“Let method be the servant, not the master”

NSC Methodology Workshop in Taiwan
January 4–7

Glenn Firebaugh
Liberal Arts Research Professor, Penn State University
Visiting Scholar, 2004-05, Harvard University

"Let method be the servant, not the master"

OVERARCHING QUESTION: HOW IS IT THAT WE CAN HAVE STUDIES THAT APPEAR TO BE TECHNICALLY (STATISTICALLY) CORRECT, YET ARE WILDLY OFF IN THEIR CONCLUSIONS?

"Let method be the servant, not the master"

By "wildly off in their conclusions" I mean one or more of the following:

- <u>Description failure</u>: Failure to get the basic facts right. Example: Initial studies in sociology concluded that global income inequality is growing rapidly, when a more sober analysis using the same sample and data shows a stable or declining trend.

- <u>Replication failure</u>: Failure to meet a minimum substitutability standard. That is, you get very different results (coefficients go from big to zero, or from zero to big, or reverse sign) with minor modifications of the model or sample or method.

"Let method be the servant, not the master"

- <u>Prediction failure</u>: Failure to identify true causal connections in an "if…then" sense when applied to the real world ("If we do X, the result will be Y"). Example: We are an Eastern European country worried about below-replacement fertility. Results of a causal analysis of fertility find that income has a direct positive effect on fertility, so we recommend government subsidies to families with children – essentially the government pays couples to have more children. It doesn't work.

Where do our analyses most often go wrong, and how can we improve our methods to correct these problems? That's the underlying theme of these talks.

Seven Rules

SEVEN PRACTICAL RULES TO HELP YOU AVOID ERRORS AND REACH DEFENSIBLE CONCLUSIONS (Days 1 & 2)

Rule 1. Interpret effects in the context of models.

- Example 1 (microlevel): relative and absolute effects of income on subjective well-being
- Example 2 (macrolevel): estimating the effect of industrialization on income growth in poor countries

Seven Rules

Rule 2. Consider all sources of error: Sampling error is not the whole story.

- Omitted-variables bias, including selection bias; functional form of relationships; measurement error

Rule 3. To control for confounding factors in observational studies, try to supplement "control by regression" with techniques that simulate the effects of random assignment. In other words, when random assignment isn't possible, try to think of alternatives that would serve as the functional equivalent of random assignment.

- First-difference models, growth rate models, sibling models, natural experiments

Seven Rules

Rule 4. Focus on the size of effects. Don't worship at the altar of statistical significance.

- Ziliak-McCloskey paper on misuse of statistical methods in economics. Discuss: Is the situation better or worse in sociology?

Rule 5. You can't expect good estimates from bad models, no matter how sophisticated your estimation technique. "Your estimates are only as good as your models."

- Example: effects of foreign investment on income growth in developing countries.

Seven Rules

Rule 6. Define your terms. Seems obvious, but is often ignored, and has led to much mischief and unproductive arguments.

- Example: inequality, perhaps the most fundamental of sociological concepts. Confusion over what this term means precisely, and how it is measured, has plagued research on global income inequality, as we will see in the Case Study.

Seven Rules

Rule 7. Look for, and exploit, empirical discrepancies.

CASE STUDY: Unfounded claims about the evolution of global income inequality – illustrates how things can go badly wrong in our analyses

Social Change

II. ASSESSING SOCIAL CHANGE (Day 3)

- **Assessing social change within nations (descriptive analysis)**

1. Models for detecting change in individual-level <u>effects</u> from time 1 to time 2

- Examples: Has the effect of gender on gender role attitudes become stronger or weaker over time? Has education become more important in determining income in Taiwan?

Social Change

2. Regression standardization models for determining how change in the Xs "adds up" to change in Y from time 1 to time 2

3. Convergence models: Models for determining the convergence or divergence of trends over time.

- Example: Has satisfaction with family life changed in Taiwan? Has it changed the same for men and women?

Social Change

4. Cohort replacement models: Where does social change come from? Is the proximate source changing individual views, or population turnover?

- Examples: Changing gender roles in the family; changing attitudes toward mainland China

Change Across Nations

B. Assessing change across nations

- Problems in cross-country research
- Principles in cross-country research

Publishing Your Work

III. PUBLISHING YOUR WORK (Day 4)

- **Session 7:** How to write for leading journals – things to do, things to avoid

- **Session 8:** Ideas on editing a journal – perspectives of Editors (panel). What lies in that "black box" from submission → decision letter?

Rule1 – interpret effects in context of models

Rule #1: Interpret Effects in the Context of Models

- For example, the variable "income" might reflect one construct in one model and have quite a different meaning in another model. This might seem like a trivial point but it is not hard to find examples where research has gone astray by failure to think carefully about what the variables are actually measuring in the context of the model.

Rule1 – interpret effects in context of models

- I begin with an example of how one can exploit this principle to shed light on how income affects one's happiness, and then give a negative example (what to avoid) to show how failure to observe this rule caused mischief in studies of how industrialization affects economic growth in developing countries.

- **Reading:** G. Firebaugh and Laura Tach, "Income and happiness in the United States," unpublished manuscript.

Rule1 – interpret effects in context of models

Example #1 (micro): Relative and absolute effects.

Concept: the value or effect of something depends not only on how much of it you possess but also on how much of it is possessed by <u>others in your peer group</u>.

- Income and relative income effects, education and relative education effects, age and relative age effects…

General form of model: $Y = f(X_{ij}, X_j)$ where X_{ij} is the level of X for the i^{th} person in the j^{th} peer or reference group, and X_j measures the level of X for the j^{th} group.

Rule1 – interpret effects in context of models

Application: absolute and relative income effects on happiness:

$$H_{ij} = \alpha + \gamma X_{ij} + \delta (X_{ij} - \mu_j) + \epsilon_{ij} \qquad (1)$$

- H_{ij} is measured happiness for i_{th} person in j^{th} reference group
- X_{ij} is measured income for the i^{th} person in the j^{th} reference group
- μ_j is mean income for the j^{th} reference group
- ϵ is a random error term

Rule1 – interpret effects in context of models

$$H_{ij} = \alpha + \gamma X_{ij} + \delta (X_{ij} - \mu_j) + \epsilon_{ij} \qquad (1)$$

This model states that your happiness depends on your own income *and* on the income of others – H is a function of X and μ.

- If your income per se affects your happiness, then $\gamma > 0$ (presumably the effect is positive)
- If your *relative* income affects your happiness, net of the effect of your income per se, then $\delta > 0$.

Rule1 – interpret effects in context of models

page 20

$$H_{ij} = \alpha + \gamma X_{ij} + \delta (X_{ij} - \mu_j) + \epsilon_{ij} \qquad (1)$$

The key point is that here, in eqn. 1, the coefficient for income reflects the **absolute** effect of income – the happiness it brings because of the pleasure it can bring you (better houses, travel, etc.) regardless of the income of your peers. Suppose your income changes +Q and the average income of your peers also increases +Q – so your relative income doesn't change, but your absolute income does. If $\gamma>0$, we expect happiness to increase. (What would you find for Taiwan?)

Rule1 – interpret effects in context of models

Instead of regressing happiness on income and $(X_{ij} - \mu_j)$, as in equation 1, suppose we regressed happiness on income and mean income μ_j, as in equation 2:

$$H_{ij} = \alpha + \phi X_{ij} + \lambda \mu_j + \epsilon_{ij} \qquad (2)$$

Now how do we interpret the coefficient for income? Is it absolute income effect, or the relative effect of income, or what?

Rule1 – interpret effects in context of models

We find out by rearranging the terms in equation 1:

$$H_{ij} = \alpha + \gamma X_{ij} + \delta (X_{ij} - \mu_j) + \epsilon_{ij}$$
$$= \alpha + (\gamma + \delta) X_{ij} - \delta \mu_j + \epsilon_{ij}$$

So the income effect ϕ in equation 2 is no longer the absolute effect of income, but is the sum of the absolute and relative effect of income. This makes intuitive sense when we look at ϕ in equation 2: when your income goes up with the income of your peers constant (that is, μ_j constant), then your relative income goes up along with your absolute income; so ϕ should capture both effects.

Rule1 – interpret effects in context of models

When does this issue arise? Examples:

(1) Models of crime rates across cities. Is it because of poverty or inequality or relative deprivation? Measures that *by themselves* measure poverty or inequality might be capturing something else when put in such a model.

(2) In general, issue arises when you have similar concept measured in different ways, or same variables at different levels of aggregation (e.g., gender and gender ratio, race and % black)

Rule1 – interpret effects in context of models

page 24

General point: the interpretation of the income coefficient here depends on what other variables are in the model. Although that point is more apparent here than usual, it is not unusual for sociologists to misinterpret their models because they fail to think carefully about what their variables mean in the context of the other variables in the model. In the next session I give an example from a recent paper on the effect of industrialization on economic growth in developing countries.

// end of session 1//

Rule1 – interpret effects in context of models

Second Example of Rule #1: Industrialization and income growth

On the basis of a cross-country regression analysis of income growth for about five dozen non-Western countries from 1965 to 1998, ASB (2003) conclude that industrialization no longer promotes economic growth in poor countries. "For most [poor] countries," they write (p. 18), "industrialization [has] turned out to be an ineffectual means of economic advancement."

Source Arrighi, Silver, and Brewer. 2003. "Industrial convergence, globalization, and the persistence of the North-South divide." <u>Studies in Comparative International Development</u> 38:3-31. See also G. Firebaugh. 2004. "Does industrialization no longer benefit poor countries? A comment on Arrighi, Silver, and Brewer." <u>Studies in Comparative International Development</u> 39:99-105.

Rule1 – interpret effects in context of models

page 26

Yet, for the countries in their analysis the correlation between industrial growth and income growth is $r = .93$ for the 1965-1998 period they studied ($r = .98$ if countries are weighted by population size). Given the size of this correlation, how do we account for ASB's result? For Alice Amsden the ASB result is so puzzling that in a comment on the article she suggests that it must be due to a "bug in their program." In their reply, ASB state that they double-checked their computer program and are satisfied that their finding is not due to a programming error.

Rule1 – interpret effects in context of models

What is the problem? Violation of rule # 1 – ASB misinterpret the results of their model by failing to interpret their coefficient for industrialization in light of the other variables in their model. Here is the model they estimate, where prime denotes growth rate:

$$(Y/P)_i' = \beta_0 + \beta_1 (M/Y)_i' + \epsilon_i$$

Rule1 – interpret effects in context of models

page 28

$$(Y/P)_i' = \beta_0 + \beta_1 (M/Y)_i' + \epsilon_I$$

- Y denotes income (national output), M denotes manufacturing output, P denotes population – so $(Y/P)_i'$ is rate of growth of per capita income in the *i^th^* Third World country and $(M/Y)_i'$ is rate of growth of manufacturing output as a fraction of total output in the i^{th} Third World country.
- ASB estimate this model for 58 Third World countries for 1960-1980 and for 59 Third World countries for 1980-1998. Because the slope is zero in both regressions, they conclude that "none of the variability in Third World country income performance was predicted by variability in their industrialization performance" (pp. 12 and16).

Rule1 – interpret effects in context of models

page 29

$(Y/P)_i' = \beta_0 + \beta_1 (M/Y)_i' + \epsilon_I$

Note that the rate of growth of a ratio A/B is the rate of growth of A minus the rate of growth of B. Hence $(Y/P)_i' = Y_i' - P_i'$ and $(M/Y)_i' = M_i' - Y_i'$, and we can rewrite their equation as:

$$Y_i' - P_i' = \beta_0 + \beta_1 (M_i' - Y_i') + \epsilon_i. \qquad (2)$$

What ASB find is that $\beta_1=0$. What does $\beta_1=0$ mean?

Rule1 – interpret effects in context of models page 30

In plain English, their model states that the extent to which income growth exceeds (or lags behind) population growth is a function of the extent to which manufacturing growth exceeds (or lags behind) income growth.

So ASB did not actually estimate what they thought they had (the economic effect of industrialization in the Third World). Instead they estimated the effect of the *difference* in two growth rates -- growth rate of industrial output minus growth rate of total output -- on the difference in income growth and population growth.

Rule1 – interpret effects in context of models

Their results show that the difference between income growth rate and population growth rate is uncorrelated with the difference between manufacturing growth rate and income growth rate (growth rate of total output). *That is precisely what we would expect if industrialization boosts economic performance*. To benefit a country economically, manufacturing should be linked to other sectors of the economy. Successful industrialization means that manufacturing growth does not just boost manufacturing, but boosts other economic sectors as well.

Rule1 – interpret effects in context of models

In effect, what ASB tested is the economic effect of *segregated* manufacturing growth in the Third World. If incomes grew fastest in countries where manufacturing growth surged ahead of total income growth (that is, segregated manufacturing growth, not linked to growth in the rest of the economy) we would find a positive correlation between M'-Y' and Y' - P'. So in the case of segregated industrialization we would find the positive association ASB were looking for, but not for the reasons they think.

Rule1 – interpret effects in context of models

Note an important methodological lesson here: *Be careful about variable transformations*. To interpret properly a model with transformed variables, it is often necessary to express it in alternative (algebraically-equivalent) forms. (I could have included that as another rule but I had seven already.) Here I use this example to illustrate the rule that you need to interpret effects in the context of models. The interpretation of the effect of X_1 on Y often depends on what other Xs are in the model.

Rule 2 – Consider total uncertainty

page 34

Rule # 2. Consider all sources of error, not just sampling error

Omitted-variables bias, including selection bias; functional form of relationships; measurement error, including error-in-data and error-in-index

Rule 2 – Consider total uncertainty

page 35

It is well-established that non-sampling errors are at least as important as sampling errors, and *the relative importance of non-sampling errors increases as our samples get larger*. We all know this in principle, but in practice we often focus only on sampling errors without always thinking very hard about how non-sampling errors might be affecting our results. We might have omitted important causal variables; we might have the right variables but they are measured imperfectly; we might have the right variables measured perfectly, but we assume the relationships are linear when they are not – all sources of possible non-sampling errors.

Rule 2 – Consider total uncertainty

One of the areas that needs more work in sociology is that of modeling **total uncertainty** in our estimates. We do a good job of modeling uncertainty due to sampling error, but we know much less about how to model uncertainty aside from sampling error. Or, better still, we need to use methods designed to reduce non-sampling error – the issue in Rule #3.

Rule 2 – Consider total uncertainty

Omitted-variables bias

Begin w/ "baseline" regression model in social sciences:

$$Y_i = \mathbf{X_i}\,\beta + \mathbf{Z_i}\,\beta^* + \epsilon_i$$

Notation here and throughout:

Y_i is the value of DV for *ith* unit, Xs are measured causal variables, Zs are unmeasured causal variables, ϵ is well-behaved random error. Bolding = vector, so **X** is a row vector with elements $[1\ \ X_1\ X_2 \ldots.\ X_P]$ and β is a column vector of parameters with elements $[\beta_0\ \ \beta_1\ \ \beta_2 \ldots.\ \beta_P]$.

So $\mathbf{X_i}\,\beta = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots. + \beta_P X_{Pi}$.

Rule 2 – Consider total uncertainty

page 38

This is a familiar model. We all know, for example, that omitted-variables bias occurs in this model when the Xs and Zs are correlated. We can reduce omitted-variables bias by

- moving variables from Z to X, that is, measuring more of the causal variables or
- using research designs that reduce or eliminate the correlation between the Xs and Zs.

What I want to suggest is that we think more creatively about strategy 2.

Rule 2 – Consider total uncertainty

page 39

Usually we focus on strategy 1 – we try to measure all the causal variables we can, but of course we never know whether we have successfully identified and measured all of them. For causal analysis, then, we often would like to do experiments where we can randomly assign units to treatment and control groups to try to eliminate the correlation between the Xs and Zs.

(Note that in experiments we often want to **reduce** the number of Xs – we consider one causal variable at a time – rather than increasing the number of Xs in the model, as in strategy 1.)

Rule 2 – Consider total uncertainty

page 40

But random assignment is not possible for much of our work, so we fall back on strategy 1. In the next rule (#3) I suggest that we rethink that practice. Instead of ruling out randomization and giving up on strategy 2, we need to think of ways to mimic the effects (or some of the effects) of randomization where there is no random assignment. I suggest a few possibilities in Rule #3 in the hope that this will stimulate your thinking along these lines. Surely we can often do better than measuring all the variables we can, putting them in a regression equation, and hoping that we haven't omitted anything important.

Rule 3 – Mimic randomization

page 41

Rule # 3. To control for confounding factors in observational studies, try to supplement "control by regression" with techniques that mimic the effects of random assignment.

Examples: First-difference models, sibling models, propensity score matching

Readings: For further reading I suggest:

- First-difference models: Liker, J., S. Augustyniak, and G. Duncan. 1985. "Panel data and models of change: A comparison of first difference and conventional models." Social Science Research 14:80-101.
- Sibling models: Exchange over sibling size effects in American Sociological Review (April 1999) – Guo & VanWey; comments by Phillips and Downey et al.; reply
- Fixed-effect models in general: Halaby, Charles N. 2004. "Panel models in sociological research: Theory into Practice." Annual Review of Sociology 30:507-44.

Rule 3 – Mimic randomization

First-difference models -- Consider again the baseline regression model in the social sciences:

$$Y_i = \mathbf{X_i} \beta + \mathbf{Z_i} \beta^* + \epsilon_i$$

Suppose we measure Y and the Xs at a second point in time (that is, a two-wave panel study). Write out two regression models:

$$Y_{i2} = \mathbf{X_{i2}} \beta + \mathbf{Z_{i2}} \beta^* + \epsilon_{i2} \qquad \text{(time 2)}$$
$$Y_{i1} = \mathbf{X_{i1}} \beta + \mathbf{Z_{i1}} \beta^* + \epsilon_{i1} \qquad \text{(time 1)}$$

Rule 3 – Mimic randomization

page 43

A longitudinal data set has a crucial advantage over cross-sectional and repeated C-S designs: We can measure individual change from time 1 to time 2. Do that by subtracting Y_{i1} from Y_{i2}:

$$Y_{i2} - Y_{i1} = (\mathbf{X}_{i2}\,\beta_2 - \mathbf{X}_{i1}\,\beta_1) + (\mathbf{Z}_{i2}\,\beta_2^* - \mathbf{Z}_{i1}\,\beta_1^*) + \epsilon_i'$$
$$= (\mathbf{X}_{i2} - \mathbf{X}_{i1})\,\beta_2 + \mathbf{X}_{i1}\,(\beta_2 - \beta_1) + (\mathbf{Z}_{i2} - \mathbf{Z}_{i1})\,\beta_2^*$$
$$+ \mathbf{Z}_{i1}\,(\beta_2^* - \beta_1^*) + \epsilon_i'$$

From this equation we see that the effects of the Zs disappear in the first-difference model when

- Z is constant over time, that is, $(\mathbf{Z}_{i2} - \mathbf{Z}_{i1}) = 0$, AND
- the effect of Z is constant over time: $(\beta_2^* - \beta_1^*) = 0$

Rule 3 – Mimic randomization

page 44

Goal: Remove confounding effects of Zs.

Method 1: Random assignment.
Limitations of random assignment: often not practical.

Method 2: First-difference model. Removes confounding effects of unmeasured variables that:

- are unit-specific and time-invariant – they vary across units but are constant over time for a given unit (examples: ascribed characteristics such as race; enduring personality traits) *and*
- have constant *effects* over time

Rule 3 – Mimic randomization

page 45

Limitations of first-difference method:

- does not remove effects of Zs that change over time
- does not remove effects of constant Zs whose *effects* change over time (e.g. does not remove the effect of gender when gender effects change over time)
- see **(Xi2 - Xi1)** β**2 + Xi1** (β**2** - β**1)** -- the method removes the effects of *measured* variables that are constant over time.

Rule 3 – Mimic randomization

page 46

Example: you can't use first difference models to estimate the effect of constant variables such as country of birth or gender or parents' SES, since for those variables (**Xi2** - **Xi1**) = 0. All you can estimate is the *change* in the effect, (β**2** - β**1**).

- Note that this is another example of the importance of interpreting a variable in the context of a model. The coefficient for ethnicity in a first-difference model does *not* estimate the effect of ethnicity on Y, but *change* in the effect of ethnicity on Y.

Rule 3 – Mimic randomization

page 47

Method 3: Growth rate models.

We don't usually think of using growth rate models to eliminate confounding effects of unmeasured variables, but growth rate models bear a formal similarity to first difference models. Because $\log Y_{i2} - \log Y_{i1}$ is the rate of growth of Y (where log is the natural logarithm), growth rate models can be expressed as difference models, where the variables are logged. To simplify notation, suppose Y is caused by just two variables, X and Z, where as before X is measured and Z is not. We assume further that the effect of X and Z on Y is linear in the logs.

Rule 3 – Mimic randomization

page 48

We have:

$\log Y_{i2} = \log (\mathbf{X_{i2}}^{\beta 2}\ \mathbf{Z_{i2}}^{\beta 2^*}\ \epsilon_{i2})$

$= \beta_2 \log X_{i2} + \beta_2^* \log Z_{i2} + \epsilon_{i2}$

$\Rightarrow \log Y_{i2} - \log Y_{i1} = (\beta_2 \log X_{i2} - \beta_1 \log X_{i1}) + (\beta_2^* \log Z_{i2} - \beta_1^* \log Z_{i1}) + \epsilon'$

$= \beta_2 (\log X_{i2} - \log X_{i1}) + (\beta_2 - \beta_1) \log X_{i1} + \beta_2^* (\log Z_{i2} - \log Z_{i1}) + (\beta_2^* - \beta_1^*) \log Z_{i1} + \epsilon'$

Note the growth rates $\log Y_{i2} - \log Y_{i1}$, $\mathbf{\log X_{i2} - \log X_{i1}}$, and $\mathbf{\log Z_{i2} - \log Z_{i1}}$ (zero, for constant Zs).

Rule 3 – Mimic randomization

page 49

So by regressing growth rate of Y, $\log(Y_2/Y_1)$, on growth rates of Xs, we are removing the constant effects of constant unmeasured causes that vary across units. In cross-country research, for example, that would include the *constant* effects of a country's location, topography, climate, history, mineral resources, language, culture, political and legal systems, access to seaports, etc.

Rule 3 – Mimic randomization

page 50

Method 4: Fixed-effects models for panel data with multiple panels. Can be done by putting in a dummy variable for each unit, thereby removing the constant (but not the changing) effects of unit-specific time-invariant variables, just like the first-difference model does. Is very similar to first-difference model, and has the same limitations.

Other methods: Natural experiments or quasi-experiments (see Christopher Jencks; Tom Cook); propensity score methods.

//end of session 2//

Rule 4 – Focus on size of effects

page 51

Rule # 4. Focus on the size of effects. Don't worship at the altar of statistical significance.

- **Reading:** Stephen Ziliak and Deirdre McCloskey. 2004. "Size matters: The standard error of regressions in the American Economic Review." Econ Journal Watch 1 (August): 331-354. Retrievable at www.econjournalwatch.org.

- **Discuss:** The Ziliak-McCloskey article is very critical of statistical practices in economics. Is the situation better or worse in sociology?

Rule 5 - Avoid estimation fallacy

page 52

Rule # 5. You can't expect good estimates from bad models.

- Estimation fallacy: The assumption that good estimation methods can compensate for faulty measures and models. In fact, your estimates are only as good as your models. Sophisticated estimation methods won't bail you out. If your model is flawed, sophisticated methods will just give you precise estimates of the wrong thing.

- **Example: Effect of foreign investment on economic growth in poor countries -- Reading:** G. Firebaugh. 1992. "Growth Effects of Foreign and Domestic Investment." American Journal of Sociology 98 (July):105- 130.

Rule 5 - Avoid estimation fallacy

page 53

During the 1980s and 1990s dependency theory dominated sociological discussion of why some countries are so rich and others so poor. According to the theory the prosperity of some countries and the poverty of others are opposite sides of the same coin: Economic exchange between rich "core" countries and poor "peripheral" countries is inherently one-sided so the core countries benefit from the exchange and poor countries are harmed by the exchange.

- Critical empirical prediction of theory: Other things equal, poor countries engaging heavily in economic exchange with core countries should grow more slowly than poor countries that are not so heavily engaged.

Rule 5 - Avoid estimation fallacy

page 54

- Results of empirical tests: In the 1980s and early 1990s, many articles appeared in U. S. sociology journals testing the dependency prediction. At first the articles focused on the effects of (1) trade and (2) foreign investment. The usual finding, however, was that countries that traded more grew faster, contrary to the dependency prediction, so the literature shifted to an almost exclusive focus on foreign investment effects.

- Literally dozens of articles used cross-country regressions with income growth rate regressed on measures of foreign investment flow, foreign investment stock, and control variables.

Rule 5 - Avoid estimation fallacy

page 55

- The studies *consistently* reported a positive coefficient for investment flow and a negative coefficient for investment stock.
- Unanimous conclusion of the dependency literature: Foreign investment from core countries has short-term positive effects (effects of investment "flow" are positive) but harmful long-term effects ("stock" effects) on income growth in poor countries.

Rule 5 - Avoid estimation fallacy

page 56

The models estimated in this dependency research are based on panel data for countries at two points in time, generally several decades apart. Using the example of 1965 and 1990 data, the models have this general form:

$$(Y_{i90}/Y_{i65})' = \beta_0 + \beta_1 (KFor_{i90} - KFor_{i65}) + \beta_2 KFor_{i65} + \beta_3 (KDom_{i90}/KDom_{i65})' + \text{control variables} + \epsilon_i$$

Here Y_i is per capita income for country *i*, KFor is foreign investment, and KDom is domestic investment.

$(Y_{i90}/Y_{i65})'$ is rate of growth of per capita income from 1965 to 1990 – similarly for $(KDom_{i90}/KDom_{i65})'$

Rule 5 - Avoid estimation fallacy

page 57

$(Y_{i90} / Y_{i65})' = \beta_0 + \beta_1 (KFor_{i90} - KFor_{i65}) + \beta_2 KFor_{i65} + \beta_3 (KDom_{i90}/KDom_{i65})' + \text{control variables} + \epsilon_i$

The dozens of studies using this model uniformly report these results:

- $\beta 1 > 0$, "positive 'flow' effect of foreign investment"
- $\beta 2 < 0$, "harmful long-run ('stock') effect of foreign investment on economic growth in poor countries." The claims of dependency theory rested on this result.
- $\beta 3 > 0$, "positive effect of domestic investment"

Rule 5 - Avoid estimation fallacy

page 58

$$(Y_{i90} / Y_{i65})' = \beta_0 + \beta_1 (KFor_{i90} - KFor_{i65}) + \beta_2 KFor_{i65} + \beta_3 (KDom_{i90}/KDom_{i65})' + \text{control variables} + \epsilon_i$$

What's wrong with this picture? Note that *growth rate* is flow/stock (e.g., population growth rate is population change ["flow"] divided by initial population). So if the effect of investment rate is positive, we expect the following if we split investment rate into its two constituent parts, "flow" (the numerator) and "stock" (the denominator):

Rule 5 - Avoid estimation fallacy

page 59

$$(Y_{i90} / Y_{i65})' = \beta_0 + \beta_1 (KFor_{i90} - KFor_{i65}) + \beta_2 KFor_{i65} + \beta_3 (KDom_{i90}/KDom_{i65})' + \text{control variables} + \epsilon_i$$

- for a given initial level of investment ("stock"), the greater the flow the *higher* the investment rate => coefficient for flow should be positive, controlling for stock
- for a given level of investment flow, the greater the stock the *lower* the investment rate => coefficient for stock should be *negative*, controlling for flow

Rule 5 - Avoid estimation fallacy

page 60

$(Y_{i90}/Y_{i65})' = \beta_0 + \beta_1 (KFor_{i90} - KFor_{i65}) + \beta_2 KFor_{i65} + \beta_3 (KDom_{i90}/KDom_{i65})' + \text{control variables} + \epsilon_i$

(Note that our example here also illustrates Rule # 1, that the effect of a variable must be understood in the context of other variables in the model – the interpretation of the "stock" effect is conditioned by the presence of "flow" variables in the model, and vice versa.)

So the finding appears to be a methodological artifact of the way foreign investment is measured. We can check this by measuring foreign and domestic investment *the same way.*

Rule 5 - Avoid estimation fallacy

page 61

- When we measure foreign investment as rate of investment (flow/stock) we expect to find a positive effect.

- When we measure domestic investment as two separate components – flow and stock – we expect a positive effect for flow and a negative effect for stock

Rule 5 - Avoid estimation fallacy

page 62

We find both: (1) Rate of foreign investment has a positive effect of economic growth, just like rate of domestic investment rate does; and (2) domestic capital "stock" (net of domestic flow) has a substantial negative "effect" on economic growth.

The results for (2) are especially telling: It is ludicrous to interpret the negative slope for domestic stock as indicating a long run harmful effect due to a country's domestic investment!

Rule 5 - Avoid estimation fallacy

page 63

Lessons to be learned:

- Sophisticated analysis does not compensate for wrongheaded models. There is little value in precise estimates of the wrong thing.

- Strive for measurement consistency across measures in your research. Be wary of hybrid models in particular – they can be hard to interpret. The dependency investment model was a bad model because it contained inconsistent measures of foreign and domestic investment. A lot of research effort (and good journal space) could have been saved had researchers used similar measures for foreign and domestic investment.

Rule 6 – Define your terms

page 64

Rule # 6. Define your terms

- Seems obvious, but is often ignored, and has led to much mischief and unproductive arguments.
- Example: inequality, perhaps the most fundamental of sociological concepts. Confusion over what this term means precisely, and how it is measured, has plagued research on global income inequality, as we will see in the Case Study later.

Rule 6 – Define your terms

page 65

- *Income inequality* refers to the disproportionate distribution of income across units (e.g. individuals). It is based on income *ratios* – inequality does not change when everyone's income grows at the same rate (so income ratios are constant).
- *Global* income inequality – the disproportionate distribution of income over all the world's people (each individual counts the same)
- *Between-nation* income inequality – the disproportionate distribution of per capita income across nations

Rule 6 – Define your terms

page 66

No one disputes that income *disparities* across countries are growing in the world – the question is whether the growing gaps reflect growing income or growing income *inequality*.

Example (constant 1990 U.S. $, from Maddison 1995):

	1950	1990
• mainland China	$614	$2700
• Rest of Asia[a]	$1081	$4745
gap:	$475	$2045
ratio:	0.57	0.57

[a] Outside India

Rule 6 – Define your terms

page 67

Taiwan and the United States (in constant 1990 U.S. $, from Maddison 1995):

	1950	1990
• Taiwan	$922	$10,324
• United States	$9573	$21,866
gap:	$8651	$11,542
ratio:	0.096	0.472

Rule 7 – exploit empirical anomalies page 68

<u>Rule # 7. Look for, and exploit, empirical discrepancies or anomalies</u>

This is often how science advances – attempting to account for anomalies

Reading: Firebaugh, G. and Brian Goesling. 2004. "Accounting for the Recent Decline in Global Income Inequality." <u>American Journal of Sociology</u> 110 (September): 283-312.

Rule 7 – exploit empirical anomalies

page 69

CASE STUDY: GLOBAL INCOME INEQUALITY – to illustrate how things can go badly wrong in our analyses

Two incongruities here (as we will see in next session):

- methodological = unweighted cross-country regressions versus weighted studies
- two substantive literatures: global inequality is increasing rapidly (see UN and World Bank) versus "Asia is catching up to the West"

//End of session 3//

Social Change – Day 3

page 70

- **Reading:** Firebaugh, G. 1997. Analyzing Repeated Surveys. Sage University Paper Series on Quantitative Applications in the Social Sciences, no. 07-115. Thousand Oaks, CA: Sage.
- **Suggested further reading**: Firebaugh, G. 1992. "Where Does Social Change Come From? Estimating the Relative Contributions of Individual Change and Population Turnover." Population Research and Policy Review vol. 11:1- 20.

Analyzing change – overview (change within nations)

- There are four important questions to ask about change over time – one about individual-level change, two about aggregate-level change, and the fourth about how individual-level change “adds up” to aggregate level change

Social Change

Assessing social change within nations

- Individual-level change (two time points): Has the effect of X on Y changed over time? **Changing-parameter Models**: (1) Panel: First-difference model with X_1s included as regressors. (2) Repeated survey: Interaction effects with *time* as a conditioning ("moderating") variable.

- How change in the Xs adds up to aggregate change in the Y (two time points). **Model:** Regression standardization applied to change in Y-mean.

Social Change

3. Aggregate (multiple T): Convergence/divergence models. Example: Has satisfaction with family life changed in Taiwan? Has it changed at the same rate for men and women? **Model:** Interaction effects with Time and Group as covariates and Time*Group interaction term (*group* as a moderating variable).
4. Aggregate 2 (multiple T): What is the *proximate cause* of the change in the Y-mean -- Is the change in Y-mean due to individuals changing or to population turnover? **Model:** Cohort replacement effects.

Social Change

The models I discuss are both simple (often can be estimated with OLS) and flexible. As we will see, to track aggregate change we generally want multiple waves of data. Let t = 1, 2, … T denote cross-sectional data at a particular point in time, and N_t denote the number of individuals in the t^{th} cross-section. In sociology generally our Ns tend to be much larger than our Ts (e.g., for the GSS in the U.S., N = 1500 per year and T = about 30 years). As we continue to cumulate time-series data, however, the aggregate-change models become more feasible.

Changing-parameter models

page 74

Individual-level change: Changing-parameter models

- The question here is whether the effect of X on Y has changed over time. For example, is education becoming more important in determining your income in Taiwan? Is the gender gap in income changing in Taiwan? What about the effect of number of children on women's sense of fulfillment – is that effect changing over time?

<u>Case 1: Panel data</u>. Suppose you have panel data for two points in time, and you are interested in the effect of a *time-varying* X. Then you can estimate the first-difference model ($\epsilon_i' = \epsilon_{i2} - \epsilon_{i1}$):

$$Y_{i2} - Y_{i1} = (\mathbf{X}_{i2}\,\beta_2 - \mathbf{X}_{i1}\,\beta_1) + (\mathbf{Z}_{i2}\,\beta_2^* - \mathbf{Z}_{i1}\,\beta_1^*) + \epsilon_i'$$

$$= (\mathbf{X}_{i2} - \mathbf{X}_{i1})\,\beta_2 + \mathbf{X}_{i1}\,(\beta_2 - \beta_1) + (\mathbf{Z}_{i2} - \mathbf{Z}_{i1})\,\beta_2^*$$

$$+ \mathbf{Z}_{i1}\,(\beta_2^* - \beta_1^*) + \epsilon_i'$$

Changing-parameter models

page 76

Recall that the Zs are unmeasured causes, so the model you actually estimate is:

$$Y_{i2} - Y_{i1} = (\mathbf{X}_{i2}\,\beta_2 - \mathbf{X}_{i1}\,\beta_1) + \epsilon$$
$$= (\mathbf{X}_{i2} - \mathbf{X}_{i1})\,\beta_2 + \mathbf{X}_{i1}\,(\beta_2 - \beta_1) + \epsilon$$

Note that the coefficients for $\mathbf{X}_{i1}$ provide the test for changing parameters – nonzero coefficients indicate that the effect of X on Y has changed.

Changing-parameter models

page 77

Case 2: Repeated cross-section data. For repeated C-S data you can test for changing parameters by including an interaction term for **X* time**. For two cross-sections of data the model is:

$$Y_{it} = \mathbf{X}_{it}\beta + \mathbf{X}_{i2}\gamma + \epsilon_{it},$$

where the subscript *t* refers to time 1 or time 2 (*t = 1, 2)* and $\mathbf{X}_{i2}$ denotes the value of the Xs at time 2 (the difference in intercepts is γ_0 -- first term in $\mathbf{X}_{i2}\gamma$). This model is the same form as ANCOVA model with interaction terms to test for parallel slopes across groups (here the "groups" are time 1 and time 2).

Changing-parameter models

page 78

The interaction term of interest is $\mathbf{X}_{i2.}$ Nonzero γ indicates changing effects. To see this more clearly, substitute t =1 and t=2 successively into the equation above: (why not just estimate the 2 eqns separately?)

$$Y_{i2} = \mathbf{X}_{i2}\,\beta + \mathbf{X}_{i2}\,\gamma + \epsilon_{i2}$$

$$= \mathbf{X}_{i2}\,(\beta + \gamma) + \epsilon_{i2} \quad \text{(note intercept in } \gamma\text{)}$$

$$Y_{i1} = \mathbf{X}_{i1}\,\beta + \epsilon_{i1}$$

By comparing the two equations we see that γ is the difference in the effect of X on Y at time 2 versus time 1.

Changing-parameter models

page 79

Case 3: Effect of time-invariant variable with panel data. What if you want to know whether the effect of gender on income has changed over time? The first-difference model "differences out" gender, so you need to use the interaction model in Case 2: you would include a dummy variable for time and gender, and an interaction term gender*time

Changing-parameter models

page 80

Extension: *Multilevel* models with *time as the context* (multiple data waves to examine the correlates of changing parameters)

- Limitations of the changing-parameter model: The model is designed to *detect* -- not explain -- change in individual-level effects. Also, the changing-parameters model requires just two surveys – an advantage if that is all the surveys you have, but a disadvantage if you have multiple surveys you want to exploit. The changing-parameters model quickly becomes unwieldy as we add waves of data. If we have (say) five Xs to test for varying effects for (say) 20 surveys, we would have 100 regression coefficients to examine. To summarize those 100 coefficients, we could correlate them with macro-level variables -- a step toward multilevel analysis.

Changing-parameter models

page 81

<u>What we want to know:</u>

What are the macro-changes that matter for the X-Y relationship, i.e., *what has changed over time to cause the X-Y relationship to change*? We might want to know, for example, why the gender gap in income is declining. Is it because more women are entering the labor force? Is it related to changes in unemployment rates? Or we might want to know why the effect of education on income has increased in the U.S.

Changing-parameter models

page 82

- These sorts of questions lead naturally to a multilevel model with *time-varying* aggregate-level variables as the contextual variables. Multilevel analysis was developed to study the effects of contexts, such as schools. To apply the method here we think of *time as a context*, so we can wheel in the multilevel machinery (e.g., HLM software) to apply to *change*. The multilevel approach inspired by the school effects literature can – with some modification (see DiPrete and Grusky AJS 1990, SM 1990) – be applied to the changing-parameter situation where it is time that forms the context.

Changing-parameter models

page 83

The General Model

- Here we focus on two levels of aggregation, where individuals are nested in time. (The model can be extended to more than two levels – for example, students nested in classrooms nested in schools nested in school districts = 4 levels. Most HLM analyses, however, use two levels.)

- The two-level model consists of a "level 1" (micro) equation and a "level 2" (macro) equation. To fix basic concepts, consider the simple case where there is one level 1 regressor (X) and two level 2 regressors (Z_1 and Z_2).

Changing-parameter models

page 84

<u>Level 1 equation.</u> The level 1 equation is (*i* indexes individual and *j* indexes group):

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + \epsilon_{ij}$$

- For βs, the first subscript identifies the type of parameter (0 for intercept, 1 for slope of #1 regressor, 2 for slope of #2 regressor, etc.) and the second subscript indexes group (here, survey). Observe that both the intercepts and the slopes have "j" subscripts, implying that both the intercepts and slopes are allowed to vary across the groups (here, this means they can vary from survey to survey). These varying intercepts and slopes form the elements to be modeled in the level 2 equations.

Changing-parameter models

page 85

- The model uses deviation scores on X to “center” on X. Since regressions go through the point X-mean, Y-mean, when X-mean = 0 the y-intercept is the Y-mean. We center the Xs on zero, then, so that β_{01} is the mean of Y for the first survey, β_{02} is the mean of Y for the second survey, and so on.

Changing-parameter models

page 86

<u>Level 2 equation.</u> We are modeling both intercepts and slopes, so we have an equation for each. For two Zs:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + \epsilon_{0j} \quad \text{(intercepts)}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}Z_{1j} + \gamma_{12}Z_{2j} + \epsilon_{1j} \quad \text{(slopes)}$$

As advertised, DVs in the level 2 equations are coefficients from the level 1 equations. The first equation models the level 1 *intercepts* as functions of the two Zs. The second equation models the level 1 *slopes* as functions of the two Zs.

Changing-parameter models

page 87

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + \epsilon_{0j} \quad \text{(intercepts)}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}Z_{1j} + \gamma_{12}Z_{2j} + \epsilon_{1j} \quad \text{(slopes)}$$

I use standard HLM notation:

- first subscript for gamma identifies the dependent variable (0 is intercept, 1 is first slope, 2 is second slope, etc.) – i.e., which level 1 coefficient is the DV?
- second subscript for gamma is the conventional notation of 0 for intercept *in that equation*, 1 for the slope Z_1, etc.

Changing-parameter models

Substantively, then:

- γ_{01} is the effect of contextual variable Z_1 on the Y-means across surveys
- γ_{11} is the effect of contextual variable Z_1 on the slopes, that is, it is Z_1's effect on the *effect of X on Y*. We want to see if variance in this effect (that is, change in this effect over time) is associated with change in macro-conditions Z over time.

The level 2 equations imply an error term that violates key assumptions of OLS, leading to biased significance tests (corrected by HLM software).

Changing-parameter models

Final note

Multilevel analysis of repeated surveys will become increasingly feasible as we continue to cumulate surveys over time. The sort of analysis describe here would not have been feasible a few decades ago, when repeated surveys did not have enough measurement points for the macro-variables. It is important to note, however, that the T is still relatively small, which limits the macro-equation to just a few variables. At the macro-level, then, our analyses still lack power.

Regression standardization

Regression Standardization Model for Decomposing Aggregate Change

- The issue here is how aggregate change– change in the mean of Y – arises from change in the *levels* and *effects* of the causes of Y. It is sometimes useful to use regression standardization to decompose the change in Y between two time points (as opposed to difference in Y across groups – the conventional use of regression standardization).

Regression standardization

page 91

- Regression standardization is commonly used to account for differences between groups – e.g., to see how much of the income difference between whites and disadvantaged groups in the U.S. can be accounted for by “resource” or “compositional” differences. Regression standardization provides a way to quantify how much of the difference in Y is due to differences in Xs.

$\Delta\beta X_1$-mean + $\beta_1 \Delta X$-mean + $\Delta\beta\Delta X$-mean + $\Delta\alpha$ page 92

- The same logic applies when we think of change in Y over time as opposed to differences in Y across groups. Using the subscripts "1" and "2" to denote *time 1* and *time 2* instead of *group 1* and *group 2*, we have the standard four-component model (for simplicity we assume a single X):

$$Y_2\text{-mean} - Y_1\text{-mean} = (\alpha_2 + \beta_2 X_2\text{-mean}) - (\alpha_1 + \beta_1 X_1\text{-mean})$$

$$= \Delta\beta X_1\text{-mean} + \beta_1 \Delta X\text{-mean} + \Delta\beta\Delta X\text{-mean} + \Delta\alpha$$

$\Delta\beta X_1$-mean + $\beta_1\Delta X$-mean + $\Delta\beta\Delta X$-mean + $\Delta\alpha$ page 93

where:

- $\Delta\beta X_1$-mean is the contribution of the change *in the effect of X*
- $\beta_1\Delta X$-mean is the contribution of *change in the level of X*
- $\Delta\beta\Delta X$-mean is the joint contribution of change *in the level and effect of X*
- $\Delta\alpha$ is the part of the change in Y that is not accounted for by change in either the level or effect of X.

$\Delta\beta X_1$-mean + $\beta_1 \Delta X$-mean + $\Delta\beta\Delta X$-mean + $\Delta\alpha$ page 94

Note: In the case where X has an arbitrary zero point – that is, where X is *not* a ratio variable – then $\Delta\alpha$ and $\Delta\beta$ cannot be uniquely separated, and the most we can conclude is that:

- $\beta_1 \Delta X$-mean is the contribution of change in the level of X
- $\Delta\beta\Delta X$-mean is the joint contribution of change in the level and effect of X
- $\Delta\beta X_1$-mean + $\Delta\alpha$ is the residual.

Convergence models

<u>Aggregate-level change: Convergence Models</u>

The question here is whether groups exhibit the same trend on some dependent variable Y over time. For example, is the trend in job satisfaction the same for men and women, for the old and the young, for salaried and nonsalaried workers? Convergence models will become increasingly useful in sociology as longitudinal data sets become more common.

Convergence models

Note possibilities:

- (1) No differences between groups--coincident trends
- (2) Stable differences -- parallel trends
- (3) Converging trends
- (4) Diverging trends
- (5) Crossover -- trend lines cross

Convergence models

Recall the changing-parameters interaction model. Like that model, here we can test for parallel slopes using an ANCOVA model with interaction terms. But the methods differ in the data setup. Here *the slopes are trends over time* (so they are based on T, not N, data points) and the groups are not time points but categories of discrete variables (men vs. women, workers vs. retirees, professional class vs. working class, etc.).

- Note: T (time) is measured using the measurement intervals in the data. Hence for an annual survey, time is in years; for a monthly survey, time is in months; and so on. When comparing trends for groups we are interested in groups that are mutually exclusive, so the i^{th} individual at time t belongs to one and only one group.

Convergence models

- **Model:** Interaction effects with Time and Group as covariates and Time*Group interaction term. We want to know whether the slope of Y over time differs across groups – in other words, will the slopes be the same when you regress Y on time within the various groups? So here group is the moderating variable, conditioning the effect of "time" on Y. Let G denote a dummy variable for group and T denote time. For M groups the model is:

$$Y_{it} = \alpha + \delta_1 G_{1it} + \delta_2 G_{2it} + \ldots + \delta_{M-1} G_{M-1,it} + \beta T_i + \gamma_1 G_{1i} x T_i + \gamma_2 G_{2i} x T_i + \ldots + \gamma_{M-1,i}\, G_{M-1,\,I}\, x T_i + \epsilon_{it}$$

Convergence models

$$Y_{it} = \alpha + \delta_1 G_{1it} + \delta_2 G_{2it} + \ldots + \delta_{M-1} G_{M-1,it} + \beta T_i + \gamma_1 G_{1i} x T_i + \gamma_2 G_{2i} x T_i + \ldots + \gamma_{M-1,i} G_{M-1,I} x T_i + \epsilon_{it}$$

- α is the y-intercept for the reference group (the mean value on Y for the reference group at $t = 0$).
- δs are differences in y-intercepts between the reference group and the other groups. **Suggestion:** Code time as $t = 0$ for the first measurement, so that the "main effects" for the groups (the δs) refer to initial group differences.
- β is the slope (here, *linear time trend*) for the reference group

Convergence models

$$Y_{it} = \alpha + \delta_1 G_{1it} + \delta_2 G_{2it} + \ldots + \delta_{M-1} G_{M-1,it} + \beta T_i + \gamma_1 G_{1i}{}^* T_i + \gamma_2 G_{2i}{}^* T_i + \ldots + \gamma_{M-1,i}\, G_{M-1,\,i}{}^* T_i + \epsilon_{it}$$

- γs are differences in linear trends between the reference group and the other groups
- γs and δs the same sign =>divergence (groups with higher intercepts also have steeper positive slopes, or less steep negative slopes; either way, initial differences are getting larger)
- γs and δ have different signs => convergence (or possibly crossover)

//End of session 5

Cohort replacment

Aggregate Level Change II: Cohort Replacement

- Where does social change come from? Here we want to know the proximate sources of aggregate social change. Does social change come about because individuals change, or because of population turnover? If we find, for example, that people in Taiwan are more likely to favor close relations with mainland China today than they were 30 years ago, is that because people have changed their minds, or because the new generations are more favorably disposed toward mainland China than their elders are?

Cohort replacement

- In usual discourse on the subject, people assume that societies change on Y because individuals in society have changed. In the U.S., for example, the "conservative turn" to Republican Party has been discussed in terms of "why are Americans becoming more politically conservative?" Social commentators ask "Why are people switching parties, from Democrat to Republican?" But that might be the wrong question: Due to cohort replacement, it's possible that % Republican increases when *no one switches political parties*.
- Often the first task, then, in trying to sort out the root causes of social change is to determine the proximate source of the change: are individuals actually changing, or is the social change due to cohort replacement?

Cohort replacement

- **Accounting Scheme** ("count your change" – James Davis)
- This scheme assumes repeated survey (*not panel*) data, so new cohorts are added to the sample as they are added to the population. With repeated survey data we can not follow individuals over time as they age, but we can follow *birth-cohorts* over time as they age. The cohort replacement accounting scheme focuses on within-cohort change (*intra*-cohort change) versus cohort differences (*inter*-cohort change). Cohort differences are necessary for cohort replacement effects: If cohorts did not differ on Y-mean, then Y-mean would not be affected as some cohort die out and are replaced by others.

Cohort replacement

- The scheme divides aggregate change into (1) the part due to cohort replacement and (2) the part due to average individual change. The part due to cohort replacement is determined by the size of cohort differences (for the cohorts rotating in versus those dying off) and the rate of cohort replacement. The part due to average individual change is determined by change over time within cohort, or the intra-cohort trend.

- We examine each component in turn.

Cohort replacement

<u>Societal change due to cohort replacement</u> (contribution of population turnover)

- Requires (1) cohort replacement (the process itself) and (2) cohort differences (if all cohorts have the same mean on Y, then cohort composition doesn't affect Y, so a change in cohort composition has no effect)
- By “cohort differences” we mean cohort differences on Y at a particular point in time. Cohort differences arise either from true cohort effects (persisting differences between birth cohorts over time), or from age effects (or both).

Cohort replacement

<u>Societal change due to individuals' change</u> (within-cohort change)

- Change in Y-mean for a cohort over time arises primarily from period effects and age effects. Change in the Y-mean within cohorts can also arise from within-cohort mortality or migration effects (for example, the mean education of a cohort could rise over time simply because the less-educated tend to die younger than the better-educated).

Cohort replacement

Linear decomposition method

- The objective is to decompose the change in Y-mean on the basis of average intra-cohort change over time and inter-cohort differences at a point in time. Suppose, as before, that we have repeated survey data, that is, T cross-sections of data, with N_t cases (individuals) at the t^{th} survey.

Cohort replacement page 108

- Step 1: Define the appropriate time interval for birth-cohort and date of survey.

- Note: We want to use a model that regresses Y on birth-cohort and date of survey, and for that regression it is convenient to use the same time interval for both. A 1-year interval is convenient, and actual birth-year provides more information than, for example, birth-cohort measured in 5-year categories. So the model below defines birth-cohort as *year of birth* and date of survey as *year of survey*. The same principles apply for narrower or wider intervals. Unless societal change is unusually rapid, though, one-year intervals should be appropriate for capturing the essence of the change.

Cohort replacement

- Step 2: Regress Y on individual's birth-year (cohort) and date of survey (year individual was surveyed):

$$Y_{it} = \beta_0 + \beta_1 \text{ Survey}_i + \beta_2 \text{ Cohort}_i + \epsilon_{it}$$

Cohort replacement

- Step 3: Use change in means of the regressors to estimate contributions of intracohort change and cohort replacement

Cohort replacement

<u>Societal change due to individuals' change</u>

- Example: net effect of party-switching on % Republican
- Note that β_1 is the average within-cohort slope: controlling for cohort (that is, for a given cohort), how much does Y change, on average, per year? If we multiply this average *annual* within-cohort change (recall that Survey is measured in years) by the number of years from the first survey to the last survey, then we have the estimated change in Y-mean due to within-cohort change. In other words:

  ΔY-mean due to within-cohort change = $\beta_1 \; \Delta$ Survey
- For example, the ΔY-mean due to within-cohort change from 1961-2000 = $\beta_1 \times 40$.

Cohort replacement

- Societal change due to cohort replacement
- Example: Effect on % Republican of old highly-Democratic cohorts being replaced by younger cohorts where % Republican is higher
- The key coefficient here is β_2, the weighted-average change in the Y-mean from one cohort to the next (the across-cohort slope, survey year constant). If we multiply this across-cohort slope by the change in average year of birth from the first to the last survey, we have the estimated change in the Y-mean due to cohort replacement (change in cohort composition):

ΔY-mean due to cohort replacement = β_2 Δ cohort mean

Cohort replacement

<u>Notes on the method</u>

- A number of studies have used the linear decomposition method with reasonable results. The key assumption is that the effects are linear and additive. We can check the plausibility of that assumption by summing the two components, $\beta_1\ \Delta$ Survey + $\beta_2\ \Delta$ cohort mean. Failure of the sum to approximate the total change in the Y-mean suggests that the assumptions of linear decomposition aren't met. (In the examples I've seen the sum is very often within $\pm$ 10% of observed aggregate change.)

Cohort replacement

- If the components do not sum approximately to the ΔY-mean, the problem could be survey-to-survey bounciness in the Y-means. In that case you might want to consider smoothing techniques for the overall Y-trend, such as smoothing by regression or by using a moving average.

- The method can be used for decomposing proportions and percentages, since proportions and percentages are means (e.g., proportions are means of variables coded as 0 and 1).

Cohort replacement

- Note that, to get the best estimates of β_1 and β_2, you use the data for all the surveys, not just for the first and last surveys. Consider, for example, this question from the U.S. GSS – "Generally speaking, men are better suited than women for politics." To decompose 1972-2004 change in Americans' attitudes on this issue, we should use GSS data for all the survey years, not just the data for 1972 and 2004.

Cohort replacement

Concluding remark about cohort replacement effects

- Where does social change come from? With repeated survey data, the first step in addressing that question very often is to decompose aggregate change into two components: the contribution of cohort replacement, and the contribution of intracohort change.

Cohort replacement page 117

- The accounting scheme here is just that – an accounting scheme. It does *not* locate the underlying causes of the change, but it *does* tell us whether the change lies in individual "conversion" (individuals actually changing) or in changing population composition (as older cohorts die off and are replaced by newer cohorts). That information is useful for suggesting where we might *look* for the underlying causes of the change, even though it does not actually identify those root causes

Cohort Analysis, 2

Extension: A Differencing Solution to Identification Problem in Cohort Analysis (works only in special cases – "natural experiments")

- Basic idea: Separate out cohort effects by looking for *cohort differences* in two groups who lived in the same place at the same time (=> we can assume common period and age effects for the two groups) *yet experienced different conditions during the formative period of their lives* (the key to cohort effects).

Cohort Analysis, 2

- Example: Cohort-specific sex differences in voting rates for people born before and after the 19th Amendment in the U.S. (to examine the permanent dampening effect of disenfranchisement on women's voting). Group comparison: men's voting rate versus women's voting rate after women are permitted to vote.

Cohort Analysis, 2

Model:

Let v_{ijk} = logit (proportion voting) at age i, in period j, from cohort k (where k = j - i). We have:

$$v_{ijk}M = \beta_i + \beta_j + \beta_k + \textit{covariates} \quad \textit{--for men}$$

$$v_{ijk}W = \beta_i + \beta_j + \beta_k + \phi_k + \textit{covariates} \quad \textit{--for women}$$

Cohort Analysis, 2

- The β's are the <u>common</u> age, period, and cohort effects for men and women; ϕ is the cohort difference in voting, women - men. This is the key to the model: The model avoids the identification problem by "differencing out" the common age, period, and cohort effects for men and women. To critics of this assumption: Except where men and women are analyzed separately, or modeled with interaction terms, that assumption is made *implicitly* anyway in cohort analyses of samples of men and women.

Cohort Analysis, 2

- Voting data are from the NES, 1952-1988. The hypothesis is that sex differences in voting will appear only for those cohorts who came of age before women were enfranchised -- so $\phi_k < 0$ for cohorts coming of age before 1920 and $\phi_k = 0$ for later cohorts.

- The dependent variable is the cohort-specific sex difference in voting:

  $v_{ijk}W - v_{ijk}M = \phi_k + covariates$

Cohort Analysis, 2

$v_{ijk}W - v_{ijk}M = \phi_k + covariates$

- We find, as expected, that women have lower voting rates than men ($\phi_k < 0$) *only for* the cohorts where men and women were socialized differently – strong evidence, we believe, for actual cohort effects.

Reference: Firebaugh, G. and Kevin Chen (1995). "Vote Turnout of Nineteenth Amendment Women: The Enduring Effect of Disenfranchisement." American Journal of Sociology 100 (January): 972-996.

Change across nations

IIB. Assessing change across nations

Problems in cross-country research (problematic for regression analysis especially): small N, unreliable data, multicollinearity, changing national boundaries (all well-known)

Change across nations

Principles in cross-country macro research

- Use regression diagnostics; consider counterfactual simulations – since one case can greatly influence results

- Countries are vastly different in population – we weight countries by their size when we want to weight each person the same, and we weight each country equally when we see each country as a "realization" of some underlying principle

Change across nations

- As we have seen, be wary of hybrid or "mixed models" (or "mixed-up models") with different types of variables (some levels, some ratios, some rates, etc.) – example of foreign investment/dependency research. **Important principle:** Rewrite your model in different algebraically-equivalent forms until you are confident that you understand it well.

- Relatively simple reduced form models might be the best we can do with cross-country regression currently (Charles Ragin will likely present another strategy).

Change across nations

<u>**Opportunities in cross-country *micro* research**</u>

Opportunity 1. There are a growing number of coordinated surveys across countries

- ISSP project linked to general social surveys in many countries (e.g., GSS in the U.S.)
- Eurobarometer surveys
- WVS (World Values Surveys)
- LIS (Luxembourg Income Survey, restricted to OECD countries)
- Others?

Change across nations

Opportunity 2. Many opportunities to use these data sets to apply the models above for assessing social change. For example:

- Compare cohort replacement effects across countries
- Compare changing-parameters models across countries
- Compare convergence of trends across countries (as more data points are collected)

Change across nations

Opportunity 3. Could use *multilevel* models with *country as the context*.

- That is, you have data for the i^{th} individual in the j^{th} country at time t. You could use multilevel methods with country as the context to determine the country-level traits that are associated with the relationship between X and Y within countries – for example, why the gender gap in income is greater in some countries. This analysis is straightforward when the relationship between X and Y is constant over time within countries. Then you have J country intercepts (β_{0j}) and JxP country slopes (a slope for each of the P regressors [Xs] at the individual level) to model in the level 2 equations.

Change across nations

- If the relationship between X and Y has changed, then you need to decide “the size of the relationship between X and Y” *when?* (probably at the initial point in time). Note that with J countries you are limited to J contexts. So, just like cross-country regression analysis, you need to worry about small N. Generally, then, you can use only a sparse model, with few variables at the country level.

Change across nations

- The analysis is more complicated when the relationship between X and Y has changed over time within countries. We might want to know, for example, why the effect of education on income has risen faster in some countries than in others. Here we want to know what country traits are associated with *changes* in the slopes. In that case we must model the J country intercepts, the JxP country slopes, and the JxP differences in country slopes. Unless we can disaggregate our countries into smaller regional units that make sense, probably the best we can hope to do is examine just P = 1 or 2 individual-level regressors.
- Note: there may be a problem of unbalanced datasets – countries might differ in the number of surveys they did, and *when* they did them.

Change across nations

Opportunity 4. As before, could use *multilevel* models with *time as the context* – extension of the changing-parameters model.

- This strategy becomes more feasible as we collect more waves of data (that is, as T grows). Here the objective is to explain the cross-section relationship between X and Y, and why it has changed (if it has). So we use time-varying macro-conditions or events to model the T intercepts and the TxP slopes (one for each X at each point in time).

Change across nations

- This differs from the multilevel time-context model discussed earlier in that here we have data for more than one country. This provides more power and flexibility, since we have more data. But it could lead also to unwieldy models if the relationship between X and Y differs across countries.

// End of session 6

Publishing your work (Day 4)

How to write for leading journals – things to do, things to avoid

- "The *American Sociological Review* publishes *original* (not previously published) works of interest to the discipline in general, new theoretical developments, results of qualitative or quantitative research that advance our understanding of fundamental social processes, and important methodological innovations. All areas of sociology are welcome. Emphasis is on exceptional quality and general interest." (from inside front cover of the *ASR*).
- The key words here are "quality" and "interest," and those are the concepts I will focus on here.

Publishing your work

Common Problems of Manuscripts Submitted to ASR

- This list is based on my own impressions based on what reviewers said and on my own reading of manuscripts. I focus here on the decisive problems – the sorts of problems that undermine a paper's publication prospects.

(I am interested in what others have to say as well.)

Publishing your work

page 136

1. **Framing (or lack thereof)**: The failure to tie research into a more general conversation on a topic of interest to sociologists.

- The problem could be that the issue really is unimportant – or not of general interest.
- In many instances, though, the author simply hasn't framed the issue properly. Framing often is decisive in whether a paper is published in a top-tier journal. It is the responsibility of the author – *not* the reader – to locate the research in an on-going academic discussion that people are interested in. *As an author, you must answer the "so what?" question.*

Publishing your work

page 137

- Example 1: I once received a manuscript at *ASR* on the symbolic significance of Christmas trees in America. I rejected the paper – not because it was about Christmas trees, but because it did not link that topic to any more general issue of sociological concern. *ASR* would be happy to publish an article on Christmas trees if that article successfully linked that topic to an important issue of on-going concern for sociologists.

Publishing your work

page 138

- Example 2: The inheritance of religiosity in America – grad student found that children of religious parents tended to be more religious themselves (as measured by church attendance, etc.) than children of non-religious parents. Not very surprising or interesting. But let's frame the issue this way: (1) Religiosity, like class, is inherited -- links religious inheritance to a huge ongoing literature on class inheritance. (2) Apply the mobility table methods we use for class to analysis of religious inheritance. (3) If we could come up with an appropriate metric, we could compare inheritance of religion with inheritance of class – which tends to be stronger? The finding no longer seems trivial, and it appeared in *ASR*.

Publishing your work

page 139

2. **Disjuncture between theory section and analysis.** Often the analysis appears to have little relevance at all to the issues raised in the introductory parts of the paper. You should be very specific about which results (e.g., coefficients) apply to which propositions in the first part of the paper. Sometimes hypotheses help with this – I don't think all research papers need a hypothesis section, but writing out one's hypotheses can be good discipline for you as author, even if you don't include them in your final draft. If you can't spell out your hypotheses, that indicates your thoughts are still fuzzy.

Publishing your work

page 140

3. Paper has no coherent point – it tells no story.

Sounds obvious, but some papers seem to have no point – or so many points that you can't keep track of them. Remember that you might have been thinking about your data and analysis for months or years, but your readers have not been. If you can't summarize your article in a paragraph or two, then you have more work to do. Perhaps your paper is too "busy" – it tries to do too much (reviewers very often told me that papers needed to be "streamlined").

Publishing your work

page 141

4. Poor writing.

I'm not referring here to bad grammar or poor English or Chinese – the sorts of things copyeditors can correct fairly easily. I refer instead to instances where individual sentences are hard to understand – if I need to read a given sentence more than once or twice to understand it, the author has done a poor job of writing, in my view. Sometimes individual sentences are crystal clear, but it isn't clear where the paper is going – there is no logical order to the presentation. The best articles appear to be effortless.