

## Bootstrap standard errors for indices of inequality: INEQERR

Dean Jolliffe, Center for Economic Research and Graduate Education, Czech Republic,  
dean.jolliffe@cerge.cuni.cz

Bohdan Krushelnyskyy, Center for Economic Research and Graduate Education, Czech Republic,  
bohdan.krushelnyskyy@cerge.cuni.cz

### Description

This insert provides a program for estimating three indices of inequality – the Gini, Theil, and Variance of Logs – and bootstrap estimates of their sampling variances. The program offers three variations of the bootstrap variance estimates. The first is the standard bootstrap which assumes that the sample was selected using a simple random design. The second is a bootstrap estimate which assumes that the sample was selected in two-stages, both stages being simple random draws. The third bootstrap estimate replicates a fairly standard sample design for household survey data in which primary sampling units (PSUs) are selected with probability proportional to population (PPP) in the first stage and then in the second stage the ultimate sampling units (USUs) are selected in a simple random draw.

While there are several other programs which provide measures of inequality indices (for example, Jenkins, 1999), there are no Stata ado files which provide estimates of standard errors for the Gini, Theil, and Variance of Logs. It is only with estimates of the sampling variance that one can answer the important policy questions of whether inequality has changed over time or differs over regions. Mills and Zandvakili (1997) provide an interesting example of the importance of testing for the significance of differences in estimated inequality indices. Using data from the Panel Study of Income Dynamics, they show the somewhat troublesome result that the Gini and Theil indices are greater for post-tax income than for pre-tax income. Upon constructing bootstrap standard errors for the indices, they determine that none of the observed differences are statistically significant. In their paper they also show that the bootstrap estimated standard errors for the Gini are similar to the asymptotic estimates.

An advantage of the two-stage bootstrap estimates available in *ineqerr* is that if the sample was collected using a two-stage process, then the estimated standard errors will be robust to this design effect. Kish (1995) and Cochran (1997) show the importance of correcting mean values for design effects. Scott and Holt (1982) show that the magnitude of the bias for the estimated variance-covariance matrix for ordinary least squares estimates can be quite large when it is erroneously assumed that the data were collected using a simple random sample if in fact a two-stage design had been used. While there is no literature we are aware of which directly discusses the assumption of design effects in estimating standard errors for inequality indices, our empirical work suggests that correcting inequality estimates for design effects is also important. (See for example, Datt, Jolliffe and Sharma, 1998.)

### Syntax

**ineqerr** *varlist* [*weight*] [*if exp*] [*in range*] [, **reps**(#) **psu**(*varname*) **psuwt**(*varname* / *expression*) ]

fweights and aweights are allowed.

## Options

**reps**(#) specifies the number of bootstrap replications to be performed. The default value is 100.

**psu**(*varname*) specifies the variable identifying the primary sampling unit. If no variable is specified, then the bootstrap replication is a single-stage, simple random draw on the sample.

**psuwt**(*varname* / *expression*) specifies the weight to be used for the first-stage selection process. If, for example, the population of the primary sampling unit is specified, then the bootstrap replicates a random draw with probability proportional to population. Both *psuwt* and the *weight* options accept either a variable name or an expression, where for example an expression might be the product of two variables. The *psuwt* option can only be used if the primary sampling unit is specified. If no weighting variable is specified for the first-stage but the *psu* is specified, the bootstrap replication is two stages of a simple random draw on the sample.

## Examples

To illustrate the use of *ineqerr*, we use data from the 1997 Egypt Integrated Household Survey (EIHS). The variable *pcexp\_r* is a household-level measure of per capita consumption, which is adjusted to control for spatial price variation. *Wt96ind* is a weighting variable which is the product of household size and strata weights. When we issue the *ineqerr* command, the following results:

```
ineqerr pcexp_r [w=wt96ind]
```

```
pcexp_r ----- Real Per Capita Expenditure  
(obs=2449)
```

```
Bootstrap statistics
```

Variable	Reps	Observed	Bias	Std. Err.	[95% Conf. Interval]		
Gini	100	.3464858	.0006895	.0071348	.3323289	.3606428	(N)
					.3345012	.361028	(P)
					.3327775	.3595716	(BC)
Theil	100	.2232024	.0011754	.0134943	.1964268	.2499781	(N)
					.2036364	.2518677	(P)
					.1975445	.2516475	(BC)
Varlogs	100	.3603729	.0019088	.0128569	.334862	.3858838	(N)
					.3408104	.3913696	(P)
					.3341884	.3833428	(BC)

```
-----  
N = normal, P = percentile, BC = bias corrected
```

In this case, no sample design information is passed to *ineqerr* and the program calls Stata's *bsample* utility to re-sample the data. In order to maintain the same sample size in each bootstrap re-sample, *ineqerr* ignores observations where *pcexp\_r* or *wt96ind* is missing. Since the number of replications is not specified, the default value of 100 is used. The results from *bsample* are then passed to the *bstat* command to generate the standard Stata bootstrap output. For more information about the normal, percentile, and bias-corrected percentile confidence intervals, see *bstrap* in the Stata manuals. For an introduction to the bootstrap principle, see Efron and Tibshirani (1993). In order to reproduce results from *ineqerr* it is necessary to set the random number seed first. (See *generate* in the Stata reference manuals for more information.)

The reported standard errors above will be correct if the sample comes from a simple random

draw. This is not the case with the EIHS data, which was collected using a stratified, two-stage design. *Ineqerr* can generate bootstrap estimates of the standard errors which are robust to the two-stage design by passing the information about the primary sampling unit to *ineqerr*. So, for example, we correct the standard errors above for this aspect of the sample design by issuing the following command. (We now also specify the number of replications to be 50.)

```
ineqerr pcexp_r [w=wt96ind], reps(50) psu(psu)

pcexp_r ----- Real Per Capita Expenditure
(obs=2449)
Bootstrap statistics
```

Variable	Reps	Observed	Bias	Std. Err.	[95% Conf. Interval]		
Gini	50	.3464858	.0005401	.0123159	.3217361	.3712356	(N)
					.3233636	.3683307	(P)
					.3233636	.3732372	(BC)
Theil	50	.2232024	.0017971	.0211976	.1806042	.2658007	(N)
					.1872635	.2664682	(P)
					.1872635	.2664682	(BC)
Varlogs	50	.3603729	.0000892	.0238775	.3123893	.4083565	(N)
					.32027	.4053129	(P)
					.32027	.4121322	(BC)

-----  
N = normal, P = percentile, BC = bias corrected

Note that the point estimates for the inequality indices are unchanged but the estimated standard errors have all increased. If we consider the case of the Gini coefficient, the standard error increases by 73 percent when we correct for the two-stage nature of the sample design. It is worth noting that this program does not correct for stratification, and the reported standard errors are likely to be somewhat too large as the typical effect of stratification is to slightly improve the precision of the sample estimates.

As a final example, we consider a case which is not completely appropriate for the EIHS data, but may be of use when there is more complete information on the sample design. An important intuition behind the bootstrap is that the re-sampling of the data should replicate the way in which the data was originally collected. A fairly standard design for many household surveys is to select PSUs with probability proportional to population, and then select the USUs with a simple random draw. *Ineqerr* with the *psuwt* (PSU weight) option can replicate this design if the user specifies the population estimates that were used to select the PSUs. In the case of the EIHS data, this information is not available. The EIHS data do provide PSU population estimates from the rural questionnaire, and to illustrate this feature, we treat this information as a proxy for the weights used in selecting the PSUs. The procedure used to do this is described in section on methods and formulas. The syntax used to implement this feature follows:

```
ineqerr pexp_r [w=wt96ind] if rural==1, reps(50) psu(psu) psuwt(psupop)
```

```
pexp_r ----- Real Per Capita Expenditure
(obs=1326)
```

```
Bootstrap statistics
```

Variable	Reps	Observed	Bias	Std. Err.	[95% Conf. Interval]	
Gini	50	.316077	-.0081707	.0136645	.2886171	.3435368 (N)
					.2858742	.3353139 (P)
					.3005531	.3383057 (BC)
Theil	50	.1836162	-.0053397	.0272215	.1289126	.2383198 (N)
					.1397549	.238112 (P)
					.1469196	.252798 (BC)
Varlogs	50	.3101612	-.0142084	.021793	.2663666	.3539558 (N)
					.2552118	.3348132 (P)
					.2766019	.3368765 (BC)

N = normal, P = percentile, BC = bias corrected

## Methods and Formulas

The Gini is perhaps one of the most widely used indicators of inequality, and can be written as:

$$G = I + \frac{I}{H} - \frac{2}{H^2 m} \sum_{h=1}^H r_h M_h$$

where H is the sample size,  $r_h$  is the rank of the observations ranging from 1 to H with the richest observation having the rank of one ( $r_1 = 1$ ),  $\mu$  is average value of M, and M is the measure of welfare which is sorted in descending order so that  $M_1$  is the richest individual and  $M_H$  is the poorest individual.

When weights are introduced, we follow Deaton (1997) who shows that (non-negative) analytical weights can be treated just as frequency weights. For the purposes of exposition, we consider the case where the weights are household size so that we are adjusting our measure to reflect inequality of individuals and not households. In this case we convert the ranks to reflect household size. We set  $r_1 = 1$  and then  $r_2 = 1 + w_1$ , where  $w_1$  is the size of the first household or the weight assigned to the first household. More generally:

$$r_{h+1} = r_h + w_h$$

and the average rank of all the individuals in household h can be written as:

$$\bar{r}_h = r_h + 0.5(w_h - 1)$$

The weighted Gini coefficient is then:

$$G = I + \frac{I}{N} - \frac{2}{N^2 m_w} \sum_{h=1}^H w_h \bar{r}_h M_h$$

where N is the weighted sample size (or the number of individuals in the sample when the

weight is household size),  $M$  is sorted as above, and  $\mu_w$  is the weighted average value of  $M$ .

The Gini index satisfies the Pigou-Dalton principle of transfers, that is a transfer from a richer person to a poorer person decreases the index. The magnitude of the decline, though, is determined by the difference in income rank between the two individuals and not the difference in incomes. Another characteristic of the Gini is that it is not generally decomposable.

The Theil index of income inequality is defined as follows:

$$T = (1/H) \sum_h (M_h / \bar{M}) \cdot \log(M_h / \bar{M})$$

where  $H$  is again sample size and  $\bar{M}$  is mean income. Foster (1983) shows that the Theil index satisfies several properties, including: decomposability, principle of transfers, symmetry, and income scale independence. The Theil also has the somewhat more attractive characteristic (relative to the Gini) that the magnitude of the decline in the index resulting from a transfer from a richer person to a poorer person is determined by the difference in the log of incomes.

The Variance of Logs is defined as follows:

$$V = \frac{1}{H} \sum_h [\ln M_h - \overline{\ln M}]^2$$

where terms are defined as above, except the mean is now of the log of income. It is perhaps worth noting that the Variance of Logs index does not satisfy the transfer principle, when the transfer is between two particularly rich observations.

The three bootstraps are implemented as follows. For the simple random sample (srs) we simply use Stata's *bsample* utility to bootstrap the three inequality indices. The srs, two-stage bootstrap follows this process: In the first stage it counts the number of unique PSUs, say  $k$ , and then using Stata's *uniform* function, randomly selects with replacement  $k$  (not necessarily unique) PSUs. At this point it counts the number of times each PSU has been selected and this is stored for later use. To implement the second stage, the program first counts the number of USUs, say  $m$ , in each selected PSU and then randomly selects  $m$  USUs from each selected PSU. If a PSU is selected more than once, say  $a$  times, then in the second stage the program randomly selects  $am$  USUs from the selected PSU.

The third variant of the bootstrap is the procedure in which PSUs are first selected with probability proportional to population (or whatever weight is specified in the *psuwt* option) and then the second stage is the same as above. The first-stage selection is then a weighted, random draw. This selection is implemented by creating a new variable which is the running sum of the PSU population or weight. For the last listed PSU this variable takes the value of total population (or sum of weights) of the USUs, say  $N$ . Again assuming there are  $k$  PSUs, the first stage begins by randomly selecting (using the *uniform* function)  $k$  numbers ranging from 1 to  $N$ . Each of these numbers is then associated with the PSU it represents and this is then the population-weighted, randomly selected PSU. To illustrate this, consider the following table with 4 PSUs:

PSU	Population	Cumulative Population	Random Number [1, N]	Selected PSU
1	100	100	633	4
2	200	300	305	2
3	300	600	585	2
4	100	700	22	1

Assume the first randomly selected number is 633. The fourth PSU contains USUs ranging from 601 to 700, and so USU number 633 comes from this PSU. In the next case, suppose the randomly selected number is 305. Again note that the 305<sup>th</sup> population USU resides in the second PSU, and so this PSU is selected. Following this methodology, the resulting selected PSUs are a randomly selected with probability proportional to population. (For more details on this see for example, Cochran, pp. 250 - 251.)

### References

- Cochran, William, *Sampling Techniques*, New York: Wiley, Third Edition, 1997.
- Datt, Gaurav; Jolliffe, Dean and Manohar Sharma, "A Profile of Poverty in Egypt - 1997." 1998. Food Consumption and Nutrition Division Discussion Paper No. 49. International Food Policy Research Institute. Washington, DC.
- Deaton, Angus, "Welfare, Poverty, and Distribution" in *The Analysis of Household Surveys*, Baltimore: Johns Hopkins, 1997, Chapter 3.
- Efron, Bradley and Robert Tibshirani, *An Introduction to the Bootstrap*, Monographs on Statistics and Applied Probability, 57, New York: Chapman Hall, 1993.
- Foster, James, "An axiomatic characterization of the Theil measure of income inequality," *Journal of Economic Theory*. 1983, 31: 105-21.
- International Food Policy Research Institute. "Egypt Integrated Household Survey (EIHS) 1997: Data and Documentation," IFPRI, Washington, DC 1998.
- Jenkins, Stephen. 1999. "Analysis of Income Distributions." *Stata Technical Bulletin* 48: 4-18.
- Kish, Leslie, *Survey Sampling*, New York: Wiley Classics Library Edition, 1995.
- Mills, Jeffrey and Sourushe Zandvakili, "Statistical Inference via Bootstrapping for Measures of Inequality," *Journal of Applied Econometrics*, 12: 133-150, 1997.
- Scott, A.J. and Holt, D. 1982. "The Effect of Two-Stage Sampling on Ordinary Least Squares Methods." *Journal of American Statistical Association* 77(380): 848-854.