Marginal Chi² Analysis:

Beyond Goodness of Fit for Logistic Regression Models

Quantitative Financial Risk Management Centre

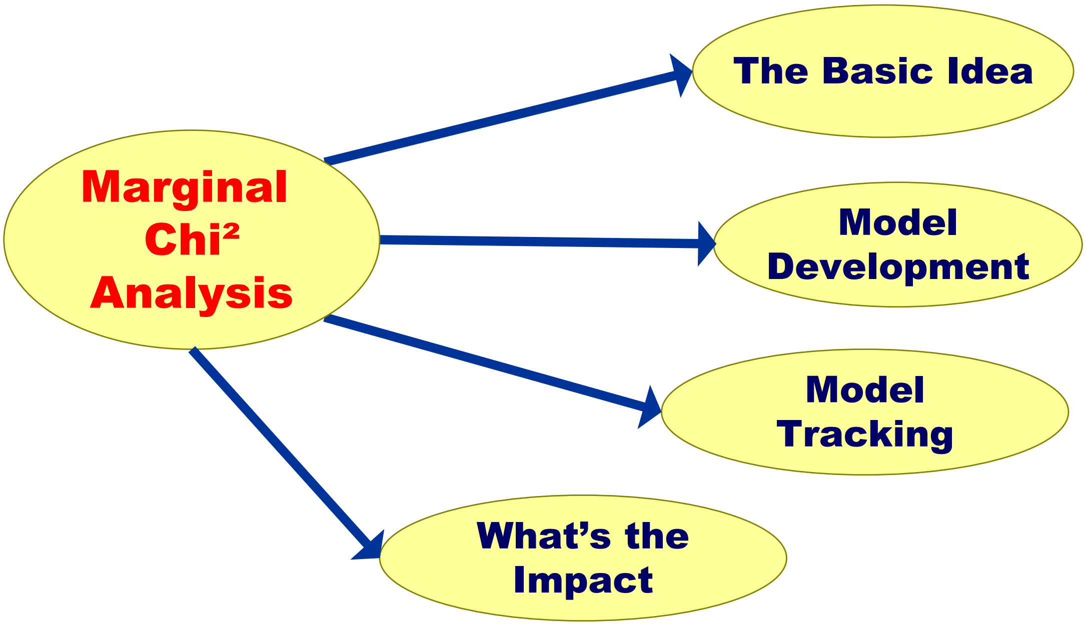
Conference on Risk Management in the Retail Financial Services Sector

London – 22-23 January 2009

Gerard.Scallan
@scoreplus.com

→ data → information → profit

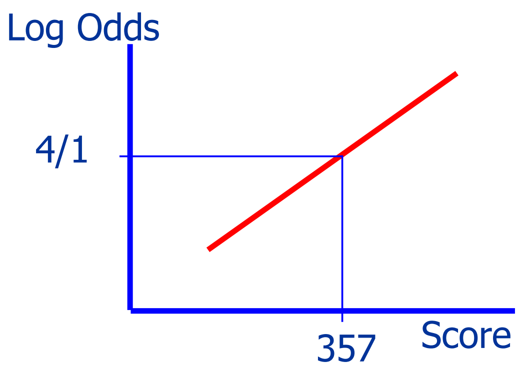
Structure of Presentation



→ data → information → profit

Logistic Regression:
Two basic ideas

Score = Log (Odds)



Actual = Expected

- For each categorical variable in model:
  - e.g. residential status
- Actual Goods in Attribute
  = Expected Goods in Attribute
- Actual Bads in Attribute
  = Expected Bads in Attribute
- Direct consequence of maximum likelihood equations
- Analogous result on averages for continuous predictors

Model correctly estimates "average" risk for each group

Actual = Expected Equations
… equivalent to Maximum Likelihood

Problem : estimate scorecard β from sample of Goods (G) and Bads (B)

For case i : $\Pr_\beta(i \in G) = \dfrac{e^{x_i'\beta}}{1+e^{x_i'\beta}}$ $\qquad \Pr_\beta(i \in B) = \dfrac{1}{1+e^{x_i'\beta}}$

Likelihood Function : $L(\beta) = \prod_{i \in G} \dfrac{e^{x_i'\beta}}{1+e^{x_i'\beta}} \times \prod_{i \in B} \dfrac{1}{1+e^{x_i'\beta}}$

$$\ln L(\beta) = \sum_{i \in G} x_i'\beta - \sum_{i \in G \cup B} \ln(1+e^{x_i'\beta})$$

Maximise by setting partial derivatives w.r.t. each component j of β to zero :

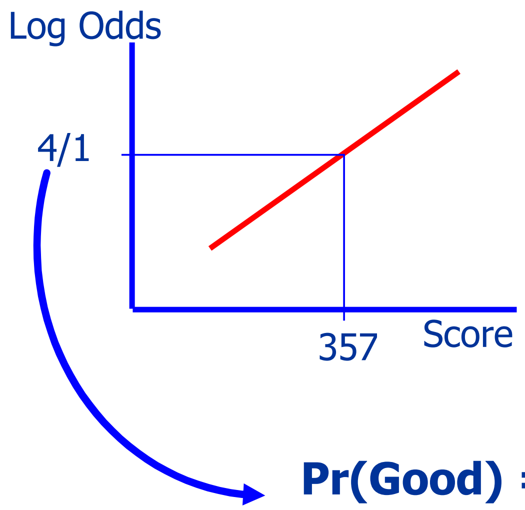
$$\frac{\partial \ln L(\beta)}{\partial \beta_j} = \sum_{i \in G} x_{ij} - \sum_{i \in G \cup B} \frac{e^{x_i'\beta} x_{ij}}{1+e^{x_i'\beta}} = \sum_{i \in G} x_{ij} - \sum_{i \in G \cup B} x_{ij} \Pr_\beta(i \in G) = 0$$

Let $x_{ij} = 1$ if i is in category A_j, $x_{ij} = 0$ otherwise :

$$\left\| A_j \cap G \right\| = \sum_{i \in A} \Pr_\beta(i \in G)$$

Actual Goods = Expected Goods

What is "Expected"?



Pr(Good) = 0.8
Pr(Bad) = 0.2

Model implies "expected" outcome for each sample point

Characteristic in model
(Categorical variables)

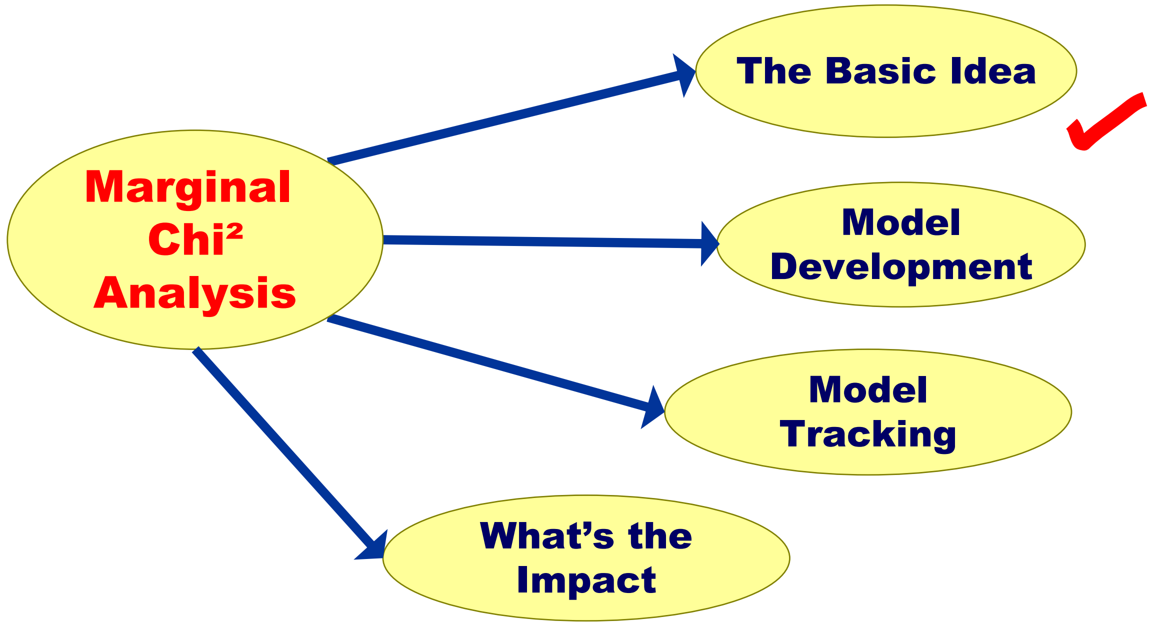
Subpopulation: **High Risk Population - Model Build Sample**
Characteristic **C007 CTO**
Current Credit Turnover

Attribute Description		**ACTUALS** Goods Count	Bads Count	Total Count	Weight of Evidence	**EXPECTED (by score)** Goods Count	Bads Count	Total Count	Weight of Evidence
1.	<= 500	761	415	1176	-1.503	761.0	415.0	1176.0	-1.503
2.	<= 1500	197	57	254	-0.875	197.0	57.0	254.0	-0.875
3.	<= 2500	148	35	183	-0.678	148.0	35.0	183.0	-0.678
4.	<= 3000	99	23	122	-0.666	99.0	23.0	122.0	-0.666
5.	<= 4000	253	46	299	-0.413	253.0	46.0	299.0	-0.413
6.	<= 5000	369	52	421	-0.157	369.0	52.0	421.0	-0.157
7.	> 5000	5896	3[illegible]9	6205	0.838	5896.0	30[illegible].0	6205.0	0.838
	Total (Valid)	7723	937	8660		7723.0	937.0	8660.0	

	Actuals		Expected (by score)	
Information value:	0.940		0.940	
Gini Coefficent	47.7%			
Likelihood Ratio Chi² (G²):	814.210	6 D.F.	814.210	6 D.F.
Chi Square Significance:	**0.00000%**		**0.00000%**	

Exact Equality: maximum likelihood equations

Structure of Presentation



Marginal Chi²:
Characteristic not (yet) in model

- Null Hypothesis: Existing score accurately estimates probabilities
  - Probabilities generate "expected" values in each cell

Debit Turnover	OBSERVED			EXPECTED			
	Goods	*Bads*	*Total*	*Goods*	*Bads*	*Total*	
<= 1000	436	**174**	610	487.7	**122.3**	610	*overscored*
1000 <= 2000	178	**38**	216	184.6	**31.4**	216	*overscored*
2000 <= 2500	84	**17**	101	86.2	**14.8**	101	*overscored*
2500 <= 3500	263	**46**	309	263.1	**45.9**	309	*ok*
> 3500	6240	**618**	6858	6179.4	**678.6**	6858	*underscored*
Total	7201	**893**	8094	7201	**893**	8094	
	Chi² =	33.06	D.F. =	4	p-value	0.00012%	

- Calculated on model build sample:
  - Intercept term in model guarantees actual = expected for total sample
- Use Log-Likelihood Chi² - a matter of taste!

Observed pattern not explained by model estimates
=> score is not a sufficient statistic for risk

Chi² Measure - Pros and Cons

Pros

- Identify candidates for entry to model
- For many potential predictors, expected converges to actual rapidly
  - As terms added to model
  - Indicates common information content
  - Gives understanding of collinearity structure
- Highlights "incremental" information

Cons

- Lots of very significant misfits
- Chi² measures certainty – not distance
  - 0.0000009% vs. 0.0000007% meaningless
- Ambiguity in degrees of freedom
  - Classed characteristics
- Chi² statistic proportional to sample size
  - Hinders learning across samples
- Beware of false positives!

Right idea – wrong packaging

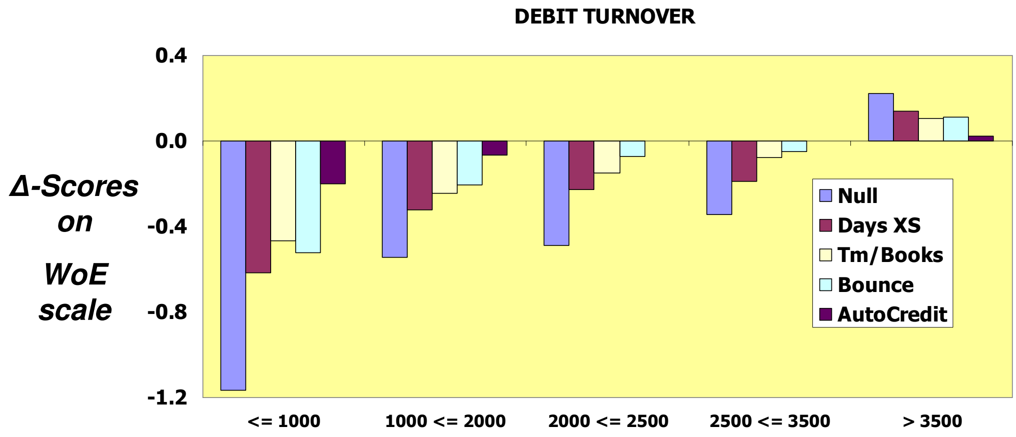
Marginal Information and Delta Scores

Debit Turnover	OBSERVED Goods	OBSERVED Bads	OBSERVED WoE	EXPECTED Goods	EXPECTED Bads	EXPECTED WoE	Δ-score
<= 1000	436	174	-1.17	487.7	122.3	-0.70	-0.46
1000 <= 2000	178	38	-0.54	184.6	31.4	-0.32	-0.23
2000 <= 2500	84	17	-0.49	86.2	14.8	-0.33	-0.16
2500 <= 3500	263	46	-0.34	263.1	45.9	-0.34	0.00
> 3500	6240	618	0.22	6179.4	678.6	0.12	0.10
Total	7201	893	0.00	7201	893	0.00	0.00
	Chi² =	33.06	D.F. =	4	p-value	0.00012%	
	Marginal Information Value					0.086	

- Weight of Evidence (WoE) = log (Attribute Odds) – log (Population Odds)
  - One-dimensional score coefficients
- Delta Score = Observed WoE – Expected WoE
  - **Approximation** to score coefficients needed to line up expected with observed
- Marginal Information Value = $\text{Avg}_{\textbf{Good}}$(Delta Score) - $\text{Avg}_{\textbf{Bad}}$(Delta Score)
  - Similar to Kullback-Liebler Information Value
  - Increased spread between average score of goods and bads
  - … if this characteristic brought into model

Measuring Collinearity
Overlaps in predictive power

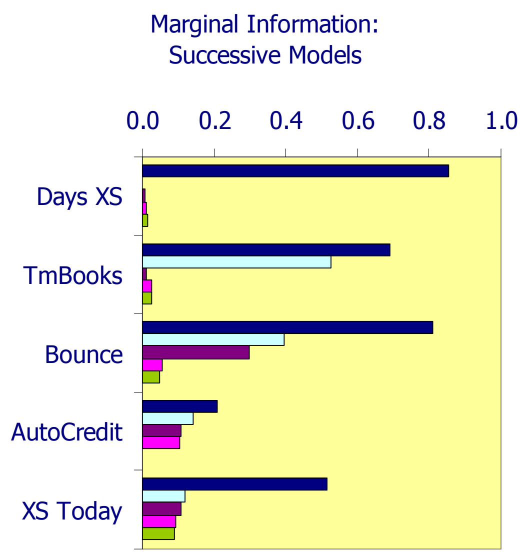
- Most information is not unique to a single characteristic
- Delta scores reduce in magnitude as "correlated" variables enter model



Small Delta Scores => Information already covered by other characteristics in model

Selection of Model Characteristics
Marginal IV

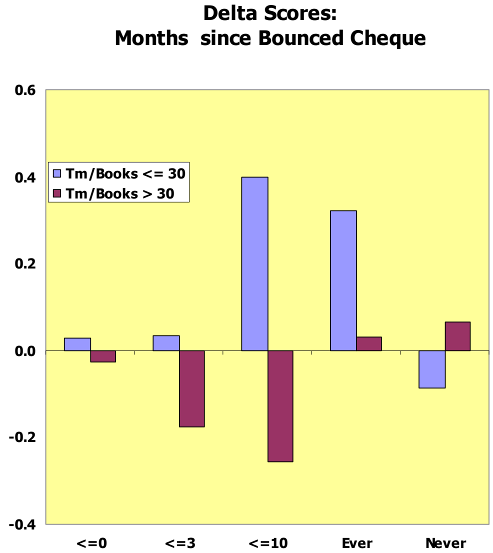
- Marginal IV is best indicator of potential contribution to model
- Choose the largest Marginal IV
- Provided "significant" Marginal Chi²
  - Problem with degrees of freedom
- Better approach than Stepwise
- Negative Marginal IVs indicate possible over-fitting
- Rule of Thumb:
  - **-.020 < MIV < .020**



Zero Marginal Information = Sufficient Statistic

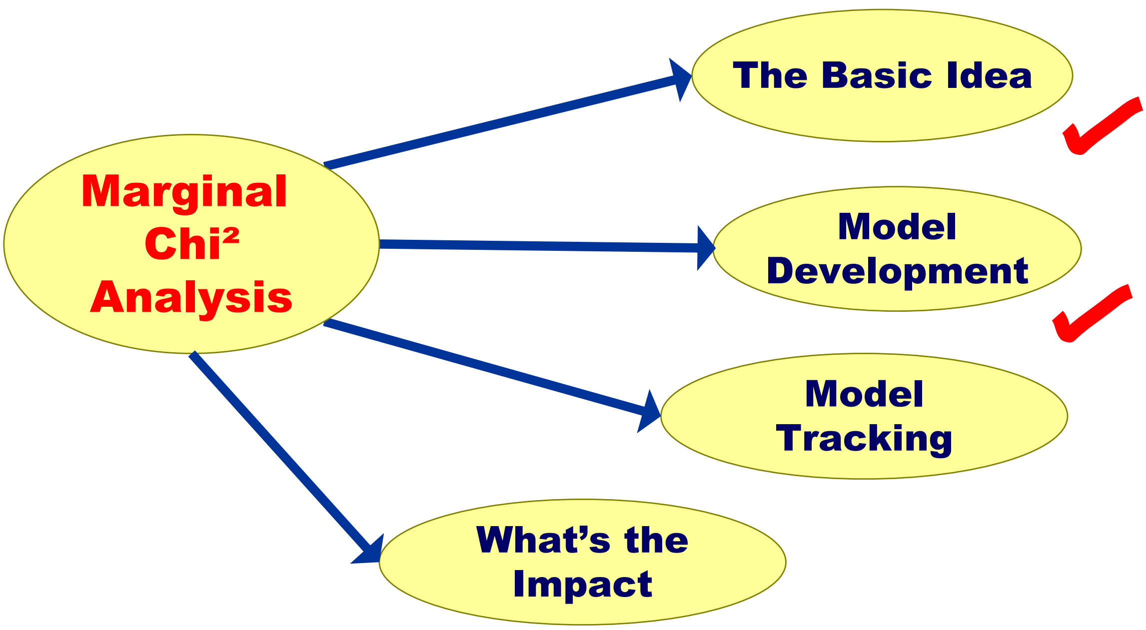
Model Segmentation
Testing for Interactions

- Characteristic interactions
  - => Multiple models
  - e.g. Delinquency - Time on books
- Test for Actual = Expected on each subpopulation
  - For each predictive characteristic
  - Enables systematic screening for interactions
- Small samples => Statistics matter!
- Shows many splits unnecessary



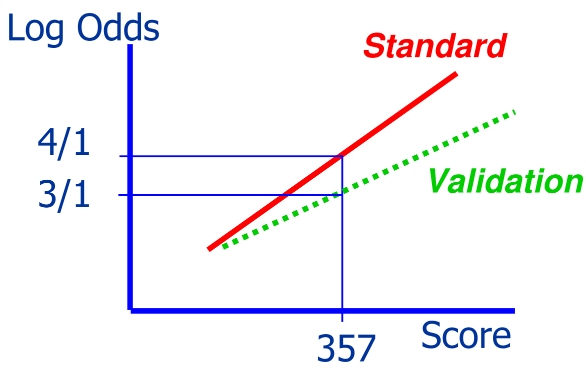
Clear conceptual framework (and algorithm) for tough problem

Structure of Presentation



Tracking Approach
(and model validation!)

Score = Log (Odds)

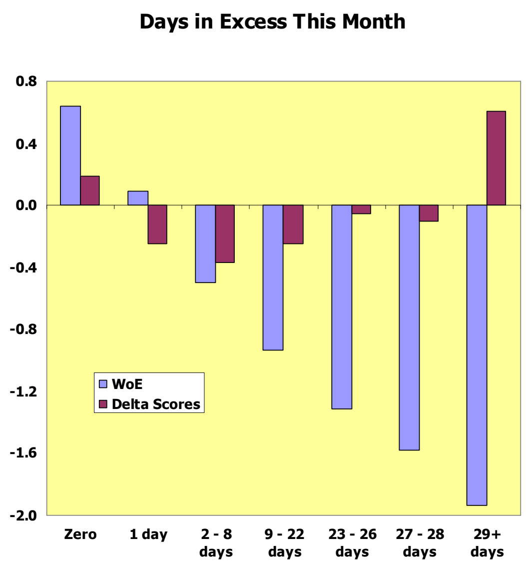


Validation Process

- Key business decisions based on assumed score-risk relationship
  - Basis for strategies
  - Requires management assumptions on PIT parameters
- Fit logistic regression on validation population
  - Evaluates overall performance of model
  - Ensures Actual = Expected for total population
- Starting point for Marginal Chi² analysis

Marginal Chi² reports should be part of regular monitoring

Change in Behaviour?
Example of tracking analysis



- Clear WoE pattern
- IV: 0.62 Marginal IV: 0.07
  - But some negative contributions
- The Δ-scores show that scorecard "exaggerates"
  - Worst not as bad as scores suggest
- Why? Change in treatment of Excess?
- Zero excess (2/3 of population) is under-rated

Use statistics to tell the business story

Assessing Branch Performance
Adding business value

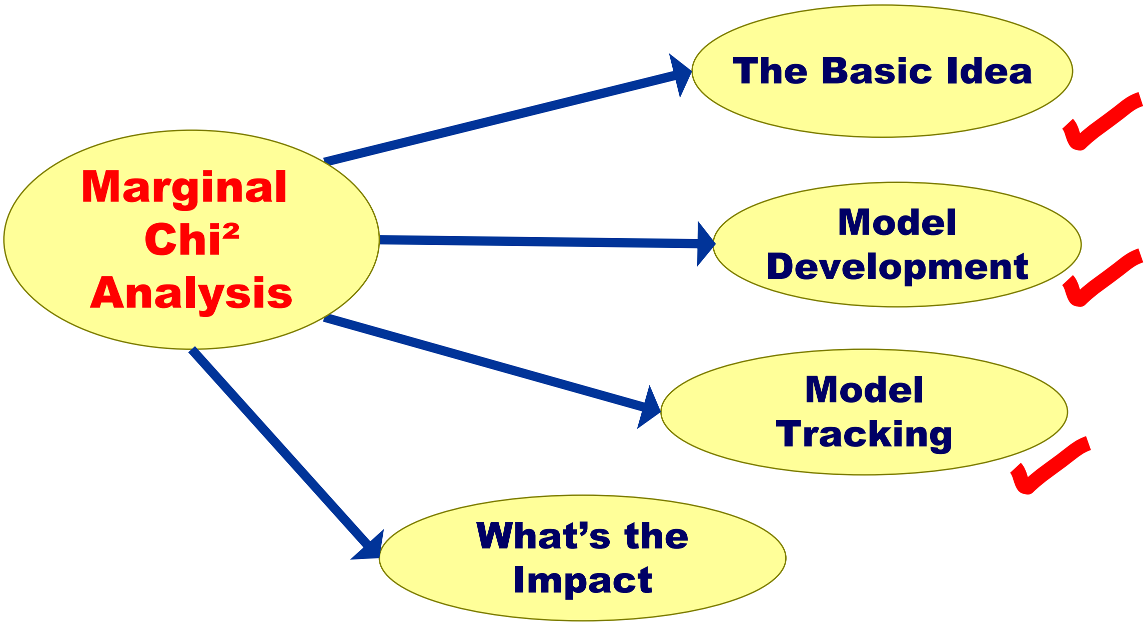
ACCOUNT OPENINGS 2008/Q2

Store	Default Rate			Performance	
	Budget	At Opening	at 9 months	Absolute	Relative
Paris	3.5%	3.7%	3.9%	Good	Poor
Lille	5.0%	5.2%	4.8%	Poor	Good
Lyon	4.0%	3.9%	3.6%	Good	Good
Marseille	4.8%	5.0%	5.3%	Poor	Poor
Total	4.1%	4.2%	4.2%		

- "At opening" figures derived from scores on account opening time
  - Profile of applicants different from budget expectations
- Isolate departures from expectations
  - Take account of differing potential
- Can be extended to policy rules, marketing campaigns, collections strategies, …

The power of sufficient statistics …

Structure of Presentation



Basel: Litmus test for rating systems

Basel Requirements

- Banks must use "all relevant and material information in assigning ratings" (Basel Accord, para. 411)
- Validation must show outcomes are in line with model expectations
- Management must show understanding of rating systems

Marginal Chi² Approach

- … allows rigourous verification that rating systems are "sufficient statistics"
- … identifies any departures from model predictions
  - … and suggests fixes
- … provides understandable interpretation of ratings:
  - Actual = Expected

Use Basel infrastructure to improve business decisions

The Credit Crunch
… and Marginal Chi²

Principles

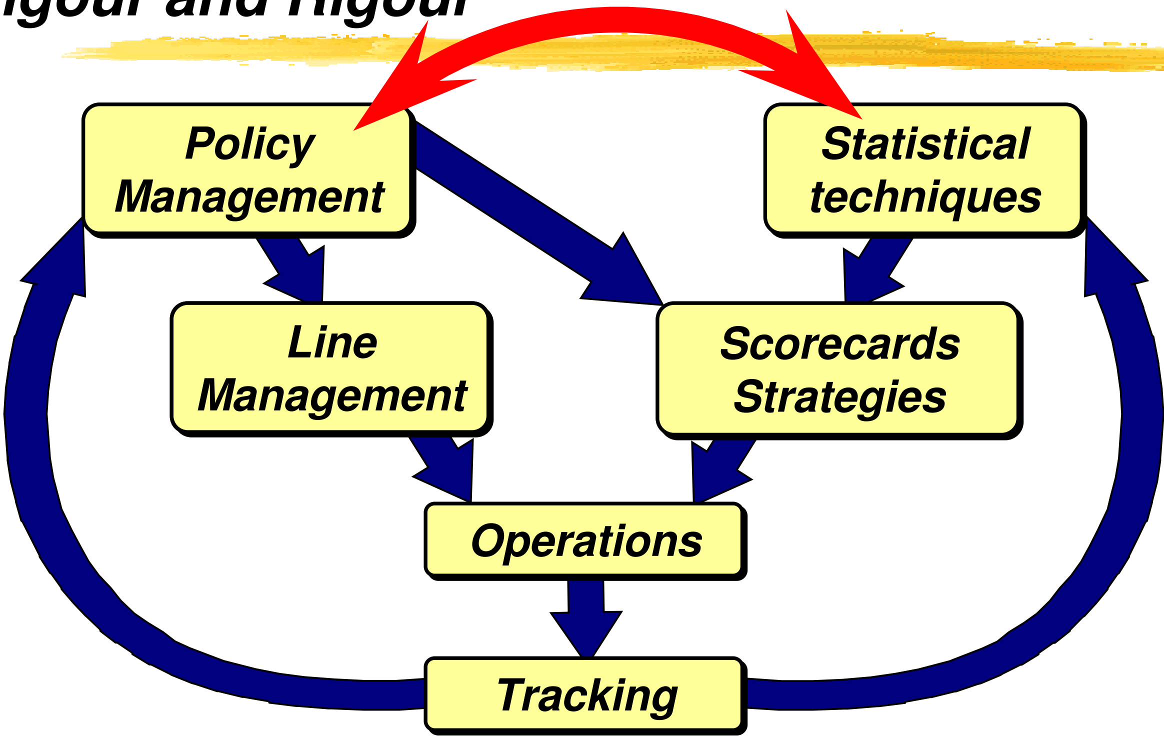
- Identify emerging variables
  - E.g. balance building
  - Potential additions to model
- Works with small samples
  - Chi² measures reliability
  - Useful results from 50-100 bads
- Works fast
  - 3-4 months after scoring
- Indicates quick (and dirty) corrective action
- Spots emergence from recession
  - Segments outperforming
  - Best time to be in business

Practice

- Cheque Account
- Emerging market
- Mild excess more likely to deteriorate
- Strong vintage effect
  - short time on books
- Amount of excess balances matters more
- "Invulnerable" accounts unaffected
- Worst accounts don't deteriorate proportionately
  - "Permanent recession"

Makes scoring models more transparent to ordinary people

The Management Link:
Vigour and Rigour



Key Management Consequences
Accountability

- Fast recognition of changes in risk
  - … and business consequences
  - … and suggests what to do about it
- Accountability for performance
  - E.g. risk performance of marketing campaigns
  - What is changing and why?
- Better business integration
  - Blurs line between model development and management
  - Aligns risk feedback loop (nearly) to marketing cycle

Makes scoring models more transparent to ordinary people

Open questions

- Continuous predictors
  - Analogue of Marginal IV
- Probabilities not homogeneous
  - Is Chi^2 still robust?
- Alternative definitions of Δ-scores
  - 1^{st} iteration of Newton-Raphson
- Variance of Δ-scores
  - Variance of expected WoE?
  - Use of re-sampling techniques
- Translate from log-odds language to PDese
- Sequential testing
  - Information from consistency of results over time?
- Extend to models other than Logistic Regression
  - Survival analysis
  - Balance and revenue models

Some trivial – others not