STRATEGIC ANALYTICS

Specification Errors in Retail Lending Stress Test Models

Joseph L. Breeden
Strategic Analytics Inc.
breeden@strategicanalytics.com

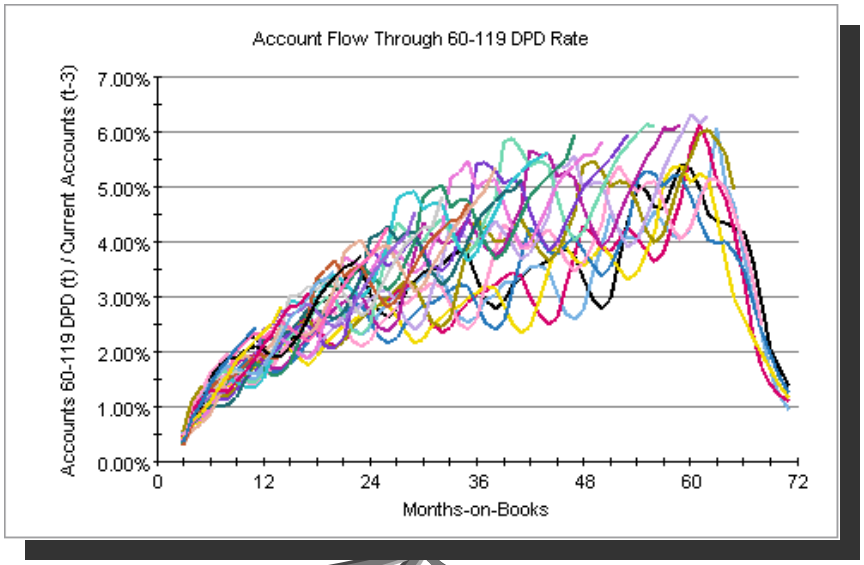
Lyn Thomas
University of Southampton
L.Thomas@soton.ac.uk

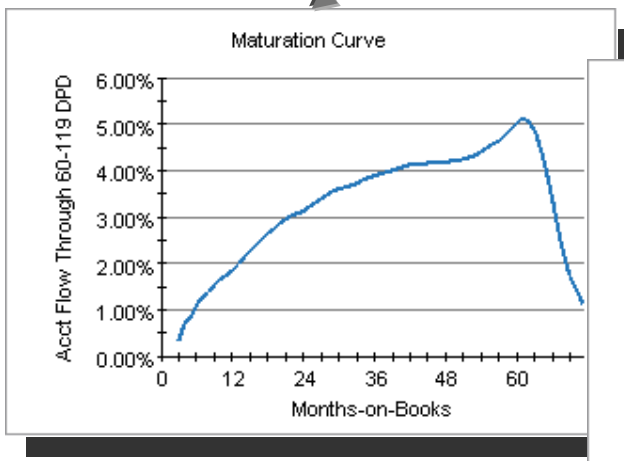
Key Questions

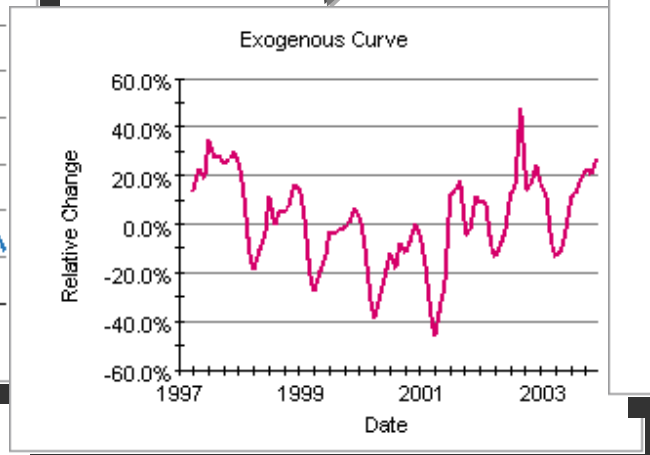
- Can we uniquely measure macroeconomic impacts in a retail lending stress test model?
- Is the answer model-specific?
- Are we helped by including macroeconomic factors and credit score attributes?

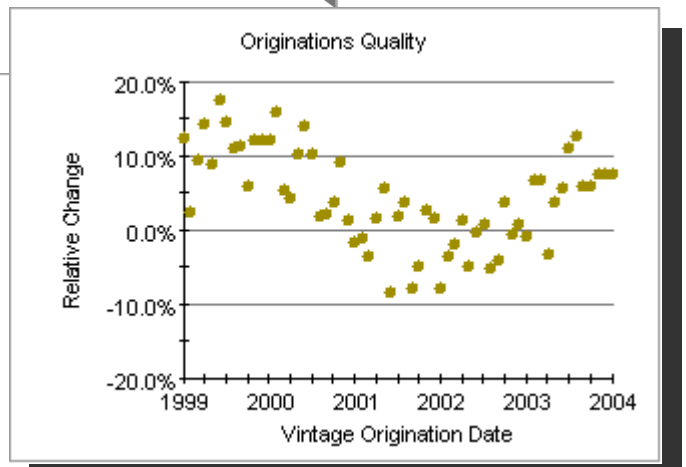
Decomposing Retail Lending Performance

- **Several different techniques can be used to decompose retail lending data into key drivers**









A Basic Model

- Whether we use
  - Dual-time Dynamics,
  - Survival Models / Proportional Hazards Models,
  - Age Period Cohort Models, or
  - Panel Data Methods,

the models can all be expressed as:

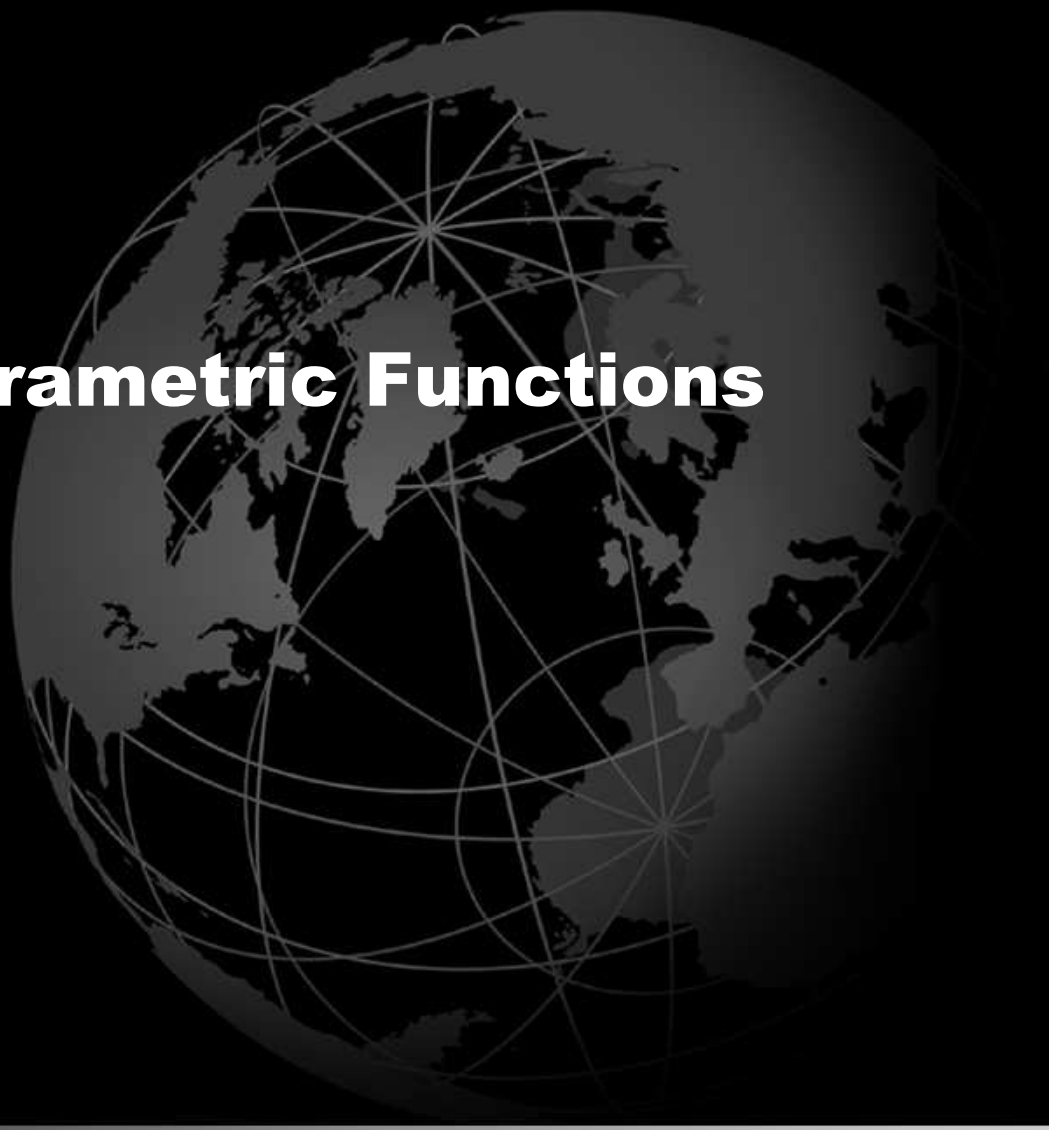
$$rate(v,a,t) = \text{lifecycle}(a) \cdot \text{environment}(t) \cdot \text{credit quality}(v)$$

for manipulation, we will express this as

$$r(v,a,t) = e^{f_m(a)} \cdot e^{f_g(t)} \cdot e^{f_Q(v)} \cdot e^{\varepsilon(v,a,t)}$$

with lifecycle, $e^{f_m(a)}$, environment, $f_g(t)$, and quality, $f_Q(v)$.

STRATEGIC ANALYTICS



Using Nonparametric Functions

A Basic Model

- Regardless of how these functions are estimated, for illustration we can take the log of the expression as

$$\ln r(v,a,t) = f_m(a) + f_g(t) + f_Q(v) + \varepsilon(v,a,t)$$

- These functions are measured on a discrete number of months. Therefore, they can be represented precisely with a polynomial through those points.

- We can separate the constant, linear, and nonlinear terms as

$$f_m(a) = \alpha_0 + \alpha_1 a + f_m'(a)$$
$$f_g(t) = \beta_0 + \beta_1 t + f_g'(t)$$
$$f_Q(v) = \gamma_0 + \gamma_1 v + f_Q'(v)$$

Constant Linear Nonlinear

The Constant Terms

- Substituting the polynomial forms into the original expression, we get

$$\ln r(v,a,t) = \alpha_0 + \alpha_1 a + f'_m(a) + \beta_0 + \beta_1 t + f'_g(t) + \gamma_0 + \gamma_1 v + f'_Q(v) + \varepsilon(v,a,t)$$

- This shows that we have three constant terms, where only one can be estimated uniquely. This can be remedied by the following definition:

$$\alpha'_0 = \alpha_0 + \beta_0 + \gamma_0$$

$$\beta'_0 = 0$$

$$\gamma'_0 = 0$$

- Rewriting gives:

$$\ln r(v,a,t) = \alpha'_0 + \alpha_1 a + f'_m(a) + \beta_1 t + f'_g(t) + \gamma_1 v + f'_Q(v) + \varepsilon(v,a,t)$$

Constant Term Interpretation

- Two of the three functions must have zero intercepts. They are measured relative to a base function.
- The simplest approach is to scale the lifecycle function to match the performance rate, and measure environment and quality relative to that rate.

The Linear Terms

- Because of the relationship between age, vintage, and time:

$$a = t - v \qquad \text{or equivalently} \qquad v = t - a$$

- We can write

$$\ln r(v,a,t) = \alpha_0' + \alpha_1 a + f_m'(a) + \beta_1 t + f_g'(t) + \gamma_1 (t-a) + f_Q'(v) + \varepsilon(v,a,t)$$

$$\ln r(v,a,t) = \alpha_0' + (\alpha_1 - \gamma_1) a + f_m'(a) + (\beta_1 + \gamma_1) t + f_g'(t) + f_Q'(v) + \varepsilon(v,a,t)$$

- With the definition

$$\alpha_1' = \alpha_1 - \gamma_1$$
$$\beta_1' = \beta_1 + \gamma_1$$
$$\gamma_1' = 0$$

this becomes

$$\ln r(v,a,t) = \alpha_0' + \alpha_1' a + f_m'(a) + \beta_1' t + f_g'(t) + f_Q'(v) + \varepsilon(v,a,t)$$

Linear Interpretation

- Because of the relationship between age, time, and vintage, we cannot have three unique linear terms, only two.
- This is the key finding of the Age Period Cohort literature, and it means that we can never be certain of the magnitude of the trends in lifecycle, environment, or credit quality.
- This specification error in the linear terms can be solved via assumptions that may not be justified, such as assuming that the environment has no net trend. Although true in the long run, data sets less than 7 years in length can easily show a trend.

The Nonlinear Terms

- Substituting expressions for the nonlinear terms, we get

$$f_m'(a) = \sum_{i=2}^{N-1} \alpha_i a^i$$

$$f_g'(t) = \sum_{i=2}^{N-1} \beta_i t^i$$

$$f_Q'(v) = \sum_{i=2}^{N-1} \gamma_i v^i$$

$$\ln r(v,a,t) = \alpha_0' + \alpha_1' a + \beta_1' t + \sum_{i=2}^{N-1} \alpha_i a^i + \sum_{i=2}^{N-1} \beta_i t^i + \sum_{i=2}^{N-1} \gamma_i v^i + \varepsilon(v,a,t)$$

This can be rewritten as

$$\ln r(v,a,t) = \alpha_0' + \sum_{i=1}^{N-1} \left(\alpha_i + (-1)^i \gamma_i\right) a^i + \sum_{i=2}^{N-1} (\beta_i + \gamma_i) t^i + \sum_{i=2}^{N-1} \sum_{j=2}^{N-1} (-1)^i \delta_{ij} a^i t^j + \varepsilon(v,a,t)$$

where δ_{ij} are the cross-term coefficients.

Nonlinear Interpretation

- The cross-terms between a and t cannot be rewritten purely as functions of age or time.
- Therefore, we cannot eliminate the nonlinear function in vintage from the equation.
- The nonlinear functions *are* unique.
- We can confirm via a couple of different methods that the two excess constant terms and the third linear term are all of the model specification errors in this system.

Account-level Models

- The previous result is generic whether the model is account-level or vintage level.

  $p(i,a,t) = e^{f_m(a)} e^{S(i)} e^{f_g(t)} e^{\varepsilon(i,a,t)}$ where $S(i)$ is a score

- With an account-level model, we can still get trends in the origination quality with time if we look at the average residual by vintage.
- All of the same model specification errors occur in account-level models, if those models include lifecycle, credit quality, and environment.

STRATEGIC ANALYTICS

Using Factor-based Functions

Specific Economic Factors

- If the nonparametric function for the environment is replaced with specific macroeconomic factors, the account-level formula can be rewritten as

$$p(i,a,t) = e^{f_m(a)+S(i)+\mathbf{C}\cdot\mathbf{E}(t)+\varepsilon(i,a,t)}$$

- If we do a search across all available economic factors, $\mathbf{C}\cdot\mathbf{E}(t)$ can take on any trend, reintroducing the specification error in trend.
- If we assume that we know exactly the right macroeconomic factors so that no spurious trends are introduced, we still have the problem that management-specific factors versus time are usually present.

Trends in Error Terms

- Unexplained management actions can introduce temporal structure in the error term.
- We can express this as $\varepsilon(v,a,t) = \varepsilon(t) + \varepsilon'(v,a,t)$

  where $\varepsilon'(v,a,t)$ is IID and $\varepsilon(t)$ is the net error as a function of time.
- The temporal error can be written as a combination of linear and nonlinear components:

$$\varepsilon(v,a,t) = et + \varepsilon'(t) + \varepsilon'(v,a,t)$$

Error Term Trends

- Substituting the error term back into the forecast equation gives

$$p(i,a,t) = e^{f_m(a)+S(i)+\mathbf{C}\cdot\mathbf{E}(t)+et+\varepsilon'(t)+\varepsilon'(i,a,t)}$$

- Again using the expression $a = t - v$ we get

$$p(i,a,t) = e^{f_m(a)-ea+S(i)+ev+\mathbf{C}\cdot\mathbf{E}(t)+\varepsilon'(t)+\varepsilon'(i,a,t)}$$

which we can rewrite as

$$p(i,a,t) = e^{f'_m(a)+S'(i)+\mathbf{C}\cdot\mathbf{E}(t)+\varepsilon'(t)+\varepsilon'(i,a,t)}$$

$$f'_m(a) = f_m(a) - ea$$

$$S'(i) = S(i) + ev$$

Factor Model Interpretation

- If a search is conducted over many possible economic variables and trends thereof, spurious explanations of residual trends can occur, meaning that we still have a model specification error.
- Anything less than a perfect factor model will produce residual trends that can be confused with trends in maturation or score.
- The same problem exists for Scoring models. Residual trends are easily explained with spurious loan or consumer attributes. A perfect score would leave an unexplained residual that still creates a trend ambiguity.

STRATEGIC ANALYTICS

Models without Specification Errors

Simplified Models

- Models that do not include all three components (lifecycle, environment, and quality), do not have a specification error.
- BUT, they achieve uniqueness at the cost of severe assumptions.
- e.g. A typical Logistic Regression model includes neither lifecycle nor environment.
  - Not including those terms implicitly sets their trends to 0.
  - Therefore, all trends are forced into the scoring model.
  - This is one contributor to the need to frequently recalibrate the scoring model.

Conclusion

- Any model that does not include one of the three components (lifecycle, environment, or quality) is setting that component's trend to zero, to be absorbed by the other components.
- From an estimation perspective, the model is "unique", but under an unjustifiable assumption for nearly all data sets.
- Models that do include all components simply make the specification errors explicit.

Remedies

- Even with very long data sets, 15 years or more, the economic impacts should be flat, but management policies and account features definitely are not, so we cannot assume a flat environmental trend.
- The best approach is to measure maturation curves from very long data sets by each product sub-type and hold them fixed. Models that do not include lifecycle cannot correctly measure macroeconomic sensitivity.
- Otherwise, we can never be absolutely certain of the sensitivity of the portfolio to macroeconomic or scoring factors.