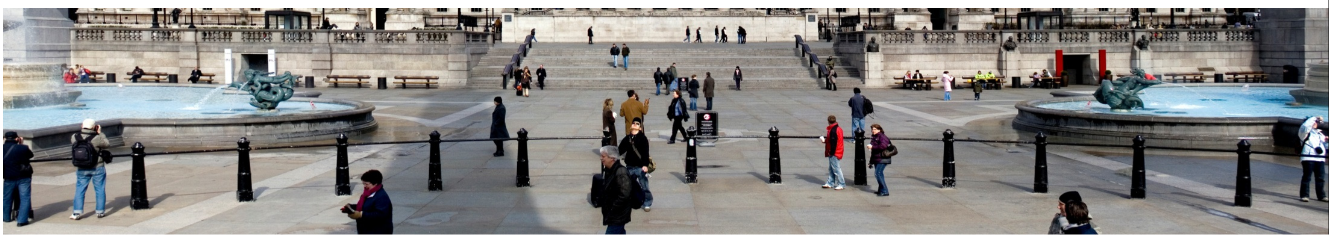
# *Class(ic) Scorecards*

Selecting Characteristics and Attributes in Logistic Regression
Edinburgh Credit Scoring Conference - 25 August 2011

***Gerard Scallan***
***gerard.scallan@scoreplus.com***

Score Plus
→ data → information → profit



# Class(ic) Scorecards
## *Using the Statistics!*

→
- What's the Problem?
- Nested Dummy Variables
- Stepwise Method
- Selecting Characteristics
- Lessons Learned





Score Plus
→ data → information → profit

# Example: Age Characteristic

## *Typical Analysis Layout*

**CHARACTERISTIC: AGE** 0.5

| Attribute | SAMPLE COUNTS Goods | Bads | Total | COLUMN % Goods | Bads | Total | WEIGHT OF EVIDENCE | INFORMATION VALUE | GLOBAL CHI² |
|---|---|---|---|---|---|---|---|---|---|
| **TOTAL** | **3608** | **1018** | **4645** | **100.0%** | **100.0%** | **100.0%** | **0.000** | **0.373** | **334.61** |
| 18 | 12 | 11 | 23 | 0.3% | 1.1% | 0.5% | -1.182 | 0.009 | 7.62 |
| 19 | 22 | 19 | 41 | 0.6% | 1.9% | 0.9% | -1.122 | 0.014 | 12.18 |
| 20 | 25 | 19 | 44 | 0.7% | 1.9% | 0.9% | -0.997 | 0.012 | 10.14 |
| 21 | 24 | 29 | 53 | 0.7% | 2.8% | 1.1% | -1.451 | 0.032 | 27.17 |
| 22 | 26 | 29 | 55 | 0.7% | 2.8% | 1.2% | -1.372 | 0.029 | 25.10 |
| 23 | 32 | 31 | 63 | 0.9% | 3.0% | 1.4% | -1.234 | 0.027 | 22.96 |
| 24 | 34 | 26 | 60 | 0.9% | 2.6% | 1.3% | -1.001 | 0.016 | 14.01 |
| 25 | 44 | 29 | 73 | 1.2% | 2.8% | 1.6% | -0.854 | 0.014 | 12.18 |
| …. | …. | …. | …. | …. | …. | …. | …. | …. | …. |
| 66+ | 18 | 1 | 19 | 0.5% | 0.1% | 0.4% | 1.247 | 0.005 | 4.30 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Information Value: | 0.373 | Chi² | 334.61 | DF | 47 | p-level | 5.04938E-45 | |

**WoE = LnOdds(attr) – LnOdds(popn)**

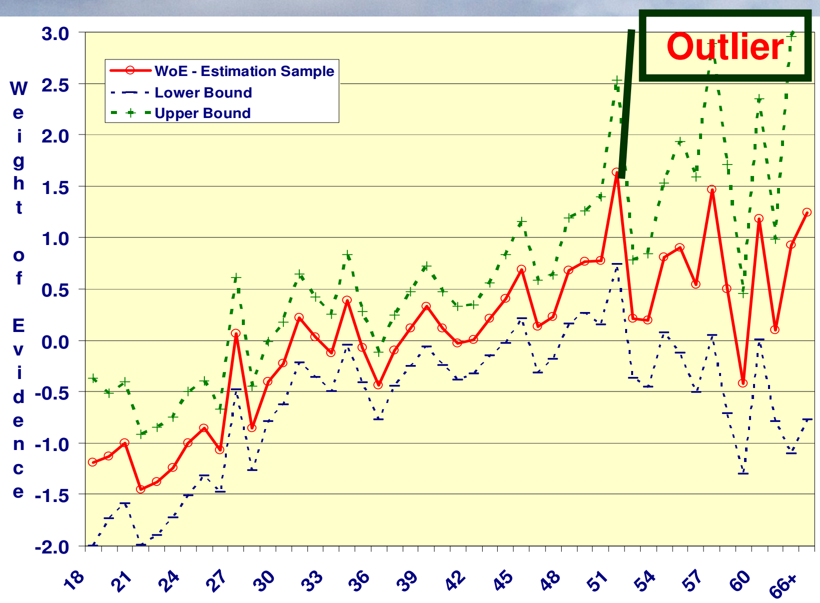
**IV = $Avg_G$(WoE) – $Avg_B$(WoE)**

**Goal of Classing → Maximise predictive power**





# WoE Graph: Show overall picture




**Var($WoE_i$) = $1/G_i + 1/B_i - 1/G_{total} - 1/B_{total}$**

Poor man's hypothesis test!

$$|WoE_i - WoE_{i+1}| \geq 2\,StDev$$

→ Reject equal risk

→ Separate classes

Equivalent to 2 x 2 Chi²

But "real" hypothesis is not equality …

**Problem: Testing Wrong Hypothesis**

# Current Practice: Classing

## Current Practice

- "Fine" breakdowns on each predictive characteristic
- Manual or Automatic Classing
  - Based on Information Value
  - or $Chi^2$ measure
- 1 dummy variable per class
- Select model variables using stepwise Logistic Regression

## And what's wrong with it

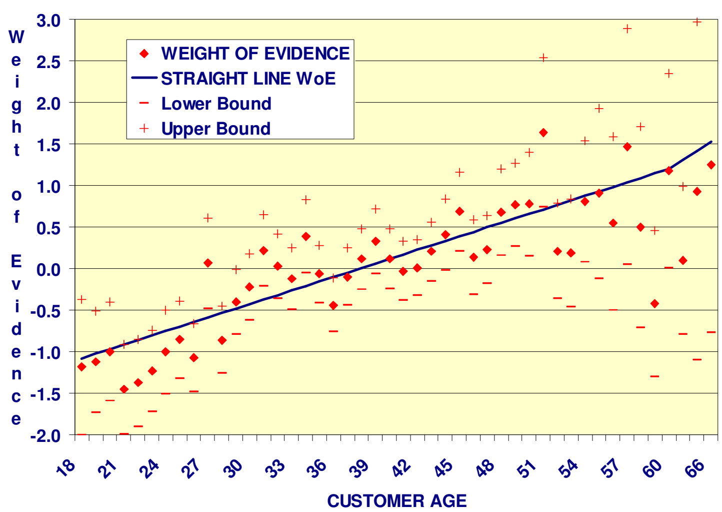
- One characteristic at a time
  - Anomalies in one characteristic often explained by another
- Lots of predictors → Lots of time
  - 700 chars x 3 mins. = 35 hours
- Variable selection in model at attribute level
  - "gap toothed" models
  - Age 18-21, Age 25-29 in model
  - Age 22-24 not in model
- Stepwise measures certainty
  - Not distance

**Good technical solutions – but wrong problem**





# Solution 1: Continuous Variables
## *Risk improves continuously with Age*

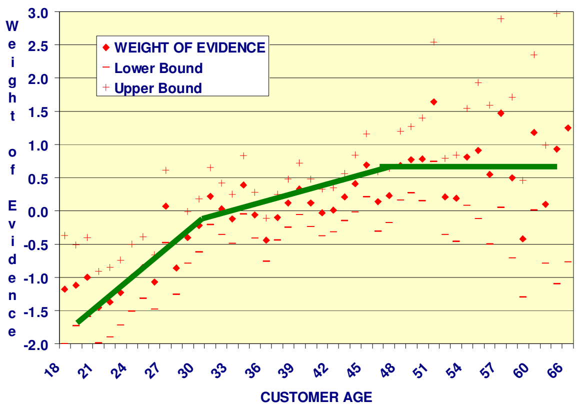



- Simpler Hypothesis
  - 1 parameter vs.15+
- Data do not contradict the linear hypothesis
  - In most cases
- But sample sliced into many small categories
  - Combine categories
  - → More reliable tests
- Slope changes ~ age 30
  - Again ~ age 50?

**Better Starting Point**

# Why Discretise?

## Non-Linearities




- Slope changes ~ age 30
  - Again ~ age 50 ?

**Not quite discrete …**

## Tradition – 1960s

- Scores calculated by hand
  - No pocket calculators
- Multiplication less reliable than addition
- Coefficients – 2 digit integers

**No longer justified**



# Class(ic) Scorecards
## *Using the Statistics!*

✓ ◆ What's the Problem?

→ ◆ Nested Dummy Variables

◆ Stepwise Method

◆ Selecting Characteristics

◆ Lessons Learned

# Partition Variables
## *a.k.a. Nested Dummy Variables*

| Variable | Age 18 | Age 19 | Age 20 | Age 21 | Age 22 | Age 23 | … |
|---|---|---|---|---|---|---|---|
| P18 | 1 | 1 | 1 | 1 | 1 | 1 | |
| P19 | 0 | 1 | 1 | 1 | 1 | 1 | |
| P20 | 0 | 0 | 1 | 1 | 1 | 1 | |
| P21 | 0 | 0 | 0 | 1 | 1 | 1 | |
| P22 | 0 | 0 | 0 | 0 | 1 | 1 | |
| P23 | 0 | 0 | 0 | 0 | 0 | 1 | |
| … | | | | | | | |
| … | | | | | | | |

- Partition variable for each fine class
- P18 = intercept – will not enter model
- Score for 22 year old =
  P18 + P19 + P20 + P21
- Coefficient P22 = incremental change for Age 22 compared to Age 21
- Partition model gives same score to each individual as Attribute model
- Partition and Attribute variables = two bases for same linear space
- Monotone increasing ↔
  Partition Coefficients > 0

**Different coding – Same model**





# Variance of Coefficients and Significance Testing

**MODEL 1 - TmBooks, DaysXS, Bounce, Autocredit**

| No. | Characteristic | Variable | Estimate | Std. Error | z-value | Pr(>\|z\|) | Significance | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|---|---|---|
| 0 | (Intercept) | | 0.54343 | 0.17772 | 3.058 | 0.00223 | *** | 0.19510 | 0.89176 |
| 1 | TmBooks | 2y6m+ | 0.82928 | 0.09972 | 8.316 | < 2e-16 | *** | 0.63383 | 1.02473 |
| 2 | TmBooks | 7y1m+ | 0.68709 | 0.12361 | 5.558 | 2.72E-08 | *** | 0.44481 | 0.92937 |
| 3 | TmBooks | 14y1m+ | 0.56779 | 0.1673 | 3.394 | 0.000689 | *** | 0.23988 | 0.89570 |
| 4 | DaysXS | Any | -0.68069 | 0.13247 | -5.138 | 2.77E-07 | *** | -0.94033 | -0.42105 |
| 5 | DaysXS | 11+ | -0.45509 | 0.18657 | -2.439 | 0.01472 | * | -0.82077 | -0.08941 |
| 6 | DaysXS | 16+ | -0.08821 | 0.17508 | -0.504 | 0.614396 | | -0.43137 | 0.25495 |
| 7 | DaysXS | 61+ | -0.45783 | 0.11057 | -4.141 | 3.47E-05 | *** | -0.67455 | -0.24111 |
| 8 | Bounce | 1m+ | 0.39119 | 0.13214 | 2.96 | 0.003072 | ** | 0.13220 | 0.65018 |
| 9 | Bounce | 41m+ | -0.06494 | 0.28124 | -0.231 | 0.81739 | | -0.61617 | 0.48629 |
| 10 | Bounce | Never | 1.02127 | 0.28125 | 3.631 | 0.000282 | *** | 0.47002 | 1.57252 |
| 11 | AutoCredit | Any | 0.41368 | 0.12795 | 3.233 | 0.001225 | ** | 0.16290 | 0.66446 |
| 12 | AutoCredit | 4000+ | 0.44995 | 0.11074 | 4.063 | 4.84E-05 | *** | 0.23290 | 0.66700 |

| | | | | | |
|---|---|---|---|---|---|
| LogLikelihood | -2206.8 | Deviance | 4413.6 | DF model | 13 |
| AIC | 4439.6 | BIC | 4530.6 | DF residual | 8081 |
| Number of Fisher Scoring Iterations | | | 3 | | |

- Maximum Likelihood Estimates
- Std. Error from Covariance Matrix of Estimates
- Z-value = Estimate/Std. Error
- OR Wald Statistic = $Z^2$

# Z-test and Wald Chi² Test: Is this variable necessary?

## Z-test = Wald Chi² Test

**Z-test**

- Z-value = Estimate/Std. Error
- If "true" value of Coefficient = 0
  - Null Hypothesis
- then sample value of Z has Normal distribution
  - Mean = 0, Variance = 1
- (From theory of Max Likelihood)
- If Null Hypothesis is true, then unlikely to get this big |z| OR
- If |z| is "large", data are not consistent with NH

**Wald Chi² Test**

- $Z^2$ = Estimate²/Variance
- Under Null Hypothesis $Z^2$ has Chi² Distribution w/ 1 DF
  - Square of N(0,1)
- Same test!
- Test at 10%, 5%, 1%, .1%
  - *** p < 0.1%
  - ** p < 1%
  - * p < 5%
  - . p < 10%

**Large sample approximation – easy to apply**





# Hypothesis Tests with Partition Variables

**Attribute Dummy Variables**

- "Reference Attribute" on every characteristic
  - Receives 0 score
  - Avoids linear indeterminacy
  - Usually last attribute
  - E.g. Age 60+
- Coefficient = 0 ↔ Risk same as Reference Attribute
- E.g. Risk on Age 22-25 = Risk on Age 60+
- Useless hypothesis

**Partition Dummy Variables**

- Coefficient = 0 ↔ Risk same as neighbour to left
- E.g. No difference in risk between Age 22-25 and Age 20-21
- What are key turning points in risk pattern?

**Ignore statistics**

**Key information**

# Automated classing
## *Provisional Solution*

### Algorithm

- Partition Vars. for “fine” classes
  - Must be ordered “sensibly”
  - Natural order or WoE
  - Possibly 20-30 variables/ characteristic
  - All characteristics in model
- Candidates in stepwise Logistic
- Stepwise algorithm identifies “significant” breakpoints
  - Partition variable enters iff “significant” difference between neighboring attributes

### Advantages

- Less work for analyst!
- Classing adapts to sample size
  - Small sample → Coarser
  - Large sample → Finer
- Accounts for interactions between characteristics
  - Fewer classes/characteristic
  - Multivariate approach
- Equivalent to systematic use of Marginal Chi²
  - But approximations are better!
- Avoids gap-toothed scorecards

**Get minimal classing needed for predictive structure**



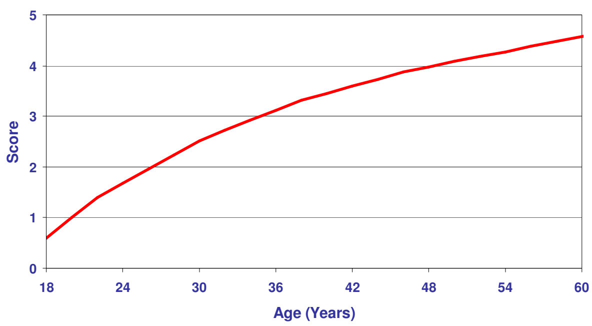


# Continuous Variables
# Piecewise Linear

### Idea

- Analogous idea for continuous predictors
- Family of spline variables
- E.g. Age
  - $(\text{Age} - 20)_+ = \max(0, \text{Age}-20)$
  - $(\text{Age} - 22)_+ = \max(0, \text{Age}-22)$
  - $(\text{Age} - 24)_+ = \max(0, \text{Age}-24)$
  - … etc.
- Candidates in stepwise Logistic
- Terms entering correspond to significant changes in slope
- a.k.a. MARS
  - Multivariate Adaptive Regression Splines

### Example




$$\begin{aligned}\text{Score} = \; & .2 \times \text{Age} \\ & -.06 \times (\text{Age} - 22)_+ \\ & -.04 \times (\text{Age} - 30)_+ \\ & -.03 \times (\text{Age} - 38)_+ \\ & -.02 \times (\text{Age} - 46)_+\end{aligned}$$

# Class(ic) Scorecards
## *Using the Statistics!*

✓ ◆ What's the Problem?

✓ ◆ Nested Dummy Variables

→ ◆ Stepwise Method

◆ Selecting Characteristics

◆ Lessons Learned



# Stepwise Approach

**3 variants**

- Forward Selection
  - Start with null model
  - Add variables
  - Until no further variable adds significant predictive power
- Backward Elimination
  - Start with all variables
  - Drop variable which makes least contribution to likelihood
  - Until no further variable can be dropped without significant loss of predictive power
- Bidirectional
  - Start with null model
  - Add variables
  - At each step, check to see if variables can be dropped
  - Then check to see if any variable can be added
  - Until no variable to be dropped AND
  - No variable to be added

**Computation: Forward < Backward < Bidirectional**

# What's wrong with Stepwise?

**"If this method had just been proposed … it would most likely be rejected because it violates every principle of statistical estimation and hypothesis testing"**
**– Harrell 2001 "Regression Modeling Strategies", p. 56**

- Parameters estimates too large
  - Selects "overestimated" coefficients
- Overestimates precision
  - Because underestimates variance
- Collinearity makes variable selection arbitrary
- Lots of candidates → Lots of noise in model

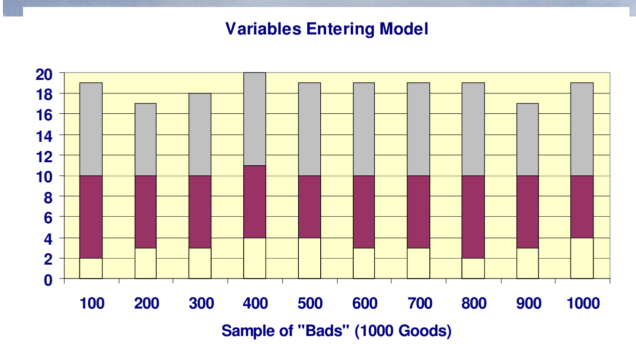
**"It allows us not to think about the problem"**

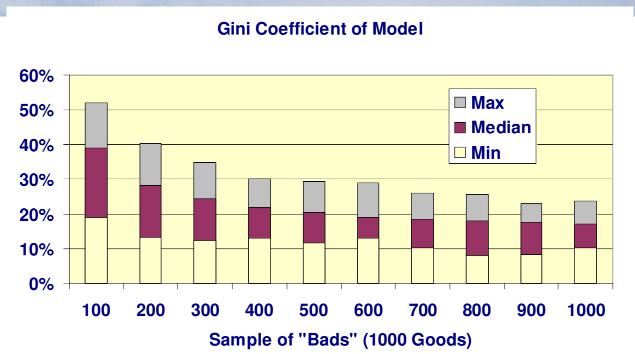




# Stepwise Logistic on Random Numbers

## *Simulated Example*




- Similar to Flom & Cassell (2007)
- 1000 Goods
- Bads from 100 to 1000
- 100 candidate variables
- All "white noise"
  - Random from Normal Distribution
  - Real predictive power = 0
- 100 replications for each sample size
- Entry/Exit criterion: $p < 0.1$

- Results on estimation sample
- Won't validate (we hope!)
- All models have Deviance statistics w/ p-level < 0.1%
- 2/3 of variables significant at 5% p-level

**Adds noise to model**

# Class(ic) Scorecards
## *Using the Statistics!*

✓ ◆ What's the Problem?

✓ ◆ Nested Dummy Variables

✓ ◆ Stepwise Method

→ ◆ Selecting Characteristics

◆ Lessons Learned






# Goal: Minimal Sufficient Model

- Bring in enough variables to explain the variation in outcome across the sample
- But no more …
- Tell a (sensible) story

**End point: predictive power of sample is exhausted**

*Edinburgh Credit Scoring Conference 2011*
*© ScorePlus SARL 2011*

Score Plus
→ data → information → profit

# Marginal Information and Delta Scores

| Debit Turnover | OBSERVED Goods | Bads | WoE | EXPECTED Goods | Bads | 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 coeffts needed to line up expected with observed
- Marginal Information Value = $\text{Avg}_{\text{Good}}$(Delta Score) - $\text{Avg}_{\text{Bad}}$(Delta Score)
  - Similar to Kullback-Liebler Information Value
  - Increased spread between average score of goods and bads
  - … if this characteristic brought into model



# Selecting Scorecard Characteristics

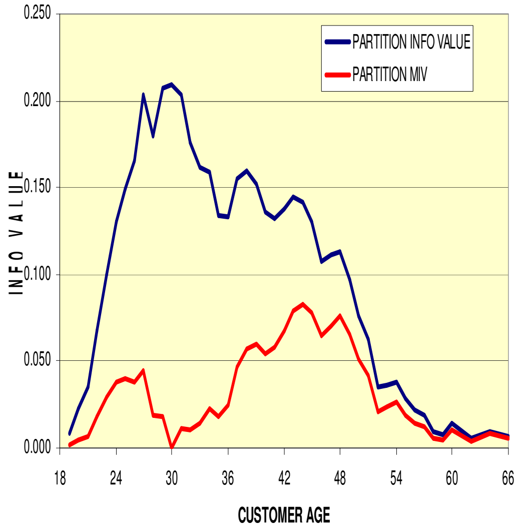
| *Characteristic* | *IV* | DaysXsL6m *Score1* | ToB *Score2* | SinceDish *Score3* | AutoCr *Score4* | CurDaysXs *Score5* |
|---|---|---|---|---|---|---|
| CurBal | 0.032 | 0.019 | 0.017 | 0.013 | 0.010 | 0.008 |
| CurCTO | 0.185 | 0.121 | 0.086 | 0.089 | 0.007 | 0.006 |
| CurDaysXs | 0.616 | 0.125 | 0.113 | 0.106 | 0.094 | 0.021 |
| CurDTO | 0.215 | 0.117 | 0.087 | 0.093 | 0.026 | 0.025 |
| CurValXs | 0.515 | 0.121 | 0.110 | 0.093 | 0.090 | 0.007 |
| ToB | 0.692 | 0.526 | 0.010 | 0.026 | 0.025 | 0.025 |
| MthsInact | 0.012 | 0.005 | 0.001 | 0.004 | -0.002 | -0.003 |
| MthsNoCTO | 0.077 | 0.066 | 0.043 | 0.045 | 0.001 | 0.000 |
| NetTO | 0.074 | 0.028 | 0.007 | 0.010 | 0.002 | 0.000 |
| DaysDbL3m | 0.055 | 0.008 | 0.013 | 0.008 | 0.005 | 0.004 |
| DaysXsL6m | 0.856 | 0.000 | 0.008 | 0.011 | 0.015 | 0.012 |
| CurMxBal | 0.033 | 0.015 | 0.018 | 0.013 | 0.005 | 0.003 |
| DishL1m | 0.291 | 0.090 | 0.084 | -0.006 | -0.008 | -0.010 |
| DishL3m | 0.292 | 0.081 | 0.077 | 0.005 | 0.011 | 0.011 |
| SinceDish | 0.810 | 0.397 | 0.299 | 0.057 | 0.050 | 0.051 |
| InterCTO | 0.017 | 0.004 | -0.003 | -0.004 | -0.001 | -0.001 |
| InterDTO | 0.003 | 0.001 | 0.000 | 0.000 | -0.002 | -0.002 |
| AutoCr | 0.209 | 0.143 | 0.108 | 0.106 | 0.005 | 0.004 |
| ValDishL6m | 0.468 | 0.145 | 0.137 | -0.001 | -0.001 | 0.003 |

- Rank characteristics by Marginal IV
- Characteristic with maximum MIV enters model …
- … i.e. partition variables become candidates for entry to model

**Continue until no SIGNIFICANT MIV left**

# Marginal IV and Collinearity

- As each variable enters MIV on remaining characteristics reduces
- Reduction measures collinearity
  - "overlap" in predictive power
  - Improperly called "correlation"
- Understand relationships between characteristics through MIV decay
- Frequently identify "families"
  - Or "Factors"
  - If one member enters model,
  - MIV drops severely on other members
- Choice of member is arbitrary



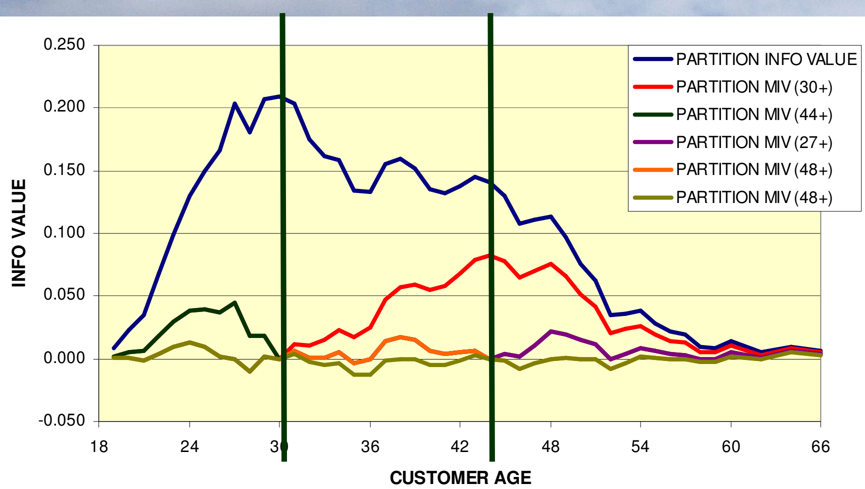

**Zero Marginal Information = Sufficient Statistic**





# Automated Classing with Marginal IV

## *Customer Age Example*

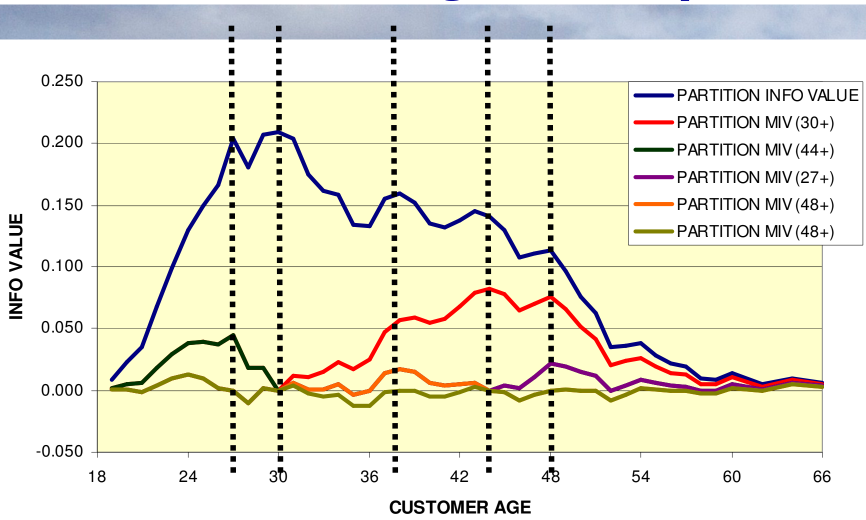



Variable 1: 30+
Variable 2: 44+

- Compute Marginal Info Value for each partition
- Select partition with max. MIV
- Check Significance → Deviance Test
- Rebuild model w/ new variable
- Re-estimate MIVs
- Continue until no significant MIV left
- All characteristics processed simultaneously

# Automated Classing with Marginal IV

## *Customer Age Example - Completion*




| Max. MIV | Variable |
|---|---|
| 0.209 | 30+ |
| 0.083 | 44+ |
| 0.045 | 27+ |
| 0.022 | 48+ |
| 0.018 | 38+ |

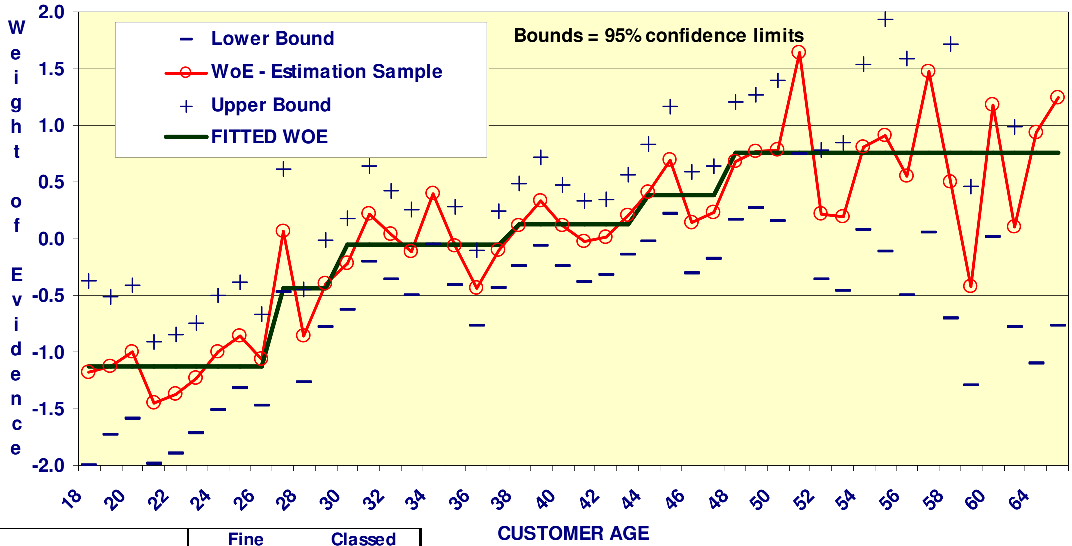
- Continue until all MIVs < .020
- 5 variables – 6 classes
- -ve MIVs → Wrong direction
- In real life, do all chars simultaneously

**End of process: "Zero" Marginal Information**





# Actual vs. Fitted WoE




| | Fine | Classed |
|---|---|---|
| Nb Attributes | 48 | 6 |
| Information Value | 0.373 | 0.303 |
| Chi² | 334.61 | 280.99 |
| p-level | 5.04938E-45 | 1.21876E-58 |

- "Few" significant differences between fitted and actual
- Differences in neighbouring groups all significant at 95%

# Triple Test
# Bottom Line

- Marginal Information Value = **Importance**
  - Distance measure
  - Rule of Thumb: -.020 < MIV > +.020
  - Negative value indicates over-fitting
  - Re-examine history of MIV to drop variable from model
- Marginal Chi² = **Reliability**
  - Measure of certainty
  - Thousands of tests - beware of false positives
  - Sensitive to classing used for analysis
  - More robust to use Stepwise approach for classing
- Business sense = **Coherence**
  - Does characteristic tell a believable story?
  - Does the model make sense

**Model complete when no further variable satisfies these 3 criteria**





# Class(ic) Scorecards
## *Using the Statistics!*

✓ ◆ What's the Problem?

✓ ◆ Nested Dummy Variables

✓ ◆ Stepwise Method

✓ ◆ Selecting Characteristics

→ ◆ Lessons Learned

# Conclusions

- Standard statistical tools can be used better
  - Corollary: We don't need lots of special-purpose analysis software
- No statistical tool can take over the burden of sense-checking models





Score Plus
→ data → information → profit

# Outstanding Issues
## *Topics for Research*

### Marginal Analysis

- Confidence intervals on
  - Delta scores (easy)
  - Marginal Information values (hard)
- Re-design characteristic analysis to focus on partition variables
- Characteristic Analysis for Continuous Characteristics
  - Splines
  - Cf. Ross Gayler

### Scorecard Estimation

- "Stepwise" type algorithm using Marginal IV
  - rather than Deviance measures
  - but also using significance checks
- Logistic Regression with constraints
  - Monotonicity ↔ Sign constraint
  - Would eliminate much over-fitting through stepwise

**MORE POWER FROM STANDARD TOOLS**
**USE THE STATISTICS!**





Score Plus
→ data → information → profit

# References


- Frank HARRELL (2001) “Regression Modeling Strategies” (Springer, 2nd edition)
- Peter L. FLOM, David L. CASSELL (2007) “Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use” (NESUG 2007 – North Eastern SAS User Group)
- Alan AGRESTI (2012) “Categorical Data Analysis” (Wiley, 3rd edition – forthcoming). See Chapter 15.
- Gerard SCALLAN (2009) “Marginal Chi² Analysis: Beyond Goodness of Fit for Logistic Regression Models” (http://www.scoreplus.com/ref/001.pdf)
- Gerard SCALLAN (2011) “Building Better Scorecards” (Scoreplus, Course Notes, 2011 edition – Sections 5, 7, 8; Sections 8, 11 in older editions)