THE ADVANTAGES OF USING GROUP MEANS IN ESTIMATING THE LORENZ CURVE AND GINI INDEX FROM GROUPED DATA

MERRITT LYON, LI C. CHEUNG, AND JOSEPH L. GASTWIRTH

1. WEB APPENDIX

1.1. Estimating the Lorenz Curve using the Linear Interpolation Method.

Consider two cases for the integral $\int_0^p \hat{F}^{-1}(t)dt$.

Case 1: $p < \hat{F}_{J-1}$

If $p \in [\hat{F}_{j-1}, \hat{F}_j)$, $j = 1, \dots, J-1$, then

$$\begin{aligned} L(p) = \sum_{i=1}^{j-1} & \left[-\frac{\hat{\alpha}_i}{\hat{\beta}_i} \hat{f}_i + \frac{1}{3\hat{\beta}_i^2} \left([\hat{C}_i + 2\hat{\beta}_i \hat{F}_i]^{3/2} - [\hat{C}_i + 2\hat{\beta}_i \hat{F}_{i-1}]^{3/2} \right) \right] \\ & - \frac{\hat{\alpha}_j}{\hat{\beta}_j} (p - \hat{F}_{j-1}) \\ & + \frac{1}{3\hat{\beta}_j^2} \left([\hat{C}_j + 2\hat{\beta}_j p]^{3/2} - [\hat{C}_j + 2\hat{\beta}_j \hat{F}_{j-1}]^{3/2} \right). \end{aligned}$$

Case 2: $p \geq \hat{F}_{J-1}$

$$\begin{aligned} L(p) = \sum_{j=1}^{J-1} & \left[-\frac{\hat{\alpha}_j}{\hat{\beta}_j} \hat{f}_j + \frac{1}{3\hat{\beta}_j^2} \left([\hat{C}_j + 2\hat{\beta}_j \hat{F}_j]^{3/2} - [\hat{C}_j + 2\hat{\beta}_j \hat{F}_{j-1}]^{3/2} \right) \right] \\ & + a_{J-1} (p - \hat{F}_{J-1}) \\ & - \hat{\lambda} \left[\hat{\eta} \left(1 - \frac{p - \hat{F}_{J-1}}{\hat{\eta}} \right) - \hat{\eta} \left(1 - \frac{p - \hat{F}_{J-1}}{\hat{\eta}} \right) \ln \left(1 - \frac{p - \hat{F}_{J-1}}{\hat{\eta}} \right) \right]. \end{aligned}$$

1.2. Estimating the Gini Index using the Linear Interpolation Method.

Proof. Details of calculating the integrals $\hat{I}_j = \int_{a_{j-1}}^{a_j} x \hat{F}(x) d\hat{F}(x)$ to obtain the estimate of the Gini index using the linear interpolation method.

For each finite interval $j < J$,

$$\begin{aligned}
\hat{I}_j &= \int_{a_{j-1}}^{a_j} x \hat{F}(x) d\hat{F}(x) \\
&= \int_{a_{j-1}}^{a_j} x \left\{ \hat{F}_{j-1} + \hat{\alpha}_j(x - a_{j-1}) + \frac{\hat{\beta}_j}{2} [x^2 - (a_{j-1})^2] \right\} (\hat{\alpha}_j + \hat{\beta}_j x) dx \\
&= \frac{\hat{\alpha}_j}{2} [(a_j)^2 - (a_{j-1})^2] \left[\hat{F}_{j-1} - \hat{\alpha}_j a_{j-1} - \frac{\hat{\beta}_j (a_{j-1})^2}{2} \right] \\
&\quad + \frac{1}{3} [(a_j)^3 - (a_{j-1})^3] \left[\hat{F}_{j-1} \hat{\beta}_j + \hat{\alpha}_j^2 - \hat{\alpha}_j \hat{\beta}_j a_{j-1} - \frac{\hat{\beta}_j^2 (a_{j-1})^2}{2} \right] \\
&\quad + \frac{3\hat{\alpha}_j \hat{\beta}_j}{8} [(a_j)^4 - (a_{j-1})^4] + \frac{\hat{\beta}_j^2}{10} [(a_j)^5 - (a_{j-1})^5].
\end{aligned}$$

For the final unbounded interval,

$$\begin{aligned}
\hat{I}_J &= \int_{a_{J-1}}^{\infty} x \hat{F}(x) d\hat{F}(x) \\
&= \int_{a_{J-1}}^{\infty} x \left[\hat{F}_{J-1} + \hat{\eta} \left(1 - e^{-\frac{x-a_{J-1}}{\hat{\lambda}}} \right) \right] \frac{\hat{\eta}}{\hat{\lambda}} e^{-\frac{x-a_{J-1}}{\hat{\lambda}}} dx \\
&= \hat{f}_J \bar{x}_J - \frac{\hat{f}_J^2}{4} (\bar{x}_J + a_{J-1}).
\end{aligned}$$

The estimated mean difference is thus

$$\hat{\Delta} = 4 \left[\sum_{j=1}^{J-1} \hat{I}_j + \hat{f}_J \bar{x}_J - \frac{\hat{f}_J^2}{4} (\bar{x}_J + a_{J-1}) \right] - 2\bar{x}.$$

□

1.3. Proof of Lemma 2.

Proof. $\widehat{GL}_r \leq \widehat{GL}_{r+1}$; $\widehat{GU}_r \geq \widehat{GU}_{r+1}$

First consider the lower bound $\widehat{GL}_r$. Without loss of generality, examine a partition $[a_{j-1}, a_j)$ and assume the group is split by a break point x^* with $a_{j-1} < x^* < a_j$ and corresponding quantiles $p_j^- < p^* < p_j^+$.

The lower bound for the Gini index is obtained from the upper bound for the Lorenz curve L_r . For the original partition $[a_{j-1}, a_j)$ for $p_j^- < p < p_j^+$ the upper bound of the Lorenz curve is the chord above the Lorenz curve connecting $L(p_j^-)$ and $L(p_j^+)$:

$$L_r^U(p) = L(p_j^-) + \frac{p - p_j^-}{p_j^+ - p_j^-} [L(p_j^+) - L(p_j^-)].$$

For the new partition with x^* there are two chords, one connecting $L(p_j^-)$ and $L(p^*)$:

$$L_{r+1}^U(p) = L(p_j^-) + \frac{p - p_j^-}{p^* - p_j^-} [L(p^*) - L(p_j^-)]$$

and another connecting $L(p^*)$ and $L(p_j^+)$:

$$L_{r+1}^U(p) = L(p^*) + \frac{p - p^*}{p_j^+ - p^*} [L(p_j^+) - L(p^*)].$$

It will be shown that the new bounds lie within the bounds of the original. The quantile p^* can be expressed as a convex combination of p_j^- and p_j^+ for some $t \in [0, 1]$, and by the convexity of the Lorenz curve, it follows that $L(p^*) \leq tL(p_j^-) + (1 - t)L(p_j^+)$. This implies for $p < p^*$:

$$\begin{aligned} L_{r+1}^U(p) &= L(p_j^-) + \frac{p - p_j^-}{p^* - p_j^-} [L(p^*) - L(p_j^-)] \\ &\leq L(p_j^-) + \frac{p - p_j^-}{p^* - p_j^-} [tL(p_j^-) + (1 - t)L(p_j^+) - L(p_j^-)] \\ &= L(p_j^-) + \frac{p - p_j^-}{(1 - t)(p_j^+ - p_j^-)} (1 - t) [L(p_j^+) - L(p_j^-)] \\ &= L_r^U(p). \end{aligned}$$

An analogous argument using the right endpoints for the second chord for $p \geq p^*$ shows $L_{r+1}^U(p) \leq L_r^U(p)$ for $p_j^- < p < p_j^+$. It follows that

$$\int_{p_j^-}^{p_j^+} L_{r+1}^U(p) dp \leq \int_{p_j^-}^{p_j^+} L_r^U(p) dp \iff - \int_{p_j^-}^{p_j^+} L_{r+1}^U(p) dp \geq - \int_{p_j^-}^{p_j^+} L_r^U(p) dp.$$

As $\widehat{GL}_r = 1 - 2 \int_0^1 L_r^U(p) dp$ and $L_r^U(p) = L_{r+1}^U(p)$ outside of $[p_j^-, p_j^+]$, we have

$$\widehat{GL}_r = 1 - 2 \int_0^1 L_r^U(p) dp \leq 1 - 2 \int_0^1 L_{r+1}^U(p) dp = \widehat{GL}_{r+1}.$$

Next, consider the upper bound $\widehat{GU}_r$. Now examine a partition $[a_{j-1}, a_j]$, and assume the group is split by an additional break point x^* with $a_{j-1} < x^* < a_j$ and corresponding quantiles $p_j^- < p^* < p_j^+$.

The upper bound for the Gini index is derived from the lower bound for the Lorenz curve L_r . For the original partition $[a_{j-1}, a_j]$ for $p_j^- < p < p_j^+$ the lower bound of the Lorenz curve is the curve formed by the intersection of two support tangent lines at $L(p_j^-)$ and $L(p_j^+)$, that is

$$L_r^L(p) = L(p_j^-) + \frac{F^{-1}(p_j^-)}{\mu} (p - p_j^-)$$

for $p < \bar{p}$ and

$$L_r^L(p) = L(p_j^+) + \frac{F^{-1}(p_j^+)}{\mu} (p - p_j^+)$$

for $p \geq \bar{p}$ where $\bar{p} \in (0, 1)$ is the quantile at which the tangent lines intersect.

For the new partition with x^* there is an additional support curve at p^* that will intersect $L_r^L(p)$ at p' and $L_r^L(p)$ at p'' . For the new partition of $[p_j^-, p_j^+]$ with $p_j^- < p' < p^* < p'' < p_j^+$ the new bounds for the Lorenz curve are:

$$L_{r+1}^L(p) = L(p_j^-) + \frac{F^{-1}(p_j^-)}{\mu} (p - p_j^-)$$

for $p < p'$,

$$L_{r+1}^L(p) = L(p') + \frac{F^{-1}(p^*)}{\mu} (p - p')$$

for $p' < p < p''$, and

$$L_{r+1}^L(p) = L(p_j^+) + \frac{F^{-1}(p_j^+)}{\mu} (p - p_j^+)$$

for $p > p''$.

It will be shown that the new bound lies above the original. The support curves are identical for $p < p'$ and $p > p''$, so consider $p' < p < p''$. First taking the case

$p' < p < p^*$, it follows that

$$\begin{aligned}
L_{r+1}^L(p) &= L(p') + \frac{F^{-1}(p^*)}{\mu} (p - p') \\
&\geq L(p') + \frac{F^{-1}(p_j^-)}{\mu} (p - p') \\
&= L(p_j^-) + \frac{F^{-1}(p_j^-)}{\mu} (p' - p_j^-) + \frac{F^{-1}(p_j^-)}{\mu} (p - p') \\
&= L(p_j^-) + \frac{F^{-1}(p_j^-)}{\mu} (p - p_j^-) \\
&= L_r^L(p).
\end{aligned}$$

This arises from the fact that F^{-1} is non-decreasing and the support curves intersect at p' . An analogous convexity argument using the right-hand endpoints applied to the second chord for $p^* < p < p''$ shows $L_{r+1}^L(p) \geq L_r^L(p)$ for $p_j^- < p < p_j^+$. It follows that

$$\int_{p_j^-}^{p_j^+} L_{r+1}^L(p) dp \geq \int_{p_j^-}^{p_j^+} L_r^L(p) dp \iff - \int_{p_j^-}^{p_j^+} L_{r+1}^L(p) dp \leq - \int_{p_j^-}^{p_j^+} L_r^L(p) dp.$$

As $\widehat{GU}_r = 1 - 2 \int_0^1 L_r^L(p) dp$ and $L_r^L(p) = L_{r+1}^L(p)$ outside of $[p_j^-, p_j^+]$, we have

$$\widehat{GU}_{r+1} = 1 - 2 \int_0^1 L_r^L(p) dp \leq 1 - 2 \int_0^1 L_{r+1}^L(p) dp = \widehat{GU}_r,$$

as required. $\square$

1.4. Comparison of the Estimates to the Reported Gini index for the Historical Data. Web Appendix Table 1 compares the estimates from the various methods to the reported Gini index in the historical income data, adjusted for inflation, for the 1967-2013 time period. Notice that the estimates derived from the procedures using the group means are closer to the Census Bureau's reported values than the original T&L method in every year and usually are at least as close as our improved T&L method, which fits an exponential distribution to the open-ended interval.

Web Appendix Table 1. Estimated Gini index from historical six income groups
(quintiles and top 5% adjusted to 2013 dollars) from 1967-2013.

Year	Gini	T&L	T&L	SH	SH	Linear	$\frac{1}{3}\widehat{GL}$
	Index	UB=\$500k	Exp. Tail	UB=\$500k	Exp. Tail	Intrpl.	$+\frac{2}{3}\widehat{GU}$
2013	0.481	0.493	0.464	0.475	0.476	0.475	0.476
2012	0.477	0.492	0.462	0.476	0.477	0.476	0.477
2011	0.477	0.492	0.462	0.476	0.477	0.477	0.478
2010	0.470	0.489	0.459	0.469	0.469	0.469	0.470
2009	0.468	0.485	0.456	0.467	0.467	0.467	0.468
2008	0.466	0.483	0.453	0.465	0.465	0.465	0.466
2007	0.463	0.480	0.452	0.462	0.462	0.462	0.463
2006	0.470	0.481	0.454	0.469	0.469	0.469	0.470
2005	0.469	0.481	0.453	0.468	0.469	0.468	0.470
2004	0.466	0.480	0.450	0.466	0.466	0.466	0.467
2003	0.464	0.480	0.451	0.463	0.464	0.464	0.465
2002	0.462	0.478	0.448	0.462	0.462	0.462	0.463
2001	0.466	0.477	0.449	0.465	0.466	0.465	0.467
2000	0.462	0.473	0.444	0.461	0.461	0.461	0.462
1999	0.458	0.475	0.447	0.457	0.457	0.457	0.458
1998	0.456	0.475	0.443	0.456	0.456	0.456	0.457
1997	0.459	0.478	0.443	0.458	0.458	0.458	0.459
1996	0.455	0.479	0.441	0.454	0.454	0.454	0.455
1995	0.450	0.477	0.436	0.450	0.450	0.450	0.451
1994	0.456	0.484	0.443	0.455	0.456	0.455	0.456
1993	0.454	0.483	0.439	0.453	0.454	0.453	0.454
1992	0.433	0.480	0.434	0.434	0.434	0.433	0.434
1991	0.428	0.477	0.430	0.428	0.428	0.428	0.428
1990	0.428	0.474	0.429	0.428	0.428	0.428	0.429
1989	0.431	0.472	0.429	0.431	0.431	0.431	0.432
1988	0.426	0.473	0.427	0.427	0.427	0.427	0.427
1987	0.426	0.474	0.426	0.426	0.426	0.426	0.427
1986	0.425	0.475	0.426	0.425	0.425	0.425	0.426
1985	0.419	0.475	0.421	0.420	0.420	0.420	0.421
1984	0.415	0.477	0.421	0.416	0.415	0.415	0.416
1983	0.414	0.478	0.419	0.415	0.415	0.414	0.415
1982	0.412	0.478	0.418	0.413	0.413	0.413	0.413
1981	0.406	0.475	0.412	0.407	0.407	0.407	0.407
1980	0.403	0.472	0.409	0.404	0.404	0.404	0.404
1979	0.404	0.468	0.408	0.405	0.405	0.405	0.405
1978	0.402	0.467	0.405	0.403	0.403	0.403	0.403
1977	0.402	0.470	0.405	0.404	0.403	0.403	0.404
1976	0.398	0.469	0.401	0.399	0.399	0.399	0.399
1975	0.397	0.470	0.399	0.398	0.398	0.398	0.399
1974	0.395	0.465	0.398	0.396	0.396	0.396	0.396
1973	0.400	0.465	0.402	0.401	0.401	0.400	0.401
1972	0.401	0.467	0.401	0.402	0.402	0.402	0.402
1971	0.396	0.469	0.396	0.397	0.397	0.397	0.397
1970	0.394	0.466	0.394	0.396	0.395	0.395	0.396
1969	0.391	0.462	0.388	0.392	0.392	0.391	0.392

Source: U.S. Census Bureau, Current Population Survey, Annual Social and Economic Supplement: Table H-1, All Races, adjusted to 2013 dollars
<http://www.census.gov/hhes/www/income/data/historical/household/> . The Gini index 0.470 for 2010 in this table incorporates the results of the 2010 Census, so it differs slightly from the original published value of 0.469.

1.5. R Code for computing Gini Index from Grouped Data. The following code produces estimates of the Gini index derived from the following methods:

- (1) The linear interpolation method that uses an exponential tail.
- (2) The modified split histogram using an exponential tail.
- (3) $\frac{1}{3}$ the Gini lower bound + $\frac{2}{3}$ the Gini upper bound.

In addition, the Gini lower and upper bounds are estimated.

```
#### Clear Workspace + Load Packages
rm(list=ls(all=TRUE))
library(reshape)
library(caTools)

#####
#### Note: Input data GR_DATA must have the format below
#####
#### Create Summary Dataset
GR_DATA=data.frame(#### READ YOUR DATA ####)
attach(GR_DATA)
#> GR_DATA
#  GROUP    GLB    GUB    GAVG FREQ
#1     0      0  20900  11651.0 0.20
#2     1  20900  40187  30509.0 0.20
#3     2  40187  65501  52322.0 0.20
#4     3  65501 105910  83519.0 0.20
#5     4 105910 196000 139493.7 0.15
#6     5 196000    NA 322343.0 0.05

#### Check for 1/3 + 2/3 bounds
LCHECK = (2/3)*GLB + (1/3)*GUB
UCHECK = (1/3)*GLB + (2/3)*GUB
GOOD= (LCHECK<=GAVG & GAVG<=UCHECK)
if (sum(!GOOD,na.rm=TRUE)>0)
{
  cat("CAUTION: there exists a mean outside of the [1/3,2/3] bound: \n
      USE SPLIT HISTOGRAM! \n")
}

##### Average Income #####
AVE_INCOME_MEAN=sum(GAVG*FREQ)

##### Gini Lower and Upper bounds #####
#calculate Gini index lower bound
```

```

FCROSS=outer(FREQ,FREQ,'*')
DCROSS=abs(outer(GAVG,GAVG,"-"))
G_LB=sum(FCROSS*DCROSS)/(2*AVE_INCOME_MEAN)

# Calculate Gini Index Upper Bound - Gastwirth (1972) page 310 (18 + 19)
GLEN <- GUB-GLB
n_bndgrps = nrow(GR_DATA)-1
G_UB_LAST = FREQ[nrow(GR_DATA)]^2*(GAVG[nrow(GR_DATA)]-GLB[nrow(GR_DATA)])
G_UB = G_LB + (sum(FREQ[1:n_bndgrps]^2*(GUB[1:n_bndgrps]-GAVG[1:n_bndgrps])*
(GAVG[1:n_bndgrps]-GLB[1:n_bndgrps])/GLEN[1:n_bndgrps])+G_UB_LAST)/AVE_INCOME_MEAN

#####
##### LINEAR DENSITY coefficients
#####
D1=GUB-GLB
D2=GUB^2-GLB^2
D3=GUB^3-GLB^3
AB=matrix(rep(0,2*nrow(GR_DATA)),nrow=nrow(GR_DATA))
##### A/B coefficients f(x)=a+bx

for (i in 1:nrow(GR_DATA)){
  M=matrix(c(2*D1[i],D2[i],3*D2[i],2*D3[i]),nrow=2,byrow=TRUE)
  y=matrix(c(2*FREQ[i],6*GAVG[i]*FREQ[i]),nrow=2)
  AB[i,]=solve(M,y)
}

#### If using Exp Tail
AB[nrow(GR_DATA),]=c(FREQ[nrow(GR_DATA)],GAVG[nrow(GR_DATA)]-GLB[nrow(GR_DATA)])
##### Add names
rownames(AB)=GROUP
colnames(AB)=c("a","b")

#####
##### Split Histogram with Exp. Tail
#####
#### Split Histogram Breaks
GR_DATA$SP = cumsum(c(0,GR_DATA$FREQ[-nrow(GR_DATA)]))+GR_DATA$FREQ*
(GR_DATA$GUB-GR_DATA$GAVG)/(GR_DATA$GUB-GR_DATA$GLB)

SH=matrix(rep(0,2*nrow(GR_DATA)),nrow=nrow(GR_DATA))

for (i in 1:( nrow(GR_DATA) - 1)){
  SH[i,1]= ( FREQ[i] / GLEN[i] ) * (GUB[i] - GAVG[i]) / (GAVG[i] - GLB[i])
  SH[i,2]= ( FREQ[i] / GLEN[i] ) * (GAVG[i] - GLB[i]) / (GUB[i] - GAVG[i])
}
E=nrow(GR_DATA)
SH[E,]=AB[E,]
rownames(SH)=GROUP

```



```
#####
#### Quantile Functions
#####
F=c(0,cumsum(FREQ));
SPLIT = GR_DATA$SP
alpha=seq(.000001,.9999999,length=30000)
alpha_lin=rep(0,length(alpha))
alpha_SH=rep(0,length(alpha))

for (i in 1:length(alpha))
{
  findex=match(TRUE,F>alpha[i])
  if ( (findex-1)<nrow(GR_DATA) )
  {
    #### Lin Den
    C = alpha[i]- F[findex-1]+AB[findex-1,1]*GLB[findex-1]+(AB[findex-1,2]/2)*
      GLB[findex-1]^2
    Root=polyroot(c(-C,AB[findex-1,1],AB[findex-1,2]/2))
    alpha_lin[i]=Re( Root[Re(Root)<=GUB[findex-1] & Re(Root)>=GLB[findex-1]] )
  }
  else #### Final Piece
  {
    alpha_lin[i]= GLB[findex-1]-AB[findex-1,2]*log(1-(alpha[i]-F[findex-1])/
      AB[findex-1,1])
  }

  ##### S&H
  if ( (findex-1)< (nrow(GR_DATA)) )
  {
    if (alpha[i] < SPLIT[findex-1])
    {
      sprop = (SPLIT[findex - 1] - F[findex - 1])
      alpha_SH[i]=GLB[findex-1]+(GAVG[findex-1]-GLB[findex-1])*
        (alpha[i]-F[findex-1])/sprop
    }
    else
    {
      sprop = (F[findex ] - SPLIT[findex - 1] )
      alpha_SH[i]=GAVG[findex-1] + (GUB[findex-1]-GAVG[findex-1])*
        (alpha[i] - SPLIT[findex - 1])/sprop
    }
  }
  else #### Final Piece
  {
    alpha_SH[i]= GLB[findex-1] - SH[findex-1,2]*log(1-(alpha[i]-F[findex-1])/
      SH[findex-1,1])
  }
}
}
```

```
#####
### Lorenz CURVES
#####
L_Lin=rep(0,length(alpha));
L_SH=L_Lin;

for (i in 1:length(alpha))
{
  L_Lin[i]=trapz(alpha[1:i],alpha_lin[1:i])/AVE_INCOME_MEAN;
  L_SH[i]=trapz(alpha[1:i],alpha_SH[1:i])/AVE_INCOME_MEAN;
}

#####
### Calculate Ginis
#####
Int_L = trapz(alpha[-1],L_Lin[-1])
G_Lin=1-2*Int_L   ### Linear Interpolation estimate of the Gini Index
Int_SH = trapz(alpha[-1],L_SH[-1])
G_SH=1-2*Int_SH   ### Split histogram estimate of the Gini Index
G_BND_EST = G_LB/3+2*G_UB/3   ### Gini lower bound + 2/3 Gini upper bound

cat("Linear Interpolation: ",G_Lin,"\n")
cat("Split Histogram: ",G_SH,"\n")
cat("1/3 GL + 2/3 GU :",G_BND_EST,"\n")
cat("Gini Lower Bound (GL) :",G_LB,"\n")
cat("Gini Upper Bound (GU) :",G_UB,"\n")
```