Using survival models for profit and loss estimation

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

Credit Scoring and Credit Control XIII conference
August 28-30, 2013
Motivation

• There is a need to move beyond models of credit risk ranking and probability of default (PD), to use model structures that give estimates of profit and loss based on individual characteristics and at portfolio level.

• Survival models are dynamic models that can provide an estimate of PD over the *lifetime* of a credit product, enabling profit/loss estimates to be computed over a period of time.

• In this presentation, we review an experiment to use survival models as the basis of profit/loss forecasts on a portfolio of credit cards.
Points of investigation

Special consideration is given to addressing these questions:

1. How is the profit/loss estimate constructed, and what model components are required.
2. How accurate are the profit/loss estimates?
3. How well-calibrated are the estimates, by risk grade?
4. How sensitive are the estimates to different model components?
5. Do the models give us any useful information about different groups of obligors?
6. Can we use this approach for stress testing?
7. Is the survival model approach really necessary? Will simpler modelling approaches suffice?
Profit formula - Gain

- Consider a credit card account with monthly account records over months 1 to \( N \) where
  
  - \( r_i \) is the monthly interest rate;
  - \( b_i \) is the interest bearing account balance for month \( i \);
  - \( t \) is the month the account defaults; let \( t = \infty \) if it does not default;

- Then formulae for *gain* from interest payments over the period 1 to \( N \) for the account is given as

\[
G_N = \sum_{i=1}^{\min(N,t-1)} r_i b_i
\]
Profit formula - Loss

- Let $l_t$ be the loss given default (LGD) if the account defaults in month $t$.

- Then, *loss* due to default over the period 1 to $N$ is given by

$$L_N = \begin{cases} l_t b_t & \text{if } t \leq N \\ 0 & \text{if } t > N \end{cases}$$

and overall profit is given as

$$P_N = G_N - L_N$$

- Note, for simplicity, no other aspects of gain or loss are considered (eg transaction fees or cash reward payments), although in principle this should not be difficult to incorporate.
Expected value of Profit/Loss

- Treat time to default as a random variable $T \in \mathbb{N}^+$ governed by a discrete distribution $f$ with cumulative distribution $F$.

- Define the survival function $S(t) = 1 - F(t)$, where default is the failure event.

- Then expected values of gain and loss are derived as:

  $$E_T(G_N) = \sum_{i=1}^{N} S(i)r_ib_i$$

  $$E_T(L_N) = \sum_{i=1}^{N} f(i)l_ib_i$$

- Hence, $E_T(P_N) = E_T(G_N) - E_T(L_N)$.

These formulae are based on the approach given in Lyn Thomas, *Consumer Credit Models* (2007), section 4.6.
To use these formulae, an estimate of the survival function is needed. A discrete survival model is used to do this since credit accounting data is discrete:

\[ p_{jt} = P(Y_{jt} = 1|Y_{js} = 0 \text{ for } s < t, x_{jt}) = F(\beta_0 + \Phi^T \phi(t) + \beta^T x_{jt}) \]

is the PD for account \( j \) at \textit{duration} time \( t \), conditional on not defaulting before, where

- \( Y_{jt} \) indicates a default event for account \( j \) at time \( t \);
- \( x_{jt} \) are a vector of predictor variables, possibly varying over time;
- \( \phi \) is a transformation of \( t \), allowing a parametric baseline \( \Phi \) on hazard probability;
- \( F \) is a link function; in this experiment, logit.
Including the discrete survival estimate

- Let \( \tau_j \) be the duration (age) of account \( j \) at the beginning of the profit calculation period.

- Then we have

\[
S(t) = \prod_{i=1}^{t} \left(1 - p_{j(\tau_j+i)}\right)
\]

and

\[
f(t) = p_{j(\tau_j+t)}S(t - 1)
\]

- These estimates can be used directly in the profit/loss formulae.
Predictor variables for the discrete survival model

- A wide variety of predictor variables can be included in the discrete survival model:
  
  1. Application variables;
  2. Credit bureau data;
  3. Behavioural (card usage) data;
  4. Vintage effect (fixed effects);
  5. Calendar time (fixed) effect;
  6. Macroeconomic variables.

- Variables of type 5 and 6 are useful to forecast ahead with economic scenarios in mind; in particular, for stress testing.

- In these experiments, variables of types 1 and 5 have been included.
The focus of this presentation is the discrete survival model, but several other model components are needed:

<table>
<thead>
<tr>
<th>Term estimated</th>
<th>Note</th>
<th>Model</th>
<th>Model fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f(t)$ and $S(t)$</td>
<td></td>
<td>Discrete survival model</td>
<td>OK</td>
</tr>
<tr>
<td>$b_i$</td>
<td>Unconditional balance used in the calculation of gain.</td>
<td>Two-stage cross-sectional model: 1. Logistic regression for balance=0; 2. OLS for when balance&gt;0.</td>
<td>Very good</td>
</tr>
<tr>
<td>$l_t$</td>
<td>LGD model.</td>
<td>OLS regression</td>
<td>Poor</td>
</tr>
<tr>
<td>$b_t$</td>
<td>Balance conditional on default; EAD.</td>
<td>OLS regression</td>
<td>Good</td>
</tr>
</tbody>
</table>
Correlations between PD, LGD and EAD

• It is known that there are correlations between PD, LGD and EAD. This leads to problems when we want to compute an estimate of expected value of loss. In essence, what we want is

\[ E(PD \times LGD \times EAD) \]

But what we are computing is

\[ E(PD) \times E(LGD) \times E(EAD). \]

If there is a correlation, this will give the wrong result.

• In order to adjust for the correlation between components, we include a multiplicative constant \( \lambda \) which can be estimated on training data:

\[
E_T(L_N) = \sum_{i=1}^{N} \lambda f(i) l_i b_i
\]

• This seems to work well, but there is probably a better way to do this… (further investigation).
Experiments on credit card data

- UK credit card data set is used with 39,800 accounts spanning a period from July 2008 to June 2011.

- Training data is taken from July 2008 to February 2010 (19 months).
- Test data is taken from March 2010 to February 2011 (12 months).
- Hence, this experiment simulates forecasting ahead.

- Default is defined as 3 months of consecutive missed payments.

- Note that the last 4 months of data since LGD needs to be observed for at least up to that period.

- This data is confidential, hence absolute figures for default rates and profits and losses will not be shown.
Results: Estimated baseline hazard

- The baseline hazard probability is computed from the parametric estimate $\Phi^T \phi(t)$ in the survival model as:
Results: Estimated calendar-time effect

- Calendar-time fixed effects are estimated as follows.
- Higher values indicate greater risk.

This graph demonstrates that default risk was relatively high during the credit crisis period of 2008, but has reduced over 2009.

For forecasting, the value of the fixed effect at the last day of the training data (February 2010) is used.
Results: Forecast default rate by calendar month

- The discrete survival model is used to estimate monthly default rates (DR) over the forecast period.

  Estimated default rates (DR) follow observed DR, but are somewhat overestimated towards the end of the forecast period.

- DRs are generally falling over time, but this is a consequence of not considering new accounts being added to the portfolio over time (but this could be done using a simulation approach).
Results: Gain, loss and profit estimation

- Correlation between estimated and observed profit (gain-loss):
  - Linear correlation: \( \rho = 0.160 \) (Pearson’s correlation coefficient);
  - Rank concordance: \( \tau = 0.223 \) (Kendall’s correlation coefficient).

- Profit estimation (normalized) across whole portfolio:

<table>
<thead>
<tr>
<th></th>
<th>Gain (G)</th>
<th>Loss (L)</th>
<th>Profit (G-L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>2.324</td>
<td>1.324</td>
<td>1.000</td>
</tr>
<tr>
<td>Estimated</td>
<td>2.315</td>
<td>1.314</td>
<td>1.001</td>
</tr>
</tbody>
</table>

[ Note: This is a good estimate of Profit, but the high precision of the estimate is partly down to luck! This will be clear when we consider adjusting for model estimation error. ]
Results: Profit estimates by risk bucket

- Accounts are arranged into risk buckets according to their default risk over the 12 month forecast period.
Results: Profit forecasts by groups

- The profit estimates is used to forecast expected profits within different groups. In this example, employment status:
Adjusting credit score by systemic factor

- The calendar time effect is a *systemic* effect that will affect all forecasts during the same period in the same way.
- By default, it is set to 0 (the effect in the last training month, but different values can be selected to correspond with alternative scenarios.

* Loosely 95%CI based on +/-2 s.d. from distribution of estimated coefficients.
Results: Forecasts after adjusting systemic factor

<table>
<thead>
<tr>
<th></th>
<th>Gain (G)</th>
<th>Loss (L)</th>
<th>Profit (G-L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>2.324</td>
<td>1.324</td>
<td>1.000</td>
</tr>
<tr>
<td>Point estimate</td>
<td>2.315</td>
<td>1.314</td>
<td>1.001</td>
</tr>
<tr>
<td>Plausible range</td>
<td>(2.308, 2.322)</td>
<td>(1.379, 1.253)</td>
<td>(0.929, 1.069)</td>
</tr>
<tr>
<td>Prediction interval “95%”</td>
<td>(2.284, 2.341)</td>
<td>(1.589, 1.083)</td>
<td>(0.695, 1.258)</td>
</tr>
<tr>
<td>Stress test</td>
<td>2.264</td>
<td>1.759</td>
<td>0.505</td>
</tr>
</tbody>
</table>

- The plausible range shows the range of plausible profit estimates given a moderate change in the systemic effect. E.g., could be due to estimation error on systemic coefficient.

- The “95%CI” includes all observed values of gain, loss and profit, but yields a broad interval.
Sensitivity of profit forecasts to model components

- We consider how forecasts change when weaker versions of model components are included in the model.

<table>
<thead>
<tr>
<th>Results</th>
<th>Correlation $\tau$</th>
<th>Gain (G)</th>
<th>Loss (L)</th>
<th>Profit (G-L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>n/a</td>
<td>2.324</td>
<td>1.324</td>
<td>1.000</td>
</tr>
<tr>
<td>Point estimate</td>
<td>0.223</td>
<td>2.315</td>
<td>1.314</td>
<td>1.001</td>
</tr>
<tr>
<td>No adjustment for PD, EAD, LGD correlation</td>
<td>0.104</td>
<td>2.315</td>
<td>1.687</td>
<td>0.628</td>
</tr>
<tr>
<td>Weak PD model *</td>
<td>0.203</td>
<td>2.295</td>
<td>1.533</td>
<td>0.761</td>
</tr>
<tr>
<td>Weak model of balance *</td>
<td>-0.174</td>
<td>2.546</td>
<td>1.155</td>
<td>1.391</td>
</tr>
<tr>
<td>Weak EAD model **</td>
<td>0.227</td>
<td>2.315</td>
<td>1.102</td>
<td>1.214</td>
</tr>
<tr>
<td>Weak LGD model **</td>
<td>0.215</td>
<td>2.315</td>
<td>1.354</td>
<td>0.961</td>
</tr>
</tbody>
</table>

* Duration only predictor variables;
** Mean values from training data only.
Comparison with simpler approaches

- Is the lifetime survival profit estimates more complex than necessary?
- To find out, contrast against two simpler approaches:
  1. Direct OLS model of Profit;
  2. Build logistic regression PD model for 12 month period and compute estimates as:
     - Gain = mean gain (from non-defaults in training data) × (1-PD)
     - Loss = mean loss (from defaults in training data) × PD

<table>
<thead>
<tr>
<th>Results</th>
<th>Correlation $\tau$</th>
<th>Gain (G)</th>
<th>Loss (L)</th>
<th>Profit (G-L)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>n/a</td>
<td>2.324</td>
<td>1.324</td>
<td>1.000</td>
</tr>
<tr>
<td>Lifetime survival estimate</td>
<td>0.223</td>
<td>2.315</td>
<td>1.314</td>
<td>1.001</td>
</tr>
<tr>
<td>1. Direct OLS model</td>
<td>0.177</td>
<td>n/a</td>
<td>n/a</td>
<td>0.524</td>
</tr>
<tr>
<td>2. PD logistic model with mean gain and loss</td>
<td>0.203</td>
<td>2.295</td>
<td>1.533</td>
<td>0.761</td>
</tr>
</tbody>
</table>
Results: Profit forecasts by risk bucket for simpler approaches

1. Direct OLS model
2. Logistic regression PD with mean gain and loss

- It is clear from these graphs that the simpler models are unable to model the shape of profit over the different risk buckets.
Conclusion

• In this presentation, we have seen how to use discrete survival models as the basis of profit and loss calculations.

• Our experiment using real credit card data shows that these profit forecasts are accurate at aggregate levels.

• The model is able to reproduce the profit profile shown by risk bucket.

• Including a calendar time fixed effect enables a systemic effect that can, in principle, be used to construct confidence intervals and for stress testing.

• This method is superior to some simpler alternatives that might also be considered.
Further work

• Further work is required:-

1. On all the component models, but especially PD and LGD;

2. Estimating and incorporating the correlation between PD, LGD and EAD in the calculation of expected loss.

3. A more rigorous approach to extrapolating systemic effect forward for forecasting and stress testing (e.g., use of macroeconomic variables in estimate)

4. Further experiments using other credit data sets.
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