Straightforward (yet Novel) Methodology for Inequality: Conditional Lorenz Curves

Duke University
Conference on Social Determinants of Health Disparities
August 2011

Kobi Abayomi[1]

1: Asst. Professor, ISyE, Statistics Group, Georgia Institute of Technology

Motivation

Constrained Sum Data

Inequality as a Measurement

- Partition Inequality
  - Group-wise
  - Contribution-wise

Motivation

Constrained Sum Data

Inequality as a Measurement

- Partition Inequality
  - Group-wise
  - Contribution-wise
- Statistically Specify Inequality
  - As data
  - From some 'Random' process
  - for tests of significant differences

GOAL: Straightforward (Easy) Conditional/Groupwise Estimates of Inequality, with Probability Intervals

Just a little notation

Brief Notation

Brief Notation

- $\mathbf{y} = (y_1, ..., y_N) \leftarrow$ data, y some 'good', $i = 1, ..., N$ people, say.

Brief Notation

- $\mathbf{y} = (y_1, ..., y_N) \leftarrow$ data, y some 'good', $i = 1, ..., N$ people, say.
- $\mathbb{1}_{[y_i \leq y]} \leftarrow$ Indicator function. Say $y = 5$ and $y_1 = 3$, $y_2 = 7$ then $\mathbb{1}_{[y_1 \leq y]} = 1$ but $\mathbb{1}_{[y_2 \leq y]} = 0$

Brief Notation

- $\mathbf{y} = (y_1, ..., y_N) \leftarrow$ data, y some 'good', $i = 1, ..., N$ people, say.
- $\mathbb{1}_{[y_i \leq y]} \leftarrow$ Indicator function. Say $y = 5$ and $y_1 = 3$, $y_2 = 7$ then $\mathbb{1}_{[y_1 \leq y]} = 1$ but $\mathbb{1}_{[y_2 \leq y]} = 0$
- $\sum_{i=1}^{n} apple_i \leftarrow$ add up apples 1 through N.

Just a little notation

Brief Notation

- $\mathbf{y} = (y_1, ..., y_N) \leftarrow$ data, y some 'good', $i = 1, ..., N$ people, say.
- $\mathbb{1}_{[y_i \leq y]} \leftarrow$ Indicator function. Say $y = 5$ and $y_1 = 3$, $y_2 = 7$ then $\mathbb{1}_{[y_1 \leq y]} = 1$ but $\mathbb{1}_{[y_2 \leq y]} = 0$
- $\sum_{i=1}^{n} apple_i \leftarrow$ add up apples 1 through N.
- Empirical distribution function (ecdf)

$$F_Y^n(y) = \sum_{i=1}^{n} \mathbb{1}_{[y_i \leq y]} \tag{1}$$

Brief Notation

- $\mathbf{y} = (y_1, ..., y_N) \leftarrow$ data, y some 'good', $i = 1, ..., N$ people, say.
- $\mathbb{1}_{[y_i \leq y]} \leftarrow$ Indicator function. Say $y = 5$ and $y_1 = 3,\; y_2 = 7$ then $\mathbb{1}_{[y_1 \leq y]} = 1$ but $\mathbb{1}_{[y_2 \leq y]} = 0$
- $\sum_{i=1}^{n} apple_i \leftarrow$ add up apples 1 through N.
- Empirical distribution function (ecdf)

$$F_Y^n(y) = \sum_{i=1}^{n} \mathbb{1}_{[y_i \leq y]} \tag{1}$$

The ecdf in this context is just the proportion of people with a less or equal amount y of the 'good'

US Income Data - ecdf

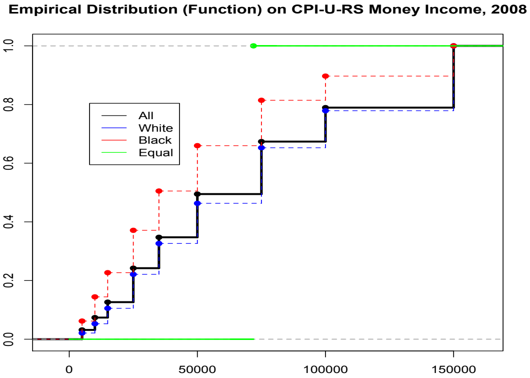


Figure: Graph of empirical cumulative distribution function (ecdf) of Money Income of Households — Consumer Price Index Research Series Using Current Methods, CPI-U-RS

US Income Data - L-curve

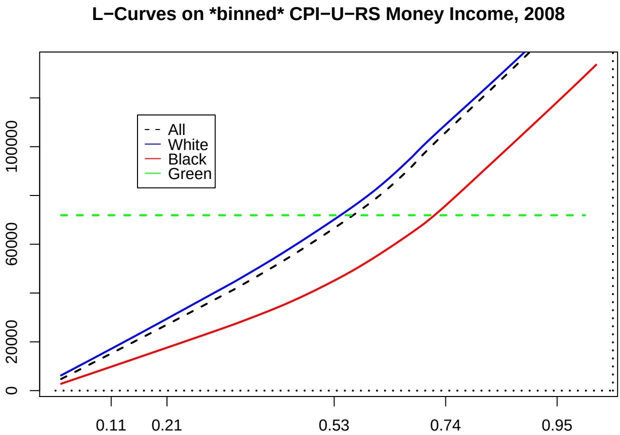


Figure: Illustration of L-curves calculated on US Census CPI-U-RS money income in 2008.

US Income Data - Narrative

The median household net worth for white Americans is $113,149, and for blacks it's $5,677. **That's not a misprint or a misunderstanding; the median white household is 20 times richer than the median black household.**

Figure: Powerful Words

Measuring Inequality

Essentially all functions of ecdf

'Information' based
Theil Index:

$$T = N^{-1} \sum_{i=1}^{N} r_i \; log \; r_i = \sum_{j=1}^{m} \pi_j \; r_j \; log_b \; r_j + \sum_{j=1}^{m} \pi_j \; r_j \; T_j \tag{2}$$

$r_i = y_i / \overline{y}$,
$\pi_j \leftarrow$ relative size of group j,
$T_j \leftarrow$ fix group j.

Measuring Inequality

Essentially all functions of ecdf

'Mean Absolute Deviation'
Gini Index:

$$G = \frac{\binom{n}{2}^{-1}}{2} \sum_{i<j} |y_i - y_j| \tag{3}$$

Measuring Inequality

Measuring Inequality

- Theil often used for within vs. across inequality

Measuring Inequality

- Theil often used for within vs. across inequality
  - Misspecified function of ecdf

Measuring Inequality

- Theil often used for within vs. across inequality
  - Misspecified function of ecdf
  - Log base $\rightarrow$ across and within 'partitions' not directly comparable

Measuring Inequality

- Theil often used for within vs. across inequality
  - Misspecified function of ecdf
  - Log base $\rightarrow$ across and within ‘partitions’ not directly comparable
  - Range of index dependent upon total group size, partitioned group sizes...

Measuring Inequality

- Theil often used for within vs. across inequality
  - Misspecified function of ecdf
  - Log base $\rightarrow$ across and within 'partitions' not directly comparable
  - Range of index dependent upon total group size, partitioned group sizes...
  - ...there are ways to correct [2]
- Gini is popular but...

Measuring Inequality

- Theil often used for within vs. across inequality
  - Misspecified function of ecdf
  - Log base $\rightarrow$ across and within ‘partitions’ not directly comparable
  - Range of index dependent upon total group size, partitioned group sizes...
  - ...there are ways to correct [2]
- Gini is popular but...
  - ...not immediately apparent how to partition it, though

Measuring Inequality

- Theil often used for within vs. across inequality
  - Misspecified function of ecdf
  - Log base $\rightarrow$ across and within 'partitions' not directly comparable
  - Range of index dependent upon total group size, partitioned group sizes...
  - ...there are ways to correct [2]
- Gini is popular but...
  - ...not immediately apparent how to partition it, though
  - desirably scaled between 0 and 1

Measuring Inequality

- Theil often used for within vs. across inequality
  - Misspecified function of ecdf
  - Log base $\rightarrow$ across and within 'partitions' not directly comparable
  - Range of index dependent upon total group size, partitioned group sizes...
  - ...there are ways to correct [2]
- Gini is popular but...
  - ...not immediately apparent how to partition it, though
  - desirably scaled between 0 and 1
  - properly a function of ecdf

Lorenz Curve

The beautiful Lorenz Curve

The Lorenz curve is just a list of population proportions — numbers between 0 and 1 — joined to the list of 'good' proportions,

Lorenz Curve

The beautiful Lorenz Curve

The Lorenz curve is just a list of population proportions — numbers between 0 and 1 — joined to the list of 'good' proportions,

$$L(p) = (N \cdot \overline{y})^{-1} \sum_{i=1}^{\lfloor Np \rfloor} y_{(i)} \tag{4}$$

Lorenz Curve

The beautiful Lorenz Curve

The Lorenz curve is just a list of population proportions — numbers between 0 and 1 — joined to the list of 'good' proportions,

$$L(p) = (N \cdot \overline{y})^{-1} \sum_{i=1}^{\lfloor Np \rfloor} y_{(i)} \tag{4}$$

also numbers between 0 and 1.

Brief Notation

Just a little more notation

Brief Notation

- $\overline{y} \leftarrow$ the observed mean of the 'good'

Just a little more notation

Brief Notation

- $\overline{y} \leftarrow$ the observed mean of the 'good'
- $\mathbf{y}_{()} = (y_{(1)}, ..., y_{(N)}) \leftarrow$ the sorted list of 'goods'

Brief Notation

- $\overline{y} \leftarrow$ the observed mean of the 'good'
- $\mathbf{y}_{()} = (y_{(1)}, ..., y_{(N)}) \leftarrow$ the sorted list of 'goods'
- $F_N^{-1}(p) \leftarrow$ the observed *pth* quantile, the quantity of the 'good' that p% of the people have less than (or equal to).

Brief Notation

- $\overline{y} \leftarrow$ the observed mean of the 'good'
- $\mathbf{y}_{()} = (y_{(1)}, ..., y_{(N)}) \leftarrow$ the sorted list of 'goods'
- $F_N^{-1}(p) \leftarrow$ the observed *pth* quantile, the quantity of the 'good' that $p\%$ of the people have less than (or equal to).
- Lorenz Curve

$$L(p) = (N \cdot \overline{x})^{-1} \sum_{i=1}^{\lfloor Np \rfloor} F_N^{-1}(i/N) \tag{5}$$

Brief Notation

- $\overline{y} \leftarrow$ the observed mean of the 'good'
- $\mathbf{y}_{()} = (y_{(1)}, ..., y_{(N)}) \leftarrow$ the sorted list of 'goods'
- $F_N^{-1}(p) \leftarrow$ the observed *pth* quantile, the quantity of the 'good' that p% of the people have less than (or equal to).
- Lorenz Curve

$$L(p) = (N \cdot \overline{x})^{-1} \sum_{i=1}^{\lfloor Np \rfloor} F_N^{-1}(i/N) \tag{5}$$

The Lorenz curve is just the sorted, cumulative list of 'good' shares by population proportion.

Lorenz $\rightarrow$ Gini

The Gini coefficient is a function of the Lorenz curve...

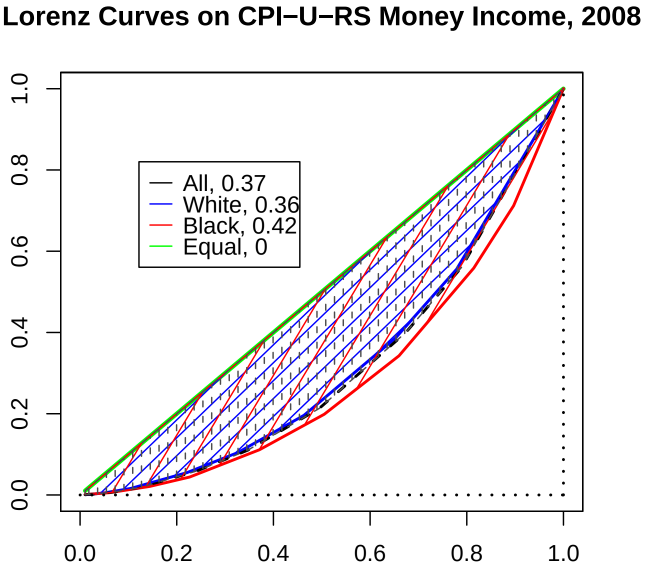
The Gini coefficient is a function of the Lorenz curve...

$$G = \frac{\frac{1}{2} - \sum_{p=1/N}^{N} \frac{1}{N} L(p)}{1/2} = 1 - 2\,\frac{1}{N} \sum_{p=1/N}^{N} L(p) \tag{6}$$

The Gini coefficient is a function of the Lorenz curve...

$$G = \frac{\frac{1}{2} - \sum_{p=1/N}^{N} \frac{1}{N} L(p)}{1/2} = 1 - 2\,\frac{1}{N} \sum_{p=1/N}^{N} L(p) \tag{6}$$

...the scaled difference between the area under the observed Lorenz and equality


Lorenz Curves on CPI−U−RS Money Income, 2008
All, 0.37
White, 0.36
Black, 0.42
Equal, 0
0.0
0.2
0.4
0.6
0.8
1.0

The trick is to see covariates as 'conditional information'

Aaberge et al [1] define pseudo-Lorenz regression curve as a function, in the presence of covariates $\mathbf{x}$ for y, such that

$$E[\Lambda(p|\mathbf{x})] = L(p) \tag{7}$$

The trick is to see covariates as 'conditional information'

Aaberge et al [1] define pseudo-Lorenz regression curve as a function, in the presence of covariates $\mathbf{x}$ for y, such that

$$E[\Lambda(p|\mathbf{x})] = L(p) \tag{7}$$

e.g. that the conditional curves should 'sum' to the original curve

Conditional Lorenz Curve

This is just the law of iterated expectation...

This is just the law of iterated expectation...

for discrete, i.e. categorical, covariates, this is easy

This is just the law of iterated expectation...

for discrete, i.e. categorical, covariates, this is easy

$$L(p) = \sum_{j=1}^{m} \pi_j \, \Lambda(p|\mathbf{x} \in C_j) \tag{8}$$

This is just the law of iterated expectation...

for discrete, i.e. categorical, covariates, this is easy

$$L(p) = \sum_{j=1}^{m} \pi_j \, \Lambda(p|\mathbf{x} \in C_j) \tag{8}$$

and setting

$$\Lambda(p|C_j) = \frac{\overline{y}_j}{\overline{y}} \cdot n_j \, L(F_j(F^{-1}(p))|C_j) \tag{9}$$

guarantees that the overall Lorenz curve will be the weighted sum of conditional 'pseudo'-Lorenz curves.

Just a little more notation

More Brief Notation

More Brief Notation

- $\pi_j = \frac{\overline{y}_j}{\overline{y}} \cdot n_j \leftarrow$ the proportional size of group j

More Brief Notation

- $\pi_j = \frac{\overline{y}_j}{\overline{y}} \cdot n_j \leftarrow$ the proportional size of group j
- p the proportion of the population

More Brief Notation

- $\pi_j = \frac{\overline{y}_j}{\overline{y}} \cdot n_j \leftarrow$ the proportional size of group j
- p the proportion of the population
- $F_N^{-1}(p) \leftarrow$ the observed *pth* quantile *of overall* $\mathbf{y}$

Just a little more notation

More Brief Notation

- $\pi_j = \frac{\overline{y}_j}{\overline{y}} \cdot n_j \leftarrow$ the proportional size of group j
- p the proportion of the population
- $F_N^{-1}(p) \leftarrow$ the observed *pth* quantile *of overall* $\mathbf{y}$
- $F_j(F^{-1}(p)) \leftarrow$ the observed proportion of population in group j at the *pth* quantile of the *overall* distribution

More Brief Notation

- $\pi_j = \frac{\overline{y}_j}{\overline{y}} \cdot n_j \leftarrow$ the proportional size of group j
- p the proportion of the population
- $F_N^{-1}(p) \leftarrow$ the observed *pth* quantile *of overall* $\mathbf{y}$
- $F_j(F^{-1}(p)) \leftarrow$ the observed proportion of population in group j at the *pth* quantile of the *overall* distribution
- $L(F_j(F^{-1}(p))|C_j) \leftarrow$ the Lorenz curve of group j on the observed proportion of population in group j at the *pth* quantile of the *overall* distribution

In Layman's terms...

A simple algorithm

1. Sort all the data; Generate the *pth* quantiles of the unconditioned distribution.$\rightarrow F_N, F^{-1}(p)$

A simple algorithm

1. Sort all the data; Generate the *pth* quantiles of the unconditioned distribution.$\rightarrow F_N, F^{-1}(p)$
2. Sort the data within each group; Generate the ecdf for each group (conditional distribution) at the *pth* quantiles, *of the original distribution*.$\rightarrow F_j(F^{-1}(p))$

A simple algorithm

1. Sort all the data; Generate the *pth* quantiles of the unconditioned distribution.$\rightarrow F_N, F^{-1}(p)$
2. Sort the data within each group; Generate the ecdf for each group (conditional distribution) at the *pth* quantiles, *of the original distribution*.$\rightarrow F_j(F^{-1}(p))$
3. Join the *pth* proportions for each group F_j with the cumulative proportion of income at each group.$\rightarrow L(F_j(F^{-1}(p))|C_j)$

A simple algorithm

1. Sort all the data; Generate the *pth* quantiles of the unconditioned distribution.$\rightarrow F_N, F^{-1}(p)$
2. Sort the data within each group; Generate the ecdf for each group (conditional distribution) at the *pth* quantiles, *of the original distribution*.$\rightarrow F_j(F^{-1}(p))$
3. Join the *pth* proportions for each group F_j with the cumulative proportion of income at each group.$\rightarrow L(F_j(F^{-1}(p))|C_j)$
4. Compute the contribution to the overall Lorenz curve, at each *pth* proportion.$\rightarrow \frac{\overline{y}_j}{\overline{y}} \cdot n_j \; L(F_j(F^{-1}(p))|C_j)$

A simple algorithm

1. Sort all the data; Generate the *pth* quantiles of the unconditioned distribution.$\rightarrow F_N, F^{-1}(p)$
2. Sort the data within each group; Generate the ecdf for each group (conditional distribution) at the *pth* quantiles, *of the original distribution*.$\rightarrow F_j(F^{-1}(p))$
3. Join the *pth* proportions for each group F_j with the cumulative proportion of income at each group.$\rightarrow L(F_j(F^{-1}(p))|C_j)$
4. Compute the contribution to the overall Lorenz curve, at each *pth* proportion.$\rightarrow \frac{\overline{y}_j}{\overline{y}} \cdot n_j \; L(F_j(F^{-1}(p))|C_j)$

Example

Consider this data

```
g1<-c(1,5,5,1) g2<-c(3,3,3,3) g3<-c(1,1,1,9)
```

Example

Consider this data

```
g1<-c(1,5,5,1) g2<-c(3,3,3,3) g3<-c(1,1,1,9)
```

Sort all the data

```
sort(c(g1,g2,g3))
```

Example

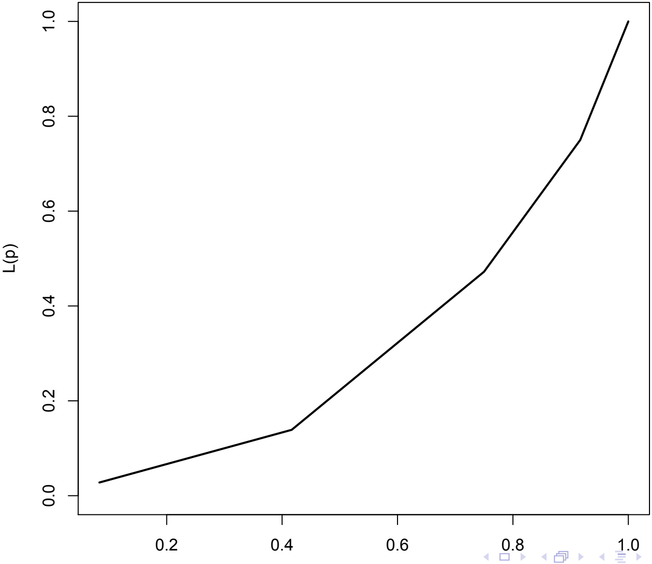
Consider this data

```
g1<-c(1,5,5,1) g2<-c(3,3,3,3) g3<-c(1,1,1,9)
```

Sort all the data

```
sort(c(g1,g2,g3))
1 1 1 1 1 3 3 3 3 5 5 9
```

Simple Example



Example

Illustrate the conditional lorenz curves for each group

```
lnew1<-lorenz(g1); lnew2<-lorenz(g2); lnew3<-lorenz(g3)
```

Example

Illustrate the conditional lorenz curves for each group

`lnew1<-lorenz(g1); lnew2<-lorenz(g2); lnew3<-lorenz(g3)`

The function to compute the lorenz curve is sooooo easy

Conditional Lorenz Curve

Example

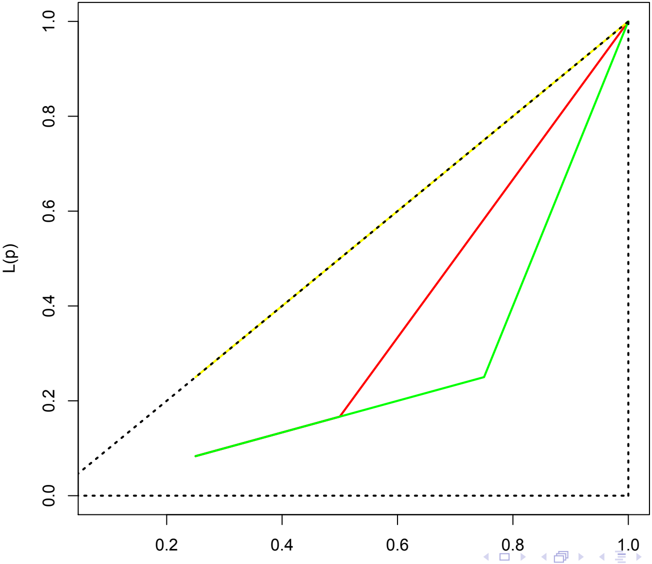
Illustrate the conditional lorenz curves for each group

```
lnew1<-lorenz(g1); lnew2<-lorenz(g2); lnew3<-lorenz(g3)
```

The function to compute the lorenz curve is sooooo easy

```
lorenz function(x)
y<-sort(x)
m<-mean(y); s<-sum(y)
l<-cumsum(y)/s
l
```

Simple Example



Example

Generally the ‘resolution’ can be set ‘arbitrarily’. (but it’s easy to set it at fewest group)

Example

Generally the ‘resolution’ can be set ‘arbitrarily’. (but it’s easy to set it at fewest group)

```
> lresg [1] 4
```

Example

Generally the ‘resolution’ can be set ‘arbitrarily’. (but it’s easy to set it at fewest group)

```
> lresg [1] 4
```

And compute the multipliers for each of the groups

Example

Generally the ‘resolution’ can be set ‘arbitrarily’. (but it’s easy to set it at fewest group)

```
> lresg [1] 4
```

And compute the multipliers for each of the groups

```
meanratiosg<-c(mg1,mg2,mg3)/mgall
[1] 0.6859177 0.8883197 5.2228916 0.8163842
```

Example

Generally the ‘resolution’ can be set ‘arbitrarily’. (but it’s easy to set it at fewest group)

```
> lresg [1] 4
```

And compute the multipliers for each of the groups

```
meanratiosg<-c(mg1,mg2,mg3)/mgall
[1] 0.6859177 0.8883197 5.2228916 0.8163842
groupsizesg<-c(4,4,4)/12 [1] 0.3333333 0.3333333 0.3333333
```

Example

Essentially the contribution to the overall lorenz curve is calculated pointwise

Conditional Lorenz Curve

Example

Essentially the contribution to the overall lorenz curve is calculated pointwise

```
for(pg in uppsg)
lpg<-c(lnew1[pg],lnew2[pg],lnew3[pg])
conditionallorenzedg[pg]<-
as.double(sum(meanratiosg*lpg*groupsizesg))
conditionallorenzedg
[1] 0.1388889 0.2777778 0.5277778 1.0000000
```

Example

Essentially the contribution to the overall lorenz curve is calculated pointwise

```
for(pg in uppsg)
lpg<-c(lnew1[pg],lnew2[pg],lnew3[pg])
conditionallorenzedg[pg]<-
as.double(sum(meanratiosg*lpg*groupsizesg))
conditionallorenzedg
[1] 0.1388889 0.2777778 0.5277778 1.0000000
```

For instance at $p = .50$, the conditional lorenz curves are

```
[1] 0.1666667 0.5000000 0.1666667
```

Example

Essentially the contribution to the overall lorenz curve is calculated pointwise

```
for(pg in uppsg)
lpg<-c(lnew1[pg],lnew2[pg],lnew3[pg])
conditionallorenzedg[pg]<-
as.double(sum(meanratiosg*lpg*groupsizesg))
conditionallorenzedg
[1] 0.1388889 0.2777778 0.5277778 1.0000000
```

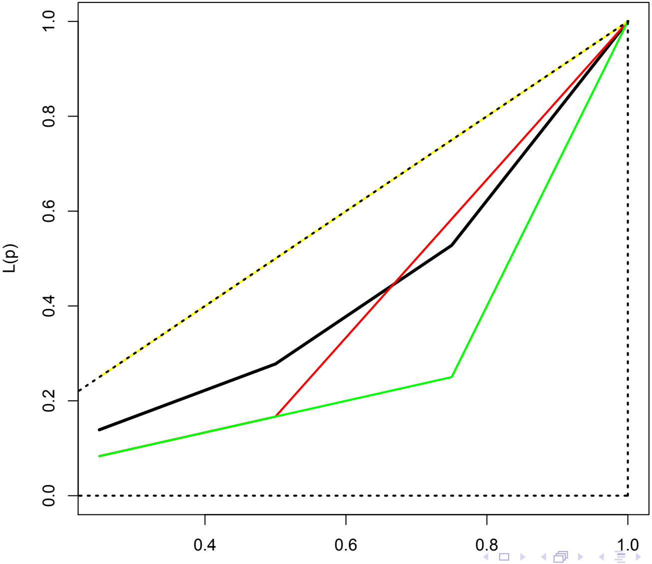
For instance at $p = .50$, the conditional lorenz curves are

```
[1] 0.1666667 0.5000000 0.1666667
```

And their contributions to the overall lorenz curve are

```
[1] 0.3333333 0.3333333 0.3333333
```

Simple Example



Example

And the Gini's are easy to compute

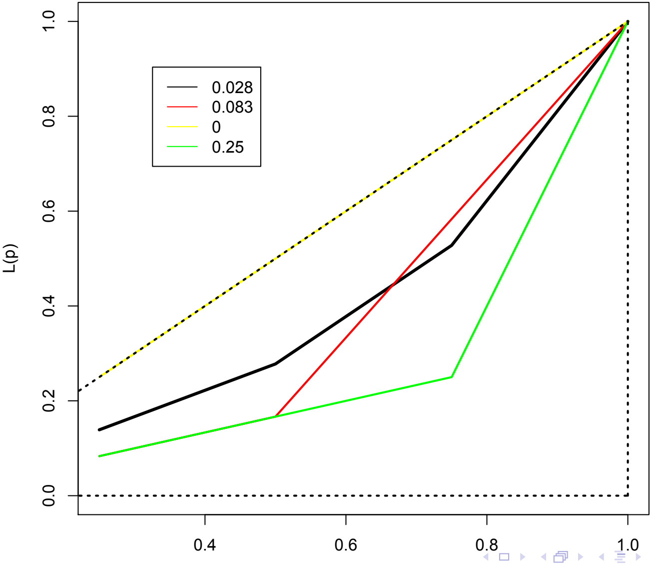
```
ginioverall<-1-2*sum(conditionallorenzedg)/4
```

Example

And the Gini's are easy to compute

```
ginioverall<-1-2*sum(conditionallorenzedg)/4
0.02777778
```

Simple Example



Final (Important) Comments

Effect of group membership on overall inequality

Like in Linear Regression we want effect of covariate (here c_j) on response (here Lorenz/Gini)

Final (Important) Comments

Effect of group membership on overall inequality

Like in Linear Regression we want effect of covariate (here c_j) on response (here Lorenz/Gini)
Mathematically this is

$$\frac{\partial L(p)}{\partial C}\bigg|_{C=c_j}[\sum_{j=1}^{m}\pi_j\,\frac{\overline{y}_j}{\overline{y}}\cdot n_j\;L(F_j(F^{-1}(p))|C_j)] \tag{10}$$

Final (Important) Comments

Effect of group membership on overall inequality

Like in Linear Regression we want effect of covariate (here c_j) on response (here Lorenz/Gini)
Mathematically this is

$$\frac{\partial L(p)}{\partial C}\bigg|_{C=c_j} [\sum_{j=1}^{m} \pi_j \, \frac{\overline{y}_j}{\overline{y}} \cdot n_j \, L(F_j(F^{-1}(p))|C_j)] \tag{10}$$

But if we remember the definition of the derivative, and that the categorical covariate is 'singular', this is just

$$L(p)\bigg|_{C_{-j}} - L(p)\bigg|_{C} \tag{11}$$

Just the difference between the overall (conditionally defined) lorenz curve without and with the *jth* group.

Final (Important) Comments

Statistical significance

We can test for statistical significance using exploiting the duality between the Lorenz curve and the ecdf

Final (Important) Comments

Statistical significance

We can test for statistical significance using exploiting the duality between the Lorenz curve and the ecdf
since

$$F_N(t) \sim N(F(t), F(t)[1 - F(t)]) \tag{12}$$

Final (Important) Comments

Statistical significance

We can test for statistical significance using exploiting the duality between the Lorenz curve and the ecdf
since

$$F_N(t) \sim N(F(t), F(t)[1 - F(t)]) \tag{12}$$

Then

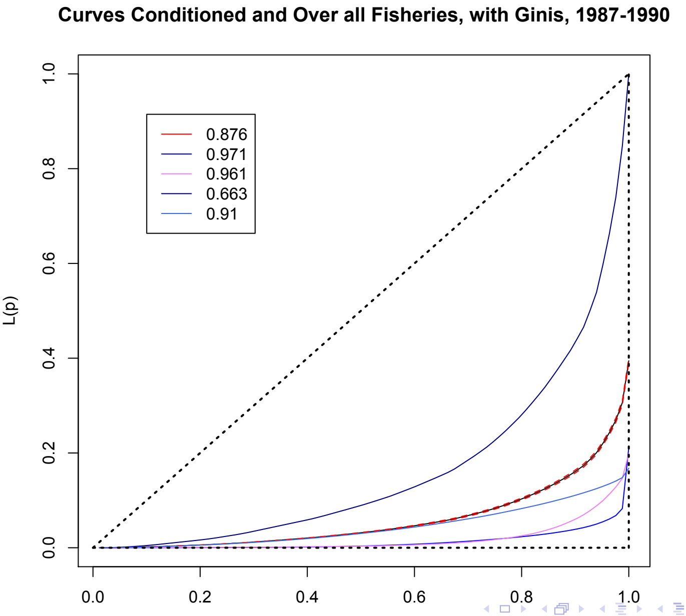
$$L_N(p) \sim N(L(p), \frac{L(p)[1 - L(p)]}{N}) \tag{13}$$

and we can use normal confidence bounds (pointwise), or at least the Kolomorogov-Smirnov (KS) test for differences in distributions to test for significant effects. See [3].

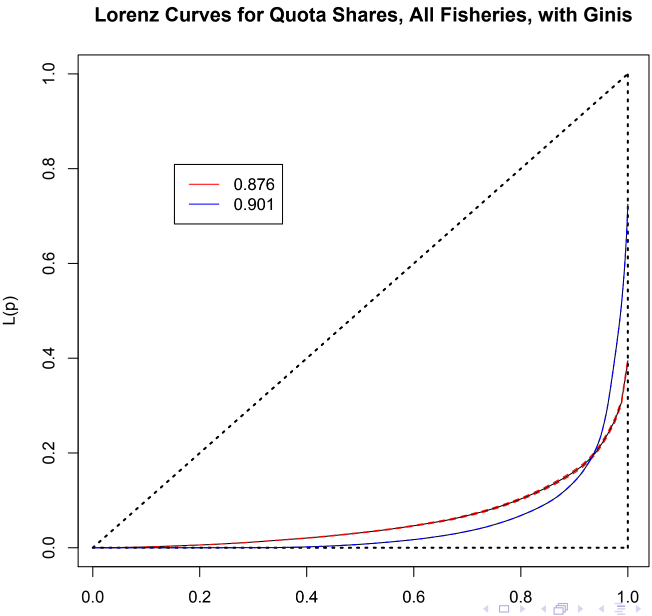
> *We must be careful not to confuse data with the abstractions we use to analyze them.*

-William James

Actually Done



Actually Done



References I

Rolf Aaberge, Steinar Bjerve, and Kjell Doksum.
Decomposition of rank-dependent measures of inequality by subgroups.
Metron - International Journal of Statistics, 63(3):493–503, 2005.

Kobi Abayomi and William Darity Jr.
A friendly amendment to the theil index.
Working paper, 2010.

Kobi Abayomi and Tracy Yandle.
A novel method of measuring consolidation, using conditional lorenz curves to examine itq consolidation in new zealand commercial fishing.
Marine Resources Research, 2011.