Forecasting and Stress Testing Credit Card Default Using Dynamic Models

Dr Tony Bellotti
Department of Mathematics
Imperial College London
a.bellotti@imperial.ac.uk

Prof Jonathan Crook
University of Edinburgh Business School
j.crook@ed.ac.uk

Overview

- Motivation
- Survival Models of Default
- Credit and Macroeconomic Data
- Forecasting
- Stress Testing
- Using Macroeconomic Factors
- Conclusions

Motivation

Build models of **probability of default** (PD):

Probability of a borrower missing payments on a loan.

Why?

1. Forecasting default at individual level:
   - Credit application decision;
   - Response to behaviour of existing borrowers.

2. Calculate **Expected Loss** on a portfolio of loans.

3. Regulatory requirement for PD estimates (Basel II Accord).

4. "Unexpected" Loss on portfolio:
   - Value at Risk (VaR) and Expected Shortfall;
   - Downturn conditions and Stress testing.

Background

- Traditional default models are *static*.

- Typically logistic regression models are used to model default given borrower characteristics and past credit history.

- However, we may want to include behavioural or macroeconomic variables, both of which are *time varying*.

- Traditionally, behavioural data is included in a static model as aggregates (eg *maximum* monthly spend over the last year).

- However, a more principled approach is to use a *dynamic* model that allows for time varying data.

Why include macroeconomic variables?

1. It is natural to hypothesize that borrower behaviour changes with the economic climate.

    - *Very crudely, during a recession borrowers are more likely to default*.

2. Therefore including macroeconomic variables in the model may improve PD estimates.

3. Furthermore, the inclusion of macroeconomic conditions enables a stress test of loan portfolios by observing how the estimated default rate would change with different macroeconomic scenarios.

Survival Model

We take the first time a borrower defaults as the failure event.

If an account is closed then this the observation is right-censored.

Importantly, time varying covariates (TVCs) can be included in the survival model.

The classic Cox proportional hazard (PH) model:

$$h(t, \mathbf{x}_t, \boldsymbol{\beta}) = h_0(t)\exp(\boldsymbol{\beta} \cdot \mathbf{x}_t)$$

where

- $h(\cdot)$ is the hazard of default at time t for a borrower;
- $\mathbf{x}_t$ is a vector of covariates which are possibly varying with time;
- $h_0(t)$ is a non-parametric baseline hazard rate.

Discrete survival model of default

The Cox PH model is suitable for continuous time data.

However, we have discrete time (ie monthly account data and monthly macroeconomic data).

Hence, a discrete survival model is more appropriate.

We want to model PD P_{it} for each account i at time t.

We model for duration time:

- t is the number of months since an account was opened.

Let d_{it} indicate whether account i defaults at time t after account opening (0=non-default, 1=default).

Then unconditional PD is simply $P(d_{it}=1)$.

However, we will model PD conditional on the following covariates:

- $\mathbf{w}_i$ is a vector of static application variables (AV);
- $\mathbf{x}_{it}$ is a vector of behavioural variables (BV) collected across the lifetime of the account.
- $\mathbf{z}_{it}$ is a vector of macroeconomic variables (MV) which are the same for each account on the same date;
  - that is, for any two accounts i, j having records for duration times t and s respectively, if $a_i + t = a_j + s$ then $\mathbf{z}_{it} = \mathbf{z}_{js}$;
  - where a_i is the date that account i was opened;
- The survival model assumption is that default on an account i is conditional on no previous defaults: $d_{is}=0$ for all $s<t$.

This leads to the following conditional probability:

$$P_{it} = P\left(d_{it}=1 \mid d_{is}=0 \text{ for all } s<t, \mathbf{w}_i, \mathbf{x}_{it}, \mathbf{z}_{ia_i+t}, k, l\right)$$

with fixed lags k and l on BVs and MVs respectively.

This is modelled using a logistic regression structure:

$$\begin{aligned} P_{it} &= P\left(d_{it}=1 \mid d_{is}=0 \text{ for all } s<t, \mathbf{w}_i, \mathbf{x}_{it}, \mathbf{z}_{ia_i+t}, k, l\right) \\ &= F\left(\alpha+\boldsymbol{\varphi}(t)^T\boldsymbol{\beta}_1+\mathbf{w}_i{}^T\boldsymbol{\beta}_2+\mathbf{x}_{i(t-k)}{}^T\boldsymbol{\beta}_3+\mathbf{z}_{i(a_i+t-l)}{}^T\boldsymbol{\beta}_4\right) \end{aligned}$$

where

- F is the logit link function $F(x)=1/\left(1+e^{-x}\right)$;
- $\boldsymbol{\varphi}$ is a vector transformation function of duration that is used to build a parametric survival model. In particular, we use the transformation: $\boldsymbol{\varphi}(t)=\left(t,t^2,\log t,(\log t)^2\right)$
- α is an intercept, to be estimated;
- $\boldsymbol{\beta}_1,\boldsymbol{\beta}_2,\boldsymbol{\beta}_3,\boldsymbol{\beta}_4$ are vectors of coefficients to be estimated.

Survival probability

Note that once we have this model then the survival probability of an account i over time t is given by

$$\hat{S}_i(t)=\prod_{s=1}^{t}\left(1-P_{is}\right)$$

and the PD ***within time*** t is given by the failure probability $1-\hat{S}_i(t)$.

- Note also that this formula does *not* assume independence between observations over time for the same account, because of the survival model condition.

Model comparisons

We want to test the full model with behavioural and macroeconomic data against the simple application model.

We consider four nested model specifications:

1. Duration only: fix $\boldsymbol{\beta}_2,\boldsymbol{\beta}_3,\boldsymbol{\beta}_4$ to zero.
2. Application variables only: fix $\boldsymbol{\beta}_3,\boldsymbol{\beta}_4$ to zero.
3. Application and behavioural variables only: fix $\boldsymbol{\beta}_4$ to zero.
4. Application, behavioural and macroeconomic variables: all coefficients are estimated.

Performance Measure

We are interested in how well the models predict default on a data set.

Predictive performance is measured based on the log-likelihood function for logistic regression.

Each account i contributes linearly to the log-likelihood function with

$$L_i = \sum_{s=1}^{t_i^*} d_{is} \log P_{is} + (1 - d_{is}) \log(1 - P_{is})$$

where

- t_i^* is the last observation available for account i.

Imposing the survival model condition, this leads to a **log-likelihood residual**:

$$-L_i = r_{Ci} - \delta_i \log\left(P_i^* / (1 - P_i^*)\right)$$

where

- $P_i^* = P_{it_i^*}$ denotes the PD of the last observation;
- $\delta_i = d_{it_i^*}$ indicates whether it failed;
- $r_{Ci} = -\log \hat{S}_i(t_i^*)$ is the Cox-Snell residual.

Predictive performance on a test data set of accounts i=1 to n is then given as the sum

$$-\sum_{i=1}^{n} L_i$$

Credit Card Data

Three large data sets of UK credit card data covering a period from January 1999 to June 2006 comprising over 750,000 accounts:

- All data sets include application variables taken at time of application (eg age, income, employment, credit bureau score); and
- monthly account behavioural variables (ie card usage, repayment history and missed payments).

Lag structure: Behavioural variables are lagged 3, 6, 9 or 12 months. Clearly, the older the lag the more useful the model (ie it can forecast further into the future).

Default: We define **default** on a credit card as three consecutive missed payments.
- This is a typical industry definition.

Macroeconomic data

Macroeconomic variable	***Description***	***Source***
IR	UK bank interest rates	ONS
Unemp	UK unemployment rate (in '000s) SA	ONS
Prod	UK production index (all)	ONS
RS	Retail sales value	ONS
FTSE	FTSE 100 all share index	FTSE
HP	Halifax House Price index	LBG
RPI	Retail price index (all items)	ONS
Earnings	Earnings (log) all including bonus	ONS
CC	Consumer confidence index	EC

Sources: UK Office of National Statistics (ONS), Lloyds Banking Group (LBG) and the European Commission (EC). Data is monthly and may be seasonally adjusted (SA).

MVs are included in the model as differences over 12 months.

Forecasting procedure

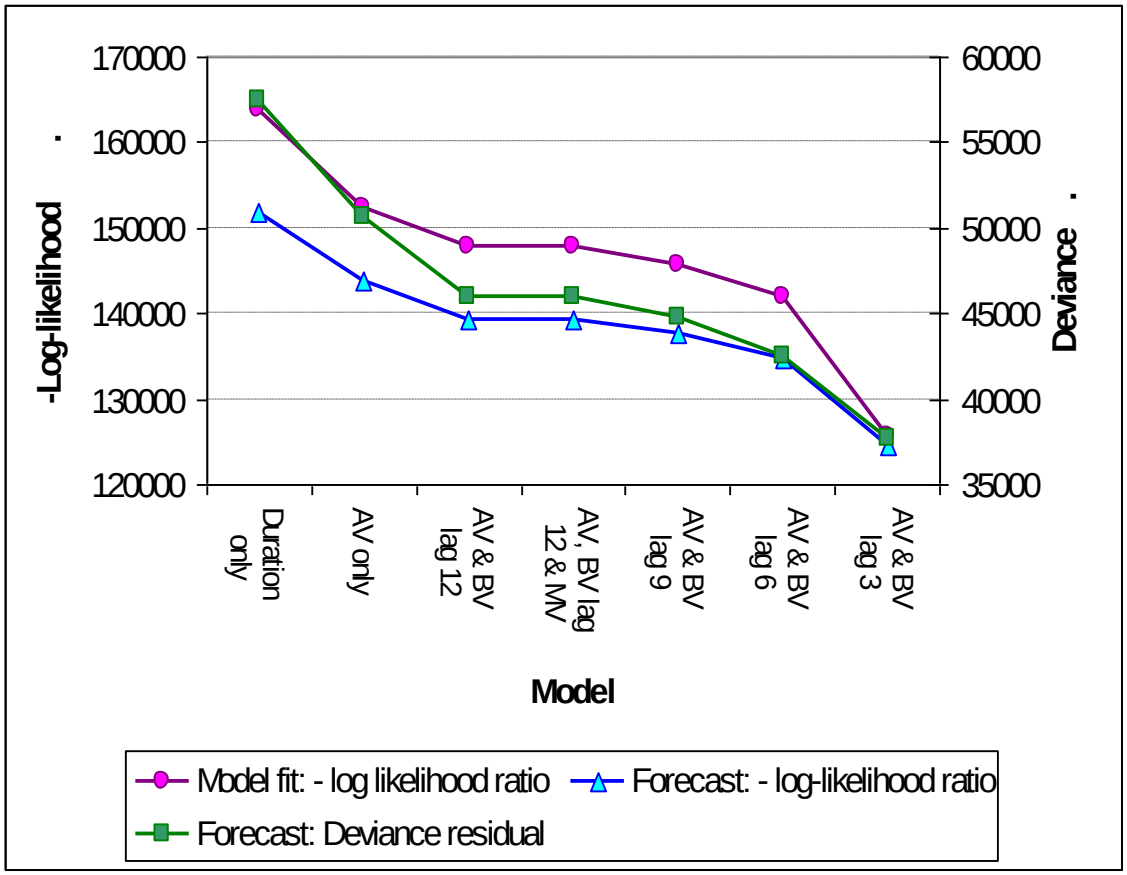
Use out-of-sample, out-of-time hold-out sample:

1. Different accounts in training/test samples;

2. Set observation date to 1 January 2005.
   - Then, training data runs from 1999 to 2004;
   - Test data runs from 2005 to June 2006.

This gives over 400,000 and 150,000 accounts in the training and test sets respectively.

Results 1: Forecasting



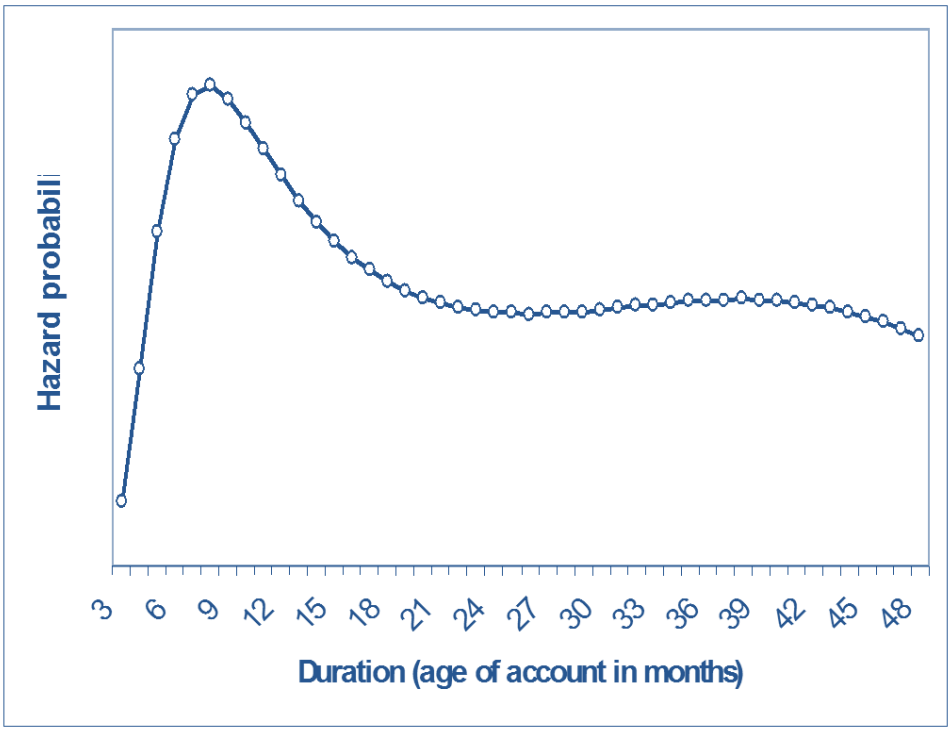
Model fit and forecast results for several model specifications

Key:

AV	application variables
BV	behavioural variables
MV	macroeconomic variables

Hazard Probability

Shape of $\varphi(t)$: typical of credit card default.



Scale on y-axis is not shown for reasons of commercial confidentiality.

Coefficient estimates

We only report coefficient estimates on macroeconomic variables, since these are the main focus of our interest.

Macroeconomic variables, lag 3 months	
Bank interest rate	0.113**
Unemployment rate	0.000672**
Production index	-0.0101
FTSE all 100 (log)	0.0591
Earnings (log)	1.57
Retail sales	0.00929
House price (log)	-0.218
Consumer confidence	-0.00217
Retail price index (RPI)	-0.0298

** Statistically significant at 0.01 level.

However, how stable are these coefficient estimates?
What about multicollinearity between macroeconomic variables?
We will return to these questions later.

Forecasting Default Rates

Observed default rate (DR) is the proportion of accounts that actually default.

For an aggregate of *N* accounts in a particular calendar month c this is

$$D_c = \frac{1}{N}\sum_{i=1}^{N} d_{i(c-a_i)}$$

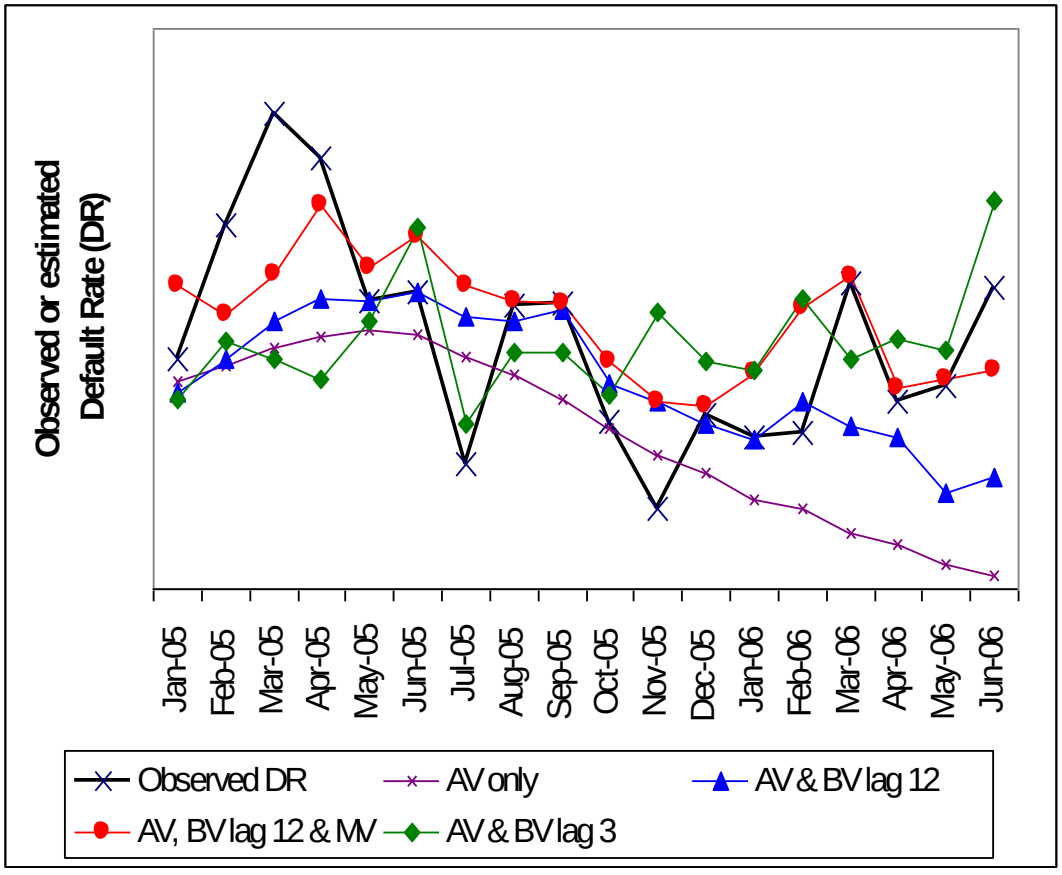
Expected DR forecast is then given by a particular model as

$$E(D_c) = \frac{1}{N}\sum_{i=1}^{N} P_{i(c-a_i)}$$

The difference between expected and observed DR gives a measure of performance for forecasts at the aggregate (or portfolio) level.

Results 2: Forecasting Default Rates

Table of results over the whole period
(Jan 2005-June 2006)



Scale on y-axis is not shown for reasons of commercial confidentiality.

Summary of results over whole period:

Model	**MAD**
AV only	0.087
BV lag 12	0.058
BV lag 12 & MV lag 3	**0.049**
BV lag 9	0.062
BV lag 6	0.070
BV lag 3	0.068

MAD = Mean absolute difference between estimated and observed DR

Scenario-based stress testing

We can specify macroeconomic conditions to see how the credit portfolio behaves in different conditions such as a recession scenario.

Using the latent variable model of logistic regression we simulate default rate (DR) for some calendar time period c, given

1. a model with MVs,
2. a vector of MVs $\mathbf{z}$, representing an economic scenario, and
3. a vector of N independent residual terms $\mathbf{e} = \left(e_{(1)}, \cdots, e_{(N)}\right)$, each drawn from F,

as

$$\hat{D}_c\left(\mathbf{z}, \boldsymbol{\varepsilon}\right) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{I}\left(\hat{\alpha} + \boldsymbol{\varphi}\left(t\right)^T \hat{\boldsymbol{\beta}}_1 + \mathbf{w}_i{}^T \hat{\boldsymbol{\beta}}_2 + \mathbf{x}_{i(t-k)}{}^T \hat{\boldsymbol{\beta}}_3 + \mathbf{z}^T \hat{\boldsymbol{\beta}}_4 + e_{(i)} > 0\right)$$

where $\mathrm{I}(\cdot)$ is the indicator function.

Monte Carlo simulation

Scenarios can be generated by simulation and a loss distribution of DR computed across a range of *plausible* economic conditions.

So extreme values are computed as **Value at Risk** (VaR) and **expected shortfall** for a percentile q and approximated as

$$P\left(D_C \le V_q\right) = q/100 \quad \Rightarrow \quad V_q \approx \hat{D}_c\left(\mathbf{z}'_{\lceil qm \rceil}, \mathbf{e}'_{\lceil qm \rceil}\right)$$

$$S_q = E\left(D_C \mid D_C \ge V_q\right) \approx \frac{1}{qm}\sum_{j=1}^{\lceil qm \rceil} \hat{D}_c\left(\mathbf{z}'_j, \mathbf{e}'_j\right)$$

respectively, where, for j=1 to m, each $\mathbf{z}'_j$ is generated by macroeconomic simulation and $\mathbf{e}'_j$ are generated randomly from F^N and both are indexed so that the simulated DRs are in descending order;

ie for all $h \leq j$, $\hat{D}_c(\mathbf{z}'_h, e'_h) \geq \hat{D}_c(\mathbf{z}'_j, e'_j)$.

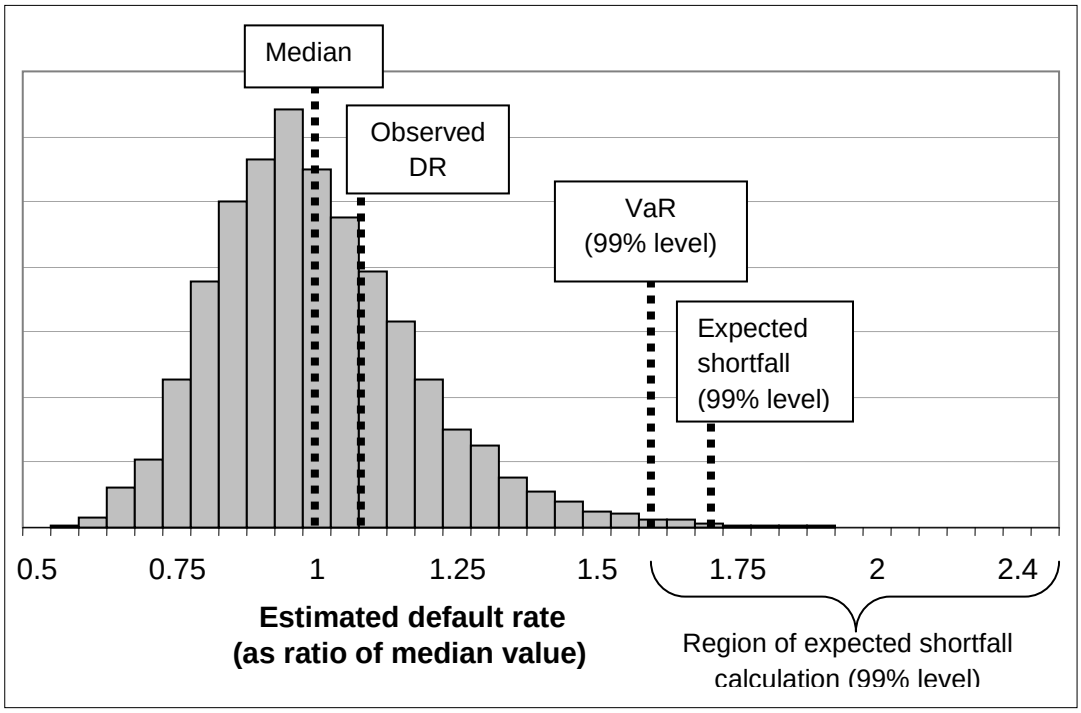
Macroeconomic simulations

To preserve the covariance structure amongst macroeconomic variables, **Cholesky decomposition** is conducted as part of the Monte Carlo simulation based on the covariance of macroeconomic variables over the training period:

- If $\mathbf{V}$ is a covariance matrix for historic macroeconomic data then it is decomposed by a lower triangular matrix $\mathbf{L}$ such that $\mathbf{V} = \mathbf{L}\mathbf{L}^T$.
- Then, if $\mathbf{u}_j$ is a sequence of independently generated values from the standard normal distribution, $\mathbf{z}^*_j = \mathbf{L}\mathbf{u}_j$ will follow the covariance structure of $\mathbf{V}$ and so can be used as plausible economic simulations.

Cholesky decomposition assumes the variables are normally distributed. However, this is not always the case for macroeconomic variables and so we apply a **Box-Cox transformation** if required, prior to simulation.

Simulation results



The distribution is based on simulation of economic scenarios for credit card accounts during December 2005, based on a model with MVs trained on data prior to January 2005, shown as a histogram. The observed DR for the test data set is shown along with Value at Risk (VaR) and expected shortfall at 99% probability. All values are expressed as a ratio of the median estimated DR.

Comparisons

Study	**Result**
Federal Reserve stress testing exercise (2009)	Between 20% and 55% rise for US credit cards, contrasting "baseline" and "more adverse" conditions.
Rösch & Scheule (2004) *Asset correlation model*	99% VaR on US credit cards: 2.3 times the mean.
Our study	Expected shortfall (99% level) on UK credit cards: 75% rise.

Rösch D and Scheule,T. (2004). Forecasting Retail Portfolio Credit Risk. *Journal of Risk Finance*, Winter/Spring, pp16-32.

Stability of models with macroeconomic conditions

Including raw macroeconomic variables in models presents a problem:

- Macroeconomic effects are different for different products – lack of stability.
- Suspect multicollinearity – problem for explanation, not forecasting (eg correlation coefficient between IR and Unemp = -0.6 and between Prod and Unemp = -0.58 over period of analysis).

To solve this, we try using **macroeconomic factors**, using factor analysis on vectors of macroeconomic variables.

- Use principal components analysis (PCA), selecting factors with eigenvalues>1.
- Precedent: *Chicago Fed National Activity Index* is a highly regarded and reliable factor representing the US economy, based on factor analysis of many macroeconomic conditions.

Results 3: Macroeconomic Factors

		Macroeconomic factors							
	EE	**MF1**	**EE**	**MF2**	**EE**	**MF3**	**EE**	**MF4**	**EE**
Eigenvalue		2.47		1.83		1.44		1.08	
Variables:									
IR	+	0.80	+	0.21	+	-0.03	–	-0.30	–
Unemp	+	-0.85	–	0.33	+	0.02	+	0.22	+
Prod	–	0.57	–	-0.42	+	0.40	–	-0.24	+
FTSE		0.01		-0.11		0.84		0.01	
Earnings	–	0.36	–	0.60	–	0.33	–	0.48	–
House price	–	0.64	–	-0.13	+	-0.36	+	0.28	–
Retail sales		0.36		-0.26		-0.35		0.58	
RPI	+	0.34	+	0.85	+	0.19	+	0.08	+
Cons conf		-0.05		-0.56		0.42		0.48	

EE=Expected effect on default

Effect in model with macroeconomic factors for two distinct products.

Model	**Factor**	**Product A**			**Product B**		
		Estimate	**Chi-sq**		**Estimate**	**Chi-sq**	
All MF	MF1	-0.0325	1.1		0.00931	0.1	
	MF2	0.1429	15.3	**	0.1659	20.1	**
	MF3	0.0329	5.3		0.0711	23.3	**
	MF4	0.0494	6.0		0.0473	4.8	

** = statistically significant at 0.01% level.

Conclusions

1. Including behavioural variables as TVCs in discrete survival models improves estimate of PD.

2. Including macroeconomic variables only minimally improves estimate of PD *at individual level*.

3. Including macroeconomic variables allows better forecasts of portfolio-level default rate (DR).

4. Including macroeconomic variables enables stress testing through realistic estimates of DR distributions.

5. Using macroeconomic *factors* may lead to more stable explanatory models of effect of the economy on consumer default.