

ASSIGNMENT 4 Due Monday May 5th by 5:00 pm

You have decided to study the causes of state-level income inequality, measured by the gini coefficient of family income. (*Note:* The gini is an aggregate summary statistic measuring how much inequality there is in a group relative to the maximum possible inequality. A gini value of .33 indicates that inequality is one-third as large as the maximum possible inequality).

In particular you are interested in trying to explain why income inequality is higher in the Southern region of the United States than it is elsewhere. Based on your reading and past research, you argue that differences in human capital, economic structure, and family structure across states can explain this regional difference. That is, you are trying to explain away the bivariate effect of Southern region on income inequality:

- H1:** The (positive) effect of Southern region on income inequality should NOT be significant once controls are introduced for interstate differences in human capital, economic structure, and family structure.
- H2:** Higher levels of human capital in a state (measured by average levels of education) should be associated with lower levels of income inequality.
- H3:** Economic structure in a state will influence inequality. Specifically, a greater prevalence of small businesses (measured by proportion of labor force in establishments employing 19 or less workers) should be associated with higher levels of income inequality.
- H4:** A greater prevalence of single parent households (measured by the proportion of households headed by single mothers) should be associated with higher levels of income inequality.

Using the data described on the codebook (other side), test these hypotheses using the appropriate OLS or EGLS results (you need not test if there is a bivariate effect of region on inequality but may correctly assume it does exist). **Only include as predictors those variables specified by the hypotheses.** Report and discuss any relevant diagnostics and tests that you performed to determine the absence, presence, or form of heteroskedasticity. Also present and discuss results from a standard regression table: present and interpret the meaning of the unstandardized and standardized coefficients, t-values [or standard errors], and summary statistics. *Write up your discussion and construct your table(s) as you would for a professional paper. The best way to organize the discussion of regression results is to discuss **all** the numbers/information for a predictor together rather than discussing the interpretation of b for all the predictors followed by the interpretation of the standardized coefficients for all predictors, etc.*

NOTE: If you find heteroskedasticity, you should assume that EGLS is the best estimator and do not need to compare it to Omega ($\hat{\Omega}$) corrected OLS or OLS with robust standard errors to determine which to use.

CODEBOOK

<u>Variable</u>	<u>Record</u>	<u>Columns</u>	<u>Content</u>
State	1	1-13	Two digit State identification code, 11 to 93
Region	1	14-26	Census Region 1 = North East, 2 = North Central, 3 = South, 4 = West
South	1	27-39	Southern Dummy 1 = South, 0 = other regions
Gini	1	40-52	Gini coefficient of family income in state
Mean_Ed	1	53-65	Mean of years of school completed in state
Manufacturing	1	66-78	Proportion of labor force employed in manufacturing industries
Service	2	1-13	Proportion of labor force employed in service industries
Small_Business	2	14-26	Proportion of labor force employed in businesses which employ fewer than 20 workers
Single_Head	2	27-39	Proportion of families headed by single mothers
Sample_size	2	40-52	Size of <i>CPS</i> sample in state

STATA Data Input Note: To read in multi-record data files, use the following syntax for **infix**. Note that you specify how many lines/records to read immediately after **infix** by specifying **nn lines** then use **record#:** followed by the variables and their column locations for each record. This example reads 3 lines (records) of data

```
infix 3 lines 1: var1 1-6 var2 7-9 ... ///
      2: var7 1-2 var8 3-10 ... ///
      3: var18 1-9 var19 10-14 using ...
```

The data file for 48 states (excluding Hawaii & Alaska) is *s:\703 Kaufman\hetdat* in the SIL and *s:\fac\rlk\703\hetdat* in the SRL. They can also be downloaded from the class website. These data were constructed from the March 1990 Current Population Survey (*CPS*) and are made available through the courtesy of Dr. Hyun-Song Lee. To ensure that you have correctly read the data, your descriptive statistics should reproduce:

<u>Variable</u>	<u>Mean</u>
South	.33333
Gini	.37010
Mean Ed	12.483
Manufacturing	.17773
Service	.69025
Small Business	.28448
Single Headed	.17038
Sample Size	3188.5