Summarizing Insurance Scores Using a Gini Index

Edward W. (Jed) Frees

Joint work with Glenn Meyers and Dave Cummins

University of Wisconsin – Madison and Insurance Services Office

May 25, 2010
Outline

2 The Ordered Lorenz Curve

3 Insurance Scoring

4 Effects of Model Selection
   - Under- and Over-Fitting
   - Non-Ordered Scores
   - Gini Coefficients for Rate Selection

5 Statistical Inference
   - Estimating Gini Coefficients
   - Comparing Gini Coefficients
Would like to consider the degree of separation between insurance losses $y$ and premiums $P$

- For typical portfolio of policyholders, the distribution of premiums tends to be relatively narrow and skewed to the right.
- In contrast, losses have a much greater range.
- Losses are predominantly zeros (about 93% for homeowners) and, for $y > 0$, are also right-skewed.
- Difficult to use the squared error loss - mean square error - to measure discrepancies between losses and premiums.

We are proposing several new methods of determining premiums (e.g., instrumental variables, copula regression).

- How to compare?
- No single statistical model that could be used as an “umbrella” for likelihood comparisons.

Want a measure that not only looks at statistical significance but also monetary impact.
We consider methods that are variations of well-known tools in economics, the *Lorenz Curve* and the *Gini Index*.

A Lorenz Curve
- is a plot of two distributions
- In welfare economics, the vertical axis gives the proportion of income (or wealth), the horizontal gives the proportion of people
- See the example from Wikipedia
The Gini Index

- The 45 degree line is known as the “line of equality”
  - In welfare economics, this represents the situation where each person has an equal share of income (or wealth)
- To read the Lorenz Curve
  - Pick a point on the horizontal axis, say 60% of households
  - The corresponding vertical axis is about 40% of income
  - This represents income inequality
  - The farther the Lorenz curve from the line of equality, the greater is the amount of income inequality
- The Gini index is defined to be (twice) the area between the Lorenz curve and the line of equality
The Ordered Lorenz Curve

- We consider an “ordered” Lorenz curve, that varies from the usual Lorenz curve in two ways
  - Instead of counting people, think of each person as an insurance policyholder and look at the amount of insurance premium paid
  - Order losses and premiums by a third variable that we call a relativity

- Notation
  - Let $x_i$ be the set of characteristics (explanatory variables) associated with the $i$th contract
  - Let $P(x_i)$ be the associated premium
  - Let $y_i$ be the loss (often zero)
  - Let $R_i = R(x_i)$ be the corresponding relativity
The Ordered Lorenz Curve

- **Notation**
  - \( x_i \) - explanatory variables, \( P(x_i) \) - premium, \( y_i \) - loss, \( R_i = R(x_i) \), \( I(\cdot) \) - indicator function, and \( E(\cdot) \) - mathematical expectation

- **The Ordered Lorenz Curve**
  - **Vertical axis**
    \[
    F_L(s) = \frac{E[yI(R \leq s)]}{E(y)} = \frac{\sum_{i=1}^{n} y_i I(R_i \leq s)}{\sum_{i=1}^{n} y_i}
    \]
    that we interpret to be the *market share of losses*.
  - **Horizontal axis**
    \[
    F_P(s) = \frac{E[P(x)I(R \leq s)]}{E(P(x))} = \frac{\sum_{i=1}^{n} P(x_i)I(R_i \leq s)}{\sum_{i=1}^{n} P(x_i)}
    \]
    that we interpret to be the *market share of premiums*.

- **The distributions are unchanged when we**
  - rescale either (or both) losses \( y \) or premiums \( P(x_i) \) by a positive constant
  - transform relativities by any (strictly) increasing function
Example

Suppose we have only \( n = 5 \) policyholders

<table>
<thead>
<tr>
<th>Variable</th>
<th>( i )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loss ( y_i )</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>Premium ( P(x_i) )</td>
<td>4</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>8</td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>Relativity ( R(x_i) )</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Lorenz Curve](image)

![Ordered Lorenz Curve](image)
Another Example

- Here is a graph of \( n = 35,945 \) contracts, a 1 in 10 random sample of an example that will be introduced later.
- To read the Lorenz Curve:
  - Pick a point on the horizontal axis, say 60% of premiums.
  - The corresponding vertical axis is about 50% of losses.
  - This represents a profitable situation for the insurer.
- The “line of equality” represents a break-even situation.
- Summary measure: the Gini coefficient is (twice) the area between the line of equality and the Lorenz Curve.
  - It is about 6.1% for this sample, with a standard error of 3.7%.
Policies are profitable when expected claims are less than premiums.

Expected claims are unknown but we will consider one or more candidate insurance scores, \( S(x) \), that are approximations of the expectation.

- We are most interested in polices where \( S(x_i) < P(x_i) \).

One measure (that we focus on) is the relative score

\[
R(x_i) = \frac{S(x_i)}{P(x_i)},
\]

that we call a *relativity*.

- This is not the only possible measure. Might consider

\[
R(x_i) = S(x_i) - P(x_i).
\]
Ordered Lorenz Curve Characteristics

Additional notation: Define $m(x) = E(y|x)$, the regression function. Recall the distribution functions

$$F_L(s) = \frac{E[yI(R \leq s)]}{E[y]} \quad \text{and} \quad F_P(s) = \frac{E[P(x)I(R \leq s)]}{E[P(x)]}$$

1. Independent Relativities. Relativities that provide no information about the premium or the regression function
   - Assume that $\{R(x)\}$ is independent of $\{m(x), P(x)\}$.
   - Then, $F_L(s) = F_P(s) = Pr(R \leq s)$ for all $s$, resulting in the line of equality.

2. No Information in the Scores
   - Premiums have been determined by the regression function so that $P(x) = m(x)$.
   - Scoring adds no information: $F_P(s) = F_L(s)$ for all $s$, resulting in the line of equality.
A Regression Function is a Desirable Score.

- Suppose that \( S(x) = m(x) \),
- Then, the ordered Lorenz curve is convex (concave up).
- This means that it has a positive (non-negative) Gini index.
Regression Bound

- Suppose that \( S(x) = m(x) \),
- and total premiums equals total claims. Then

\[
F_L(s) \leq s F_P(s).
\]

- The curve \((F_P(s), s F_P(s))\) is labeled as a “regression bound.”

![Ordered Lorenz Curve Characteristics](image-url)
Additional Explanatory Variables Provide More Separation

- Suppose that \( S_A(x) = m(x) \) is a score based on explanatory variables \( x \).
- Consider additional explanatory \( z \) with score \( S_B(x) = m(x, z) \).
- Then, the ordered Lorenz Curve from Score \( S_B \) is “more convex” than that from Score \( S_A \).
  - For a given share of market premiums, the market share of losses for the score \( S_B \) is at least as small when compared to the share for \( S_A \).
The Gini coefficient is a measure of association between losses and premiums.

- When the insurance score is a regression function, the more explanatory information, the smaller is the association between losses and premiums.
- In this sense, the Gini coefficient can be viewed as another goodness of fit measure from a regression analysis.

To see how the Gini performs in different situations, we conduct a simulation study where the amount of fit is known.

We consider 5,000 contracts with expected claims:
The regression scores are given by:

\[ m(x) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2). \]

We compare this to an underfit score

\[ S_{Under}(x) = \exp(\beta_0 + \beta_1 x_1) \]

and an overfit score

\[ S_{Over}(x) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3). \]

Here, each \( x_j \) was generated from a chi-square distribution with 20 degrees of freedom, rescaled to have a zero mean and variance 1/10.

Consider 3 cases for premiums \( P(x) \)

- Constant premiums (constant exposure),
- Premiums “close to” the regression function, and
- Premiums “very close to” the regression function
Case 1. Substantial Opportunities for Risk Segmentation

By controlling the beta parameters, we have the following relationships among scores, summarized by Spearman correlations

<table>
<thead>
<tr>
<th></th>
<th>$S_{Under}$</th>
<th>$m(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m(x)$</td>
<td>0.444</td>
<td></td>
</tr>
<tr>
<td>$S_{Over}$</td>
<td>0.439</td>
<td>0.973</td>
</tr>
</tbody>
</table>

Interpret this to mean

- If the insurer uses the conservative score $S_{Under}$, substantial opportunities are missed.
- There is little penalty for being over-aggressive; the score $S_{Over}$ is similar to the regression function $m(x)$.  

Case 1. Substantial Opportunities for Risk Segmentation

- Each panel gives a Lorenz curve for an **under-fit score**, an **over-fit score**, a score using the regression function and a **constant score**.
Case 1. Substantial Opportunities for Risk Segmentation

Table: Gini Coefficients

<table>
<thead>
<tr>
<th>Score</th>
<th>Premiums Close to Regression Function</th>
<th>Premiums Very Close to Regression Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Under-fit Score</td>
<td>9.60</td>
<td>-5.69</td>
</tr>
<tr>
<td>Regression Function</td>
<td>20.76</td>
<td>14.62</td>
</tr>
<tr>
<td>Over-fit Score</td>
<td>20.38</td>
<td>14.04</td>
</tr>
<tr>
<td>Constant Score</td>
<td>0.06</td>
<td>-14.62</td>
</tr>
</tbody>
</table>
Case 1. Substantial Opportunities for Risk Segmentation

The regression function has the largest Gini for each of the 3 premium cases:

- Use of this as a score yields the most separation between losses and premiums
- The Over-fit score is a close second
- Both the under-fit and constant scores perform poorly

**Table: Gini Coefficients**

<table>
<thead>
<tr>
<th>Score</th>
<th>Premiums</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Close to Regression Function</td>
</tr>
<tr>
<td>Under-fit Score</td>
<td>9.60</td>
</tr>
<tr>
<td>Regression Function</td>
<td>20.76</td>
</tr>
<tr>
<td>Over-fit Score</td>
<td>20.38</td>
</tr>
<tr>
<td>Constant Score</td>
<td>0.06</td>
</tr>
</tbody>
</table>
The (Spearman) correlation coefficients are

<table>
<thead>
<tr>
<th></th>
<th>$S_{\text{Under}}$</th>
<th>m(x)</th>
</tr>
</thead>
<tbody>
<tr>
<td>m(x)</td>
<td>0.879</td>
<td>.</td>
</tr>
<tr>
<td>$S_{\text{Over}}$</td>
<td>0.534</td>
<td>0.592</td>
</tr>
</tbody>
</table>

Interpret this to mean

- In this case, if the insurer uses the conservative score $S_{\text{Under}}$, few opportunities are missed.
- By being over-aggressive, the use of the score $S_{\text{Over}}$ means using a very different measure than the regression function $m(x)$. 

Case 2. Few Opportunities for Risk Segmentation

Table: Gini Coefficients

<table>
<thead>
<tr>
<th>Score</th>
<th>Premiums</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Close to Regression Function</td>
</tr>
<tr>
<td>Underfit Score</td>
<td>9.18</td>
</tr>
<tr>
<td>Regression Function</td>
<td>10.24</td>
</tr>
<tr>
<td>Overfit Score</td>
<td>6.50</td>
</tr>
<tr>
<td>Constant Score</td>
<td>-0.15</td>
</tr>
</tbody>
</table>

- Again, the regression function has the largest Gini, the constant score the lowest, for each of the 3 premium cases.
- The under-fit score outperforms the over-fit score.
- The separation among Gini coefficients decreases as the premium becomes closer to the (optimal) regression function.
Case 3. Effects of Non-Ordered Scores

- Return to the Case 1 design where \( S_{\text{Over}} \) performs well and \( S_{\text{Under}} \) performs poorly.
- Define two new scores

\[
S_1(x) = \begin{cases} 
S_{\text{Over}}(x) & \text{if } m(x) < \tau \\
S_{\text{Under}}(x) & \text{if } m(x) \geq \tau 
\end{cases}
\]

and

\[
S_2(x) = \begin{cases} 
S_{\text{Under}}(x) & \text{if } m(x) < \tau \\
S_{\text{Over}}(x) & \text{if } m(x) \geq \tau 
\end{cases}
\]

- We use \( \tau = 2.5 \times E m(x) \).

- Idea: we consider scores that do well in one domain and not well in others.
Case 3. Effects of Non-Ordered Scores

- No score dominates the other, crossing patterns are evident.
- The left-hand panel shows $S_1$ outperforming $S_2$ for small market shares and $S_2$ outperforming $S_1$ for large market shares.
Case 3. Effects of Non-Ordered Scores

Table: Gini Coefficients

<table>
<thead>
<tr>
<th>Score</th>
<th>Premiums</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Constant</td>
</tr>
<tr>
<td>$S_1$ Score</td>
<td>16.07</td>
</tr>
<tr>
<td>Regression Function</td>
<td>20.76</td>
</tr>
<tr>
<td>$S_2$ Score</td>
<td>13.64</td>
</tr>
<tr>
<td>Constant Score</td>
<td>0.06</td>
</tr>
</tbody>
</table>

Score performance depends on the premium as well as the level of expected claims.

- $S_1$ outperforms $S_2$ when premiums are constant,
- $S_2$ outperforms $S_1$ when premiums are very close to the regression function and
- their performance is similar when premiums are close to the regression function.
We have shown how to use the Lorenz curve and associated Gini coefficient for risk segmentation.

By identifying unprofitable blocks of business, the risk manager can introduce loss controls, underwriting and risk transfer mechanisms (such as reinsurance) to improve performance.

Further, the Gini coefficient can be viewed as a goodness of fit measure.

As such, it is natural to use this measure to select an insurance score.

The Gini coefficient measures the association between losses and premiums.

This association implicitly depends on the ordering of risks through the relativities.

It also depends on the premiums.
Case 4. A Volatile Market

- Consider “a volatile market.”
  - The variable $x_2$ adds little to the regression function
  - $x_3$ provides substantial extraneous information
- The (Spearman) correlation coefficients are:

\[
\begin{array}{c|cc}
 & S_{\text{Under}} & m(x) \\
\hline
m(x) & 0.115 & . \\
S_{\text{Over}} & 0.106 & 0.781 \\
\end{array}
\]
- With the conservative score $S_{\text{Under}}$, substantial opportunities are missed.
- The over-aggressive score $S_{\text{Over}}$ is more useful but still deviates from the true regression function
- Instead of having externally available premiums $P(x)$, we let each score to serve as the premium.
### Case 4. A Volatile Market.

**Gini Coefficients for “Champion-Challenger” Competition**

<table>
<thead>
<tr>
<th>Premiums</th>
<th>Underfit Score</th>
<th>True Regression Function</th>
<th>Overfit Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Underfit Score</td>
<td>0.19</td>
<td>18.73</td>
<td>15.65</td>
</tr>
<tr>
<td>Overfit Score</td>
<td>7.79</td>
<td>13.89</td>
<td>-0.01</td>
</tr>
</tbody>
</table>

- **First row, the underfit score = premium base, our “champion.”**
  - The “challenger” scores are used to create the relativities.
  - When both the true regression function and the overfit score are used, there is substantial separation between losses and premiums.

- **Second row, the overfit score is our “champion.”**
  - When the true regression function is used for scoring there is substantial separation between losses and premiums.
  - Also substantial separation between losses and premiums when the underfit score is used to create relativities.
  - By design, there is substantial deviation between the score $S_{Over}$ and expected claims.
  - This deviation can still be detected even when using only a mildly informative score such as $S_{Under}$ to create relativities.
Let \( \{(x_1, y_1), \ldots, (x_n, y_n)\} \) be an i.i.d. sample of size \( n \).

Let \( \hat{Gini} \) be the empirical Gini coefficient based on this sample. We have the following results:

- The statistic \( \hat{Gini} \) is a (strongly) consistent estimator of the population summary parameter, \( Gini \).
- It is also asymptotically normal, with asymptotic variance denoted as \( \Sigma_{Gini} \).
- We can calculate a (strongly) consistent estimator of \( \Sigma_{Gini} \).

For these results, we assume a few mild regularity conditions. The most onerous is that the relativities \( R \) are continuous.

These three results allow us to calculate standard errors for our empirical Gini coefficients.
Simulation Study: Estimating Gini Coefficients

- Return to the Case 1 design where $S_{\text{Over}}$ performs well and $S_{\text{Under}}$ performs poorly
- For each expectation, generate 10 independent losses from a Tweedie distribution
- This results in a sample size of $n = 50,000$

Table: Gini Coefficients with Standard Errors

<table>
<thead>
<tr>
<th>Score</th>
<th>Premiums</th>
<th>Constant</th>
<th>Close to Regression Function</th>
<th>Very Close to Regression Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Underfit Score</td>
<td></td>
<td>10.69 (1.78)</td>
<td>-4.76 (2.58)</td>
<td>-4.19 (2.61)</td>
</tr>
<tr>
<td>Regression Function</td>
<td></td>
<td>19.99 (1.32)</td>
<td>13.88 (1.58)</td>
<td>5.15 (1.96)</td>
</tr>
<tr>
<td>Overfit Score</td>
<td></td>
<td>19.55 (1.34)</td>
<td>13.29 (1.61)</td>
<td>4.37 (2.02)</td>
</tr>
<tr>
<td>Constant Score</td>
<td></td>
<td>-0.78 (2.34)</td>
<td>-13.88 (3.02)</td>
<td>-5.15 (2.67)</td>
</tr>
</tbody>
</table>

Notes: Standard errors are in parens.
Comparing Estimated Gini Coefficients

- Consider two Gini coefficients with common losses and premiums.
- Let $\hat{Gini}_A$ be the empirical Gini coefficient based on relativity $R_A$ and $\hat{Gini}_B$ be the empirical Gini coefficient based on relativity $R_B$
  - From the prior section, each statistic is consistent
  - We show that they are jointly asymptotically normal, allowing us to prove that the difference is asymptotically normal
  - We can also calculate standard errors
- This theory allows us to compare estimated Gini coefficients and state whether or not they are statistically significantly different from one another
Concluding Remarks

- The ordered Lorenz curve allows us to visualize the separation between losses and premiums in an order that is most relevant to potential vulnerabilities of an insurer’s portfolio.
  - The corresponding Gini index captures this potential vulnerability.
- When regression functions are used for scoring, the Gini index can be viewed as a goodness-of-fit measure.
  - Premiums specified by a regression function yield $Gini = 0$.
  - Scores specified by a regression function yield desirable Gini coefficients.
  - More explanatory variables in a regression function yield a higher Gini.
- We have introduced measures to quantify the statistical significance of empirical Gini coefficients.
  - The theory allows us to compare different Ginis.
  - It is also useful in determining sample sizes.