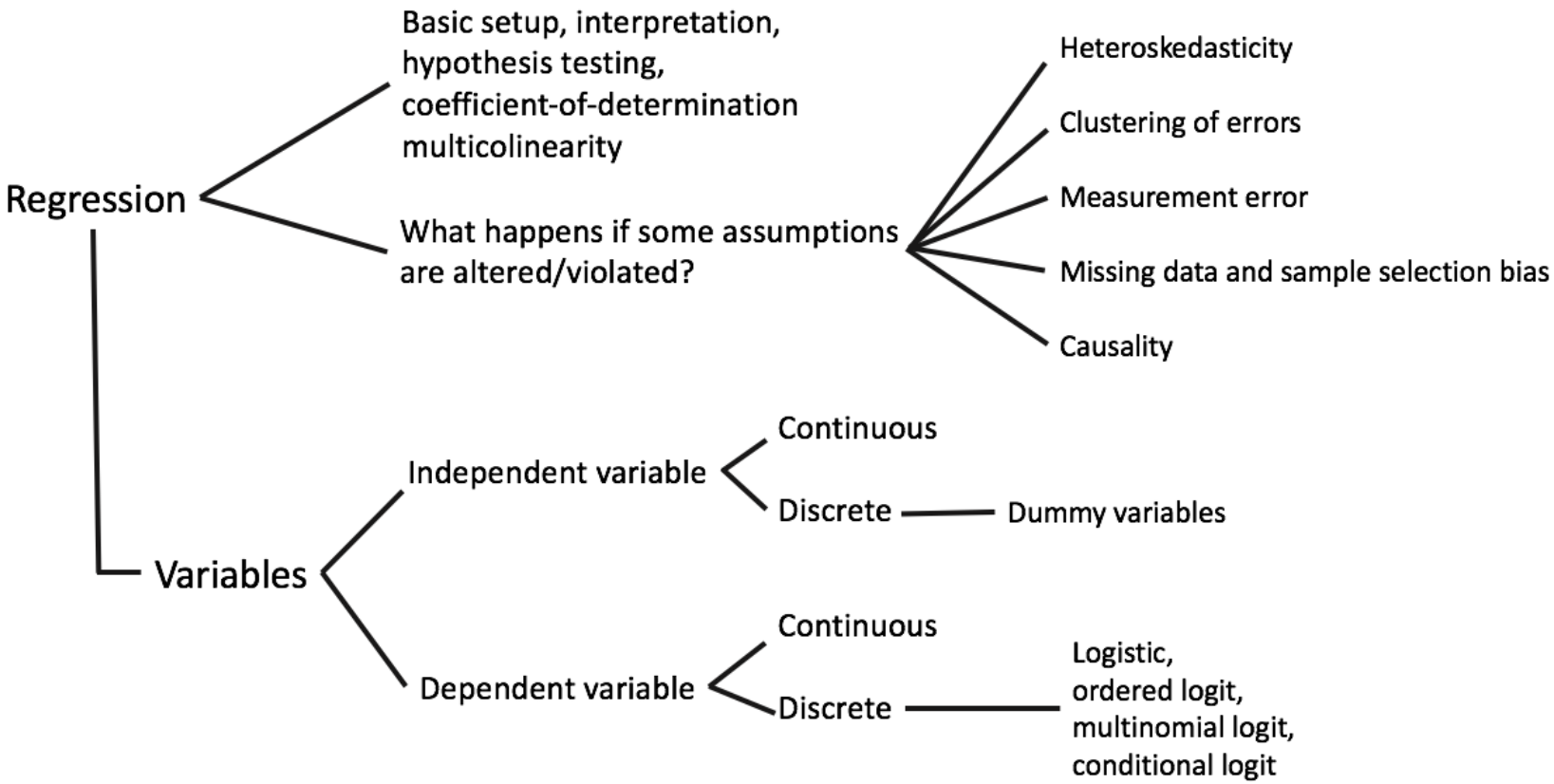
Introduction to Statistical Methods
SOC-GA 2332

Lecture 7: Regression Extended II

Siwei Cheng

NEW YORK UNIVERSITY

Roadmap


Regression
Basic setup, interpretation, hypothesis testing, coefficient-of-determination multicolinearity
What happens if some assumptions are altered/violated?
Heteroskedasticity
Clustering of errors
Measurement error
Missing data and sample selection bias
Causality
Variables
Independent variable
Continuous
Discrete
Dummy variables
Dependent variable
Continuous
Discrete
Logistic, ordered logit, multinomial logit, conditional logit

Lecture Outline

- Heteroskedasticity
- Clustering of errors
- Measurement error
- Missing data and sample selection bias

Heteroskedasticity and Clustering of Errors

- Consider again the multivariate regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \epsilon. \quad (1)$$

- When we set up our multivariate regression models, we made the assumption that the error terms are **identically** and **independently** distributed.
- Now, we are going to discuss what happens if these assumptions are violated.
- Heteroskedasticity occurs when the variance of the error term changes across different values of the explanatory variables (i.e. not identical). Clustering of errors is a special case of correlated errors (i.e. not independent).

Heteroskedasticity and Correlated Errors

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \epsilon. \quad (2)$$

- Under **homoskedasticity and no correlated errors**, the variance-covariance matrix of the residuals looks like this:

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 & ... & 0 \\ 0 & \sigma^2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \sigma^2 \end{bmatrix} = \sigma^2 I \quad (3)$$

Heteroskedasticity and Correlated Errors

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \epsilon. \tag{2}$$

- Under **homoskedasticity and no correlated errors**, the variance-covariance matrix of the residuals looks like this:

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 & ... & 0 \\ 0 & \sigma^2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \sigma^2 \end{bmatrix} = \sigma^2 I \tag{3}$$

- Under **heteroskedasticity**, it turns into:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & ... & 0 \\ 0 & \sigma_2^2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & & \sigma^2 \end{bmatrix} \tag{4}$$

Heteroskedasticity and Correlated Errors

- Under **heteroskedasticity and correlated errors**, it turns into:

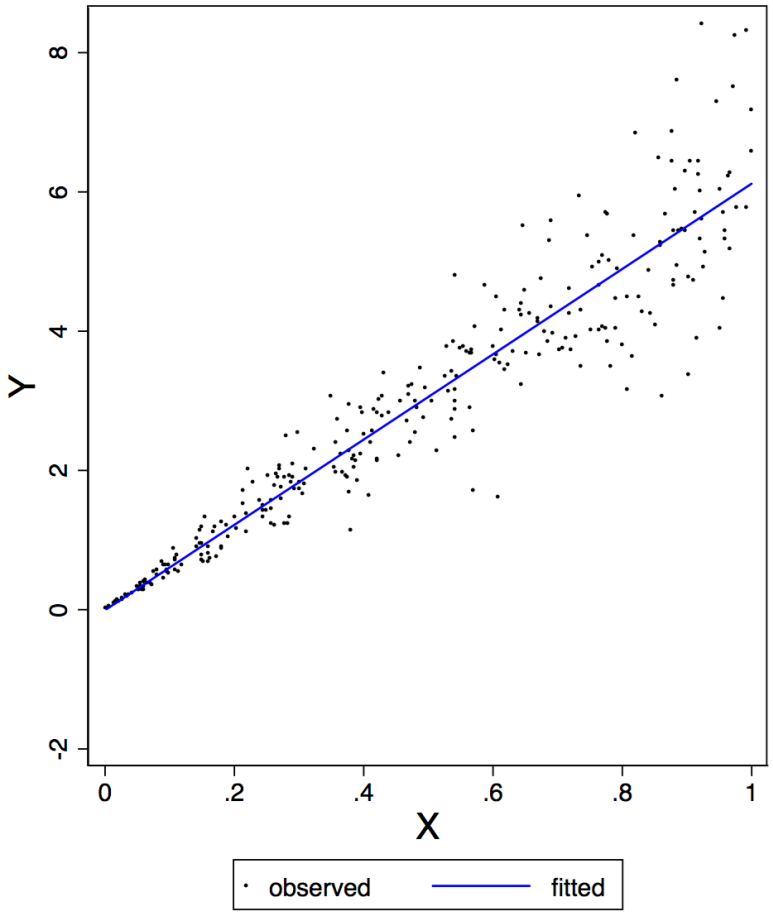
$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{1,2} & \dots & \sigma_{1,n} \\ \sigma_{2,1} & \sigma_2^2 & \dots & \sigma_{2,n} \\ \dots & \dots & \dots & \dots \\ \sigma_{n,1} & \sigma_{n,2} & \dots & \sigma_n^2 \end{bmatrix} \tag{5}$$

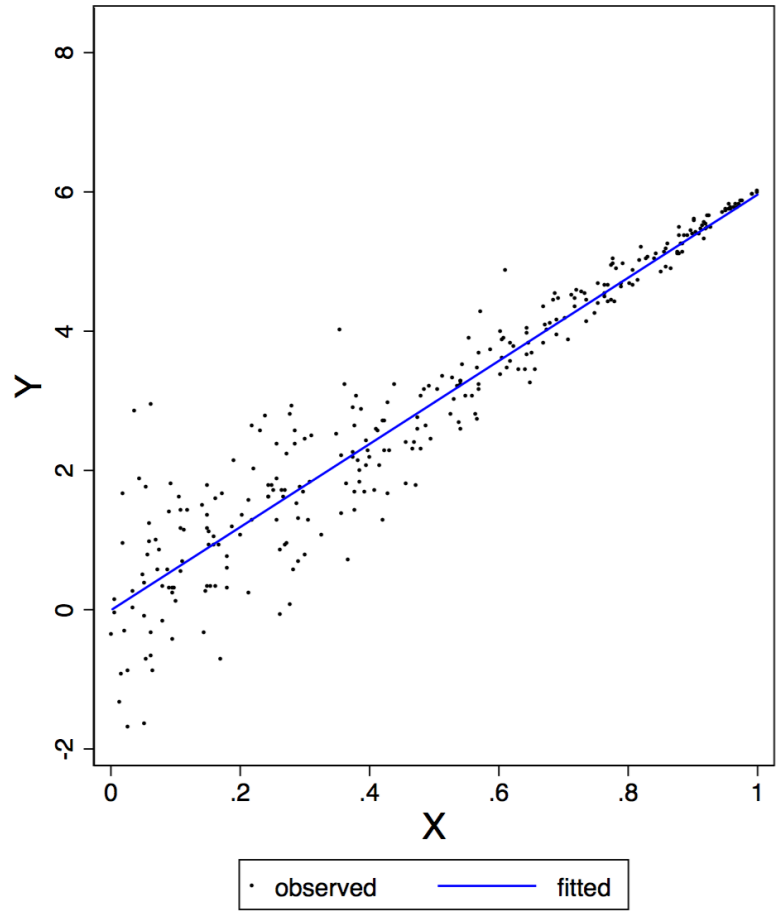
Heteroskedasticity

- Let's first focus on heteroskedasticity.
- In the basic assumptions of OLS, we assume that the variance of the error terms is constant across explanatory variables.
- However, in some cases, there are reasons to believe that the variance changes as the value of X changes.
- For example:
  1. If we regress wage on firm size, there are reasons to suspect that wages in larger/smaller firms are more volatile and thus have larger residual variance.
  2. If we regress income on age, the residual variance in income is typically larger among older workers than among younger workers.

Heteroskedasticity

- A graphic illustration of heteroskedasticity:





Heteroskedasticity

- How does heteroskedasticity affect our estimates from the regression model?
- **Good news** - heteroskedasticity does not cause bias or inconsistency in the OLS estimators of the $\beta's$.
- Quick question: Do you still remember what does "bias" mean in an estimator? What do we mean when we say an estimator is consistent?

Heteroskedasticity

- How does heteroskedasticity affect our estimates from the regression model?
- **Good news** - heteroskedasticity does not cause bias or inconsistency in the OLS estimators of the $\beta's$.
- Quick question: Do you still remember what does "bias" mean in an estimator? What do we mean when we say an estimator is consistent?
- **Another good news** - heteroskedasticity does not affect R^2. The intuition is that R^2 is about the "unconditional" variance in the population, and thus is unaffected by how the variance of the error term depends on X.

Heteroskedasticity

- So, why do we care about heteroskedasticity?
- Because heteroskedasticity will affect our estimated standard errors of β coefficients:
- In the presence of heteroskedasticity, the usual OLS t statistics do not follow t distributions, and the F statistics are no longer F distributed.
- In addition, the estimated standard errors may not be accurate - it can be an overestimate or underestimate.
- This means that our hypothesis testing and constructed confidence intervals may not be accurate.

Heteroskedasticity

- Diagnosing heteroskedasticity:
  1. Graphical approach: plot Y against X, or plot regression residuals $(y - \hat{y})$ against X.
  2. Statistical approach (for which we are skipping the details in this course)
- How do we deal with heteroskedasticity?
- There are statistical methods developed to adjust standard error to account for heteroskedasticity (in STATA, use the "robust" command; in R, try the sandwich package).

Heteroskedasticity

- But if we know the robust standard errors are valid more often than the usual OLS standard errors, why do we bother with the usual standard error at all?
- This is because when the homoskedasticity assumption holds and the errors are normally distributed, then the OLS t statistics will have (almost) exact t distributions, regardless of sample size.
- The robust standard errors are valid only when the sample size is large.

Heteroskedasticity

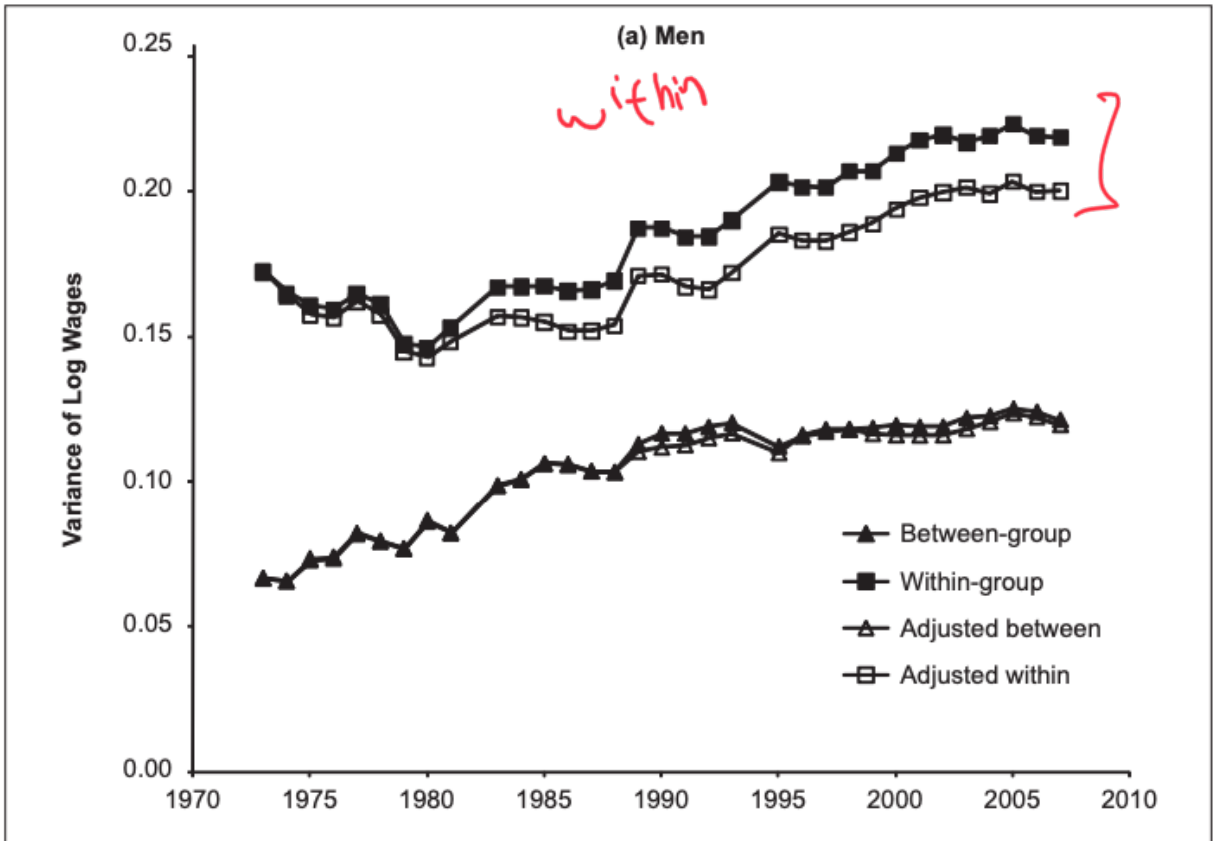
- In addition to its relevance to statistical inference, the pattern of heteroskedasticity itself can also be of interest to social scientists.

Heteroskedasticity

- In addition to its relevance to statistical inference, the pattern of heteroskedasticity itself can also be of interest to social scientists.
- Let's look at an example from a published research by Western et al. (2011). The authors examines the effect of unionization on wage inequality between union members and non-members, and also the extent to which the decline of unions explains the rise of wage inequality from 1973 to 2007.
- Their outcome variable is wage. We can think of "**within-group inequality**" as the variance of residuals in the wage distribution.
- They found that within-group inequality is greater among non-union members. → higher residual variance

Heteroskedasticity

- Below is a panel from their Figure 6 (fixing union membership at year 1973):




©Warren Photographic

Clustering of Errors

Clustering of Errors

- One of the basic OLS assumptions is that the error term is independently distributed across observations. i.e.:

$$Corr(\epsilon_i, \epsilon_j) = 0 \tag{6}$$

- But in the real world, we can think of cases in which the error terms are correlated within clusters of observations.
- For example:
  - repeated observations of the same individual
  - samples of husbands and wives
  - samples containing siblings
  - samples containing multiple respondents' from the same neighborhood/school/company

Clustering of Errors

- Like heteroskedasticity, clustered error terms will not cause the estimated coefficients to be biased.
- However, when the errors are correlated but we do not take this into account, then our OLS standard errors of the coefficients are typically too small.
- Intuitively, if the error terms are correlated, then each observation in our sample contains **less unique information** (because observations in the same cluster "share" some common variation).
- Hence, if we use the OLS standard errors to make the inference, then we are more likely to reject the null hypothesis than we should be.

Clustering of Errors

- So, how to diagnose clustered errors?
- Typically, we simply consider the possibility of clustered errors and take steps to adjust for it, as opposed to actually trying to diagnose it.
- It is possible to calculate what is called the "intra-cluster correlation," which tells the proportion of overall variance that is within clusters.
- To see this, let's consider a simple linear regression model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{7}$$

Clustering of Errors

- Suppose we know that the error term for each individual ϵ_i is "nested" within their schools. Now, we capture this "school effect" by decomposing ϵ_i into two parts: a school-specific effect η_s and a "pure" individual effect u_{si}.
- This is equivalent to writing Equation (7) as:

$$y_i = \beta_0 + \beta_1 x_i + \underbrace{\eta_s + u_{si}}_{\epsilon_i}, \tag{8}$$

where:

$$\epsilon_i = \eta_s + u_{si}. \tag{9}$$

Clustering of Errors

- Writing the equation this way will help us decompose the variance in ϵ_i into two parts:

$$Var(\epsilon_i) = Var(\eta_s) + Var(u_{si}). \tag{10}$$

- The intra-cluster correlation is defined as:

$$\rho_s = \frac{Var(\eta_s)}{Var(\eta_s) + Var(u_{si})}. \tag{11}$$

- ρ_s captures the ratio of the covariance between two individuals in the same school and each individual's total variance.
- Intuitively, we can think of this intra-class correlation as capturing how much of the variation in the error term on the individual level can be explained by the fact that they come from different schools.

Measurement Error

- Next, we discuss dealing with measurement errors in the dependent and independent variables.
- Let's start with the measurement error in the **dependent** variable Y.

Measurement Error in the Dependent Variable

- The measurement error in Y is defined as the difference between the observed value (Y) and the actual value (Y^*):

$$e_y = Y - Y^* \tag{12}$$

- Because the relationship between two variables should be relationship about the "true values," we know that the "true equation" should be:

$$Y^* = \beta_0 + \beta_1 X + \epsilon. \tag{13}$$

- In reality we only observe Y. Plug in Equation (12) we have:

$$Y = \beta_0 + \beta_1 X + \epsilon + e_y. \tag{14}$$

Measurement Error in the Dependent Variable

$$Y = \beta_0 + \beta_1 X + \underbrace{\epsilon + e_y}_{\epsilon^*}. \tag{15}$$

- Now we see that the measurement error gets "absorbed" in the new residual ϵ^*. So we can just ignore the fact that Y is an imperfect measure of Y and proceed as usual.
- As long as the measurement error in Y is not related to the explanatory variable X, the measurement error in Y will not affect the unbiasedness and consistency of the OLS estimator.
- But because the error term contains a "measurement error of Y" part, we expect that large measurement error in Y will lead to **large standard errors** in the estimated coefficients.

Measurement Error in the Independent Variable

- Let's move on the measurement error in an explanatory variable.
- In fact,measurement errors in explanatory variables cause a much more important concern in regression analysis.
- Let's consider a "true" regression model:

$$Y = \beta_0 + \beta_1 X^* + \epsilon. \tag{16}$$

- Let's assume that there is a measurement error in X:

$$e_x = X - X^* \tag{17}$$

- Let's further assume that the measurement error is uncorrelated with the true value of X:

$$Cov(e_x, X^*) = 0 \tag{18}$$

- It thus follows that $Var(X) = Var(X^*) + Var(e_x)$.

Measurement Error in the Independent Variable

- Now, let's plug Equation (17) into Equation (16):

$$Y = \beta_0 + \beta_1(X - e_x) + \epsilon. \tag{19}$$
$$= \beta_0 + \beta_1 X + (\epsilon - \beta_1 e_x) \tag{20}$$

- As you can see, the measurement error becomes part of the error term in the equation.

Measurement Error in the Independent Variable

- What would become the OLS estimator for β_1 in this case?

$$\begin{aligned}\hat{\beta}_1 = \frac{Cov(Y,X)}{Var(X)} &= \frac{Cov(\beta_0 + \beta_1 X + (\epsilon - \beta_1 e_x), X)}{Var(X)} \\ &= \frac{\beta_1 Var(X) - \beta_1 Cov(X, e_x)}{Var(X)} \\ &= \beta_1 \cdot (1 - \frac{Cov(X, e_x)}{Var(X^*) + Var(e_x)})\end{aligned} \tag{21}$$

- Now, recall that $Cov(X, e_x) = Cov(X^* + e_x, e_x) = Var(e_x)$.
- It thus follows that:

$$\hat{\beta}_1 = \beta_1 \cdot (1 - \frac{Var(e_x)}{Var(X^*) + Var(e_x)}) \tag{22}$$

Measurement Error in the Independent Variable

- Or, equivalently:

$$\hat{\beta}_1 = \beta_1 \cdot \frac{Var(X^*)}{Var(X^*) + Var(e_x)} \tag{23}$$

- What does this mean? This means that the "estimated" $\hat{\beta}_1$ is always closer to zero than is the true β_1!
- This is called the **"attenuation bias"** in the OLS estimator due to the measurement error in the explanatory variable.
- $\frac{Var(X^*)}{Var(X^*)+Var(e_x)}$ can be seen as the "signal-to-noise" ratio.
- If the variance of X is very large relative to the variance in the measurment error, then the inconsistency in OLS will be small.


©Warren Photographic

Measurement Error in the Independent Variable

- Let's consider a real empirical example of estimating how father's income influence son's income.

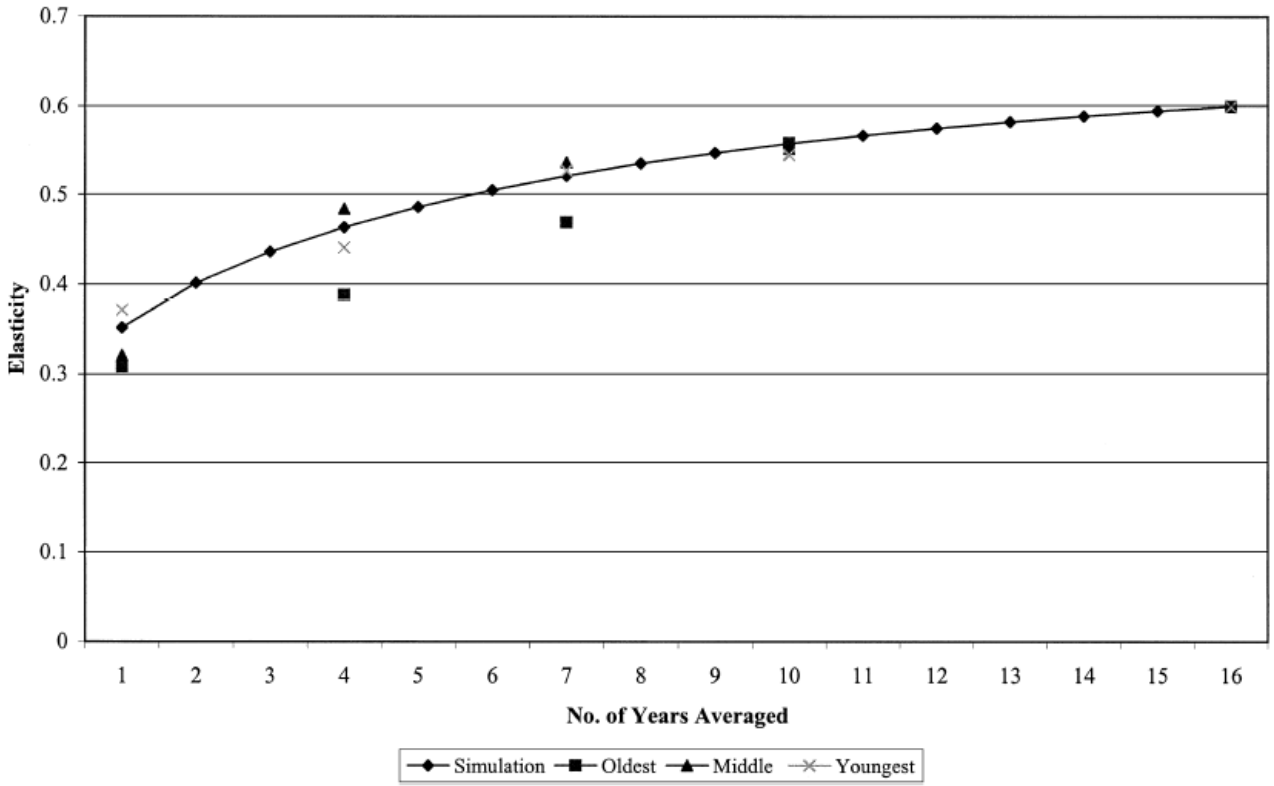
$$INC^{son} = \beta_0 + \beta_1 INC^{father} + \epsilon. \tag{24}$$

- Possible sources of measurement error in father's income:
  (1) income reporting error in survey data.
  (2) volatility in income at a single year.
  (3) income coding error by the survey interviewer.
- The larger the measurement error in INC^{father}, the greater attenuation of the estimated β_1 towards zero.

Measurement Error

- One of the strategy that researchers in this field adopt is to make use of the longitudinal income records from administrative data and take a multi-year average to reduce measurement error. See this graph in Mazumder (2005):

FIGURE 4.—SIMULATION AND ACTUAL ESTIMATES FROM AVERAGING FATHERS' EARNINGS



Missing Data and Sample Selection Bias

Missing Data and Sample Selection Bias

- A missing data problem can arise when one or some of the variables are missing for some units in a sample.
- Missing data may result from restrictions in the data collection process, from respondents' refusal to answer, or simply because the respondent do not know the answer two a certain question.
- When we work with data, we keep track of the observations that have missing values

Missing Data and Sample Selection Bias

- For example:

earnrs	earthsun	educ	effctsup	elecfair	electron	elecvote
2	earth a…	16	.i	somewha…	false	very ho…
2	earth a…	13	.i	very fa…	true	very ho…
0	earth a…	15	.i	.i	false	.i
2	.i	12	very	very fa…	.i	neither…
0	sun aro…	14	.i	neither…	true	somewha…
0	earth a…	12	.i	.i	.d	.i
2	.i	16	a little	.i	.i	.i
1	earth a…	16	.i	very fa…	false	very ho…
1	.i	14	.i	somewha…	.i	somewha…
1	.i	13	extreme…	somewha…	.i	very di…
1	.i	14	.i	.i	.i	.i
1	earth a…	6	not at …	.i	true	.i
1	.i	12	somewhat	.i	.i	.i
2	.i	14	a little	.i	.i	.i
0	earth a…	11	.i	.i	false	.i
1	.i	16	.i	.i	.i	.i
2	sun aro…	15	very	neither…	true	somewha…
1	earth a…	12	.i	neither…	.d	somewha…
2	.d	16	.i	.i	.d	.i
1	earth a…	18	a little	very fa…	.d	very ho…
1	earth a…	20	.i	very fa…	true	somewha…

Missing Data and Sample Selection Bias

- One option is to to estimate models using observations with a complete set of data on the outcome and explanatory variables. This is called a "complete case estimator."
- In the language of the missing data literature, if the data are missing completely at random (**MCAR**), then missing data cause no statistical problems (except that the sample sizes will be smaller due to the deletion of cases).
- When the cases are not missing at random, the estimations may be biased with regard to the population parameters.

Missing Data and Sample Selection Bias

- For example, suppose those with higher income are likely to refuse to provide their income information in a survey. How will this non-random missing data affect the estimated relationship between education and income?

Sample Attrition in PSID Data

Table 4
1968 Characteristics by Attrition Status: Female Heads, Age 25–64

	Always In	Ever Out	Ever Out/ Not Dead	Ever Out/ Dead
Welfare participation (%)	4.3	10.5*	10.0	17.9
Marital status (%):				
Married	1.4	1.8	1.7	2.5
Never married	21.2	14.6*	14.7	13.1
Widowed	38.7	39.1	39.1	39.0
Divorced/separated	36.7	40.8	40.6	43.8
Percent with annual hours worked > 0	80.4	67.4*	67.0	73.7
Annual labor income	8,199	6,950	7,167	3,482
Annual labor income for those w/income > 0	10,214	10,296	10,679	4,723
Annual hours worked for those w/hours > 0	1,593	1,645	1,676	1,203
Variance of log annual labor income for those w/income > 0	1.426	1.185	1.045	1.739
Labor income quintile ratios for those w/labor income > 0:				
Quintile 20/median	.316	.424	.471	.438
Quintile 40/median	.737	.800	.838	.653
Quintile 60/median	1.163	1.178	1.178	2.483
Quintile 80/median	1.553	1.468	1.440	5.724

Sample Attrition in PSID Data

Education (%)				
< 12	45.1	49.2	46.8	88.4
12	28.3	32.4	33.7	11.6
12–15	13.8	9.6*	10.2	0.00
16+	12.8	8.8*	9.3	0.00
Race (%)				
White	80.3	76.0*	77.3	55.4
Black	18.8	23.2*	21.9	44.6
Region (%)				
Northeast	25.2	26.2	26.3	24.8
North Central	30.0	24.6*	25.6	9.3
South	25.8	27.7	25.9	57.5
West	19.0	21.4	22.2	8.4
Age	44.9	47.4*	47.2	50.4
Tenure (%):				
Own home	45.0	40.3*	40.3	40.7
Rent	50.3	55.9*	55.8	58.2
Number of children in family	1.3	1.0	1.0	1.8
Sample size	502	526	475	51

Sample Attrition in PSID Data (Dynamic Attrition)

Table 16

Dynamic Attrition Models with Focus on Lagged Earnings (Logit Coefficients)

	Males					Females
	(1)	(2)	(3)	(4)	(5)	(6)
$\bar{y}$	−.20*	−.24*	−.28*	−.26*	−.07	.23
	(.07)	(.08)	(.08)	(.08)	(.09)	(.14)
y_{t-1}	−.22*	−.17*	−.18*	−.20*	−.15*	−.11
	(.06)	(.08)	(.06)	(.06)	(.07)	(.11)
y_{t-2}	—	−.09	—	—	—	—
		(.09)				
Var (y)	—	—	.32*	.33*	.38*	−.04
			(.09)	(.09)	(.09)	(.23)
Time dummies and age	*n*	*n*	*n*	*y*	*y*	*y*
Other characteristics[a]	*n*	*n*	*n*	*n*	*y*	*y*
Pseudo R^2	.018	.017	.020	.025	.043	.018

Missing Data and Sample Selection Bias

- Missing data can also occur when we **define** certain variables in a way that render part of the sample to be "missing."
- The selection of sample will be particularly problematic when the cases that have missing values are likely to be systematically different from the non-missing cases. This is also called sample selection bias.
- Let's consider an example of "sample selection" in Western and Pettit (2005). The article addresses the problem of the "selectivity" of being employed among blacks and whites.
- A person's wage will be **missing** if the person is **not employed**. But we also know that blacks and whites have *different* employment rates.

Missing Data and Sample Selection Bias

- See the authors' illustration below:

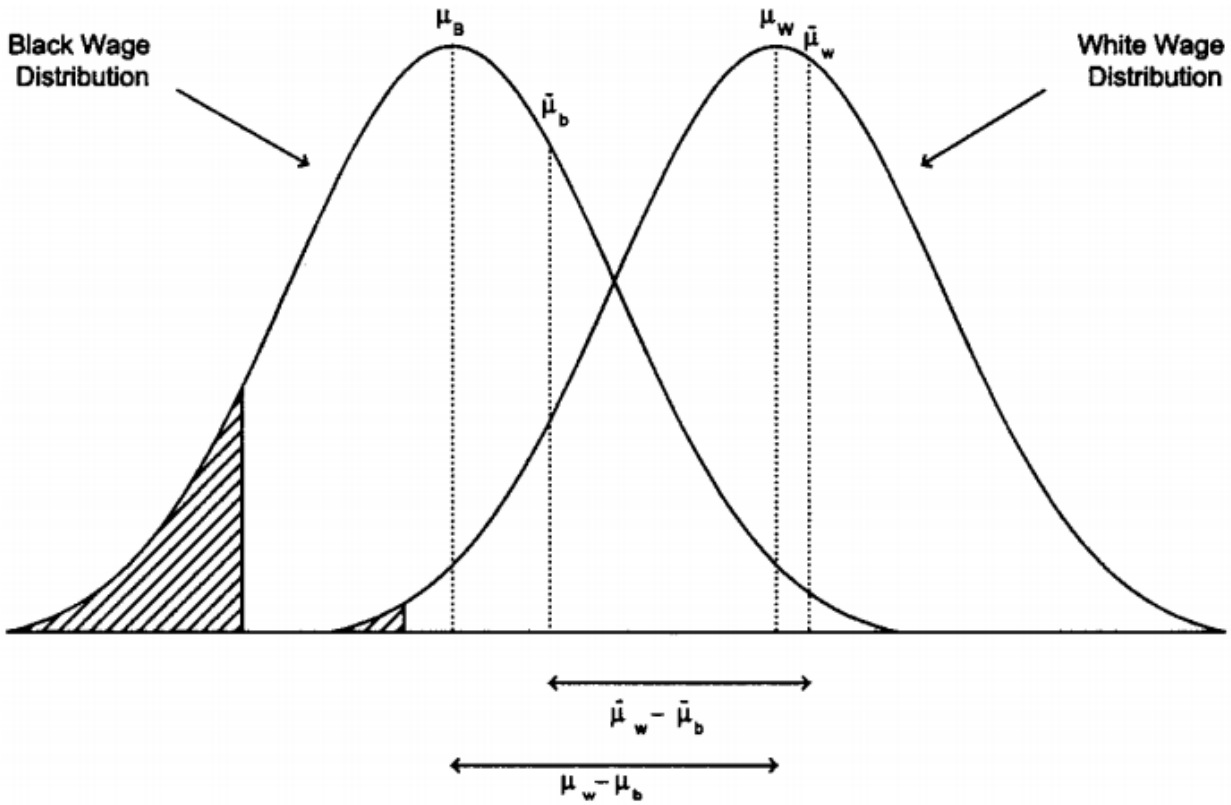


FIG. 1.—Hypothetical black and white wage distributions (shaded areas indicate wages unobserved because of joblessness; means estimated from the observed wage distribution are given by $\bar{\mu}$; means of the complete wage distributions are given by μ).

Missing Data and Sample Selection Bias

- Moreover, the selectivity of black and white workers into the labor force may have varied over time. Noting this potential problem, the researchers then corrected for the selectivity of being employed by trying to create a "counterfactual" scenario in which the jobless individuals were working:

Missing Data and Sample Selection Bias

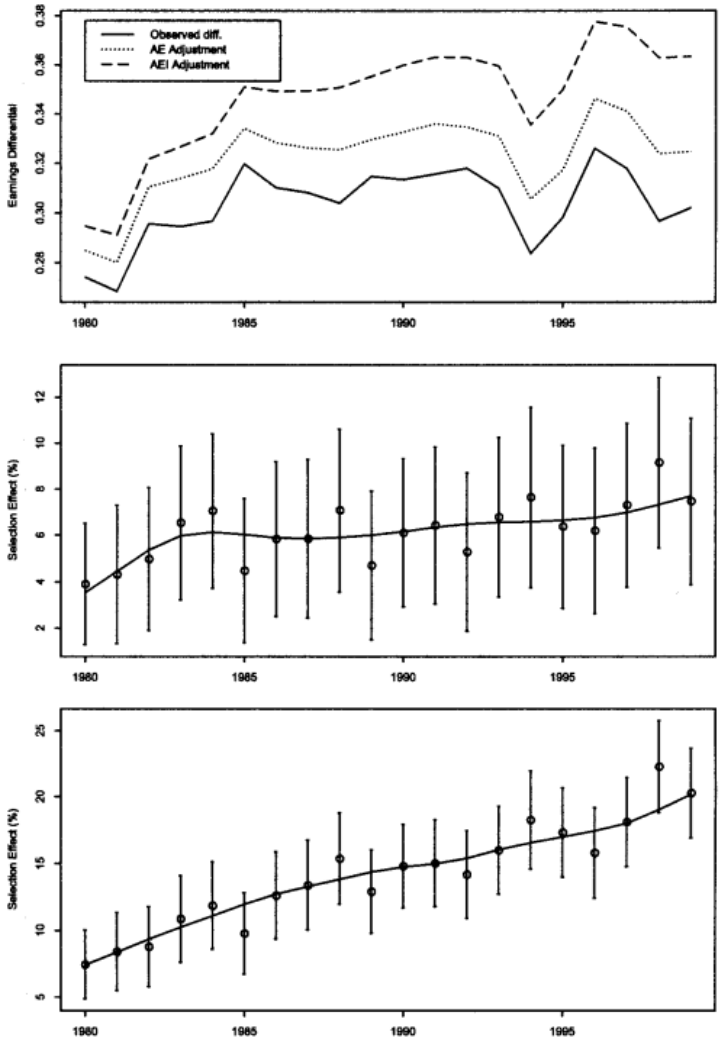


FIG. 2.—The white-black difference in mean log hourly wages, men ages 22–64, 1980–99. *Top panel:* observed earnings differential, $\bar{d}$, and adjusted differentials, $\hat{d}$; *middle panel:* selection effect using the AE adjustment; *bottom panel:* selection effect using the AEI adjustment. A smooth line indicates the trend. The selection effect and 95% confidence intervals are measured as a percentage of the observed difference, $\bar{d}$.

Missing Data and Sample Selection Bias

- More recently, Cheng (2020) proposed a relative density based methods for combining information across different data sets to address this employment selection issue: