Imperial College London

Some statistical reflections about Expected Loss

Tony Bellotti, Department of Mathematics, Imperial College London

Credit Scoring and Credit Control Conference XIV
Edinburgh, 26-28 August 2015

Imperial College London

Expected Loss

For a portfolio of loans, with $i = 1$ to n accounts, expected loss is

$$EL = \sum_{i=1}^{n} PD_i LGD_i EAD_i$$

where, for each account i,

- PD_i = probability of default;
- LGD_i= loss given default;
- EAD_i= exposure at default

Questions

This formula raises several interesting questions:

(1) Does the **correlation** between PD and LGD affect the accuracy of the EL calculation?

(2) Given that LGD and EAD models are developed based *only* on accounts in default, does this introduce a **selection bias** when applying the models to EL calculation across *all* accounts in a portfolio?

(3) **Correlations** may also exist between LGD/EAD and PD/EAD. Do these correlations affect the accuracy of the EL calculation?

(4) It is possible to **include EAD as a predictor for LGD**? Is this a good thing to do or not? In particular, EAD will not be known when the LGD model is used for forecasting LGD across all accounts.

Approach to these questions

We explore these questions in the following way:-

1. Provide a derivation of EL as an expected value, based on some underlying model structures.
2. Based on this, propose an adjustment to the EL calculation.
3. Illustrate our findings using a simulation study.
4. Show results using real UK credit card data.

Derivation of Expected Loss

Consider a single account where

- $Y \in \{0,1\}$ is default indicator; 1=default event;
- $L \in [0,1]$ is LGD;
- $E > 0$ is EAD;
- $\mathbf{X}$ is a vector of predictor variables;
- $V = YLE$ is the value of loss.

Then EL can be expressed as the conditional expectation $\mathbb{E}(V|\mathbf{X} = \mathbf{x})$ (for a single account).

Derivation of Expected Loss

Suppose the following model components:-

- $\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}) = P(\mathbf{x})$ is a PD model;
- $\mathbb{E}(L|Y = 1, \mathbf{X} = \mathbf{x}, E = e) = \phi_1(\mathbf{x}) + \phi_2(e)$ is an LGD model;
- $\mathbb{E}(E|Y = 1, \mathbf{X} = \mathbf{x}) = \lambda(\mathbf{x})$ is a model of EAD.

  - Notice that both LGD and EAD models are conditional on Y=1 (default). This is natural since only accounts in default will be used when modelling them.
  - We have imposed a structure on LGD: ie it can be linearly decomposed as a function of the predictor variables and a function of EAD.

Then EL can be written as

$$\mathbb{E}(V|\mathbf{X} = \mathbf{x}) = P(\mathbf{x})\left[\phi_1(\mathbf{x})\lambda(\mathbf{x}) + \mathbb{E}\left(E\phi_2(E)\right)\right].$$

Notice the additional expected value term. We explore this further…

Inclusion of EAD with log-transformation

Since EAD is greater then 0 and typically long tailed, it is natural to use a log transformation for the model of EAD.

Hence, for EAD e and some residual ε,

$$\log(e) = \gamma(\mathbf{x}) + \varepsilon$$

and let ϕ_2 be linear on log-EAD: ie $\phi_2(e) = \beta_E \log(e)$.

Then

$$\mathbb{E}(V|\mathbf{X} = \mathbf{x}) = P(\mathbf{x})[\phi_1(\mathbf{x}) + \beta_E \gamma(\mathbf{x}) + \beta_E r]\lambda(\mathbf{x})$$

where

$$r = \frac{\mathbb{E}(\varepsilon \exp(\varepsilon)|\mathbf{X} = \mathbf{x})}{\mathbb{E}(\exp(\varepsilon)|\mathbf{X} = \mathbf{x})}.$$

If we assume $\varepsilon \sim N(0, \sigma^2)$, then it turns out that $r = \sigma^2$ and $\lambda(\mathbf{x}) = \exp(\gamma(\mathbf{x}) + \sigma^2/2)$.

Implications of the modified EL formula

$$\mathbb{E}(V|\mathbf{X}=\mathbf{x}) = P(\mathbf{x})[\phi_1(\mathbf{x}) + \beta_E\gamma(\mathbf{x}) + \beta_E r]\lambda(\mathbf{x})$$

(1) The **correlation** between PD and LGD has no impact on the calculation of EL.

(2) There is no **selection bias** introduced when applying the models to EL calculation across *all* accounts in a portfolio.

(3) The **correlation** between PD/EAD has no impact. However, an adjustment is required to deal with correlations between LGD and EAD.

(4) It is *essential* to **include EAD as a predictor for LGD**. This allows the LGD/EAD correlation to be modelled. For forecasting, the formula allows the estimate of EAD ($\gamma(\mathbf{x})$) to be used.

Conditional and "unconditional" LGD

- In this presentation, we are using LGD, genuinely conditional on default.
- This is natural in (unsecured) retail credit, when we will model LGD based on accounts in default (not possible to get LGD for non-defaults).
- However, if we have a good measure of collateral (eg for corporates) then LGD can be calculated based on collateral.
- In this sense, LGD is "unconditional" on default.
- In particular, collateral can be different in default and non-default states.
- Some authors (eg Pykhtin 2003) discuss LGD in this unconditional sense.
- Using unconditional LGD gives different answers to our questions:

(1)Yes, PD/LGD correlation needs to be factored in;

(2)Yes, there is a selection bias effect.

Ref: Pykhtin M., Unexpected recovery risk, Risk, August 2003.

Including EAD in the LGD model

- The association of EAD with LGD (β_E) needs to be fully and accurately quantified.

- Therefore, if there are any predictor variables that are highly correlated with EAD, this will be a problem
  … this is very likely (eg past account balance).

- So propose a two-stage model of LGD l :

$$l = \beta_E \log(e) + \varepsilon_{L1}$$
$$\varepsilon_{L1} = \phi_1(\mathbf{x}) + \varepsilon_{L2}$$

for some residuals ε_{L1} and ε_{L2}.

Experiments

- Although we have the mathematical result, it is a good idea to test it with data:-

  (1)To check that the result is *really* true;

  (2)To check the *magnitude* of the problem, if we do not adjust for EAD/LGD correlation.

- For this reason, we demonstrate with

  1. **Simulation study**: allows us to control the distribution of the data to test different aspects of the EL calculation.
  2. **Real data study**: using UK credit card data, to see what the affect looks like for real data.

- In both cases, logistic regression is used for the PD model and OLS regression for both LGD and EAD models.

Imperial College London

Simulation study: Data generation

- A credit portfolio was simulated with multiple risk factors to simulate default events, LGD and EAD.

Risk factors:	X1	X2	X3	X4	X5
Default	*	*	*		
LGD	*	*		*	
EAD	*				*

- All variables are standard normally distributed,
- All variables are expressed as the sum of an observable and unobservable component; only the observable component can be used in the model built, hence simulating uncertainty.
- X1 and X2 are common to more than one component, hence inducing a correlation.

Simulation study: Data generation

Simulate n=1,000,000 accounts, for good sample size.

Summary statistics for simulated data	Value
Defaults (N, %)	49,827 (5%)
LGD (mean; %LGD=0, %LGD=1)	0.8; 9%, 65%
EAD (mean; min, Q1, median, Q3, max)	1465; 199, 1007, 1341, 1782, 7503
Loss = Default*LGD*EAD (mean)	62
Default/LGD correlation	+0.283
Default/EAD(log) correlation	+0.152
LGD/EAD(log) correlation	+0.467

Simulation study: Results

	Experiment	**Mean diff %**	**P-value**
1	EL without allowing for EAD/LGD correlation	-6.1%	<0.0001
2	EL allowing for EAD/LGD correlation, but no adjustment $+\beta_E r$ in EL calculation	-4.8%	<0.0001
3	*Proposed EL calculation with adjustment*	*+0.4%*	*0.43*
4	As experiment 3 but with a single (one-stage) LGD model	+0.5%	0.29
5	Simulate with zero EAD/LGD correlation and use EL from experiment 1 again	+0.0%	0.99

- Mean diff % = 100 (EL - Observed Loss) / (Observed Loss)
- P-value for t-test between EL and observed losses.
- All these experiments are in-sample.
- Note: experiment 5 demonstrates that estimation problem is with EAD/LGD correlation, not PD/LGD, eg.

UK credit card data study

- 30,407 accounts for UK credit card, observed during 2009-2010.
- Measure default as 3 months missed payments within 2 year period.
- Predictor variables include client and account ages, application data (employment status, tenure status, months at current address) and behavioural data (balance, utilization, past delinquency) .
- Build simple underlying models for PD using logistic regression, LGD and log-EAD using OLS linear regression.
- Use 10-fold cross-validation to test accuracy of EL calculation (the sample size is small enough that an in-sample test may give training bias).
- Correlation between EAD(log) and LGD is +0.066.

Imperial College London

UK credit card data study: Results

	Experiment	Mean diff %
1	Model without allowing for EAD/LGD correlation	-2.20%
2	Model allowing for EAD/LGD correlation, but no adjustment $+\beta_E r$ in EL calculation	-2.87%
3	*Proposed EL calculation with adjustment*	*+0.12%*
4	As experiment 3 but with a single (one-stage) LGD model	+4.51%

- Mean diff % = 100 (EL - Observed Loss) / (Observed Loss)
- Note: No statistical test of difference due to relatively low sample size.

.

.

Conclusion

- We have shown that correlation between LGD and EAD requires some careful attention, but other correlations are unproblematic (eg PD/LGD).
- We have developed an adjustment to the Expected Loss calculation to address this.
- Simulation studies have confirmed these analytic results.
- Results using a real credit card portfolio also demonstrate the bias if LGD/EAD correlation is not handled correctly.
- Shows a bias in estimate of 2-3%. The proposed adjustment corrects for this.
- How this analysis works out for other credit portfolios would be the subject of further research.
- The implementation presented here is rather simplistically based on a static model framework. Further research would extend to a segmented and dynamic framework using survival models to estimate lifetime expected loss and profits.

Some statistical reflections about Expected Loss

Thank you!

I hope you have found this presentation useful.

Any questions?

Dr Tony Bellotti
Lecturer in Statistics
Department of Mathematics
Imperial College London

a.bellotti@imperial.ac.uk

Part of the **Statistics in Retail Finance Research Group** at Imperial College London.

Research, Training, Consultancy.

ICON: **www.imperial-consultants.co.uk**