Why You Shouldn’t Use the Gini

ARCA Retail Credit Conference
20-22 November 2013

Gerard Scallan
gerard.scallan@scoreplus.com
Why You Shouldn’t Use the Gini

→ Why is the Gini so popular?
  ◆ Problems with Gini
    ◆ Extreme values
    ◆ Truncation
  ◆ How to fake it?
  ◆ How to replace it?

"The Hottest Cold Drink"
Where does Gini come from?

*Income Inequality*

- Corrado GINI (1884 – 1965)
  - Italian sociologist & statistician
  - Gini coefficient - 1912
- Income inequality
  - What proportion of population receives what proportion of national income?
- Adapted Lorenz curve
  - Frank LORENZ - 1905
- First use in credit 1985
  - Peter KOLESAR

**Australian Gini Curve 2011-12**

Gini = 32%

Receiver Operating Curve

- % Bads REJECTED (vertical) vs. % Goods REJECTED (horizontal)
- Area Under Curve (AUC) = (1 + Gini)/2

Why Shouldn’t You Use the Gini
ARCA Retail Credit Conference 2013
© ScorePlus SARL 2013
Why is Gini so popular?

- Measures discrimination
  - Not population quality
  - Facilitates comparison across portfolios
- Simple
  - Scale 0 to 100%
  - Visual
- Works with small sample sizes
  - But should look at confidence interval

Lifebuoy for non-technical managers thrown in at deep end!

Why Shouldn’t You Use the Gini
ARCA Retail Credit Conference 2013
© ScorePlus SARL 2013
# Gini Variants

## Variants
- Can show above/below diagonal
  - Above diagonal: Reject rates
  - Below diagonal: Accept rates
  - Mirror images
- Receiver Operating Curve
- Variant “Cumulative Accuracy Curve”
  - Horizontal: % TOTAL Population rejected
  - Vertical: % Bads rejected
  - A.k.a. Power Curve

## Synonyms
- Lorenz index
- Somers’ D statistic (in SAS)
- Goodman-Kruskal Gamma Statistic
- Wilcoxon Mann Whitney Statistic
- Efficiency
  - Old – to be avoided
- Accuracy Ratio
- Power Statistic
  - Mercer Oliver Wyman
- Concordance - Discordance
- some differ in treatment of ‘tied’ scores

*Why Shouldn’t You Use the Gini*
*ARCA Retail Credit Conference 2013*
© ScorePlus SARL 2013
Why Shouldn’t You Use the Gini

ARCA Retail Credit Conference 2013
© ScorePlus SARL 2013

Cumulative Accuracy Curve

- Write-offs vs. TOTAL population
- Perfect Model: accept all Goods, no Write-offs
- Graph looks different for sample and for population
- Accuracy Ratio = Power Statistic = Gini
Why You Shouldn’t Use the Gini

Why is the Gini so popular?
What’s wrong with Gini?
  ◆ Extreme Values
  ◆ Truncation
How to fake it?
How to replace it?
Gini Driven by Tails of Distribution

Gini Coefficient:
- Full dist: 53.7%
- 2 points: 47.6%

Insensitive to center of distribution
Gini Driven by Tails of Distribution
Why is this a problem?

- Far from where scorecard is used for decision making
  - e.g. Cut-off...

- Distorted by population definitions
  - E.g. Exclude Negative Credit Bureau
  - 1% of Goods, 12% of Bads
  - Gini before Exclusion: 57.8%
  - Gini after Exclusion: 53.7%

- Reject inference has a big impact on Gini:
  - Severe reject inference on low-end rejects → big Gini

Better alternative:
Kolmogorov-Smirnov – more sensitive to centre
Sensitivity to Population Distribution

Example

- Gini and K-S depend on:
  - Score-Risk Relationship
  - Population Profile
- Change in profile → Change in Gini
  - With no change in score-PD relationship
- Example
  - Fixed Score-PD relationship
  - Ln(odds) = 7 + 2 x score

<table>
<thead>
<tr>
<th>Distribution</th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form</td>
<td>Normal</td>
<td>Normal</td>
</tr>
<tr>
<td>Mean</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>StDev</td>
<td>1</td>
<td>0.8</td>
</tr>
<tr>
<td>K-S</td>
<td>49.4%</td>
<td>36.4%</td>
</tr>
<tr>
<td>Gini</td>
<td>64.1%</td>
<td>47.8%</td>
</tr>
</tbody>
</table>

Why Shouldn’t You Use the Gini
ARCA Retail Credit Conference 2013
© ScorePlus SARL 2013
Gini: A Fairy Story

Once Upon A Time ...

◆ Newly built scorecard
◆ Implementation:
  ◆ Identical population profile
  ◆ Fixed cut-off
  ◆ No overrides
◆ Performance:
  ◆ In line with estimated PDs
  ◆ If PD = 7% then ...
  ◆ Default rate = 7%

Scorecard works perfectly!

Log Odds

Pr(Good) = 0.8
Pr(Bad) = 0.2
What Happens to the Gini?
... Truncation → No Happy Ending

Gini cannot be used to validate application scores
... but OK for Basel PDs or behavioural scores
Why You Shouldn’t Use the Gini Measures for Scoring Model Discrimination

- Why is the Gini so popular?
- What’s wrong with Gini?
  - Extreme Values
  - Truncation

- How to fake it?
- How to replace it?

"The Hottest Cold Drink"
Trick #1: Indeterminates

Original Rationale
- Judgmental collections → ambiguity
  - Decentralized
  - Manual
- Performance cannot be determined
- Provide buffer between goods and bads
- → increase discriminant power

Current Position
- Indeterminates → statistical estimation more complex
  - cannot convert odds → prob.
  - complicate setting strategy
- Extra discrimination is spurious
  - Measure on uniform definition
- Models less sensitive on borderline cases

References
- LinkedIn Credit Scoring Forum
- Tebboth and Gadi (Edinburgh Conference 2009)

Would-be indeterminates: put with goods, not bads

Do not use indeterminate group
Trick #2: Build the Wrong Model

Current Account Behavioural Risk Score

- Score for line of credit offers
  - Mainly ...
- Estimates Pr(Bad)
  - regardless of prior product holdings
- Implement score
- Offer line of credit to previous non-borrowers

Change in policy → Change in customer behavior
- Result: much higher bad debt than predicted
- Conclusion: score should estimate Pr(Bad|Borrower)

\[
Pr(\text{Bad}) = Pr(\text{Borrower}) \times Pr(\text{Bad}|\text{Borrower})
\]
- e.g. no prior credit offer: \(.05 \times .20 = .01\)
- Change policy - Offer line of credit to all low-risk customers
  - e.g. pre-approved line: \(.50 \times .20 = .10\)
Consequences for Gini

DEFINITIONS

◆ D1: Pr(Bad)
  ◆ Non Borrowers + Good Borrowers vs. Bad Borrowers
◆ D2: Pr(Bad | Borrower)
  ◆ Good Borrowers vs. Bad Borrowers
◆ D2: the outcome that matters

<table>
<thead>
<tr>
<th>Gini Coefficients</th>
<th>Outcome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model D1 D2</td>
<td></td>
</tr>
<tr>
<td>75% 45%</td>
<td></td>
</tr>
<tr>
<td>65% 60%</td>
<td></td>
</tr>
</tbody>
</table>

*Made-up numbers... but plausible*

Model customer behavior – not bank policy!
Trick #2b: Credit Card Transactors

- Scorecard to estimate Pr(Bad) for credit card holders
- Decisions to be supported:
  - Credit line increases
  - Early collections actions
  - Payment holiday offers

\[
Pr(Bad) = Pr(Revolver) \times Pr(Bad|Revolver)
\]

- Transactors have little influence on profitability
  - Small contribution to profit
  - Little likelihood of credit loss
- Large numbers of transactors will dominate ‘good’ sample
- Conclusion: build scorecard to estimate Pr(Bad|Revolver)
- Good revolvers v. Bad revolvers
  - ... in the outcome period

Higher Gini → Not adapted to business problem
Trick #3: Extreme Bad Definition

Example: Basel Mortgage PD

**PROBLEM**
- Basel Default Definition:
  - 90 days arrears in 12-month outcome period
  - UK: 180 days in 12 m!
  - Reduces capital requirement
- Who goes bad?
- Borrowers already bad on all other accounts
- Typical Gini: 85%

**SOLUTION**
- Build on business definition
- e.g. 60-days arrears in 30 m.
- Re-calibrate to Basel definition
- Typical Gini on build: 60%
- Typical Gini on calibration: 82%

Higher Gini → Not adapted to business problem
Trick #4: Include “Nearly Bads”  
Example: Credit Card Behavioural Score

**QUESTION**

- Business purpose:
  - Limit increases
  - Payment holidays
  - Cross-selling
  - 1-down collections (??)
- Bads: 90 days arrears in 12-m.
- Population: all not-bads
- Should customers currently in Arrears 2 be included in population?
  - At scoring date

**ANSWER**

- What actions on Arrears 2 accounts?
  - No limit increases
  - No payment holidays
  - No cross-selling
  - Score doesn’t use collections history
- No BUSINESS value in scoring them ...
- **BUT...**
- PD ≈ 70% → extreme scores
- → big loss in Gini
- E.g. Gini with Arr 2: 68%
- Gini without Arr 2: 62%
- But Model 2 is better on “real” population

*Why Shouldn’t You Use the Gini*
*ARCA Retail Credit Conference 2013*
© ScorePlus SARL 2013
Why You Shouldn’t Use the Gini Measures for Scoring Model Discrimination

Why is the Gini so popular?

What’s wrong with Gini?
- Extreme Values
- Truncation

How to fake it?

How to replace it?
- If you must use it ...

"The Hottest Cold Drink"
Is Scorecard Performing in Line with Expectation?

- Make decisions based on score
  - Or (increasingly) on PD
  - e.g. PD = 20% → reject

- IS THIS ASSUMPTION JUSTIFIED?
- Key criterion for use of score
  - Score-Odds line
- Standard: line used to drive strategy
  - Not necessarily model development

Score – LogOdds Line

Pr(Good) = 0.8
Pr(Bad) = 0.2
Fix #1: Points to Double Odds

- But score-odds affected by factors beyond our (modeller) control
  - Price and market positioning
  - Collections performance
  - Economy and market
- Largely change intercept
- “Not our responsibility”
- Slope is best measure of “pure” scorecard performance

**CONVENTION**
Convert Slope → Points to Double Odds
Industry Standard: PDO = 20
PDO = ln 2 / slope
Slope = ln 2 / PDO

Why Shouldn’t You Use the Gini
ARCA Retail Credit Conference 2013
© ScorePlus SARL 2013
Fix #2: Switch to PDs

Get rid of scores...

- Scores complicate management
- PDs are more intuitive
- Key measure:
  - Default Rate vs. Avg PD
- Need measure for discrimination on PD scale
- Standard: Default Rate (DR) = PD
  - 45° line
- Double PD → what increase in DR?

![Graph showing comparison between Standard and Current Default Rates and PDs with a 45° line and a 7.5% increase in DR for a double PD change.]
CONCEPT

◆ Price Elasticity:
  ◆ How does demand (Q) change with price (P)?
  ◆ For 1% increase in oil price, x% decrease in volume sold

\[
\text{Elasticity} = \frac{dQ/Q}{dP/P} = \frac{P}{Q} \frac{dQ}{dP}
\]

◆ Default Rate (DR)
  ◆ What does a 1% increase in PD mean in default rate?

\[
\text{Elasticity} = \frac{dDR/DR}{dPD/PD} = \frac{PD}{DR} \frac{dDR}{dPD}
\]

FORMULA

◆ Assume Score-Logodds is linear
  ◆ Strong empirical evidence
  ◆ Scorecards are built this way!

◆ Then:

\[
\text{Elasticity} = \frac{PD_{PD}}{PD_{DR}} \times \frac{1 - PD}{1 - DR}
\]

◆ PDO = Points to Double Odds
  ◆ Varies with PD
    ◆ But not too much ...

◆ PD = 5%, DR = 6.7%,
◆ PDO: Actual = 22, Standard = 20
◆ Elasticity = 0.89
CONCLUSION

- Gini doesn’t measure what’s important
  - from business point of view
- React to wrong stimuli
  - Changes in population profile
- Measures scorecard in wrong place
  - At extremes
- Doesn’t work at all on application scores
  - Don’t use reject inference for monitoring!
- Can be replaced by better measures
  - Points to Double Odds
  - Default Rate Elasticity

Why You Shouldn’t Use the Gini

We can do better …
Do you really want to know why Gini is so popular?

Sexual content! You have been warned...
APPENDIX: TECHNICAL

- Confidence Intervals for Gini Coefficient
- Relationship between Gini and KS
- References
How Accurate is the Gini?

**PRINCIPLE**
- Gini is a population statistic calculated from sample
- Should quantify the error in sample estimate
- \( \rightarrow \) Calculate variance of Gini coefficient
  - See next slide
- Asymptotically, sample Gini has normal distribution around population Gini
  - Central Limit Theorem
- Confidence interval:
  - Gini \( \pm \alpha \sqrt{\text{Var}(\text{Gini})} \)
  - \( \alpha \) - from Normal distribution

**EXAMPLE**
- Sample Size:
  - Non-Default: 11002
  - Default: 794
- Gini Coefficient: 66.3%
- Variance: .000194
- Std. Deviation: 1.39%
- 95% Confidence Bound
  - Lower: 63.6%
  - Upper: 69.1%
- Std Deviation proportional to \( \sqrt{\text{Default Sample}} \)
- 100 Defaults \( \rightarrow \) Std Dev = 3.8%
Variance of Gini Coefficient

FORMULA FOR VARIANCE

\[
\text{var}(Gini) = \frac{1}{(N_G - 1)(N_B - 1)} \left[ (N_B - 1) \sum_s \hat{P}_G(s)(1 - 2 \hat{F}_B(s) - \hat{P}_B(s))^2 \\
+ (N_G - 1) \sum_s \hat{P}_B(s)(1 - 2 \hat{F}_G(s) - \hat{P}_G(s))^2 \\
- \sum_s \hat{P}_G(s)\hat{P}_B(s) - (N_G + N_B - 1)Gini^2 + 1 \right]
\]

NOTATION

- \( N_G, N_B \) Goods, bads in sample
- \( F_G, F_B \) Cumulative Distributions \( P_G, P_B \) Density functions

CONFIDENCE INTERVAL

- Confidence interval: \( Gini \pm \alpha \sqrt{\text{Var}(Gini)} \)
- \( \alpha \) - from Normal distribution
Gini and K-S coefficients

**Relationship 1**

- Gini and KS have similar properties
  - Independent of population odds
  - Depend on rank ordering only
  - Range [0,1]
- Close theoretical relationship
  - KS < Gini
  - Rule of Thumb: KS ~ 80% of Gini
  - KS corresponds to largest triangle inscribed within Gini curve

<table>
<thead>
<tr>
<th>Gini Coefft:</th>
<th>66.3%</th>
</tr>
</thead>
<tbody>
<tr>
<td>KS Statistic:</td>
<td>51.3%</td>
</tr>
</tbody>
</table>

K-S statistic corresponds to area of largest triangle inscribed between Gini curve and diagonal
Gini and K-S Proof

Area of triangle
= (Triangle to KS point) + (Trapezoid from KS point to 1) – (Area < diagonal)
= \( \frac{1}{2} \times F_G(s^*) \times F_B(s^*) + (1 - F_G(s^*)) \times (1 + F_B(s))/2 - \frac{1}{2} \)
= \( \frac{1}{2} \left[ F_G(s^*) \times F_B(s^*) + 1 - F_G(s^*) + F_B(s^*) - F_G(s^*) \times F_B(s^*) \right] - \frac{1}{2} \)
= \( \frac{1}{2} \left[ F_B(s^*) - F_G(s^*) \right] \)
= \( \frac{1}{2} \) K-S statistic

(s* is point at which K-S maximum is reached)

◆ → K-S is less than Gini – around 80%
◆ K-S more sensitive to centre of distribution
    ◆ Better for risk, mailing, attrition scores
◆ Gini more sensitive to tails
    ◆ Better for fraud scores
Gini and K-S: Relationship 2

- KS ranking invariant under monotone transformation of score
  - E.g. Score → PD
- Convert scores → %iles of “goods”
- Cumulative Good distribution → Diagonal!
  - By definition
- Cumulative Bads → ROC
- KS graph = ROC

KS statistic is largest vertical distance between ROC and diagonal

KS = 80% - 29% = 51%
References

◆ Peter KOLESAR and Janet SHOWERS (1985) "A Robust Credit Screening Model Using Categorical Data" Management Science, vol.31, pp 123-133