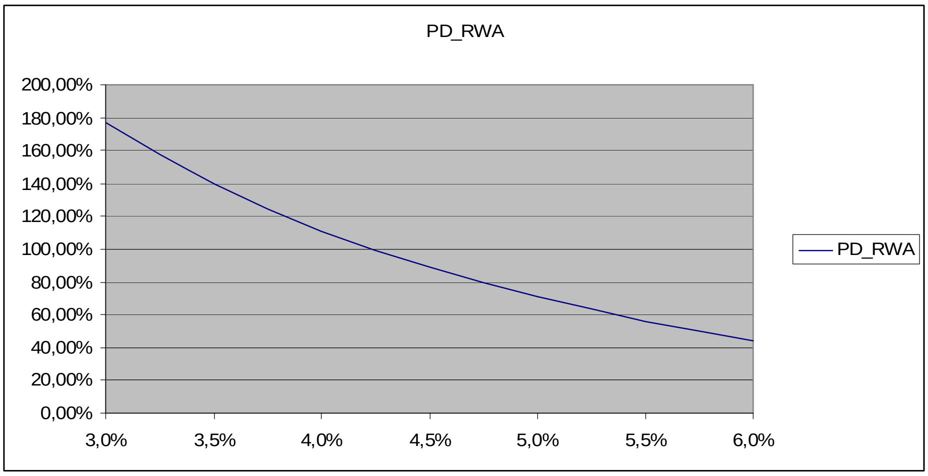


Exhibit 8 (99,9% PD-based Var Capital dependence on the definition of default)

2.6 Binomial LGD model

According to some empirical studies unimodal beta distribution does not necessarily faithfully model the distribution of recovery rates (see Schuermann (2004)). The recovery rates can be for some products bimodal – the recovery rates are either rather low or rather high. This is not surprising in particular in the case of collateralized products like mortgages. The collateral is either successfully sold and the defaulted receivables more or less paid back or there is an unexpected problem with the receivable and the recovery is low. Taking a simplified approach we can assume that that the variable LGD is binomial, i.e. there are only two possible recovery rates (and LGD) values: 0 and 1. Then (looking backward or into the future) we can distinguish just two types of defaults: full-loss-defaults and zero-loss-defaults. As there is no loss on zero-loss-defaults those cases may be forgotten and all we need to model are the full-loss defaults. The probability of full-loss-defaults is *PD*LGD* as *LGD* is the probability of full loss conditioned by a default and *PD* is the probability of default. The loss conditioned by a full-loss-default is certainly 100%, so unexpected default rate equals directly to the unexpected loss in this case. The event of a full-loss-default can be modeled using the Gaussian copula Vasicek model:

$$(20)\qquad UEL_2 = N\left(\frac{N^{-1}(PD\cdot LGD)+\sqrt{r}\cdot N^{-1}(0{,}999)}{\sqrt{1-r}}\right)\cdot EAD\,.$$

UEL_2 captures both the unexpected default rate and *LGD* given a systematic correlation r. To get LGD contribution we need to deduct the unexpected loss capturing only the unexpected default rate:

$$(21) \qquad UEL_1 = N\left(\frac{N^{-1}(PD) + \sqrt{r}\cdot N^{-1}(0{,}999)}{\sqrt{1-r}}\right)\cdot LGD\cdot EAD.$$

Unexpected loss is measured in the two formulas (20) and (21) as the loss amount out of the total portfolio outstanding before default. Hence the *LGD* economic capital as a percentage of the defaulted notional *EAD*PD* can be expressed as

$$(22) \qquad \mathrm{LGD}_{\mathrm{VaR}} = UEL_{LGD} = \frac{UEL_2 - UEL_1}{EAD\cdot PD}.$$

One may note that there is a dependence of the *LGD* economic capital on *PD* (the economic capital is allocated to a portfolio of already defaulted receivables). This problem can be solved letting *PD* converge to 1. Then we get a model for a fully defaulted portfolio and the formula (22) converges to

$$(23) \qquad LGD_{VaR} = \left(N\left(\frac{N^{-1}(LGD) + \sqrt{r}\cdot N^{-1}(0{,}999)}{\sqrt{1-r}}\right) - LGD\right).$$

There is a connection between the binomial approach and modeling *LGD* via bimodal Beta distribution with two extreme values. It can be easily empirically verified that the binomial approach is rather conservative compared the previous models.

3. Development of Scoring Functions and the Definition of Default

In the last part of the paper we analyze the question whether the definition of default impacts quality of scoring functions developed using ordinary methods (like logistic regression) modeling the probability of default. Banks could certainly use one definition of default for regulatory purposes (i.e. regulatory *PD* and *LGD* calculation) and other, more appropriate, for scoring function development. Nevertheless according to Basel II the definition default should be used consistently throughout the organization (see paragraph 456, in Basel (2006)) hence the regulatory definition should be used for scoring function development as well. As we have analyzed in the previous sections the banks are in a sense (probably unintentionally) motivated to use a softer definition of default. Our hypothesis on the other hand is that softer definition of default leads to a worse scoring function. The reasoning is that banks need to model real losses, i.e. number of clients that really do not pay their obligations to the bank. Soft definition of default (for example defined as just 30 overdue or with a very low materiality threshold) may capture a large number of clients (in particular retail) that just forgot to pay an exact amount on time but do not really intent not to pay and do have enough financial resources. It may be argued that poor payment discipline is already correlated with higher probability of default but separation of the two types of the events (poor payment discipline and real lack of resources/intention not to pay the loan back) is difficult. Hence it appears preferable to use an absorbing (hard) definition of default where clients, once marked as defaulted do not return (of course with some exceptions) to the ordinary non-defaulted status and there is a real problem. Application of a soft definition of default on the other hand means that many clients marked as defaulted become current, i.e. non-defaulted with observed *LGD*=0, moreover this situation may repeat several times for a single client.

It would be optimal to test our hypothesis on real data. Due to their lack we will use for a simulation the structural Merton - like model applied to a retail portfolio. We will model the disposable income A as a Wiener process $dA = \mu dt + \sigma\, dz$ with different initial parameters $A(0), \mu, \sigma$. The soft default event will be defined by the condition Days Past Due $DPD \geq 60$ and hard default event by $DPD \geq 180$ (relaxing the regulatory 90 DPD condition). The amount of a monthly installment will be constantly 1 and the event missing a payment at time *t* occurs if $A(t) < 1$ or when the client forgets to pay. The second independent event *Ind(t)* (with values 0 or 1) will be parameterized by a probability *p* characterizing payment discipline of a client set at the beginning of the simulation. We assume that the events "forgetting to pay in a months" are independent. Hence for example if *p*=10% then the client may with probability 1%=10%*10% forget to send two consecutive payments and be marked as (soft) defaulted even though there is no problem with his ability to pay. The probability of forgetting four consecutive payments is in this situation only 0,01% hence almost negligible. Consequently the simulated default data will contain much more "noise" when the soft definition of default is used rather than when the hard definition of default is applied.

To simulate a history of days past due it is sufficient to generate only end-of-month A(t) like in Section 2.4. The initial attributes of a client $A(0), \mu, \sigma$, *p* can be interpreted as the initial disposable income, income growth potential (e.g. due to age, education, etc.), income stability (e.g. due to marital status, number of children, etc.), and payment discipline. We shall consider the triple $A(0), \mu, \sigma$ as a proxy of full demographic data provided by a client applying for a loan. The three variables will be used as the explanatory variables of our logistic regression. Banks normally approve only the applications promising sufficient probability of repayment. Hence generating the initial data we estimate (by a separate simulation) the probability of default and accept only those with probability of default less than 15%. On the other hand we randomly modify the initial triple to capture the fact that clients are not always completely precise in their applications and moreover there is generally a time lag between the time of application and the start of our simulation.

We generate in the way described above one set of clients to be used for the scoring function development and another set of clients for testing. We firstly generate the monthly values *A(t)* and *Ind(t)* for each client. The data are then used to select soft and hard defaulted clients. The first set of data is used for a logistic regression function development with the explanatory variables $A(0), m, s$ to estimate the probability of 1) soft default 2) hard default. The regression has been performed using the application STATA and the resulting coefficients are shown in Exhibit 9. The first scoring function would be used by a bank relying on the soft definition of default while the second by a bank relying on the hard definition of default. The most important output of a scoring function is how it orders (ranks) different clients – this allows assigning ratings from bad to good to different sets of clients, and estimating their probability of default. Notice the relative differences between the soft default and hard default scoring functions coefficients (in particular comparing the coefficients of m). Consequently the functions may provide quite different ranking to identical clients. A better quality of the hard default scoring function is already indicated by the R^2 coefficient (13,7% compared to 11% of the soft default scoring function). Calculating the Gini coefficient on the testing data set we all also get a better result for the hard scoring function (72,4% compared to 71,4%).

Soft default based logistic regression

```
Logit estimates                                   Number of obs   =      10000
                                                  LR chi2(3)      =     685.72
                                                  Prob > chi2     =     0.0000
Log likelihood = -2767.2988                       Pseudo R2       =     0.1102

------------------------------------------------------------------------------
def_0_1_ |      Coef.   Std. Err.      z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
initial  |   3.514344   .2114122    16.623   0.000       3.099983    3.928704
   drift |    5.95639   .7287103     8.174   0.000       4.528045    7.384536
     std |   2.656194   .1629243    16.303   0.000        2.97552    2.336869
    cons |  -1.680595   .2871526    -5.853   0.000      -2.243407   -1.117786
```

Hard default based logistic regression

Logit estimates		Number of obs	= 10000
		LR chi2(3)	= 521.40
		Prob > chi2	= 0.0000
Log likelihood = 1641.1698		Pseudo R2	= 0.1371

[illegible]	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
initial_	4.456413	.3213701	13.867	0.000	3.826539	5.086287
drift	10.83540	1.013054	10.687	0.000	8.848550	12.82259
std	2.594334	.2060633	12.590	0.000	2.998211	2.190456
cons	2.172141	.4258281	5.101	0.000	3.006749	1.337533

Exhibit 7 (Soft and Hard default logistic regression functions)

4. Conclusion

We have offered several explanations of the observed dependence of the regulatory capital on the definition of default. It is clear that regulators should seek a remedy for the issue. One possibility is to strictly control the definition of default used by banks, another to modify the regulatory formula to capture the LGD unexpected risk as well. The latter approach is preferable as a number of studies including this one indicate that the LGD risk component is significant. The fact that the current formula does not capture unexpected LGD risk is a reason of the observed sensitivity on the definition of default. Our paper provides two alternative formulas the could be used proposed for a new regulation: a complex one decomposing and combining the PD and LGD unexpected risk, and a simplified one based on a straightforward generalization of the Vasicek formula.

The current regulatory formula (probably unintentionally) motivates banks to use a softer definition of default. Our simulation indicates that such a definition may not serve well to quality of credit risk management tools, in particular for development of scoring functions. Further research to confirm the hypothesis should be however done on real banking data.

Literature

Altman E., Resti A., Sironi A., 2004, "Default Recovery Rates in Credit Risk Modelling: A Review of the Literature and Empirical Evidence", Economic Notes by Banca dei Paschi di Siena SpA, vol.33, no. 2-2004, pp. 183-208

E. Altman – G. Fanjul, 2004, ''Defaults and Returns in the High Yield Bond Market: Analysis through 2003'', NYU Salomon Center Working Paper

Brady B., Chang P., Miu P., Ozdemir B., Schwartz D., 2006, "Discount Rate for Workout Recoveries: An Empirical Study", working paper

K. Emery, 2003, Moody's Loan Default Database as of November 2003, Moody's Investors Service.

Gupton G.M., 2005, „Advancing Loss Given Default Prediction Models: How the Quiet Have Quickened", Economic Notes by Banca dei Paschi di Siena SpA, vol.34, no. 2-2005, pp. 185-230

Kim J., Kim H., 2006, „Loss Given Default Modelling under the Asymptotic Single Risk Factor Assumption", Working Paper, MPRA

Merton, Robert C., 1974, "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates," The Journal of Finance, Vol. 29, (1974), pp. 449-470.

Schuermann T., 2004, "What Do We Know About Loss Given Default", Credit Risk Models and Management, 2nd Edition, London, Risk Books

Vasicek O., 1987, "Probability of Loss on a Loan Portfolio," KMV Working Paper

Basle Committee on Banking Supervision Paper, 2005, "Guidance on Paragraph 468 of the Framework Document"

Basel Committee on Banking Supervision, 2006, **“**International Convergence of Capital Measurement and Capital Standards, A Revised Framework – Comprehensive Version”, Bank for International Settlements, 2006

Basel Committee on Banking Supervision, 1996, Amendment to the Capital Accord to incorporate market risks, Bank for International Settlements, 1996

Basel Committee on Banking Supervision, 1998, The Framework for Internal Controls in Banking Organizations, Bank for International Settlements, 1998

JP Morgan, 1997, Gupton G.M., Finger C.C., Bhatia M., Credit Metrics – Technical Document, New York