

Robinson, "Ecological Correlations and the Behavior of Individuals"
American Sociological Review, Vol 15, June 1950, 351-357.

X_r = percentage black in region r

Y_r = percentage illiterate in region r

$$\rho(X_r, Y_r) = .946$$

For individuals: $\rho(X_i, Y_i) = .203$

For states: $\rho(X_s, Y_s) = .773$

Correlations:

$$\begin{aligned}\rho(X, Y) &= \frac{\sum_{i=1}^N (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^N (X_i - \bar{X})^2 \sum_{i=1}^N (Y_i - \bar{Y})^2} \\ &= \frac{cov(X, Y)}{\sigma_x \sigma_y}\end{aligned}$$

Table 1: Individual Correlation: Race vs Illiteracy

	Black	White	Total
Illiterate	1512	2406	3918
Literate	7780	85574	93354
Total	9292	87890	97272

Cell entries are counts (in thousands) from census data, this represents total population.

Using the cell frequencies, we could compute correlations at the individual level.

Table 2: Within Area Correlations: By Region

		Black	White	Total
New England:	Illiterate	4	240	244
	Literate	72	6386	6458
	Total	76	6626	6702
Mid-Atlantic:	Illiterate	32	719	751
	Literate	836	19,958	20,794
	Total:	868	20,677	21,545

The individual level correlation is based on **ALL THE CELLS** in Table 1 (which we just get by adding all the cells in Table 2 over all regions).

The ecological correlation is based only on the **MARGINAL FREQUENCIES** in Table 2 (extended over all regions).

Key Point: Since **many** distributions within the internal cells are consistent with the observed marginals in Table 2, there are many individual level correlations consistent with an observed ecological correlations.

Another way to say this: there is more information in the individual level data than there is in the aggregate level data. You **CANNOT** recover the individual level information from the aggregate level data.

This does not stop us from trying....

Robinson's Example:

$\rho(\textit{foreign} - \textit{born}, \textit{literacy})$:

- At the individual level (around 1920), we know this is negative.
- But, if all the foreign-born residents are in the northern states, which overall have much higher literacy rates than other states, then the “ecological correlation” across the 50 states will be positive.
- Robinson finds: $\rho(\% \textit{foreign} - \textit{born}_s, \% \textit{illiterate}_s) = -.619$
- Note the switch from literate to illiterate in that correlation. To draw the “ecological inference” that there is a negative correlation between being foreign-born and being illiterate would be nonsense.

What To Do? - Goodman's Regression

$$\text{Percent} - \text{Literate}_s = \beta_0 + \beta_1 \text{percent} - \text{black}_s + \epsilon_i$$

We could add covariates.

But, Robinson did not want the marginal effect, he wanted the bivariate relationship.

And, we still have a problem: **aggregation bias**.

Switch to King's setup.

Gary King's Generalized Method of Bounds

Gary King. *A Solution to the Ecological Inference Problem – Reconstructing Individual Behavior from Aggregate Data* (Princeton University Press, 1997).

The following illustrates how the aggregate data is usually structured, the example is voter-turnout and race.

Race of Voting-Age Person	Voting Decision		
	Vote	Not-Vote	
<u>black</u>	β_i^b	$1 - \beta_i^b$	X_i
<u>white</u>	β_i^w	$1 - \beta_i^w$	$1 - X_i$
	T_i	$1 - T_i$	

where:

- X_i, T_i each indicates a proportion out of N_i .
- β_i is the proportion who vote out of X_i . [β_i^w is the proportion of whites who vote in the i^{th} precinct; β_i^b is the proportion of blacks who vote in the i^{th} precinct.]
- The superscripts indicate the race of voting-age person.
- T_i is the proportion of people who vote.

The quantities denoted by the English letters in the table are what we observe, while those denoted by the Greek letters are what we want to find out.

Accounting Identity:

$$T_i = \beta_i^b X_i + \beta_i^w (1 - X_i) \quad (1)$$

The problem now is that for every observation i , we have two parameters to estimate: β_i^b and β_i^w .

Option 1: Assume that all β_i^b 's are the same, and all β_i^w 's are the same.

This is substantively equivalent to assuming that the proportion of whites voting is constant across precincts, and that the proportion of blacks voting is constant across precincts.

This is “ecological regression”, or “Goodman’s regression”.

Problems:

Can estimate proportions outside the bounds $[0, 1]$.

Option 2:

We can narrow down the range of β_i 's as follows:

$$\max \left(0, \frac{T_i - (1 - X_i)}{X_i} \right) \leq \beta_i^b \leq \min \left(\frac{T_i}{X_i}, 1 \right) \quad (2)$$

$$\max \left(0, \frac{T_i - X_i}{1 - X_i} \right) \leq \beta_i^w \leq \min \left(\frac{T_i}{1 - X_i}, 1 \right) \quad (3)$$

Example: If we observe a precinct that is 80% white and 20% black, and turnout is 80 there are people in the precinct. If all blacks vote, white turnout is 75% (60 out of 80 whites voted). If no blacks voted, white turnout must be 100% (80 out of 80 whites voted).

Now we can re-arrange the above equations to see that each precinct has an associated line (and associated values of β_i^b and β_i^w):

$$\beta_i^w = \left(\frac{T_i}{1 - X_i} \right) - \left(\frac{X_i}{1 - X_i} \right) \beta_i^b. \quad (4)$$

So instead of fitting a line in two dimensions, we can fit a two dimension contour to encircle the intersections of as many lines as possible.

[Figure 6.3 from *A Solution to the Ecological Inference Problem* about here.]

Like the regression lines which are the projection of a three-dimensional distribution onto a two-dimensional surface, the contour lines in the figure are also the projection of a bivariate distribution, with the parameters of the distribution printed at the top of the figure. In fact, the above figure corresponds to the graph below.

Notice:

Every line in the graph is determined by equation (2). See that they all show negative slopes.

[Figure 6.4 from *A Solution to the Ecological Inference Problem* about here.]

Assume the parameters we want to estimate, each β_i^b and β_i^w , are drawn from some common distribution.

The distribution will have a set of additional parameters that we can estimate.

Three assumptions are required by the model:

Assumption 1: β_i^b, β_i^w are generated by a truncated bivariate normal distribution conditional on X_i , i.e.,

$$P(\beta_i^b, \beta_i^w) = TN(\beta_i^b, \beta_i^w \mid \mathcal{B}, \Sigma)$$

where

$$\mathcal{B} = \begin{pmatrix} \mathcal{B}^b \\ \mathcal{B}^w \end{pmatrix} = E \begin{pmatrix} \beta_i^b \\ \beta_i^w \end{pmatrix}$$
$$\Sigma = \begin{pmatrix} \sigma_b^2 & \sigma_{bw} \\ \sigma_{bw} & \sigma_w^2 \end{pmatrix} = Var \begin{pmatrix} \beta_i^b \\ \beta_i^w \end{pmatrix}$$

Assumption 2: β_i^b, β_i^w are mean independent of X_i , i.e., completely unrelated.

Assumption 3: Values of T_i in different precincts are independent after conditioning on X_i .

