Insurance Ratemaking and a Gini Index

Edward W. (Jed) Frees

Joint work with Glenn Meyers and A. David Cummings

University of Wisconsin – Madison

Outline

Interplay between Industry and Academia

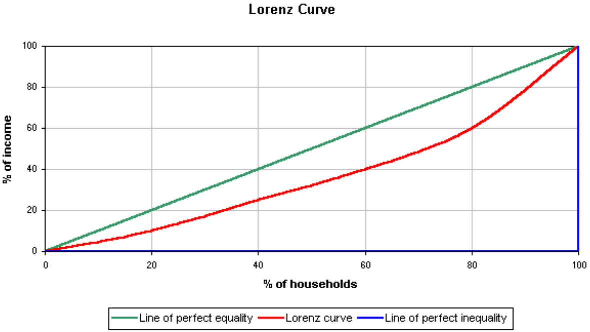
- The best actuarial example is the well-known credibility theory
- In the context of industrial workers' compensation premiums, Mowbray in **1914** introduced the idea of using a weighted average of average claims of (1) a given risk class and (2) all risk classes.
- Over 50 years, later Bühlmann in **1967** showed how to express credibility formulas in what we now call a statistical "random effects" framework.

ISO Collaboration between Industry and Academia

- We (Frees, Meyers and Cummings) wrote papers that appeared in the top statistical and actuarial journals. These were:
- Dependent Multi-Peril Ratemaking Models
  - *Astin Bulletin: Journal of the International Actuarial Association, 2010*.
- Summarizing Insurance Scores Using a Gini Index
  - *Journal of the American Statistical Association*, 2011.
- Predictive Modeling of Multi-Peril Homeowners Insurance
  - *Variance*, 2012.
- Insurance Ratemaking and a Gini Index
  - *Journal of Risk and Insurance*, 2014.

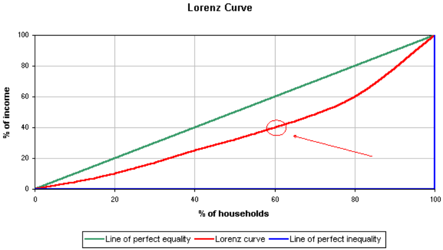
The Lorenz Curve

- 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



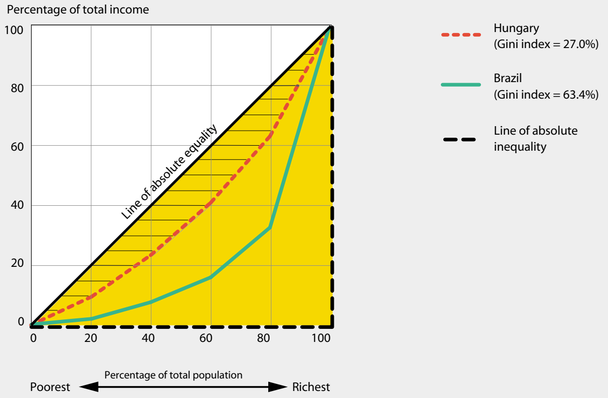
The Gini Index

- 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

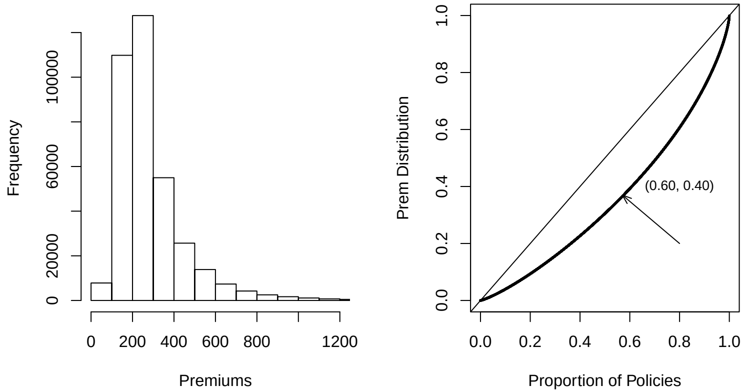


A World Bank Example

Figure 5.2 Lorenz curves and Gini indexes for Brazil and Hungary



An Insurance Example

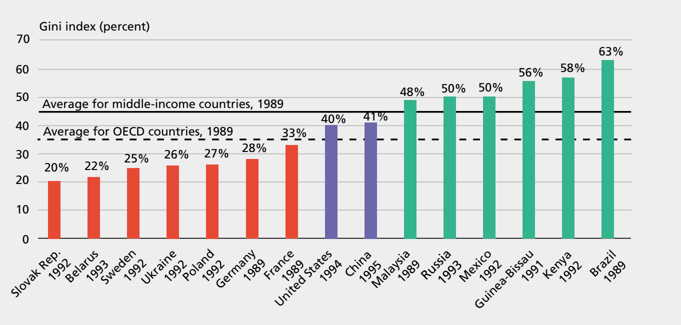


Distribution of Premiums

- The left-hand panel is a histogram of premiums from a group of 359,454 policyholders, showing a distribution that is right-skewed
- The right-hand panel provides the corresponding Lorenz curve
- The arrow marks the point where 60% of the policyholders pay 40% of premiums

More World Bank Examples

Figure 5.3 **Income inequality in selected countries, various years**



Note: An index value of 0 percent represents absolute equality in income distribution; 100 percent represents absolute inequality.

Other Applications of the Gini Index

- Gastwirth (1971, 1972) helped to emphasize the importance of the Lorenz curve and the Gini index as tools for comparing distributions, particularly in economic applications. The subsequent literature is extensive.
- Researchers have sought to understand differences in economic equality among population subgroups (e.g., Lambert and Decoster, 2005, Gastwirth, 1975).

Other Applications of the Gini Index

- Gastwirth (1971, 1972) helped to emphasize the importance of the Lorenz curve and the Gini index as tools for comparing distributions, particularly in economic applications. The subsequent literature is extensive.
- Researchers have sought to understand differences in economic equality among population subgroups (e.g., Lambert and Decoster, 2005, Gastwirth, 1975).
- Analysts have introduced weight functions into the Lorenz curve (e.g., to account for the number of publications when studying impact factors, Egghe, 2005).
- Yitzhaki (1996) describes how weighted regression sampling estimators can be of interest in welfare economic applications. Here, the idea is to adjust regression weights for social attitudes toward inequality.
- Analysts have used the Gini index for model selection in genomics (Nicodemus and Malley, 2009) and in classification trees (Sandri and Zuccolotoo, 2008).

Our Problem

- Notation
  - Let $\mathbf{x}_i$ be the set of characteristics (explanatory variables) associated with the ith contract
  - Let $P(\mathbf{x}_i)$ be the associated premium
  - Let y_i be the loss (often zero)
- y is the cost of the insurance product, P is the revenue. In a competitive market, we would like these two numbers to be close
- It is difficult for the marketplace to ensure this because
  - y is random with a distribution of outcomes
  - the distribution of y is complex, with many zeroes and when positive, right-skewed and long-tailed
  - many different sets of insureds, corresponding to a variety of $\mathbf{x}_i$'s
  - many different contract variations (deductibles, limits, coverages, riders, and so forth)

Premiums and Risk Based Scores

- One point of view is the premium should be the expected loss
- This viewpoint is supported in the context of
  - many independent contracts
  - a competitive market
- Suppose that the insurer is considering refining the classification system through the introduction of a risk based score, $S(\mathbf{x}_i)$
  - The relativity is $R(\mathbf{x}_i) = S(\mathbf{x}_i)/P(\mathbf{x}_i)$.
  - Through the relativities, we can form portfolios of policies and compare losses to premiums to assess profitability
- This is the goal of the *ordered* Lorenz curve that we introduce in this research

Example: Homeowners Insurance

Astin Bulletin: Journal of the International Actuarial Association

- We drew two random samples from a homeowners database maintained by the Insurance Services Office.
  - This database contains over 4.2 million policyholder years.
  - Policies issued by several major insurance companies in the United States, thought to be representative of most geographic areas in the US.
  - These policies were almost all for one year and so we will use a constant exposure (one) for our models.
- Our in-sample, or “training,” dataset consists of a representative sample of 404,664 records taken from this database.
  - We estimated several competing models from this dataset
- We use a held-out, or “validation” subsample of 359,454 records, whose claims we wish to predict.

$p = 9$ Perils in Homeowners Insurance

Table: Summarizing 404,664 Policy-Years

Peril (j)	Frequency (in percent)	Number of Claims	Median Claims
Fire	0.310	1,254	4,152
Lightning	0.527	2,134	899
Wind	1.226	4,960	1,315
Hail	0.491	1,985	4,484
WaterWeather	0.776	3,142	1,481
WaterNonWeather	1.332	5,391	2,167
Liability	0.187	757	1,000
Other	0.464	1,877	875
Theft-Vandalism	0.812	3,287	1,119
Total	5.889*	23,834*	1,661

Many Good Scoring Methods

- We documented many good scoring algorithms in papers that appeared in *Astin Bulletin* and in *Variance*. Here are a few:

Score	Description
	Basic, Single-peril
SP_FreqSev	Frequency and Severity model
SP_PurePrem	Pure premium Tweedie model
IND_FreqSev	Multi-peril Frequency and Severity model
	Assumes independence among perils
	Instrumental Variable Multi-peril Frequency and Severity models
IV_FreqSevA	Uses instruments for frequency component
IV_FreqSevB	Uses instruments for severity component
IV_FreqSevC	Uses instruments for frequency and severity components
	Multi-peril pure premium Tweedie models
IND_PurePrem	Assumes independence among perils
IV_PurePrem	Instrumental Variable version

Gini - Research Motivation

- We have several new methods for 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
- 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 94% 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
- Want a measure that not only looks at statistical but also monetary impact

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*
- Policies are profitable when expected claims are less than premiums
- Expected claims are unknown but we will consider one or more candidate insurance scores, $S(\mathbf{x})$, that are approximations of the expectation
  - We are most interested in polices where $S(\mathbf{x}_i) < P(\mathbf{x}_i)$
- One measure (that we focus on) is the relative score

$$R(\mathbf{x}_i) = \frac{S(\mathbf{x}_i)}{P(\mathbf{x}_i)},$$

that we call a *relativity*.

Ordered Lorenz Curve

- Notation
  - $\mathbf{x}_i$ - explanatory variables, $P(\mathbf{x}_i)$ - premium, y_i - loss, $R_i = R(\mathbf{x}_i)$, $\mathrm{I}(\cdot)$ - indicator function, and $\mathrm{E}(\cdot)$ - mathematical expectation
- The Ordered Lorenz Curve
  - Vertical axis

$$F_L(s) = \frac{\mathrm{E}[y\mathrm{I}(R \leq s)]}{\mathrm{E}\ y} \underset{empirical}{=} \frac{\sum_{i=1}^n y_i \mathrm{I}(R_i \leq s)}{\sum_{i=1}^n y_i}$$

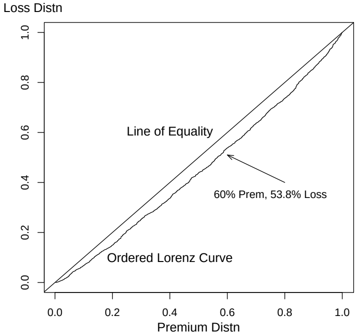
    that we interpret to be the *market share of losses*.
  - Horizontal axis

$$F_P(s) = \frac{\mathrm{E}[P(\mathbf{x})\mathrm{I}(R \leq s)]}{\mathrm{E}\ P(\mathbf{x})} \underset{empirical}{=} \frac{\sum_{i=1}^n P(\mathbf{x}_i)\mathrm{I}(R_i \leq s)}{\sum_{i=1}^n P(\mathbf{x}_i)}$$

    that we interpret to be the *market share of premiums*.
- The distributions are unchanged when we
  - rescale either (or both) losses (y) or premiums ($P(\mathbf{x}_i)$) by a positive constant
  - transform relativities by any (strictly) increasing function

Homeowners Example

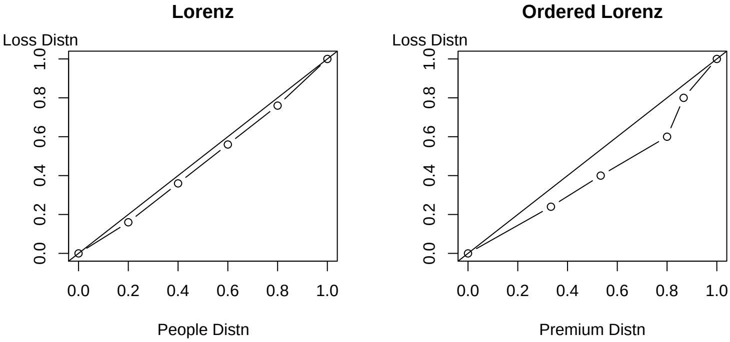
- To read the ordered Lorenz Curve
  - Pick a point on the horizontal axis, say 60% of premiums
  - The corresponding vertical axis is about 53.8% of losses
  - This represents a profitable situation for the insurer
    - Uses "SP_FreqSev_Basic" = base premium, relativity uses score "IND_FreqSev"
  - The "line of equality" represents a break-even situation
- An Ordered Lorenz Curve. For this curve, the corresponding Gini index is 10.03% with a standard error of 1.45% .



Another Example

Suppose we have only $n = 5$ policyholders

Variable	i	1	2	3	4	5	Sum
Loss	y_i	5	5	5	4	6	25
Premium	$P(\mathbf{x}_i)$	4	2	6	5	8	25
Relativity	$R(\mathbf{x}_i)$	5	4	3	2	1	



Ordered Lorenz Curve Characteristics

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

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

1. Independent Relativities. Relativities that provide no information about the premium or the regression function
   - Assume that $\{R(\mathbf{x})\}$ is independent of $\{\mathrm{m}(\mathbf{x}), P(\mathbf{x})\}$.
   - Then, $F_L(s) = F_P(s) = \Pr(R \le s)$ for all s, resulting in the line of equality.
2. No Additional Information in the Scores
   - Premiums have been determined by the regression function so that $P(\mathbf{x}) = \mathrm{m}(\mathbf{x})$.
   - Scoring adds no information: $F_P(s) = F_L(s)$ for all s, resulting in the line of equality.

Ordered Lorenz Curve Characteristics

3. A Regression Function is a Desirable Score.
   - Suppose that $S(\mathbf{x}) = \mathrm{m}(\mathbf{x})$,
   - In this case, we show that both F_P and F_L can be expressed as weighted distribution functions (cf., Furman and Zitikis, 2009)
   - Moreover, we have

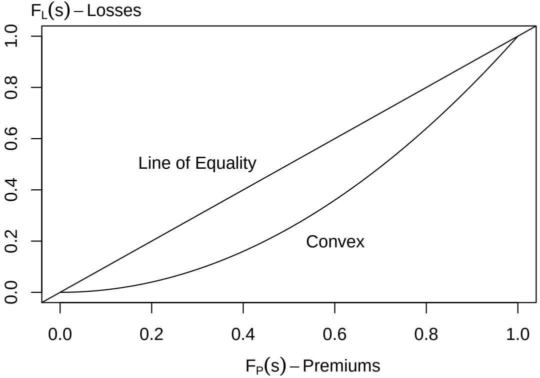
     Theorem 1. Suppose that $S(\mathbf{x}) = \mathrm{m}(\mathbf{x})$. Then, the ordered Lorenz curve may be written as a Lorenz curve. Specifically,

     $$OL(u) = \frac{\mu_P}{\mu_y} \int_0^u F_P^{-1}(z)dz = L(F_P; u).$$

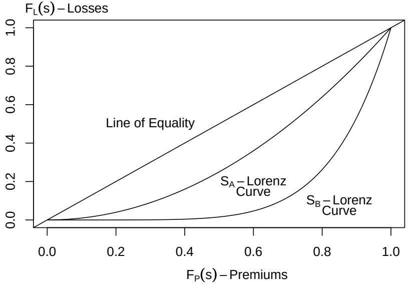
   - Then, the ordered Lorenz curve is convex (concave up).
   - This means that it has a positive (non-negative) Gini index.

Ordered Lorenz Curve Characteristics

3. A Regression Function is a Desirable Score.
   - Suppose that $S(\mathbf{x}) = \mathrm{m}(\mathbf{x})$,
   - The ordered Lorenz curve is convex (concave up).
   - This means that it has a positive (non-negative) Gini index.



4. Additional Explanatory Variables Provide More Separation
   - Suppose that $S_A(\mathbf{x}) = \mathrm{m}(\mathbf{x})$ is a score based on explanatory variables $\mathbf{x}$.
   - Consider additional explanatory $\mathbf{z}$ with score $S_B(\mathbf{x}) = \mathrm{m}(\mathbf{x}, \mathbf{z})$.
   - 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.



Interesting Special Case

Special Case: Credit Scoring

- Assume that $P(\mathbf{x}) \equiv 1$ and y is binary $(0,1)$.
  - See, for example, Gourieroux and Jasiak (2007).
  - y represents default or no default on a loan and
  - $R(\mathbf{x}) = S(\mathbf{x})$ is a credit score calculated to determine loan eligibility by a lending agency
- For this special case, we have $F_P(s) = \Pr(S \leq s)$ and

$$F_L(s) = \frac{\Pr(y=1, R \leq s)]}{\Pr(y=1)} = \Pr(S \leq s | y=1).$$

  - Gourieroux and Jasiak call the graph $(F_P(s), F_L(s))$ the "selection curve."
- Our framework permits additional potential applications in credit scoring
  - One could let y represent the *amount* of credit default (not just the event) and allow the amount charged for the loan to depend on an applicant's creditworthiness.

More Theory: Estimating Gini Coefficients

- Let $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ be an $i.i.d.$ sample of size n.
- Let $\widehat{Gini}$ be the empirical Gini coefficient based on this sample. We have the following results
  - The statistic $\widehat{Gini}$ is a (strongly) consistent estimator of the population summary parameter, $Gini$
  - It is also asymptotically normal, with asymptotic variance denoted as Σ_{Gini}
  - We can calculate a (strongly) consistent estimator of Σ_{Gini}
- For these results, we assume a few mild regularity conditions. The most onerous is that the relativities R are continuous.
- These results (based on the theory of U-statistics) allow us to calculate standard errors for our empirical Gini coefficients

Comparing Estimated Gini Coefficients

- Consider two Gini coefficients with common losses and premiums.
- Let $\widehat{Gini}_A$ be the empirical Gini coefficient based on relativity R_A and $\widehat{Gini}_B$ 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

Comparing Estimated Gini Coefficients

Table: Gini Indices and Standard Errors

Alternative Score	Gini	Standard Error	Alternative Score	Gini	Standard Error
SP_PurePrem_Basic	4.89	2.74	IV_FreqSevA	12.59	2.50
IND_PurePrem_Basic	4.01	2.77	IV_FreqSevB	10.61	2.54
IV_PurePrem_Basic	4.33	2.75	IV_FreqSevC	12.80	2.49
SP_FreqSev	11.15	2.54	DepRatio1	10.09	2.56
SP_PurePrem	9.97	2.59	DepRatio36	10.06	2.56
IND_FreqSev	10.03	2.56			
IND_PurePrem	10.96	2.57			
IV_PurePrem	11.29	2.55			

Note: Base Premium is SP_FreqSev_Basic.

Gini Indices for Ten Scores

Base Premium	Single Peril Freq Sev	Single Peril Pure Prem	IND_ Freq Sev	IND_ Pure Prem	IV_ Pure Prem	IV_FreqSev A	IV_FreqSev B	IV_FreqSev C	DepRatio 1	DepRatio 36	*Maxi-mum*
ConsPrem	28.8	28.1	28.0	28.5	28.4	29.4	28.2	29.4	28.0	28.0	29.4
SP_FreqSev	0.0	4.4	7.2	9.3	9.5	9.2	7.3	9.1	7.2	7.2	9.5
SP_PurePrem	9.1	0.0	8.6	9.7	9.5	10.3	8.8	10.5	8.6	8.6	10.5
IND_FreqSev	11.3	9.0	0.0	9.6	11.1	10.5	4.4	10.3	2.5	2.3	11.3
IND_PurePrem	8.6	6.8	4.2	0.0	3.7	7.4	4.2	7.3	4.3	4.2	8.6
IV_PurePrem	8.4	6.6	5.4	4.1	0.0	7.2	5.5	7.5	5.4	5.4	8.4
IV_FreqSevA	7.2	4.0	-2.3	4.5	5.1	0.0	-2.2	1.9	-2.2	-2.2	7.2
IV_FreqSevB	11.0	8.5	-1.6	8.9	10.3	10.1	0.0	9.9	-1.6	-1.3	11.0
IV_FreqSevC	7.4	3.9	-0.9	4.5	4.5	0.8	-1.7	0.0	-0.9	-0.9	7.4
DepRatio1	11.3	9.0	-2.3	9.5	11.0	10.4	4.4	10.2	0.0	-0.5	11.3
DepRatio36	11.2	8.9	-2.0	9.5	11.0	10.4	4.0	10.2	0.9	0.0	11.2
	All with Extended Explanatory Variables										

Gini Results from the Homeowners Example

- Standard errors are about 2.5 to 2.7 for each Gini coefficient
- When constant exposure is the base, all of the comparison scores do so well it is difficult to distinguish among them

- The relativities are based on ratios of scores
  - The two-sample test shows that relativities based on differences of scores are statistically indistinguishable - we need not consider both
- The two-sample test shows that the IVFreqSevB performs more poorly than "A" and "C" on a number of tests - not a viable candidate
- A “mini-max” strategy for selecting a score suggests that IVFreqSevA is our top performer.

Thinking About the New Gini Index

- Shlomo Yitzhaki, 1988, "More than a dozen ways of spelling Gini," *Research on Economic Inequality*, summarized several interpretations of the traditional Gini index.
  - Useful to have alternative ways to think about our new Gini index.
- Definition - The Gini as an area

$$Gini = 2\int_0^\infty \{F_P(s) - F_L(s)\}\, dF_P(s).$$

- From this, interpret the Gini index as a measure of profit

$$\frac{1}{n}\sum_{i=1}^{n}\left(\hat{F}_P(R_i) - \hat{F}_L(R_i)\right) \approx \frac{\widehat{Gini}}{2},$$

  - It is an "average profit" in the sense that we are taking a mean over all decision-making strategies, that is, each strategy retaining the policies with relativities less than or equal to R_i.
  - Insurers that adopt a rating structure with a large Gini index are more likely to enjoy a profitable portfolio.

Covariances

- After some pleasant algebra, we have

$$\widehat{Gini} = 2\widehat{\mathrm{Cov}}(y, \hat{F}_P(R)) - 2\widehat{\mathrm{Cov}}(P, \hat{F}_R) - \frac{1}{n}\widehat{\mathrm{Cov}}(y, P),$$

  - $\hat{F}_R = \mathrm{rank}(R)/n$ is the distribution function of the rank of relativities.
  - For large sample sizes n, the third term on the right-hand side is small and can be ignored.

Other things being equal:

1. We interpret a low relativity means that a policy is highly profitable and a good candidate to retain.
2. Under the relativity ordering, a large covariance between losses (y) and the proportion of premiums retained ($\hat{F}_P(R)$) implies a high Gini index.
3. A large negative covariance between premiums (P) and relativities ($\hat{F}_R$) implies a high Gini index. Stated differently, low relativities associated with high premiums implies a high Gini index.

Interesting Special Case

- Suppose that premiums (exposure) is constant over policies. Because of our rescaling, this means $P_i \equiv 1$.
- The Gini index reduces to

$$\widehat{Gini} = \frac{2}{n}\widehat{\mathrm{Cov}}(y, Rank(S)).$$

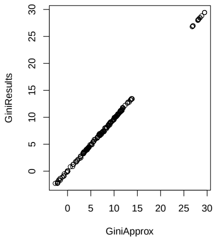
  - It is proportional to the covariance between losses and the rank of scores.
  - It is not a Pearson correlation between losses and scores, nor is it a Spearman correlation (the correlation between ranks of losses and ranks of scores).
- This statistic seems to have been first proposed by Durbin (1954) who proposed it as an instrumental variable estimator in an errors-in-variables regression problem.
  - Durbin argued that using the rank of an explanatory variable may be helpful in explaining the behavior of y when values of the explanatory variable are mis-measured.

An Approximate Gini

- Approximate the weighted premium distribution $\hat{F}_P(R)$ with the unweighted distribution of relativities $\hat{F}_R$. With this, define

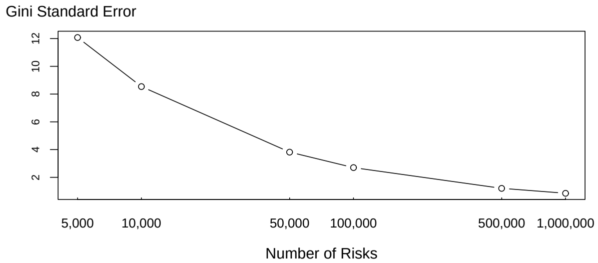
$$\widehat{Gini}_{Approx} = \frac{2}{n}\widehat{\mathrm{Cov}}\left((y-P), rank(R)\right).$$

- Think about $P-y$ as the "profit" associated with a policy.
- This approximate Gini index is proportional to the negative covariance between profits and the rank of relativities.
  - If policies with low profits$\sim$high relativities and high profits$\sim$low relativities, then the Gini index is positive and large.
- Gini Indices and an Approximation.



Additional Findings

- In a paper, we have included simulation studies that show how the Gini works under different situations
- We have also documented the effect of sample size, to give insurers a sense of how large a data set that they need to analyze in order to hope to come up with meaningful results
  - A sample size of $n = 30$ is not useful although $n = 50,000$ seems to be a good threshold number
- Effect of Sample Size on Gini Approximate Standard Errors



Summary

- The ordered Lorenz curve allows us to capture 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 view as 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

Summary

- When regression functions are used for scoring
  - These curves enjoy a partial ordering on the space of distribution functions known as a “Lorenz ordering.” (cf., Denuit and Vermandele, 1999)
  - The ordered Lorenz curves in terms of weighted distribution functions.
  - These connections may provide other researchers with motivation to enhance our understanding of characteristics of ordered Lorenz curves.
- We have provided a few alternative ways to think about our new Gini index, e.g., as an area, profit measure, .
- In particular, interpret this index as proportional to the correlation between a policy’s “profit” $(P-y)$ and the rank of the relative premium $(rank(S/P))$. Very nice intuition.

Concluding Remarks

- The Gini index is a little like a hypothesis test in that one identifies a “null hypothesis” - this is the base score in the relativity
  - There is an asymmetry in the treatment of scores
- It gives an economically meaningful way to assess out-of-sample fit
- It provides a tool for “portfolio management” - identification of good and bad risks in a portfolio (this is a little different than pricing at contract initiation or renewal)