The Journal of Risk and Insurance

© *The Journal of Risk and Insurance*, 2013, Vol. 81, No. 2, 335–366
DOI: 10.1111/j.1539-6975.2012.01507.x

# Insurance Ratemaking and a Gini Index

Edward W. (Jed) Frees
Glenn Meyers
A. David Cummings

## Abstract

Welfare economics uses Lorenz curves to display skewed income distributions and Gini indices to summarize the skewness. This article extends the Lorenz curve and Gini index by ordering insurance risks; the ordering variable is a risk-based score relative to price, known as a relativity. The new relativity-based measures can cope with adverse selection and quantify potential profit. Specifically, we show that the Gini index is proportional to a correlation between the relativity and an out-of-sample profit (price in excess of loss). A detailed example using homeowners insurance demonstrates the utility of these new measures.

## Introduction

In today's world of readily available information and interest in business analytics, insurers seek to utilize nontraditional information in their ratemaking and underwriting processes. Table 1 shows some nontraditional rating variables that are being considered and used by some insurers.

To see how insurers can use new information in the rate making process, we first consider the traditional premium. To set notation, we wish to compare the distribution of an insurance loss $y$ to a price $P$ in a portfolio of risks. For each risk in the portfolio, we assume that the analyst has available a set of known exogenous risk characteristics $\mathbf{x}$ upon which both the loss and price distributions depend. We emphasize the dependence of the price on the risk characteristics $\mathbf{x}$ through the notation $P = P(\mathrm{x})$.

Ignoring expenses, in an efficient marketplace, the price $P$ will be close to the cost $y$. This is a difficult proposition for the marketplace to ensure because:

- $y$ is random with a distribution of outcomes.
- The distribution of $y$ is complex. To illustrate this complexity, in typical homeowners insurance data described in the "Homeowners Example" section, 94 percent of

Edward W. (Jed) Frees is at the University of Wisconsin and ISO Innovative Analytics. Glenn Meyers is at ISO Innovative Analytics. A. David Cummings is at ISO Innovative Analytics. The authors can be contacted via e-mail: jfrees@bus.wisc.edu, ggmeyers@metrocast.net, and DCummings@iso.com, respectively.

**TABLE 1**
Nontraditional Rating Variables by Line of Business

| Type of Insurance | Nontraditional Characteristics |
|---|---|
| Personal automobile | Credit score, homeownership, prior bodily injury limits |
| Homeowners | Insurance credit score, prior loss information, age of home |
| Workers compensation | Safety programs, number of employees, prior loss information |
| Commercial gen liability | Insurance credit score, years in business, number of employees |
| Medical malpractice | Patient complaint history, years since residency |
| Commercial automobile | Driver tenure, average driver age, earnings stability |

*Source:* Werner and Modlin (2010).

the losses are zeros (corresponding to no claims) and when losses are positive, the distribution tends to be right skewed and thick tailed.
- There are many different sets of insureds, corresponding to a variety of $\mathbf{x}_i$s.
- There are many different contract variations, corresponding to different deductibles, limits, coverages, riders, and so forth.

One point of view is that the premium should be the expected loss. This viewpoint is supported in the context of (1) many independent contracts and (2) a competitive marketplace.

We suppose that the insurer introduces a new risk-based score, denoted by $S(\mathbf{x}_i)$. This score may be based on a new rating algorithm, for example, from a modern statistical approach such as the generalized linear model. Alternatively or additionally, it may be based on new information, for example, a new variable such as in Table 1. To assess this score, it is common (e.g., Werner and Modlin, 2010) for insurers to examine a relative premium

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

known as a *relativity*.

Suppose that the score $S$ is a desirable approximation of the expected loss. Then, if the relativity is small, we can expect a small loss relative to the premium. This is a profitable policy but is also one that is susceptible to potential raiding by a competitor (adverse selection). In contrast, if the relativity is large, then we can expect a large loss relative to the premium. This is one where better loss control measures, for example, renewal underwriting restrictions, can be helpful.

Using these relativities, this article shows how to form portfolios of policies and compare losses to premiums to assess profitability. This is the purpose of the *ordered* Lorenz curve. If the insurer can form profitable portfolios, then the competition may also be able to do so, inviting potential raiding by competing firms. To provide protection, the insurer may use the scores:

- to form the basis of a new rating algorithm (pricing), or
- as part of the underwriting guidelines, either at policy initiation or renewal.

These portfolios are natural extensions of the classic Lorenz curve and associated Gini index that we review in the following.

## Lorenz Curve and Gini Index

In welfare economics, it is common to compare distributions via the *Lorenz curve*, developed by Max Otto Lorenz (1905). A Lorenz curve is a graph of the proportion of a population on the horizontal axis and a distribution function of interest on the vertical axis. It is typically used to represent income distributions. When the income distribution is perfectly aligned with the population distribution, the Lorenz curve results in a 45 degree line that is known as the "line of equality." The area between the Lorenz curve and the line of equality is a measure of the discrepancy between the income and population distributions. Two times this area is known as the *Gini index*, introduced by Corrado Gini in 1912. See, for example, Sen and Foster (1998), for additional background on income equality. For readers interested in examining current international inequality measures, see the online resource UNU-Wider World Income Inequality Database (2008).

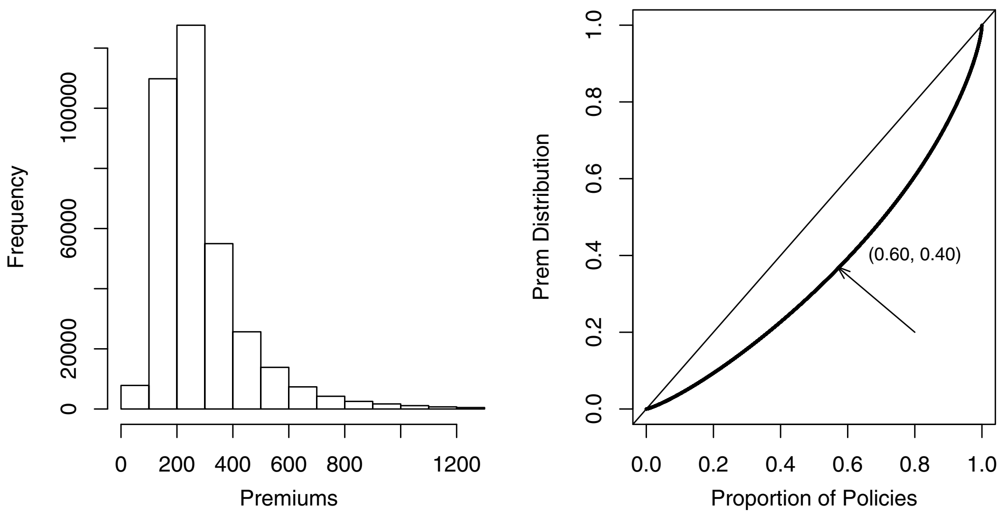
The contributions of Joseph Gastwirth in the 1970s (e.g., Gastwirth, 1971, 1972) helped to emphasize the importance of the Lorenz curve and the Gini index as tools for comparing distributions, particularly in economic applications. The subsequent literature is extensive. In one strand of the literature, researchers have sought to understand differences in economic equality among population subgroups (e.g., Gastwirth, 1975; Lambert and Decoster, 2005). In another strand, 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. In another stream of research, analysts have used the Gini index for model selection in genomics (Nicodemus and Malley, 2009) and in classification trees (Sandri and Zuccolotoo, 2008).

*Example: Distribution of Homeowners Premiums.* For an insurance example, Figure 1 shows a distribution of premiums. This figure is based on a sample of 359,454 policyholders with premiums that will be described in the "Homeowners Example" section. The left-hand panel shows a right-skewed histogram of premiums. When plotting this figure, premiums that exceeded 1,200 were ignored. The right-hand panel provides the corresponding Lorenz curve, showing again a skewed distribution. For example, the arrow marks the point where 60 percent of the policyholders pay 40 percent of premiums. The 45 degree line is the "line of equality;" if each policyholder paid the same premium, then the premium distribution would be at this line. The Gini index, twice the area between the Lorenz curve and the 45 degree line, is 29.5 percent for this data set.

## Relating Premium to Loss Distributions

From the "Lorenz Curve and Gini Index" section, the Lorenz curve is a device that is well known in welfare economics for displaying distributions. It is particularly useful for interpreting skewed distributions, a shape that insurance analysts are well acquainted with.

**FIGURE 1**
Distribution of Premiums




*Notes:* 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 percent of the policyholders pay 40 percent of premiums.

One could use classic Lorenz curves to compare a premium to a loss distribution. For example, it would be straightforward to compute Lorenz curves for premiums and for losses, and then superimpose the two curves on the same figure. However, the population distribution for each curve would be based on different sort orders (by premiums and losses, respectively), so that it would not be meaningful to compare premiums to losses for any policyholder group.

*The Role of Relativities.* As an alternative, in the following section we extend the Lorenz curve through the introduction of a third variable, the relativity. The relativity connects the losses to the premiums and is the variable that we will sort on, thus maintaining consistency between policyholder groups. In this way, we can track differences between losses and premiums and, through different sort orders, emphasize the differences between these two distributions. Because premiums (but not losses) can be influenced by insurance analysts, we will argue that this comparison provides a way to judge whether a given premium $P$ is somehow "better" than an alternative.

The idea of sorting on relativities is closely related to the notion of sorting on scores which is used, for example, in the credit scoring. As we discuss in the "Ordered Lorenz Curve and the Gini Index" section, our framework generalizes this widely used approach.

The plan for the article follows. In the "Ordered Lorenz Curve and the Gini Index"section, we develop this extension of the Lorenz curve and Gini index, providing definitions, giving an example, and focusing on a special case of interest. Theoretical properties of this Gini index were developed elsewhere (Frees, Meyers, and Cummings, 2011); to keep this article self-contained, these properties are summarized in Appendix A. The "Homeowners Example" section provides a detailed example using a sample of policies from homeowners insurance that shows how one can use the new Gini index. The "Using Covariances to Express the Gini Index" section provides additional interpretations of the Gini index, showing how it can be expressed in terms of covariance functions. Details of the covariance calculations are in Appendix B. The "Model Selection" section summarizes a small simulation study that shows how to use the Gini statistic as a tool for model selection. The "Sample Size Determination" section then describes how one can use the Gini index to suggest an appropriate sample size for pilot testing, with supporting calculations in Appendix C. The "Summary and Concluding Remarks" section closes with a summary and some additional remarks.

## ORDERED LORENZ CURVE AND THE GINI INDEX

We now introduce an *ordered* Lorenz curve, which is a graph of the distribution of losses versus premiums, where both losses and premiums are ordered by relativities. Intuitively, the relativities point toward aspects of the comparison where there is a mismatch between losses and premiums. To make the ideas concrete, we first provide some notation. We will consider $i = 1, \ldots, n$ policies. For the $i$th policy, let

- $y_i$ denote the insurance loss,
- $\mathbf{x}_i$ be the set of policyholder characteristics known to the analyst,
- $P_i = P(\mathbf{x}_i)$ be the associated premium that is a function of $\mathbf{x}_i$,
- $S_i = S(\mathbf{x}_i)$ be an insurance score under consideration for rate changes, and
- $R_i = R(\mathbf{x}_i) = S(\mathbf{x}_i)/P(\mathbf{x}_i)$ be the relativity, or relative premium.

Thus, the set of information used to calculate the ordered Lorenz curve for the $i$th policy is $(y_i, P_i, S_i, R_i)$.

### Ordered Lorenz Curve

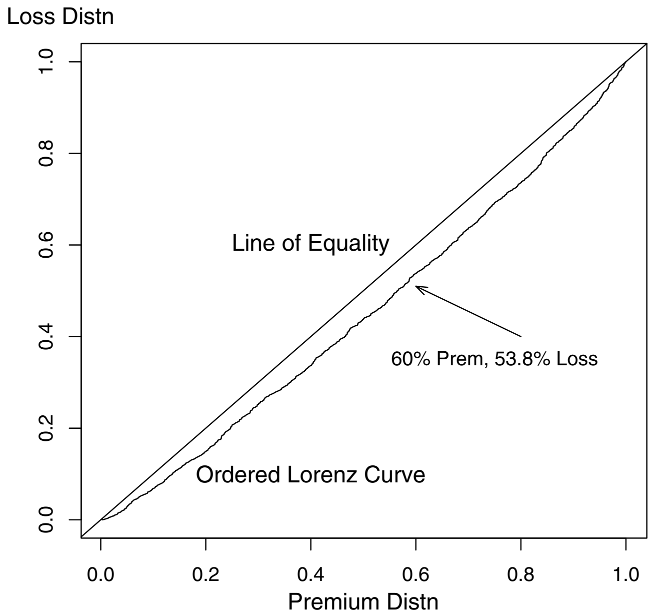
We now sort the set of policies based on relativities (from smallest to largest) and compute the premium and loss distributions. Using notation, the premium distribution is

$$\hat{F}_P(s) = \frac{\sum_{i=1}^{n} P(\mathbf{x}_i)\mathrm{I}(R_i \leq s)}{\sum_{i=1}^{n} P(\mathbf{x}_i)}, \tag{1}$$

and the loss distribution is

$$\hat{F}_L(s) = \frac{\sum_{i=1}^{n} y_i \mathrm{I}(R_i \leq s)}{\sum_{i=1}^{n} y_i}, \tag{2}$$

**FIGURE 2**
An Ordered Lorenz Curve




*Note:* For this curve, the corresponding Gini index is 10.03 percent with a standard error of 1.45 percent.

where I(·) is the indicator function, returning a 1 if the event is true and zero otherwise. The graph $\left(\hat{F}_P(s), \hat{F}_L(s)\right)$ is an *ordered Lorenz curve*.

*Example: Homeowners Loss and Premium Distributions.* As an example of an ordered Lorenz curve, Figure 2 shows a curve using our homeowners data. For this curve, the score "SP_FreqSev_Basic" was used as the base premium and the score "IND_FreqSev" was used to compute the relativities. To help interpret the curve, an arrow marks a typical point, corresponding to 60 percent of premium and 53.8 percent of losses. That is, with knowledge of relativities, an insurer could identify a portfolio that enjoys 60 percent of premiums with only 53.8 percent of losses. This is a profitable portfolio, one well worth retaining.

*The Role of Adverse Selection.* If an insurer does not adopt a refined rating plan but its competitors do, then the insurer could become victim of adverse selection. The insurer's good risks will switch to the competitor, and the insurer's remaining risks will have higher losses. The ordered Lorenz curve quantifies the extent of this potential adverse selection. Through ordering of the relativity, it summarizes the performance of portfolios that are profitable and hence subject to potential raiding by competitors. Conversely, it also summarizes the performance of poorly performing portfolios that an insurer may wish to examine for potential loss control activities, for example, tighter underwriting restrictions.

### The Gini Index

Of course, the selection of the 60th premium percentile is arbitrary. Insurers will wish to consider the profitability of different size portfolios. Thus, we summarize the curve using the Gini index, which is (twice) the area between the curve and the 45 degree line, known as "the line of equality." The line of equality can be interpreted as a "break-even" case for the insurer, where the percentage of losses equals the percentage of premiums. Curves below the line of equality represent a profitable situation for the insurer.

Specifically, the Gini index can be calculated as follows. Suppose that the empirical ordered Lorenz curve is given by $\{(a_0 = 0, b_0 = 0), (a_1, b_1), \ldots, (a_n = 1, b_n = 1)\}$ for a sample of $n$ observations. Here, we use $a_j = \hat{F}_P(R_j)$ and $b_j = \hat{F}_L(R_j)$. Then, the empirical Gini index is

$$\begin{aligned}\widehat{Gini} &= 2\sum_{j=0}^{n-1}(a_{j+1} - a_j)\left\{\frac{a_{j+1} + a_j}{2} - \frac{b_{j+1} + b_j}{2}\right\} \\ &= 1 - \sum_{j=0}^{n-1}(a_{j+1} - a_j)(b_{j+1} + b_j).\end{aligned} \tag{3}$$

As described in the "Lorenz Curve and Gini Index" section, the classic Lorenz curve shows the proportion of policyholders on the horizontal axis and the loss distribution function on the vertical axis. The "ordered" Lorenz curve extends the classical Lorenz curve in two ways: (1) through the ordering of risks and prices by relativities and (2) by allowing prices to vary by observation. We summarize the ordered Lorenz curve in the same way as the classic Lorenz curve using a Gini index, defined as twice the area between the curve and a 45 degree line. The analyst seeks ordered Lorenz curves that approach passing through the southeast corner (1,0); these have greater separation between the loss and premium distributions and therefore larger Gini indices.

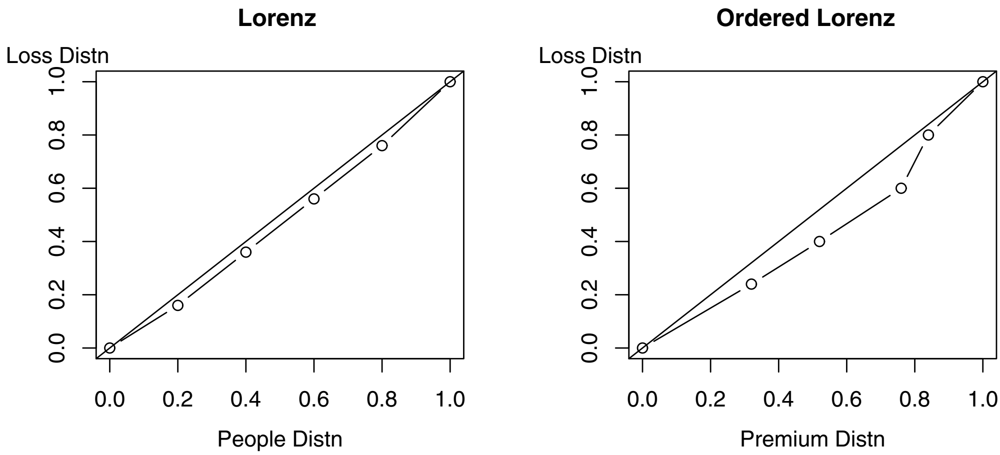
*Example.* Suppose we have only $n = 5$ policyholders with experience as:

| 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 | |

Figure 3 compares the Lorenz curve with the ordered version based on these data. The left-hand panel shows the Lorenz curve. The horizontal axis is the cumulative proportion of policyholders (0, 0.2, 0.4, and so forth) and the vertical axis is the cumulative proportion of losses (0, 4/25, 9/25, and so forth). This figure shows little separation between the distributions of losses and policyholders.

The right-hand panel shows the ordered Lorenz curve. Because observations are sorted by relativities, the first point after the origin (reading from left to right) is

**FIGURE 3**
Lorenz Versus Ordered Lorenz Curve




*Note:* The Gini index for the left-hand panel is 5.6 percent. The Gini index for the right-hand panel is 14.9 percent.

(8/25, 6/25). The second point is (13/25, 10/25), with the pattern continuing. For the ordered Lorenz curve, the horizontal axis uses premium weights, the vertical axis uses loss weights, and both axes are ordered by relativities. From the figure, we see that there is greater separation between losses and premiums when viewed through this relativity.

*Rescaling of Premiums and Losses.* From Equations (1) and (2), we see that we can arbitrarily rescale premiums and losses by any positive constant and the distribution functions remain unchanged. Thus, without loss of generality, we assume henceforth that the average loss $\bar{y}$ and average premium $\bar{P}$ are both equal to 1.

*Special Case: Credit Scoring.* Assume that $P(\mathbf{x}) \equiv 1$ and $y$ is binary $(0, 1)$. This is the credit scoring case discussed in, for example, Gourieroux and Jasiak (2007). Here, $y$ 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 (2007) call the graph $(F_P(s), F_L(s))$ the "selection curve."

As noted in Frees, Meyers, and Cummings (2011), additional potential applications in credit scoring are easy to imagine. For example, 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 applicants creditworthiness. The results presented in this article apply directly to this situation.

*Properties of the Gini Index.* Appendix A summarizes properties the Gini index, including its consistency and asymptotic normality, that are proved in Frees, Meyers, and Cummings (2011). We use this asymptotic normality extensively in this article; it is the basis for assessing the statistical significance of the Gini index.

Moreover, Frees, Meyers, and Cummings (2011) derive a result that shows that the Gini index becomes larger as one uses a "more refined" insurance score. Specifically, consider a rating plan with premiums $P$ and an insurance score $S$ that is determined by a regression function using a set of insured characteristics that produces a Gini index. Consider an alternative insurance score $S_A$ that is determined by a regression function using the base set of insured characteristics plus additional information such as more precise geographic information or credit scores. Then, the Gini index produced using this refined set of information is at least as large as the Gini index using the base information.

In this sense, the Gini index provides a summary statistic that indicates whether one insurance score outranks a competitor. Frees, Meyers, and Cummings (2011) also derive estimators to judge whether Gini indices produced by alternative scoring methods are statistically different from one another.

*The Gini Index as a Measure of Profit.* One reason that the Gini index is important is because it has a direct economic interpretation. Specifically, consider an average profit,

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{F}_P(R_i) - \hat{F}_L(R_i)) \approx \frac{\widehat{Gini}}{2}, \tag{4}$$

that can be shown to be approximately equal to the Gini index divided by two. It is an "average" 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 $R_i$. In this sense, insurers that adopt a rating structure with a large Gini index are more likely to enjoy a profitable portfolio.

Thus, we can think about the Gini index as the average expected profit to be gained by using relativities (to form portfolios). That is, the Gini index calibrates potential mismatches between losses and premiums. We change the Gini index by using different relativities; the relativities give an idea as to where potential mismatches occur. In particular, a low relativity means that a policy is highly profitable and a good candidate to retain.

## HOMEOWNERS EXAMPLE

For the "gold standard" of model validation in predictive modeling, one examines the performance of a model on an independent held-out sample (e.g., Hastie, Tibshirani, and Friedman, 2001). For this example, we used an in-sample data set of 404,664 records to compute parameter estimates. We then use the estimated parameters from the in-sample model fit as well as predictor variables from a held-out, or validation, subsample of 359,454 records, whose losses we wish to predict.

### Comparison of Scores

More details on this database and scoring methods are available in Frees, Meyers, and Cummings (2010, Forthcoming). Based on the theory developed in these two papers, we have under consideration 14 scores that are listed in second panel in Table 2. This table summarizes the distribution of each score on the held-out data. Not surprisingly, each distribution is right skewed.

For example, Table 2 also shows that the single-peril frequency severity model using the extended set of variables (SP_FreqSev) provides the lowest score, both for the mean and at each percentile (below the 75th percentile). Except for this, no model seems to give a score that is consistently high or low for all percentiles. All scores have a lower average than the average held-out actual losses (TotClaims).

### Comparing Scoring Methods With a Selected Base Premium

To compare these scoring methods, we first assume that the insurer has adopted a base premium for rating purposes; to illustrate, we use the "SP_FreqSev_Basic" for this premium. This method uses only a basic set of rating variables to determine insurance scores from a single-peril, frequency and severity model. Assume that the insurer wishes to investigate alternative scoring methods to understand the potential vulnerabilities of this premium base; Table 3 summarizes several comparisons using the Gini index. This table includes the comparison with the alternative score IND_FreqSev, shown in Figure 2, as well as 12 other scores.

The standard errors are from Appendix A (derived in Frees, Meyers, and Cummings, 2011). Thus, to interpret Table 3, one may use the usual rules of thumb and reference to the standard normal distribution to assess statistical significance. For the three scores that use the basic set of variables, SP_PurePrem_Basic, IND_PurePrem_Basic, and IV_PurePrem_Basic, the Gini indices are between 2.5 and 4 standard errors above zero, indicating statistical significance. In contrast, the other Gini indices all are more than 7 standard errors above zero, indicating that the ordering used by each score helps detect important differences between losses and premiums.

Frees, Meyers, and Cummings (2011) also derive distribution theory to assess statistical differences between Gini indices. Although we do not review that theory here, we did perform these calculations for our data. It turns out that there is no statistically significant differences among the 10 Gini indices that are based on the extended set of explanatory variables.

In summary, Table 3 suggests that there are important advantages to using extended sets of variables compared to the basic variables, regardless of the scoring techniques used.

### Comparison of Scores Using the Gini Index

As demonstrated in the preceding section, if a base premium is available, then the Gini index can be used to decide whether an alternative insurance score is useful for detecting differences between loss and premium distributions. In instances where no base premium is available, the Gini index is also useful although care is required when interpreting this measure.

**TABLE 2**
Summary Statistics of 14 Scores and Total Claims

| | | | Percentiles | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Score | Mean | Minimum | 1st | 5th | 25th | 50th | 75th | 95th | 99th | Maximum |
| SP_FreqSev_Basic | 291.10 | 20.48 | 85.00 | 120.25 | 182.74 | 240.37 | 334.62 | 618.37 | 1,025.88 | 8,856.79 |
| SP_PurePrem_Basic | 289.91 | 33.01 | 89.48 | 127.80 | 189.87 | 246.44 | 329.79 | 586.33 | 1,050.15 | 5,467.41 |
| IND_PurePrem_Basic | 290.91 | 37.49 | 92.08 | 124.04 | 182.68 | 240.30 | 328.87 | 612.47 | 1,087.06 | 13,577.91 |
| IV_PurePrem_Basic | 293.55 | 36.61 | 93.91 | 128.21 | 187.57 | 241.29 | 327.75 | 616.05 | 1,122.84 | 15,472.82 |
| SP_FreqSev | 287.79 | 8.78 | 71.55 | 105.39 | 171.55 | 237.95 | 339.40 | 631.98 | 1,039.19 | 6,864.46 |
| SP_PurePrem | 290.00 | 10.23 | 72.17 | 107.90 | 175.83 | 242.17 | 338.64 | 616.64 | 1,113.73 | 7,993.52 |
| IND_FreqSev | 294.93 | 33.05 | 97.14 | 126.61 | 185.07 | 244.99 | 333.68 | 606.03 | 1,106.17 | 22,402.49 |
| IND_PurePrem | 292.18 | 28.04 | 86.53 | 119.74 | 181.22 | 240.52 | 326.60 | 592.07 | 1,078.25 | 49,912.59 |
| IV_PurePrem | 294.06 | 12.42 | 78.41 | 113.14 | 178.62 | 240.38 | 330.21 | 614.22 | 1,095.70 | 107,158.09 |
| IV_FreqSevA | 290.91 | 23.99 | 88.70 | 121.70 | 182.29 | 241.42 | 327.81 | 606.23 | 1,096.86 | 18,102.93 |
| IV_FreqSevB | 295.32 | 28.52 | 94.58 | 124.77 | 184.29 | 245.26 | 335.38 | 606.63 | 1,100.61 | 24,394.06 |
| IV_FreqSevC | 291.17 | 20.88 | 84.78 | 118.21 | 180.63 | 241.57 | 329.92 | 608.28 | 1,098.40 | 20,046.03 |
| DepRatio1 | 301.12 | 33.38 | 98.80 | 128.95 | 188.73 | 249.97 | 340.64 | 619.79 | 1,129.96 | 23,255.94 |
| DepRatio36 | 302.39 | 33.48 | 99.27 | 129.65 | 189.87 | 251.41 | 342.30 | 620.38 | 1,132.36 | 23,092.35 |
| TotClaims | 332.89 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 660.00 | 5,916.33 | 350,000.00 |

(*Continued*)

**TABLE 2**
Continued

| Score | Interpretation |
|---|---|
| Scores using the basic set of explanatory variables | |
| SP_FreqSev_Basic | Single-peril, frequency and severity model |
| SP_PurePrem_Basic | Single-peril, pure premium model |
| IND_PurePrem_Basic | multiperil independence, pure premium model |
| IV_PurePrem_Basic | Instrumental variable multiperil pure premium model |
| Scores using the extended set of explanatory variables | |
| SP_FreqSev | Single-peril, frequency and severity model |
| SP_PurePrem | Single-peril, pure premium model |
| IND_FreqSev | multiperil frequency and severity model assuming independence among perils |
| IND_PurePrem | multiperil pure premium model assuming independence among perils |
| IV_PurePrem | Instrumental variable multiperil pure premium model. |
| Instrumental variable multiperil frequency and severity models, using the extended set of explanatory variables | |
| IV_FreqSevA | Uses instruments in frequency model |
| IV_FreqSevB | Uses instruments in severity model |
| IV_FreqSevC | Uses instruments in frequency and severity models |
| Dependence ratio multiperil frequency and severity models, using the extended set of explanatory variables | |
| DepRatio1 | Uses a single parameter for frequency dependencies |
| DepRatio36 | Uses 36 parameters for frequency dependencies |

**TABLE 3**
Gini Indices and Standard Errors

| Alternative Score | Gini | Standard Error | Alternative Score | Gini | Standard Error |
|---|---|---|---|---|---|
| SP_PurePrem_Basic | 4.89 | 1.43 | IV_FreqSevA | 12.59 | 1.39 |
| IND_PurePrem_Basic | 4.01 | 1.46 | IV_FreqSevB | 10.61 | 1.44 |
| IV_PurePrem_Basic | 4.33 | 1.46 | IV_FreqSevC | 12.80 | 1.38 |
| SP_FreqSev | 11.15 | 1.42 | DepRatio1 | 10.09 | 1.45 |
| SP_PurePrem | 9.97 | 1.42 | DepRatio36 | 10.06 | 1.46 |
| IND_FreqSev | 10.03 | 1.45 | | | |
| IND_PurePrem | 10.96 | 1.45 | | | |
| IV_PurePrem | 11.29 | 1.45 | | | |

*Note:* Base Premium is SP_FreqSev_Basic.

Table 4 summarizes the calculation of several Gini indices. Here, we allow the "base premium" $P(\cdot)$ to be each of the 14 competing scores plus the benchmark "ConsPrem," a premium that is constant over policyholders. For each base premium, Table 4 shows the Gini index for each of the thirteen competing scores. Standard errors are reported in Table 5.

From the first row of Table 4, we see that all of the Gini indices are large, indicating that any of the 14 scores considered here provide useful separation between losses and a constant premium. The next block (of four rows) consists of scores that use the basic set of explanatory variables, including SP_FreqSev_Basic. These four scores seem to perform similarly. For example, when compared with one another, the Gini indices are in the single digits. When any of the four are adopted as the base premium, double-digit Gini indices are possible using the extended set of explanatory variables.

Using the Gini measure, the dependence ratio scores advanced by Frees, Meyers, and Cummings (2010) seem to fare poorly. Almost every alternative score, except for the independence multiperil frequency and severity models upon which they are based, allows for an ordering where there is a substantial separation between premium and loss distributions.

One approach for selecting a score based on the Gini index is a "mini-max" strategy. That is, select the score that provides the smallest of the maximal Gini indices, taken over competing scores. The strategy is intuitively appealing. If one were to specify a base premium, then the maximum Gini index corresponds to the largest separation between the loss and premium distribution when considering different orderings. For example, when the base premium is SP_PurePrem_Basic, the maximum Gini index is 12.8, which is achieved when IV_FreqSevC is used to compute relativities to order distributions. For this criterion, Table 4 shows that IV_FreqSevA is the best score. It has the smallest maximum Gini index at 7.2. We interpret this to mean that this score is the least vulnerable to alternative scores.

Table 5 shows that the standard errors of the Gini indices are relatively stable across different choices of scores and premiums. We will use this observation in the

**TABLE 4**
Gini Indices for 14 Scores

| | Basic Explanatory Variables | | | | Extended Explanatory Variables | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Single Peril | | IND_ | IV_ | Single Peril | | IND_ | | IV_ | IV_FreqSev | | | DepRatio | | |
| Base Premium | Freq Sev | Pure Prem | Pure Prem | Pure Prem | Freq Sev | Pure Prem | Freq Sev | Pure Prem | Pure Prem | A | B | C | 1 | 36 | Maxi-mum |
| ConsPrem | 27.0 | 27.0 | 26.8 | 26.9 | 28.8 | 28.1 | 28.0 | 28.5 | 28.4 | 29.4 | 28.2 | 29.4 | 28.0 | 28.0 | 29.4 |
| SP_FreqSev_Basic | 0.0 | 4.9 | 4.0 | 4.3 | 11.1 | 10.0 | 10.0 | 11.0 | 11.3 | 12.6 | 10.6 | 12.8 | 10.1 | 10.1 | 12.8 |
| SP_PurePrem_Basic | 4.3 | 0.0 | 2.7 | 3.3 | 11.2 | 8.0 | 8.0 | 9.9 | 9.9 | 11.4 | 8.8 | 11.7 | 8.1 | 8.1 | 11.7 |
| IND_PurePrem_Basic | 8.1 | 7.1 | 0.0 | 3.4 | 13.4 | 11.1 | 10.0 | 11.8 | 12.2 | 13.4 | 10.6 | 13.4 | 10.0 | 10.0 | 13.4 |
| IV_PurePrem_Basic | 7.9 | 6.6 | 3.6 | 0.0 | 12.8 | 10.7 | 10.2 | 11.5 | 11.7 | 13.2 | 10.6 | 13.2 | 10.2 | 10.2 | 13.2 |
| SP_FreqSev | 1.7 | 4.0 | 4.2 | 4.9 | 0.0 | 4.4 | 7.2 | 9.3 | 9.5 | 9.2 | 7.3 | 9.1 | 7.2 | 7.2 | 9.5 |
| SP_PurePrem | 7.0 | 5.9 | 6.8 | 7.0 | 9.1 | 0.0 | 8.6 | 9.7 | 9.5 | 10.3 | 8.8 | 10.5 | 8.6 | 8.6 | 10.5 |
| IND_FreqSev | 6.8 | 6.3 | 3.2 | 4.1 | 11.3 | 9.0 | 0.0 | 9.6 | 11.1 | 10.5 | 4.4 | 10.3 | 2.5 | 2.3 | 11.3 |
| IND_PurePrem | 6.1 | 5.0 | 1.2 | 2.2 | 8.6 | 6.8 | 4.2 | 0.0 | 3.7 | 7.4 | 4.2 | 7.3 | 4.3 | 4.2 | 8.6 |
| IV_PurePrem | 6.7 | 6.4 | 3.6 | 3.5 | 8.4 | 6.6 | 5.4 | 4.1 | 0.0 | 7.2 | 5.5 | 7.5 | 5.4 | 5.4 | 8.4 |
| IV_FreqSevA | 2.8 | 1.3 | −1.1 | −0.9 | 7.2 | 4.0 | −2.3 | 4.5 | 5.1 | 0.0 | −2.2 | 1.9 | −2.2 | −2.2 | 7.2 |
| IV_FreqSevB | 6.8 | 5.9 | 3.3 | 3.9 | 11.0 | 8.5 | −1.6 | 8.9 | 10.3 | 10.1 | 0.0 | 9.9 | −1.6 | −1.3 | 11.0 |
| IV_FreqSevC | 3.4 | 1.8 | 0.0 | −0.2 | 7.4 | 3.9 | −0.9 | 4.5 | 4.5 | 0.8 | −1.7 | 0.0 | −0.9 | −0.9 | 7.4 |
| DepRatio1 | 6.8 | 6.3 | 3.2 | 4.1 | 11.3 | 9.0 | −2.3 | 9.5 | 11.0 | 10.4 | 4.4 | 10.2 | 0.0 | −0.5 | 11.3 |
| DepRatio36 | 6.8 | 6.2 | 3.2 | 4.0 | 11.2 | 8.9 | −2.0 | 9.5 | 11.0 | 10.4 | 4.0 | 10.2 | 0.9 | 0.0 | 11.2 |

**TABLE 5**
Gini Standard Errors for 14 Scores

| | Basic Explanatory Variables | | | | Extended Explanatory Variables | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Single Peril | | IND_ | IV_ | Single Peril | | IND_ | | IV_ | IV_FreqSev | | | DepRatio | |
| Base | Freq | Pure | Pure | Pure | Freq | Pure | Freq | Pure | Pure | | | | | |
| Premium | Sev | Prem | Prem | Prem | Sev | Prem | Sev | Prem | Prem | A | B | C | 1 | 36 |
| ConsPrem | 1.43 | 1.43 | 1.42 | 1.41 | 1.45 | 1.43 | 1.34 | 1.34 | 1.34 | 1.39 | 1.34 | 1.38 | 1.45 | 1.46 |
| SP_FreqSev_Basic | 0.00 | 1.43 | 1.46 | 1.46 | 1.42 | 1.42 | 1.45 | 1.45 | 1.45 | 1.39 | 1.44 | 1.38 | 1.45 | 1.44 |
| SP_PurePrem_Basic | 1.44 | 0.00 | 1.41 | 1.45 | 1.45 | 1.42 | 1.46 | 1.46 | 1.46 | 1.42 | 1.45 | 1.40 | 1.47 | 1.48 |
| IND_PurePrem_Basic | 1.44 | 1.39 | 0.00 | 1.38 | 1.40 | 1.40 | 1.44 | 1.44 | 1.44 | 1.40 | 1.42 | 1.38 | 1.34 | 1.41 |
| IV_PurePrem_Basic | 1.45 | 1.43 | 1.39 | 0.00 | 1.45 | 1.48 | 1.41 | 1.41 | 1.41 | 1.40 | 1.40 | 1.39 | 1.46 | 1.43 |
| SP_FreqSev | 1.47 | 1.49 | 1.45 | 1.50 | 0.00 | 1.44 | 1.47 | 1.47 | 1.46 | 1.45 | 1.47 | 1.45 | 1.49 | 1.50 |
| SP_PurePrem | 1.42 | 1.43 | 1.42 | 1.49 | 1.41 | 0.00 | 1.45 | 1.45 | 1.45 | 1.46 | 1.46 | 1.46 | 1.49 | 1.52 |
| IND_FreqSev | 1.47 | 1.48 | 1.48 | 1.45 | 1.45 | 1.48 | 0.00 | 1.38 | 1.40 | 1.53 | 1.30 | 1.50 | 1.52 | 1.49 |
| IND_PurePrem | 1.43 | 1.45 | 1.33 | 1.44 | 1.47 | 1.48 | 1.49 | 1.49 | 1.49 | 1.50 | 1.50 | 1.49 | 0.00 | 1.46 |
| IV_PurePrem | 1.44 | 1.47 | 1.41 | 1.42 | 1.48 | 1.51 | 1.45 | 1.45 | 1.45 | 1.49 | 1.45 | 1.50 | 1.46 | 0.00 |
| IV_FreqSevA | 1.41 | 1.43 | 1.45 | 1.43 | 1.43 | 1.49 | 1.54 | 1.54 | 1.54 | 0.00 | 1.55 | 1.34 | 1.53 | 1.52 |
| IV_FreqSevB | 1.47 | 1.47 | 1.47 | 1.44 | 1.46 | 1.49 | 1.30 | 1.30 | 1.27 | 1.55 | 0.00 | 1.53 | 1.54 | 1.49 |
| IV_FreqSevC | 1.41 | 1.42 | 1.43 | 1.43 | 1.44 | 1.49 | 1.52 | 1.51 | 1.51 | 1.34 | 1.53 | 0.00 | 1.53 | 1.53 |
| DepRatio1 | 1.47 | 1.48 | 1.48 | 1.45 | 1.45 | 1.48 | 1.38 | 0.00 | 1.32 | 1.53 | 1.30 | 1.50 | 1.52 | 1.49 |
| DepRatio36 | 1.48 | 1.48 | 1.48 | 1.45 | 1.45 | 1.48 | 1.40 | 1.32 | 0.00 | 1.53 | 1.27 | 1.50 | 1.52 | 1.49 |

"Sample Size Determination" section to propose some rules of thumb for sample size determination.

## USING COVARIANCES TO EXPRESS THE GINI INDEX

Equation (3) defines our Gini index in terms of an area associated with the ordered Lorenz curve. For the classic Lorenz curve and associated Gini index, there are several alternative (equivalent) definitions (cf. Yitzhaki, 1998). These different definitions encourage alternative interpretations of the Gini index, hence widening the scope of potential applications. As with the classic Gini, we can also provide an alternative expression for our Gini index using covariance operators.

### The Gini Index in Terms of Covariances

We use the notation $\widehat{\mathrm{Cov}}(y, P)$ to denote the (empirical) covariance between losses $y$ and premiums $P$. That is, define $\widehat{\mathrm{Cov}}(y, P) = n^{-1}(\sum_{i=1}^{n} y_i P_i - n\bar{y}\bar{P})$ (and recall that $\bar{y} = \bar{P} = 1$). Then, after some pleasant algebra (see Appendix B), we can express the Gini index as

$$\widehat{Gini} = 2\widehat{\mathrm{Cov}}(y, \hat{F}_P(R)) - 2\widehat{\mathrm{Cov}}(P, \hat{F}_R) - \frac{1}{n}\widehat{\mathrm{Cov}}(y, P), \tag{5}$$

where $\hat{F}_R = \mathrm{rank}(R)/n$ is the distribution function of the rank of relativities. For large sample sizes $n$, the third term on the right-hand side of Equation (5) is small and can be ignored.

With Equation (4), we interpret a low relativity means that a policy is highly profitable and a good candidate to retain. Additional insights arise from Equation (5). Other things being equal:

- 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.
- 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. We retain policies with a low relativity. Other things being equal, it is more profitable to retain a policy with a high premium.

For many data sets, we have found that, using Equation (5), we can approximate the weighted premium distribution $\hat{F}_P(R)$ with the unweighted distribution of relativities $\hat{F}_R$. With this, we may define

$$\widehat{Gini}_{Approx} = \frac{2}{n}\widehat{\mathrm{Cov}}((y - P), rank(R)). \tag{6}$$

Although this approximation to the Gini index provides little advantage computationally, it does give us another way to interpret the Gini index. We can think about $P - y$ as the "profit" associated with a policy. Then, we may interpret the Gini index

**TABLE 6**
Gini Indices, Approximations and Decompositions

| Score | Gini | Gini_ approx | Loss- Source | Prem- Source | Gini_ Approx2 | Gini- Reverse | Sum- Ginis |
|---|---|---|---|---|---|---|---|
| SP_PurePrem_Basic | 4.89 | 4.83 | −6.14 | −10.97 | 3.82 | 4.34 | 9.37 |
| IND_PurePrem_Basic | 4.01 | 3.98 | −3.60 | −7.58 | 3.81 | 8.08 | 12.49 |
| IV_PurePrem_Basic | 4.33 | 4.42 | −4.45 | −8.87 | 4.91 | 7.88 | 12.75 |
| SP_FreqSev | 11.15 | 11.29 | 9.25 | −2.03 | 10.79 | 1.71 | 13.18 |
| SP_PurePrem | 9.97 | 10.04 | 4.51 | −5.53 | 9.15 | 6.97 | 16.92 |
| IND_FreqSev | 10.03 | 10.31 | 0.59 | −9.72 | 10.71 | 6.85 | 17.39 |
| IND_PurePrem | 10.96 | 11.21 | 1.65 | −9.56 | 10.10 | 6.13 | 17.60 |
| IV_PurePrem | 11.29 | 11.44 | 4.28 | −7.16 | 10.56 | 6.69 | 18.26 |
| IV_FreqSevA | 12.59 | 12.86 | 5.06 | −7.80 | 12.31 | 2.79 | 15.64 |
| IV_FreqSevB | 10.61 | 10.83 | 1.54 | −9.29 | 11.18 | 6.77 | 17.76 |
| IV_FreqSevC | 12.80 | 12.99 | 5.89 | −7.10 | 12.31 | 3.44 | 16.37 |
| DepRatio1 | 10.09 | 10.36 | 0.76 | −9.60 | 10.74 | 6.83 | 17.41 |
| DepRatio36 | 10.06 | 10.34 | 0.65 | −9.69 | 10.72 | 6.77 | 17.33 |

*Note:* Base Premium is SP_FreqSev_Basic.

to be proportional to the negative covariance between profits and the rank of relativities. That is, if policies with low profits are associated with high relativities and high profits are associated with low relativities, then we have a profitable situation meaning that the Gini index is positive and large.

When "premiums" $P$ are interpreted as exposures, it is more helpful to think in terms of pure premiums. An alternative approximation is

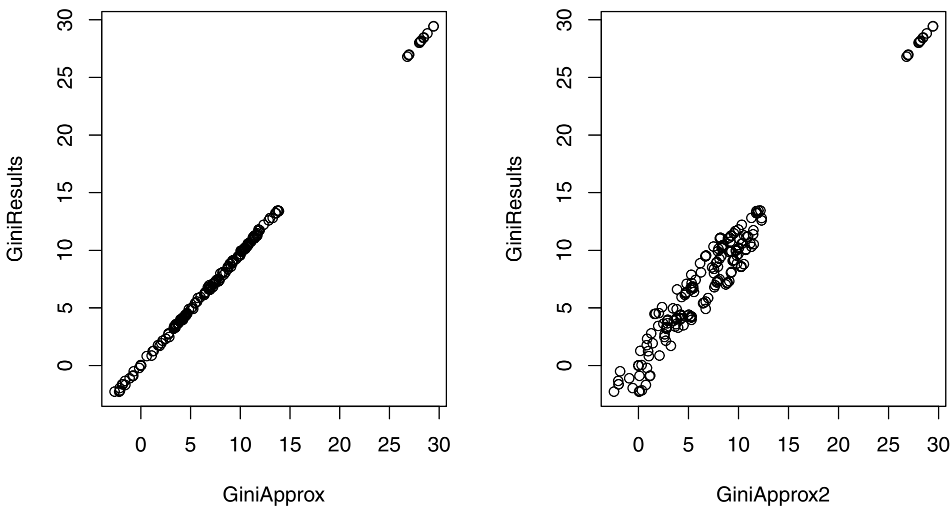
$$\widehat{Gini}_{Approx2} = \frac{2}{n}\widehat{\mathrm{Cov}}(PP, rank(R)), \tag{7}$$

where the scaled pure premium $PP$ is loss per premium $(y/P)$ rescaled (divided) by its average.

*Homeowners Example.* To understand the reliability of these approximations, we return to the homeowners example and show summaries in Table 6 and Figure 4. In Figure 4, we compare the two approximations with the Gini indices produced in Table 4. The left-hand panel shows that the approximation given in Equation (6) to be very robust over a wide range of premiums and relativities. The right-hand panel shows that the approximation given in Equation (7) to be less robust although still helpful for interpretation purposes.

In Table 6, we show the Gini indices and approximations for only those using SP_FreqSev_Basic as a base premium. We also decompose the Equation (6) approximation into "loss" and "premium" sources. Here, the loss source is given by $2\widehat{\mathrm{Cov}}\,(y, rank(R))\,/n$ and similarly for premiums. These two sources can help interpret the magnitude of the Gini index.

**FIGURE 4**
Gini Indices and Approximations.




*Note*: Observations in the upper right-hand corner correspond to using a constant premium as a base.

## Some Special Cases

*Simple Gini Index.* In some applications, one can use an exposure measure, such as car years, instead of a premium. In others, it is helpful to think about the premium and exposure as constant over policies. In this case, the relativity $R(\mathbf{x}_i) = S(\mathbf{x}_i)$ is simply an insurance score. The direct comparison of losses to scores results in what we call a "simple Gini."

When premiums are constant ($P_i \equiv 1$), interpretations (and the algebra) are simpler. In this case,the relativity is the score ($R_i = S_i$ in our notation). From Equation (5), the Gini index reduces to

$$\widehat{Gini} = \frac{2}{n}\widehat{\mathrm{Cov}}(y, Rank(S)). \tag{8}$$

Equation (8) is an exact relationship; no approximations are involved. It states that the simple Gini index is proportional to the covariance between losses and the rank of scores. Note that 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). As discussed in Frees, Meyers, and Cummings (2011), this statistic seems to have been first proposed by Durbin (1954) 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 mismeasured.

*Reverse Gini Index.* We now reverse the roles of scores $S$ and premiums $P$ and call the resulting Gini index a "reverse Gini." Returning to Equation (6), an approximation for the reverse Gini is

$$\begin{aligned} \widehat{GiniR}_{Approx} &= \frac{2}{n}\widehat{\text{Cov}}\left((y-S),(n+1-rank(R))\right) \\ &= \frac{2}{n}\widehat{\text{Cov}}((S-y),rank(R)). \end{aligned} \tag{9}$$

Moreover, suppose that the score $S$ is an unbiased estimator of the loss in the sense that $\text{E}(y|\mathbf{x}) = S$. Then,

$$\begin{aligned} \text{Cov}((S-y),rank(R)) &= \text{E}\{(S-y)\times rank(R)\} - \text{E}(S-y)\times \text{E}rank(R) \\ &= \text{E}\{(\text{E}(S-y)|\mathbf{x})\times rank(R)\} - \text{E}(\text{E}(S-y)|\mathbf{x})\times \text{E}rank(R) \\ &= 0, \end{aligned}$$

because $\text{E}(y|\mathbf{x}) = S$. This suggests that one can anticipate the reverse Gini, $\widehat{GiniR}_{Approx}$, to be zero when the model is well specified. We use the reverse Gini as another statistic to measure model fit.

*Gini Index for Refined Scores.* For another case, suppose that the score $S$ is a "more refined" version of a premium $P$. For example, $S$ may reflect information in a new rating variable (such as a credit score) or more precise geographic information. Specifically, let $S = P\exp(\mathbf{z}'\boldsymbol{\beta})$, where $\mathbf{z}$ is a vector of new variables not contained in the premium base $P$. It is helpful to think about some specific examples.

*Continuous Variable.* Suppose that we consider a single continuous variable (e.g., credit score). Then, the rank of the relative premium can be expressed as

$$rank(R) = rank(\frac{S}{P}) = rank(\exp(z\beta)) = rank(z),$$

assuming that $\beta$ is positive. Then, from Equation (6), we may interpret the Gini index to be approximately

$$\widehat{Gini}_{Approx} = \frac{2}{n}\widehat{\text{Cov}}\left((y-P),rank(z)\right);$$

that is, the Gini is approximately the covariance between the policy "profit" $P-y$ and the rank of the new variable $z$, rescaled by the constants.

*Categorical Variable.* Suppose that we consider on a single discrete variable $z$ with three possible outcomes 1, 2, and 3 (e.g., urban, suburb, and rural). Suppose we use $\mathbf{z}'\boldsymbol{\beta} = \beta_1\text{I}(z=1) + \beta_2\text{I}(z=2) + \beta_3\text{I}(z=3)$. Here, recall that $\text{I}(\cdot)$ is the indicator function. Without loss of generality, assume that $\beta_1 < \beta_2 < \beta_3$ (otherwise, simply reorder $z$). Then, $rank(R) = rank(\exp(z\beta)) = rank(z) = z$, and

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

Each level of $z$ represents a "segment" of the market that is implemented in the new score $S$ but not in the original premium base $P$. The Gini index measures the relationship between the policy "profit" $P - y$ and the market segments in $z$.

For both examples, we can think about the Gini index as summarizing the linear relationship between policy profit $P - y$ and the rank of the refinement variable, $rank(z)$. This suggests additional analyses, such as a plot of $P - y$ versus $rank(z)$, in order to understand potential nonlinear relationships.

## MODEL SELECTION

In this section, we investigate the role of the Gini index as a statistic to aid in selecting a model through a simulation study.

### Simulation Study Design

The study is designed to replicate many of the data features that we encountered when analyzing the homeowners data described in the "Homeowners Example" section.

For each scenario, we generated $n$ in-sample policyholder observations, estimated model parameters, and then calculated scores for each of $n$ out-of-sample policyholders. In our simulation, we let $n$ equal 500,000.

*Simulated Distributions.* For each policyholder, we assumed knowledge of two characteristics where each $x_j$ was generated from a chi-square distribution with 20 degrees of freedom, rescaled to have a zero mean and variance $1/10$. With these choices, the score distributions (score calculations are described below) exhibited a right-skewed distribution comparable to the premium distribution portrayed in Figure 1. The regression function was generated using a logarithmic link function, that is, $\mathrm{m}(\mathbf{x}) = \exp\left(\beta_0 + \beta_1 x_1 + \beta_2 x_2\right)$. Using the regression function as the location parameter, we generated the loss $y$ using the Tweedie distribution. Here, parameters of the Tweedie distribution were set so that the simulated distribution was comparable to the distribution of our homeowners data in the "Homeowners Example" section. We did this for $n$ in-sample and $n$ out-sample observations, respectively.

*Score Calculation.* Using the in-sample data, we estimated parameters for each of eight scores:

- $\mathrm{S}_1(\mathbf{x}) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)$, the true regression function
- $\mathrm{S}_2(\mathbf{x}) = \exp(\beta_0 + \beta_1 x_1)$, based on only $x_1$
- $\mathrm{S}_3(\mathbf{x}) = \exp(\beta_0 + \beta_2 x_2)$, based on only $x_2$
- $\mathrm{S}_4(\mathbf{x}) = \exp(\beta_0 + \beta_1 \frac{1}{x_1})$, based on an (incorrect) reciprocal transform of $x_1$
- $\mathrm{S}_5(\mathbf{x}) = \exp(\beta_0 + \beta_2 \frac{1}{x_2})$, based on an (incorrect) reciprocal transform of $x_2$

- $S_6(\mathbf{x}) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 \frac{1}{x_2})$, based on $x_1$ and an (incorrect) reciprocal transform of $x_2$
- $S_7(\mathbf{x}) = \exp(\beta_0 + \beta_1 \frac{1}{x_1} + \beta_2 x_2)$, based on $x_2$ and an (incorrect) reciprocal transform of $x_1$
- $S_8(\mathbf{x}) = \exp(\beta_0 + \beta_1 \frac{1}{x_1} + \beta_2 \frac{1}{x_2})$, based on (incorrect) reciprocal transforms of $x_1$ and $x_2$

Then, with the parameter estimates of the score coefficients from the in-sample data, we used the out-of-sample characteristics to generate scores. We then compared scores to the actual out-of-sample losses. We report results based on 100 simulations. With these sample and simulation sizes, it turns out that the simulation standard errors for all Gini indices are less than 0.2 percent.

Using the regression function in score 1, we generated two scenarios by varying the regression coefficients. In the first scenario, we let $\beta_1 = \beta_2 = 0.25$ so that both explanatory variables contribute equally to the regression function. In the second, we let $\beta_1 = 0.25$ and $\beta_2 = 0.05$ so that the second explanatory variable contributes little to the regression function.

For each scenario, we assume that the analyst is considering one of eight scores (with the first being the correct choice, unknown to the analyst). To measure the discrepancy between the chosen score and the true regression function, we present the Spearman correlation that we label as the "true correlation." Note that even when the analyst chooses the correct score, the Spearman correlation is less than one due to the (in-sample) estimation error in the regression coefficients in the score. If the true correlation were available, then the analyst would simply choose the model with the largest true correlation. However, this is unavailable, and so the analyst must use available statistics, including the "simple Gini" and the ratio Gini indices presented in Table 7. An analyst could select a model by searching for the largest simple Gini index. Alternatively, one could use a mini-max strategy discussed in the "Comparison of Scores Using the Gini Index" section.

### Simulation Study Results

Table 7 summarizes the results of the simulation study. Note that through our choice of scenarios we may observe outcomes over a broad range of models selected, ranging from a near perfect selection (where the correlation is 99.50 percent) to a very poor selection (where the correlation is only 15.35 percent).

The simple Gini index seems to be a desirable proxy for the true correlation. Over different scenarios and different premiums, as the simple Gini index increases, so does the true correlation. This is intuitively plausible in that the simple Gini index may be interpreted as proportional to the covariance between the insurance loss and rank of the score and whereas the "true correlation" is the correlation between the rank of the regression function and the rank of scores (a Spearman correlation). If one converts the simple Gini to a correlation it turns out to be much smaller than the true correlation. This is simply because of the noise in the loss random variable as a estimate of its expectation, the regression function.

**TABLE 7**
Gini Indices From a Simulation Study

| Scenario | Base Premium | Simple Gini | Ratio Gini Indices Score1 | Score2 | Score3 | Score4 | Score5 | Score6 | Score7 | Score8 | Maximum | True Correlation |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Explanatory | Score1 | 12.77 | 0.00 | 0.07 | 0.09 | 0.08 | 0.19 | 0.23 | −0.06 | 0.13 | 0.23 | 99.50 |
| variables | Score2 | 9.05 | 8.97 | 0.00 | 6.35 | −0.07 | 5.47 | 8.97 | 8.21 | 7.79 | 8.97 | 68.78 |
| contribute | Score3 | 8.97 | 9.06 | 6.53 | 0.00 | 5.61 | 0.23 | 8.42 | 9.05 | 8.01 | 9.06 | 68.77 |
| equally | Score4 | 9.05 | 9.83 | 3.43 | 6.41 | 0.00 | 5.44 | 9.65 | 8.97 | 8.97 | 9.83 | 68.78 |
| | Score5 | 8.97 | 9.80 | 6.52 | 3.11 | 5.56 | 0.00 | 9.05 | 9.62 | 9.05 | 9.80 | 68.77 |
| | Score6 | 12.44 | 3.11 | −1.75 | 1.50 | −1.31 | 0.09 | 0.00 | 2.35 | −0.07 | 3.11 | 96.96 |
| | Score7 | 12.40 | 3.43 | 1.62 | −1.68 | 0.07 | −1.15 | 2.80 | 0.00 | 0.22 | 3.43 | 97.15 |
| | Score8 | 12.26 | 5.08 | 0.51 | 0.41 | −1.75 | −1.68 | 3.43 | 3.12 | 0.00 | 5.08 | 94.88 |
| Second | Score1 | 8.87 | 0.00 | 0.30 | 0.09 | 0.42 | 0.13 | 0.43 | 0.18 | 0.28 | 0.43 | 98.93 |
| explanatory | Score2 | 8.80 | 1.45 | 0.00 | 0.39 | 0.17 | 0.33 | 1.32 | 0.86 | 0.75 | 1.45 | 97.71 |
| variable | Score3 | 1.44 | 8.79 | 8.73 | 0.00 | 8.59 | 0.45 | 8.78 | 8.80 | 8.74 | 8.80 | 17.66 |
| contributes | Score4 | 8.80 | 3.57 | 3.09 | −1.08 | 0.00 | −1.18 | 3.55 | 1.45 | 1.32 | 3.57 | 97.71 |
| little | Score5 | 1.29 | 8.82 | 8.74 | 0.36 | 8.63 | 0.00 | 8.79 | 8.76 | 8.80 | 8.82 | 15.35 |
| | Score6 | 8.85 | 0.38 | 0.16 | 0.17 | 0.30 | 0.09 | 0.00 | 0.34 | 0.17 | 0.38 | 98.61 |
| | Score7 | 8.74 | 3.08 | 3.12 | −1.63 | 0.30 | −1.50 | 3.16 | 0.00 | 0.42 | 3.16 | 98.08 |
| | Score8 | 8.76 | 3.22 | 3.08 | −1.48 | 0.16 | −1.63 | 3.09 | 0.38 | 0.00 | 3.22 | 97.90 |

Choosing the smallest of the maximum Gini ratios is also a viable model selection strategy. Table 7 shows a strong inverse relation between the "maximum" column and the true correlation column.

The simulation also allows us to document the "reverse Gini effect." To see this, consider the first scenario and suppose that the analyst initially chooses Score2 as the base premium. For Score3 as an alternative, the resulting ratio Gini index is 6.35 percent, suggesting that this score is preferred. However, if the analysts using Score3 as the base and Score2 as the alternative, then Table 7 shows that the resulting Gini index is 6.53 percent, suggesting that Score2 is preferred. This is the reverse Gini effect, where the Gini analysis provides seemingly contradictory advice.

However, because we generated the scores and the model, we know that neither Score2 nor Score3 represents the true outcome. Thus, the decision-making process with Gini indices suffers from the same drawback as with statistical hypothesis testing. Model A can be rejected in favor of Model B and vice versa if neither model is true. Table 7 shows that the reverse Gini effect is not present in the second scenario when one variable dominates the other.

A strong reverse Gini effect shows that there is room for improvement in the model used to create the score function. Although there is a variety of model selection techniques, one approach is to create an "ensemble" score by combining information in two scoring methods. To illustrate, Table 8 shows the results when we augmented Score2 with additional explanatory variables. Specifically, Score2 is created using $x_1$ as a predictor variable; "Score2(Score3)" is created using $x_1$ and Score3 as predictor variables in a two-stage process. That is, in stage 1 we created Score3 and in stage 2 we used both predictor variables to create this new ensemble score, "Score2(Score3)." Similar procedures were used for other ensemble scores in Table 8.

The top portion of Table 8 shows the advantage of using this approach beginning with the simple Score2. Here, we might be concerned with the reverse Gini effect (with Score3). One approach would be to use Score3 as an additional predictor. The resulting score, "Score2(Score3)," has a higher Gini index (with Score2 as the base premium) and we see no reverse Gini effect (the –0.20 index does not statistically differ from zero). One could also use the logarithmic Score3 or logarithmic Score8 as additional predictors and have the same desirable improvement in Score2.

Of course, this strategy does not work in all situations. The bottom portion of Table 8 shows the effects when using this approach beginning with the complex, yet incorrect, Score8. Here, we see that adding Score2, logScore2, or logScore3 improves the Gini index with no statistically significant reverse Gini effects. However, if one were to adopt the Score8(logScore2) as a base premium, then this score is not ideal in the sense that it suffers from a significant reverse Gini effect with Score8 (logScore3).

This simulation demonstrates that the ensemble technique for producing improved estimators has merit in certain situations. It also demonstrates that using the Gini index and reversing the roles of base premium and scores are useful devices in model selection.

**TABLE 8**
Gini Indices from Ensemble Scores

| Base | Simple | Ratio Gini Indices | | | | | | | True |
|---|---|---|---|---|---|---|---|---|---|
| Premium | Gini | Score2 | Score3 | Score8 | Score2(Score3) | Score2(logScore3) | Score2(logScore8) | Maximum | Correlation |
| Score2 | 9.05 | 0.00 | 6.35 | 7.79 | 8.97 | 8.97 | 8.08 | 8.97 | 68.78 |
| Score3 | 8.97 | 6.53 | 0.00 | 8.01 | 8.98 | 9.06 | 8.21 | 9.06 | 68.77 |
| Score8 | 12.26 | 0.51 | 0.41 | 0.00 | 5.03 | 5.08 | 1.41 | 5.08 | 94.88 |
| Score2(Score3) | 12.74 | −0.20 | 0.27 | 0.68 | 0.00 | 1.04 | 0.44 | 1.04 | 99.35 |
| Score2(logScore3) | 12.77 | 0.07 | 0.09 | 0.13 | −0.04 | 0.00 | 0.01 | 0.13 | 99.50 |
| Score2(logScore8) | 12.14 | −1.40 | 1.90 | 0.78 | 4.66 | 4.75 | 0.00 | 4.75 | 93.96 |

| Base | Simple | Ratio Gini Indices | | | | | | | True |
|---|---|---|---|---|---|---|---|---|---|
| Premium | Gini | Score2 | Score3 | Score8 | Score8(Score2) | Score8(logScore2) | Score8(logScore3) | Maximum | Correlation |
| Score2 | 9.05 | 0.00 | 6.35 | 7.79 | 8.84 | 8.87 | 8.16 | 8.87 | 68.78 |
| Score3 | 8.97 | 6.53 | 0.00 | 8.01 | 8.37 | 8.40 | 8.90 | 8.90 | 68.77 |
| Score8 | 12.26 | 0.51 | 0.41 | 0.00 | 3.40 | 3.40 | 3.07 | 3.40 | 94.88 |
| Score8(Score2) | 12.46 | −1.49 | 1.45 | −0.26 | 0.00 | 0.43 | 2.31 | 2.31 | 96.90 |
| Score8(logScore2) | 12.44 | −1.55 | 1.54 | −0.13 | 0.04 | 0.00 | 2.32 | 2.32 | 96.73 |
| Score8 (logScore3) | 12.40 | 1.69 | −1.41 | 0.28 | 2.77 | 2.82 | 0.00 | 2.82 | 97.07 |

## SAMPLE SIZE DETERMINATION

How large a sample size is required for a reliable *Gini* statistic? In this section, we show how to use results from the theoretical properties, plus some basic knowledge of the loss distribution, to provide rules of thumb that can be used to select an appropriate sample size. This procedure could be used, for example, to determine the size of a block of business when examining alternative premium structures on a trial, or "pilot," basis.

In earlier work, we showed that the distribution of the *Gini* statistic is approximately normal for large samples, (see Theorem A2 of Appendix A). The form of the large sample variance, $\Sigma_{Gini}/n$, given in Theorem A2, is complicated. However, Appendix C shows that using the assumption of independent relativities results in a much simpler expression, given as

$$\Sigma_{Gini} = \frac{\mathrm{Var}(y-P)}{3}. \tag{10}$$

This result is similar to one for the Pearson correlation, another measure of association, where the form of the variance simplifies under the independence assumption.

To illustrate, for our sample described in the "Homeowners Example" section, we have $n = 359,454$ observations. After rescaling so that premiums and losses are mean one, we have the standard deviation of losses and premiums are, respectively, $s_y = 14.79591$ and $s_P = 0.70558$. The covariance between losses and premium is $Cov_{yP} = 0.48538$. From this and Equation (10), an approximate standard error for the Gini index is

$$se(\widehat{Gini})_{Approx} = \sqrt{\frac{14.79591^2 + 0.70558^2 - 2 \times 0.48538}{3 \times 359454}} = 0.0142.$$

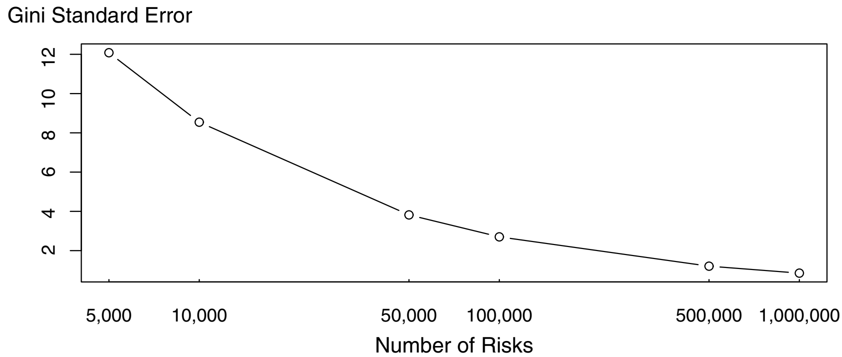
This result is close to the standard errors presented in Table 5 (calculated using the complicated, yet more precise, expression for $\Sigma_{Gini}$ given in "Appendix A: Properties of the Gini Index").

Because this approximation is valid over a range of relativities and premiums (with different dependency structures), we conclude that this approximation is helpful for determining the size of a sample to be collected for test studies. Figure 5 shows the effect of alternative sample sizes on the approximate Gini standard error for our homeowners data.

## SUMMARY AND CONCLUDING REMARKS

The Gini index is a measure of association between losses and premiums—one that has important economic content in insurance scoring applications. For a given ordering of risks, the Gini index summarizes the difference between the premium and loss distributions. An excess of premiums over losses can be interpreted to be an insurer's profit. This observation leads an insurer to seek an ordering that produces to a large Gini index. Thus, the Gini index and associated ordered Lorenz curve are useful for identifying profitable blocks of insurance business.

**FIGURE 5**
Effect of Sample Size on Gini Approximate Standard Errors




Unlike classical measures of association, the Gini index assumes that a premium base $P$ is currently in place and seeks to assess vulnerabilities of this structure. This approach is more akin to hypothesis testing (when compared to goodness of fit) where one identifies a "null hypothesis" as the current state of the world and uses decision-making criteria/statistics to compare this with an "alternative hypothesis." The purpose of this article is not to say that either hypothesis testing or goodness of fit approaches are always good or bad; rather, both have their place in statistical inference. The purpose of the article is provide another measure that can be used to augment the analyst's toolkit; we argue that this new measure provides insights that are not available from classical measures of association.

To summarize, we anticipate the Gini index being of use in at least the following three situations:

1. A premium structure $P$ is in place and we wish to assess the usefulness of a generic alternative score $S$. This is the basic scenario in which we introduced the ordered Lorenz curve and the Gini index to combat adverse selection. We also discussed the Gini index as a measure of profit in the "Ordered Lorenz Curve and the Gini Index" section. To illustrate, we demonstrated its usefulness in the homeowners example in the "Comparing Scoring Methods to a Selected Base Premium" section; analysts can supplement this analysis by looking to the reverse Gini as an additional measure of model fit.
2. A premium structure $P$ is in place and the alternative score is a refined version of the premium. Although this is the same as the basic scenario defined above, additional interpretations ("Gini Index for Refined Scores" section) are available for the relativities that are potentially helpful in model diagnostics.
3. No premium structure is in place and a number of alternative scoring methods are being considered. In this case, at least two strategies are available. One is to use the "minimax" strategy put forth in the "Comparison of Scores Using the Gini Index" section, where one chooses a score that is least vulnerable to competition

from other scores. The other strategy is to use the "simple" Gini index that has no base premium as a reference.

We motivate the introduction of relativities through the introduction of new rating variables in the "Introduction." One way for the business analyst to assess the impact of a new rating variable is through its statistical significance based on an in-sample testing procedure. As a supplement, the Gini index allows the analyst to assess its effectiveness by examining out-of-sample losses. Specifically, in Equations (5) and (6), we showed how the Gini index could be interpreted as proportional to a covariance between the relativity and an out-of-sample profit (premium in excess of loss). In predictive modeling, one validates a model by examining performance on an independent held-out sample.

To assess the reliability of the Gini index, the "Sample Size Determination" section describes principles for sample size determination. On the one hand, sample sizes required for reliable applications such as in personal lines homeowners insurance are quite large, in some cases ranging into the hundreds of thousands of observations. On the other hand, with large sample sizes available, we can enjoy reliable inferences for complex distributions that are mixtures of a large mass at zero and a right-skewed, thick-tailed positive distribution. Given the availability of large data sets in today's world, we view these sample size requirements as feasible in some important areas of applications.

## APPENDIX A: PROPERTIES OF THE GINI INDEX

The (population) Gini index is

$$Gini = 2\int_0^\infty \{F_P(s) - F_L(s)\} dF_P(s). \tag{A1}$$

Here, $F_P$ and $F_L$ are weighted distributions functions, which are the population versions of the empirical distributions given in Equations (1) and (2), respectively. As described in Frees, Meyers, and Cummings (2011), one can write the ordered Lorenz curve in the same fashion as classic Lorenz curve using weighted distribution functions in place of distribution functions. Because our work is a generalization of the idea due to Lorenz it seems appropriate to retain this name for the curve (although with the modifier "ordered").

We summarize the consistency and asymptotic normality of the empirical Gini index $\widehat{Gini}$ in the following two results.

**Theorem A1:** *Under mild regularity conditions, the Gini statistic $\widehat{Gini}$ is a consistent estimator of of the Gini index. That is, $\widehat{Gini} \rightarrow Gini$, as $n \rightarrow \infty$, with probability one.*

*For asymptotic normality, we use the projection*

$$h_1(\mathbf{x}, y) = \tfrac{1}{2}(\mu_y P(\mathbf{x}) F_L(R) + y\mu_P[1 - F_P(R)]). \tag{A2}$$

**TABLE A1**
Moment-Based Estimators for the Asymptotic Variance

| | |
|---|---|
| $\hat{h}_1(\mathbf{x}, y) = \frac{1}{2}(P(\mathbf{x})\hat{F}_L(R) + y[1 - \hat{F}_P(R)])$ | $\hat{\Sigma}_{hP} = n^{-1}\sum_{i=1}^{n} \hat{h}_1(\mathbf{x}_i, y_i)P(\mathbf{x}_i) - \bar{h}_1$ |
| $\bar{h}_1 = n^{-1}\sum_{i=1}^{n} \hat{h}_1(\mathbf{x}_i, y_i)$ | $\hat{\Sigma}_y = n^{-1}\sum_{i=1}^{n} y_i^2 - 1$ |
| $\hat{\Sigma}_h = n^{-1}\sum_{i=1}^{n} \hat{h}_1(\mathbf{x}_i, y_i)^2 - \bar{h}_1^2$ | $\hat{\Sigma}_P = n^{-1}\sum_{i=1}^{n} P(\mathbf{x}_i)^2 - 1$ |
| $\hat{\Sigma}_{hy} = n^{-1}\sum_{i=1}^{n} \hat{h}_1(\mathbf{x}_i, y_i)y_i - \bar{h}_1$ | $\hat{\Sigma}_{yP} = n^{-1}\sum_{i=1}^{n} y_i P(\mathbf{x}_i) - 1$ |

*Note:* These estimators are based on rescaling so that $\bar{y} = \bar{P} = 1$.

*Further, use the notation* $\Sigma_h = \text{Var } h_1(\mathbf{x}, y)$, $\Sigma_y = \text{Var } y$, $\Sigma_P = \text{Var } P(\mathbf{x})$, $\Sigma_{hy} = \text{Cov}(h_1(\mathbf{x}, y), y)$, $\Sigma_{yP} = \text{Cov}(y, P(\mathbf{x}))$, *and* $\Sigma_{hP} = \text{Cov}(h_1(\mathbf{x}, y), P(\mathbf{x}))$. *With these terms, we can establish:*

**Theorem A2:** *Under mild regularity conditions, the Gini statistic $\widehat{Gini}$ has an asymptotic normal distribution. Specifically, $\sqrt{n}(\widehat{Gini} - Gini) \to_D N(0, \Sigma_{Gini})$, where*

$$\Sigma_{Gini} = \frac{4}{\mu_y^2 \mu_P^2}\left(4\Sigma_h + \frac{\mu_h^2}{\mu_y^2}\Sigma_y + \frac{\mu_h^2}{\mu_P^2}\Sigma_P - \frac{4\mu_h}{\mu_y}\Sigma_{hy} - \frac{4\mu_h}{\mu_P}\Sigma_{hP} + \frac{2\mu_h^2}{\mu_y \mu_P}\Sigma_{yP}\right), \quad \text{(A3)}$$

*with* $\mu_h = \mu_y \mu_P (1 - Gini)/2$.

To estimate the asymptotic variance, Table A1 provides moment-based estimators.

**Theorem A3:** *Under mild regularity conditions, a consistent estimator of $\Sigma_{Gini}$ is*

$$\hat{\Sigma}_{Gini} = 4\big(4\hat{\Sigma}_h + \bar{h}_1^2\hat{\Sigma}_y + \bar{h}_1^2\hat{\Sigma}_P - 4\bar{h}_1\hat{\Sigma}_{hy} - 4\bar{h}_1\hat{\Sigma}_{hP} + 2\bar{h}_1^2\hat{\Sigma}_{yP}\big). \quad \text{(A4)}$$

The proofs of these results are Frees, Meyers, and Cummings (2011).

## APPENDIX B: PROOF OF EQUATION (5)

We establish Equation (5) assuming $\bar{y} = \bar{P} = 1$. First, from Equation (1), we have that $a_j - a_{j-1} = \frac{P_j}{n}$ and, from Equation (2), we have that $b_j + b_{j-1} = 2\hat{F}_L(R_j) - \frac{y_j}{n}$. Thus, with Equation (3), we have

$$\begin{aligned} \widehat{Gini} &= 1 - \sum_{j=1}^{n}(a_j - a_{j-1})(b_j + b_{j-1}) \\ &= 1 - \frac{1}{n}\sum_{j=1}^{n} P_j\left(2\hat{F}_L(R_j) - \frac{y_j}{n}\right). \end{aligned} \quad \text{(B1)}$$

Now, with Equation (2) and a change of summations, we can write

$$
\begin{aligned}
\textstyle\sum_{j=1}^{n} P_j \hat{F}_L(R_j) &= \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n} y_i P_j \mathrm{I}(R_i \le R_j) \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i \left\{\sum_{j=1}^{n} P_j \mathrm{I}(R_i \le R_j)\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i \left\{\sum_{j=1}^{n} P_j \left(1 - \mathrm{I}(R_j \le R_i) + \mathrm{I}(R_i = R_j)\right)\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i \{n - n\hat{F}_P(R_i) + P_i\}.
\end{aligned}
$$

Putting this into Equation (B1), we have

$$
\begin{aligned}
\widehat{Gini} &= 1 - \frac{2}{n^2}\sum_{j=1}^{n} y_i \{n - n\hat{F}_P(R_i) + P_i\} + \frac{1}{n^2}\sum_{i=1}^{n} y_i P_i \\
&= \frac{2}{n}\sum_{j=1}^{n} y_i \hat{F}_P(R_i) - 1 - \frac{1}{n^2}\sum_{i=1}^{n} y_i P_i \\
&= \frac{2}{n}\{n\widehat{\mathrm{Cov}}(y, \hat{F}_P(R)) + n\overline{\hat{F}_P(R)}\} - 1 - \frac{1}{n^2}\{n\widehat{\mathrm{Cov}}(y, P) + n\} \\
&= 2\widehat{\mathrm{Cov}}(y, \hat{F}_P(R)) + 2\overline{\hat{F}_P(R)} - \frac{n+1}{n} - \frac{1}{n}\widehat{\mathrm{Cov}}(y, P).
\end{aligned} \tag{B2}
$$

We now use $\widehat{\mathrm{Cov}}(P, R) = n^{-1}(\sum_{i=1}^{n} P_i \times i - n\frac{n+1}{2})$ (recall that premiums are sorted by relativities so that the rank of the $i$th relativity is $i$). To calculate the average weighted premium distribution, we have

$$
\begin{aligned}
\overline{\hat{F}_P(R)} &= \frac{1}{n}\sum_{i=1}^{n} \hat{F}_P(R_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} P_j \mathrm{I}(R_j \le R_i) \\
&= \frac{1}{n^2}\sum_{j=1}^{n} P_j (n - j + 1) \\
&= \frac{n+1}{n} - \frac{1}{n^2}\left\{n\widehat{\mathrm{Cov}}(P, R) + n\frac{n+1}{2}\right\} \\
&= \frac{n+1}{n} - \frac{n+1}{2n} - \frac{1}{n}\widehat{\mathrm{Cov}}(P, R) = \frac{n+1}{2n} - \widehat{\mathrm{Cov}}(P, F_R).
\end{aligned}
$$

Putting this into Equation (B2) yields

$$\begin{aligned}\widehat{Gini} &= 2\widehat{\mathrm{Cov}}(y, \hat{F}_P(R)) + 2\left(\frac{n+1}{2n} - \widehat{\mathrm{Cov}}(P, F_R)\right) - \frac{n+1}{n} - \frac{1}{n}\widehat{\mathrm{Cov}}(y, P)\\ &= 2\widehat{\mathrm{Cov}}(y, \hat{F}_P(R)) - 2\widehat{\mathrm{Cov}}(P, F_R)) - \frac{1}{n}\widehat{\mathrm{Cov}}(y, P),\end{aligned}$$

which is Equation (5).

## APPENDIX C: SAMPLE SIZE CALCULATIONS

The following proposition is a corollary of Theorem A2.

**Proposition:** *Assume that $R$ is independent of $(y, P)$ and the conditions of Theorem A2 hold. Then, we have $\sqrt{n}\widehat{Gini} \rightarrow_D N(0, \Sigma_{Gini})$, where*

$$\Sigma_{Gini} = \frac{\Sigma_y + \Sigma_P - 2\Sigma_{yP}}{3} = \frac{\mathrm{Var}(y - P)}{3}.$$

*For simplicity, we establish this proposition assuming by losses and premiums have been rescaled by dividing by their respective averages. Through this rescaling, we have that the mean loss is* $\mathrm{E}\ y = 1$ *and the mean premium is* $\mathrm{E}\ P = 1$.

**Proof:** Under the assumption that $R$ is independent of $(y, P)$, we first note that

$$\begin{aligned}F_P(s) &= \frac{\mathrm{E}[P(\mathbf{x})\mathrm{I}(R \le s)]}{\mathrm{E}\ P(\mathbf{x})} = \frac{\mathrm{E}[P(\mathbf{x})]\,\Pr(R \le s)}{\mathrm{E}\ P(\mathbf{x})}\\ &= \Pr(R \le s) = F_R(s),\end{aligned}$$

and similarly, $F_L(s) = F_R(s)$. Thus, using Equation (A2), we may write the projection as

$$h_1(\mathbf{x}, y) = \tfrac{1}{2}(P F_R + y[1 - F_R]).$$

Recall, for continuous relativities $R$, that $F_R$ has a uniform distribution and so $\mathrm{E}\ F_R = 1/2$, $\mathrm{Var}\ F_R = 1/12$, and $\mathrm{E}\ F_R^2 = 1/12 + (1/2)^2 = 1/3$. Now, assuming $R$ is independent of $(y, P)$, this has mean

$$\mu_h = \mathrm{E}h_1(\mathbf{x}, y) = \tfrac{1}{2}\{(1)(1/2) + (1)[1 - (1/2)]\} = \tfrac{1}{2},$$

and so $Gini = 0$. Further,

$$\begin{aligned}
\Sigma_h &= \mathrm{Var}\, h_1(\mathbf{x}, y) = \mathrm{E}\, h_1^2 - \left(\tfrac{1}{2}\right)^2 \\
&= \tfrac{1}{4}\mathrm{E}\left\{P^2 F_R^2 + y^2(1 - F_R)^2 + 2yP\left(F_R - F_R^2\right)\right\} - \tfrac{1}{4} \\
&= \tfrac{1}{4}\left((\Sigma_P + 1)\left(\tfrac{1}{3}\right) + (\Sigma_y + 1)\left(\tfrac{1}{3}\right) + 2(\Sigma_{yP} + 1)\left(\tfrac{1}{2} - \tfrac{1}{3}\right)\right) - \tfrac{1}{4} \\
&= \frac{\Sigma_y + \Sigma_P + \Sigma_{yP}}{12}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\Sigma_{hy} &= \mathrm{Cov}\; h_1(\mathbf{x}, y) = \mathrm{E}\; yh_1 - \left((1)\tfrac{1}{2}\right) \\
&= \tfrac{1}{2}\mathrm{E}\{yPF_R + y^2(1 - F_R)\} - \tfrac{1}{2} \\
&= \tfrac{1}{2}\mathrm{E}\left\{(\Sigma_{yP} + 1)\tfrac{1}{2} + (\Sigma_y + 1)\tfrac{1}{2}\right\} - \tfrac{1}{2} \\
&= \frac{\Sigma_y + \Sigma_{yP}}{4}.
\end{aligned}$$

By symmetry, we have $\Sigma_{hP} = (\Sigma_P + \Sigma_{yP})/4$.

Now, using Equation (A3), $\mu_y = 1$, $\mu_P = 1$, and $\mu_h = 1/2$, we have

$$\begin{aligned}
\Sigma_{Gini} &= 4\{4\Sigma_h + \mu_h^2\Sigma_y + \mu_h^2\Sigma_P - 4\mu_h\Sigma_{hy} - 4\mu_h\Sigma_{hP} + 2\mu_h^2\Sigma_{yP}\} \\
&= 16\Sigma_h + \Sigma_y + \Sigma_P - 8\Sigma_{hy} - 8\Sigma_{hP} + 2\Sigma_{yP} \\
&= 16\left(\frac{\Sigma_y + \Sigma_P + \Sigma_{yP}}{12}\right) + \Sigma_y + \Sigma_P - 8\left(\frac{\Sigma_y + \Sigma_{yP}}{4}\right) - 8\left(\frac{\Sigma_P + \Sigma_{yP}}{4}\right) + 2\Sigma_{yP} \\
&= \frac{\Sigma_y + \Sigma_P - 2\Sigma_{yP}}{3},
\end{aligned}$$

as required. Q.E.D.

## REFERENCES

Durbin, J., 1954, Errors in Variables, *Review of International Statistical Institute*, 22: 23-32.

Egghe, L., 2005, Continuous, Weighted Lorenz Theory and Applications to the Study of Fractional Relative Impact Factors, *Information Processing and Management*, 41: 1330-1359.

Frees, E. W., G. Meyers, and A. D. Cummings, 2010, Dependent Multi-Peril Ratemaking Models, *Astin Bulletin*, 40(2): 699-726.

Frees, E. W., G. Meyers, and A. D. Cummings, 2011, Summarizing Insurance Scores Using a Gini Index, *Journal of the American Statistical Association*, 106: 1085-1098.

Frees, E. W., G. Meyers, and A. D. Cummings, Forthcoming, Predictive Modeling of Multi-Peril Homeowners Insurance, *Variance*.

Gastwirth, J. L. 1971, A General Definition of the Lorenz Curve, *Econometrica* 39(6): 1037-1039.

Gastwirth, J. L., 1972, The Estimation of the Lorenz Curve and the Gini Index, *Review of Economics and Statistics*, 54(3): 306-316.

Gastwirth, J. L., 1975, Statistical Measures of Earnings Differentials, *The American Statistician*, 29(1): 32-35.

Gourieroux, C., and J. Jasiak, 2007, *The Econometrics of Individual Risk: Credit, Insurance and Marketing* (Princeton, NJ: Princeton University Press).

Hastie, T., R. Tibshirani, and J. Friedman, 2001, *The Elements of Statistical Learning: Data Mining, Inference and Prediction* (New York: Springer).

Lambert, P. J., and A. Decoster, 2005, The Gini Coefficient Reveals More, *Metron*, 63(3): 373-400.

Lorenz, M. O., 1905, Methods of Measuring the Concentration of Wealth, *Publications of the American Statistical Association*, 9(70): 209-219.

Meyers, G., and A. D. Cummings, 2009, "Goodness of Fit" vs. "Goodness of Lift," *The Actuarial Review: Newsletter of the Casualty Actuarial Society*, August, p. 16. World Wide Web: http://www.casact.org/newsletter/pdfUpload/ar/AR_Aug2009_1.pdf. (accessed November 23, 2011).

Nicodemus, K. K., and J. D. Malley, 2009, Predictor Correlation Impacts Machine Learning Algorithms: Implications for Genomic Studies, *Bioinformatics*, 25(15): 1884-1890.

Sandri, M., and P. Zuccolotto, 2008, A Bias Correction Algorithm for the Gini Variable Importance Measure in Classification Trees, *Journal of Computational and Graphical Statistics*, 17(3): 611-628.

Sen, A., and J. E. Foster, 1998, *On Economic Inequality* (Delhi: Oxford University Press).

UNU-WIDER World Income Inequality Database, 2008, Version 2.0c, May 2008. Sponsored by the World Institute for Development Economics Research of the United Nations University, Helsinki, Finland. World Wide Web: http://www.wider.unu.edu/research/Database/en_GB/database/. (accessed November 23, 2011).

Werner, G., and C. Modlin, 2010, *Basic Ratemaking*, 4th edition (Arlington, VA: Casualty Actuarial Society).

Yitzhaki, S., 1996, On Using Linear Regressions in Welfare Economics, *Journal of Business and Economic Statistics*, 14(4): 478-486.

Yitzhaki, S., 1998, More than a Dozen Ways of Spelling Gini, *Research on Economic Inequality*, 8: 13-30.