Evaluating Predictive Models with the Gini Index

Glenn Meyers – Introduction to Gini Index
Jed Frees – Underlying Theory
Dave Cummings – Current Applications

CAS Annual Meeting – Nov 17, 2015

Statement of the Predictive Modeling Problem

- ISO Innovative Analytics formed in 2005
- Early project – Very refined auto territories
- Available independent variables.
  - Census data – e.g. popluation density
  - Weather data – e.g. snow, wind etc.
  - Business data – e.g. schools, shopping centers, churches etc.
  - etc.
- We built a model!

Is the New Model More Valuable than the Existing Model?

- Common actuarial ratemaking practice is to balance to the same premium regardless of the class plan.
  - So what difference does it make?
- Valuable?
  - Use economic, as opposed to statistical, criteria for model selection.
- The economic rationale for risk classification is to prevent adverse selection.

An Early Attempt at an Economic Criteria The Value of Lift (VoL)

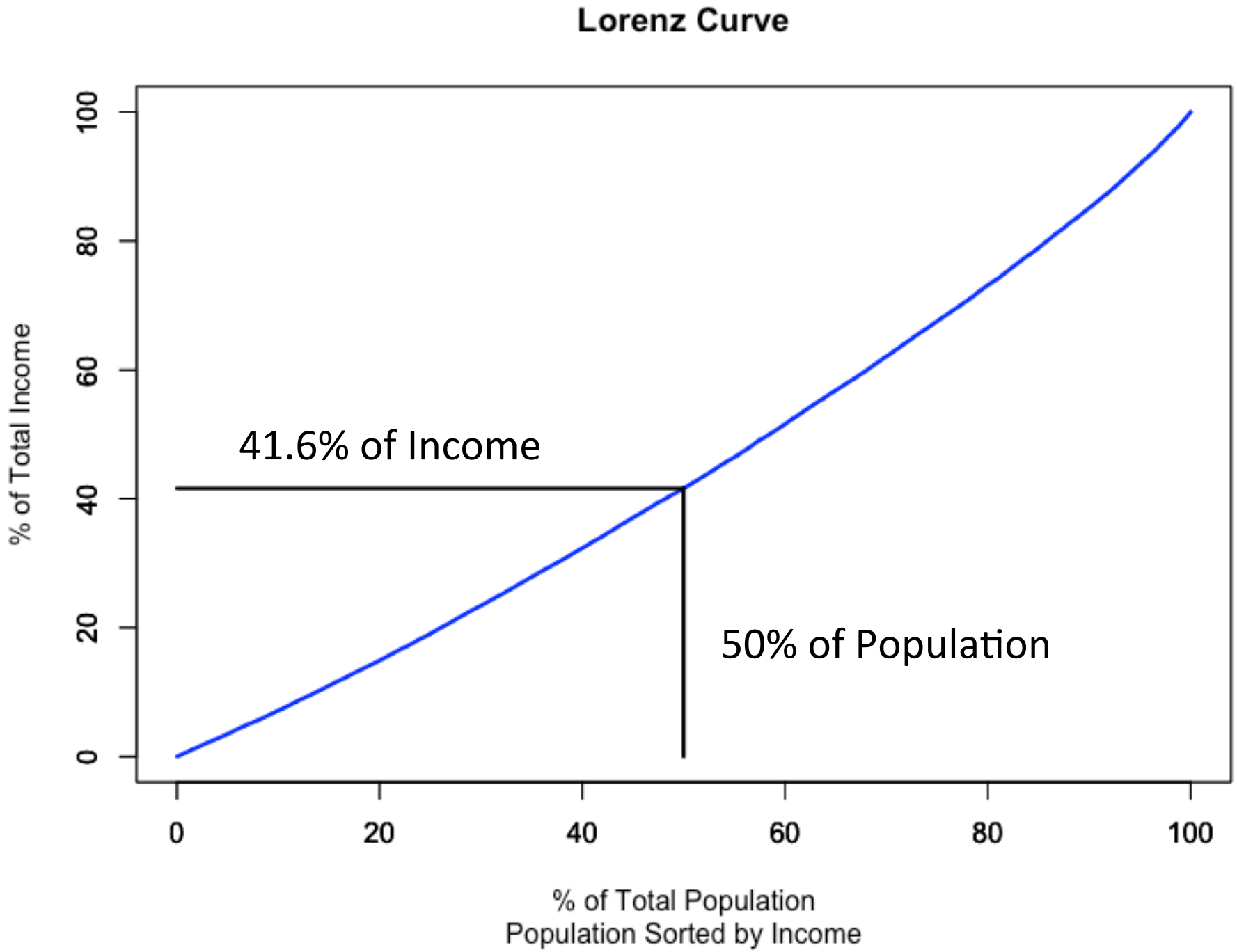
- The potential profit that could be lost to a competitor with a more accurate class plan.
- Depended on strong behavioral assumptions
  - e.g. Perfect price sensitivity
  - Ignored cost of developing and maintaining the more accurate class plan.
- We wanted a statistic that reflected economic criteria, but was less burdened by strong behavioral assumptions.
- I wrote about this in the *Actuarial Review*, February 2008.

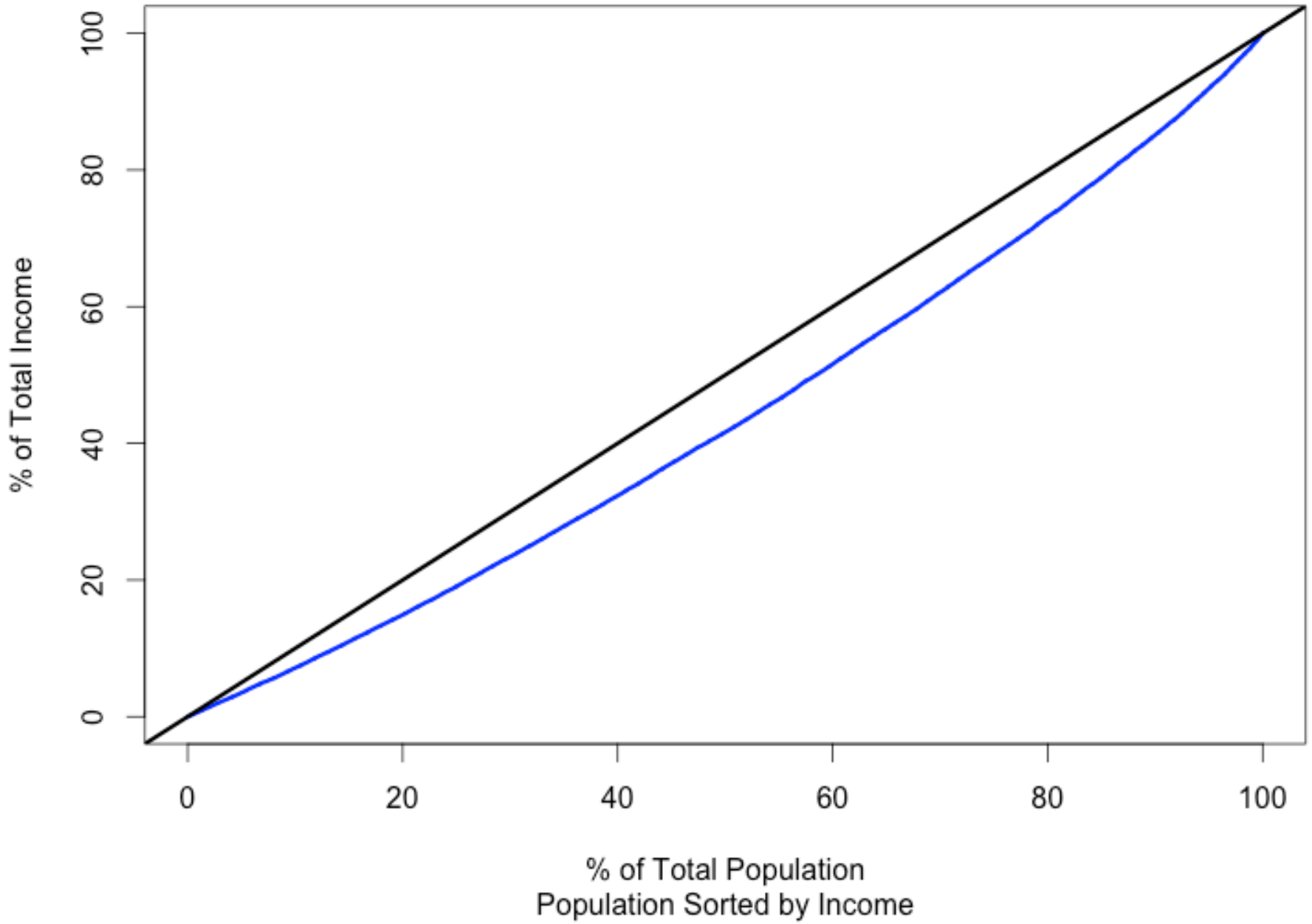
The Gini Index

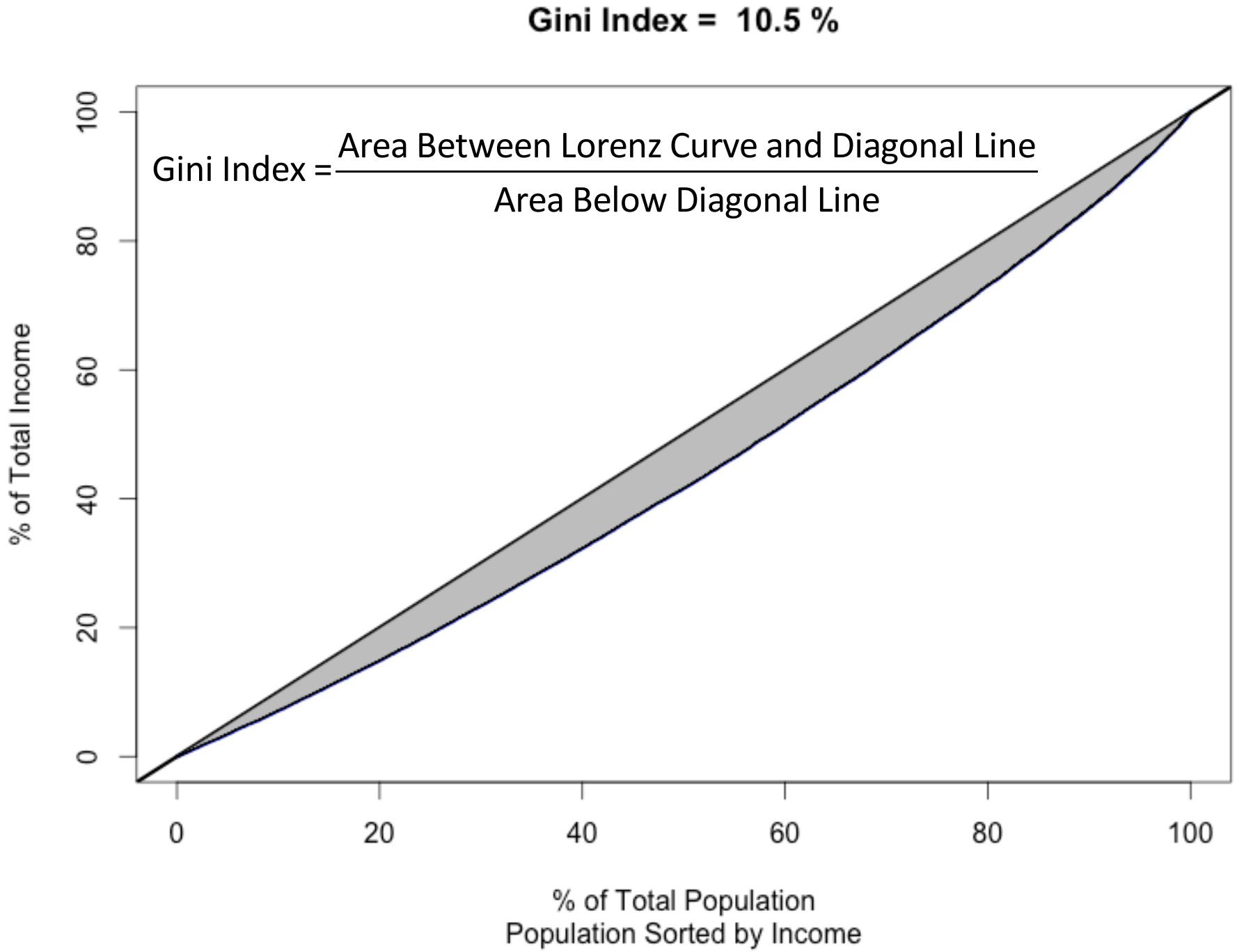
- Suggested to me by Daniel Finnegan, my boss at the time, who held a Ph.D. in Sociology from UC Berkeley.
- Proposed by Italian statistician and sociologist, Corrado Gini, in 1912 to study the distribution of income of a nation.
- It is used today to in many diverse fields such as ecology, biodiversity and business modeling.

First - The Lorenz Curve

- *Methods of measuring the concentration of wealth* - JASA 1905 by Max. O. Lorenz
  - Then a Ph.D. student at the University of Wisconsin - Madison


Lorenz Curve
% of Total Income
0
20
40
60
80
100
41.6% of Income
50% of Population
0
20
40
60
80
100
% of Total Population
Population Sorted by Income


Lorenz Curve for Equal Income - Diagonal Line
% of Total Income
0
20
40
60
80
100
0
20
40
60
80
100
% of Total Population
Population Sorted by Income


Gini Index = 10.5 %
Gini Index = Area Between Lorenz Curve and Diagonal Line / Area Below Diagonal Line
% of Total Income
0
20
40
60
80
100
0
20
40
60
80
100
% of Total Population
Population Sorted by Income

The Insurance Problem

- Suppose we have two premium calculations
  - $P_1(x_1) = \mathrm{E}[\mathrm{Loss}|x_1]$
  - $P_2(x_1,x_2)=\mathrm{E}[\mathrm{Loss}|x_1,x_2]$
- In words P_2 has a more refined classification plan than P_1.
- Is P_2 an economically significant better predictor of losses than P_1 on a holdout sample of data?

Terminology

- For risk *i* with independent variables $x_{1,i}$ and $x_{2,i}$
- Define the relativity $$R_i = \frac{P_2\left(x_{1,i}, x_{2,i}\right)}{P_1\left(x_{1,i}\right)}$$

The Lorenz Curve in an Insurance Context

Income

- X-Axis – Population
- Y-Axis – Income
- Sort order – Income

Insurance

- X-Axis – P_1
- Y-Axis – Losses
- Sort order – Relativity

- Sort order and Y-Axis variables are the same in the income context.
- Sort order and Y-Axis variables are different in the income context.
- Insurance losses are more volatile in the insurance context.

The Ordered Lorenz Curve

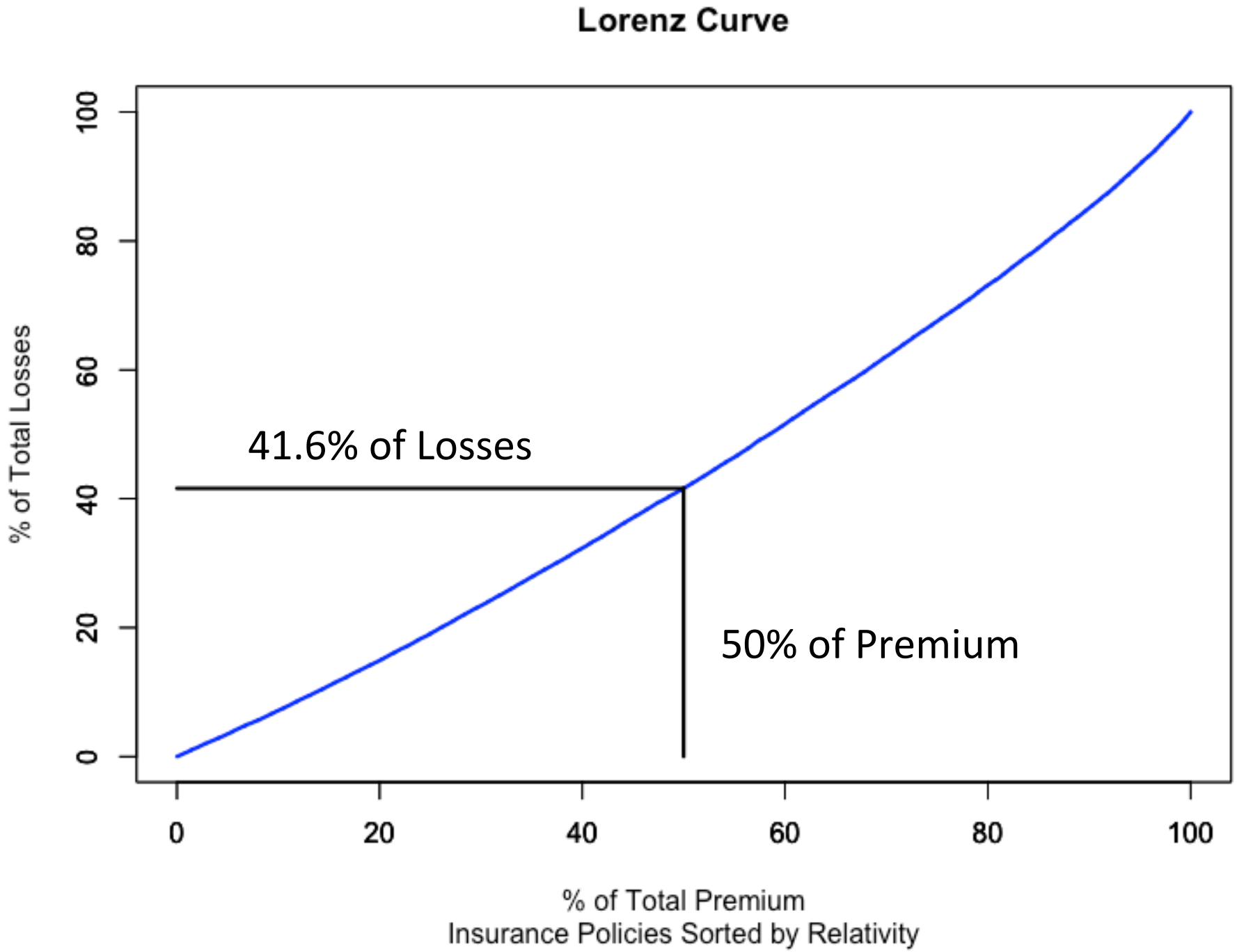
- On the horizontal axis $x_i(r) = \dfrac{\sum_{R_i<r} P_1(x_{1,i})}{\sum_{All\ i} P_1(x_{1,i})}$

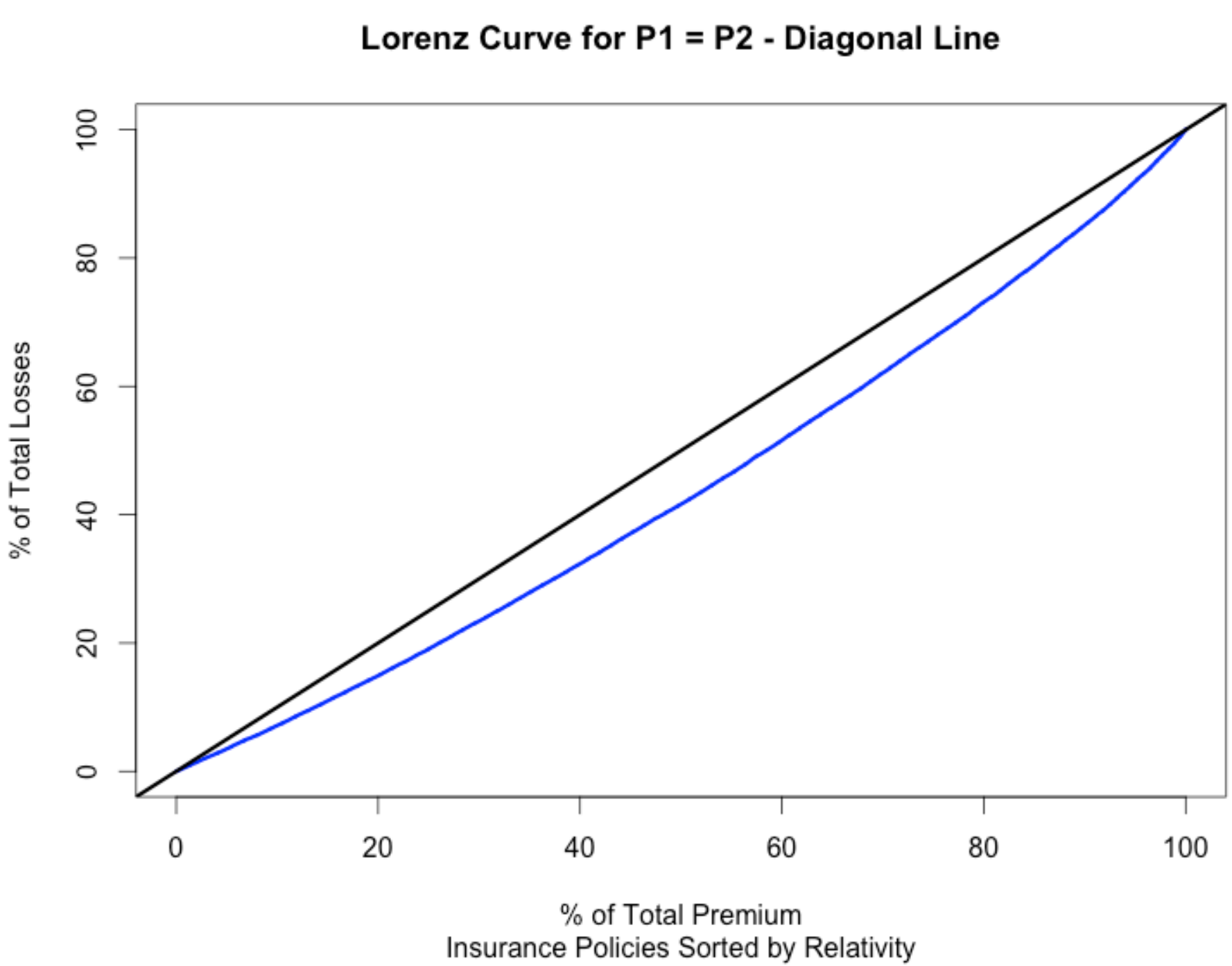
- On the vertical axis $y_i(r) = \dfrac{\sum_{R_i<r} Loss_i}{\sum_{All\ i} Loss_i}$

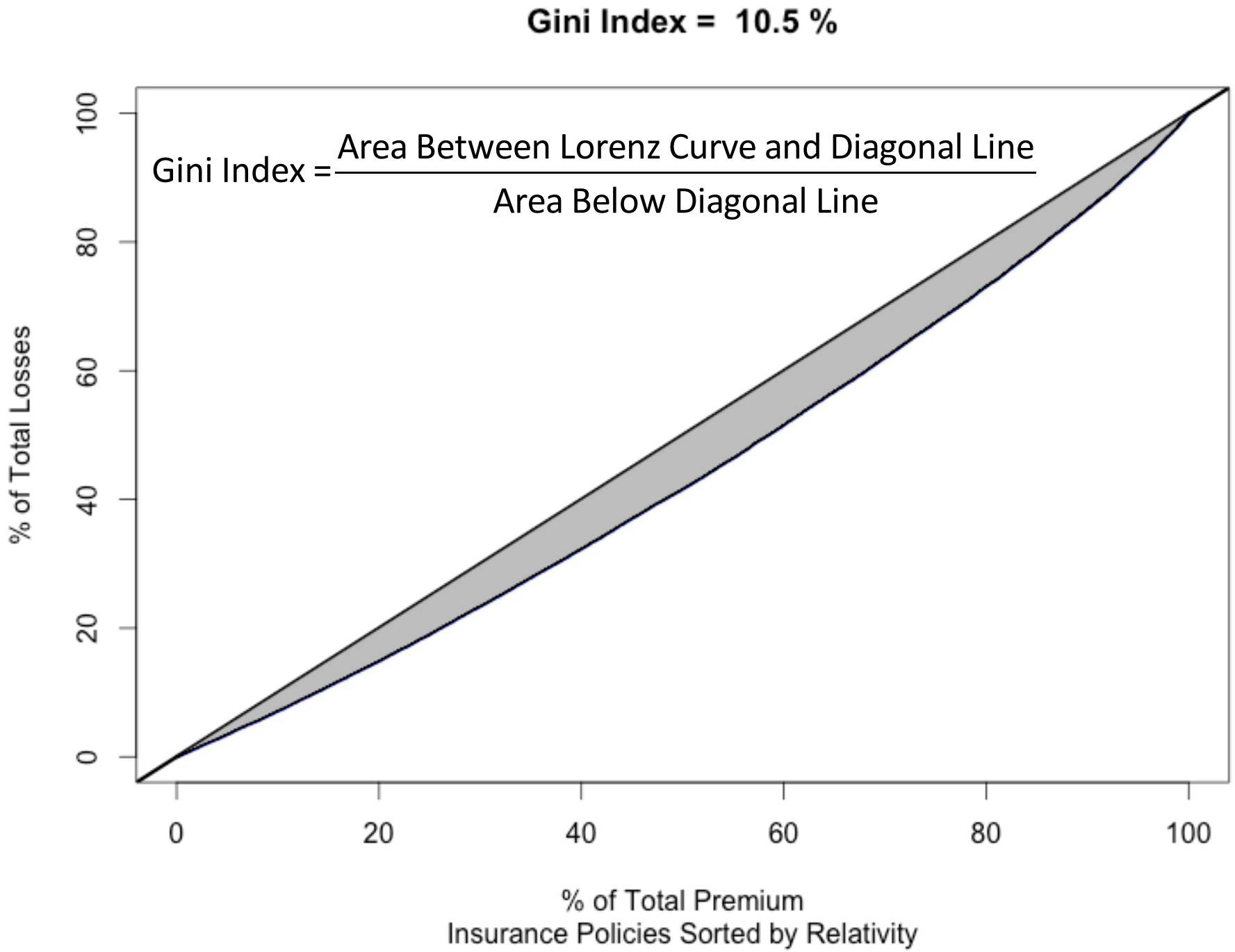
- The curve connecting all the (x_i, y_i) is called the Ordered (by Relativity) Lorenz Curve

Properties of the Ordered Lorenz Curve

- If $P_1(x_1) = P_2(x_1,x_2)$
  - x_2 adds no information to the premium calculation
  - The Lorenz curve is a straight diagonal line connecting (0,0) and (1,1)
- If $P_1(x_1) \neq P_2(x_1,x_2)$
  - Lorenz curve lies beneath the diagonal line and passes though (0,0) and (1,1)
  - The Lorenz curve is concave up
  - We have rigorous proofs of these statements.


Lorenz Curve
% of Total Losses
0
20
40
60
80
100
41.6% of Losses
50% of Premium
0
20
40
60
80
100
% of Total Premium
Insurance Policies Sorted by Relativity


Lorenz Curve for P1 = P2 - Diagonal Line
% of Total Losses
0
20
40
60
80
100
0
20
40
60
80
100
% of Total Premium
Insurance Policies Sorted by Relativity


Gini Index = 10.5 %
Gini Index = Area Between Lorenz Curve and Diagonal Line / Area Below Diagonal Line
% of Total Losses
0
20
40
60
80
100
0
20
40
60
80
100
% of Total Premium
Insurance Policies Sorted by Relativity

Statistical Inference and the Gini Index

- The Gini index is a statistic that depends upon random losses. As such, it has a distribution.
- Statistical properties will now be addressed by Jed.

The Gini Index in the Real World

- P_1 may not be derived from a regression formula.
  - Subject to competitive and regulatory pressures
  - Reflects considerations not in the data.
  - Could result in a non concave Lorenz curve and other strange behaviors.
- Will be addressed by Dave