Applied Econometrics for Practitioners:
Using Household Surveys to Inform Policy in Nepal

Dean Jolliffe, Hiroki Uematsu, Ganesh Thapa

Kathmandu, Nepal

1/19/2018

Course Description

- Topics: Review OLS, sample design and weights, limited dependent variables, instrumental variables, censored dependent variables, quantile regression, bootstrap, panel data estimation.

My Bio

- *Research Areas:* Development, Transition, U.S.
- *Research Topics*: Poverty, Inequality and Health (U.S., Egypt, Mozambique, El Salvador), Economics of Education (Ghana, Hungary, Bulgaria), Discrimination (CEE)
- *Education*: Ph.D. in Economics from Princeton
- *Professional Experience*: Economic Research Service (ERS, USDA), World Bank (LSMS), International Food Policy Research Institute (IFPRI), Center for Economic Research and Graduate Education (CERGE),

Hiroki Uematsu

- *Research Areas*: Development Economics, Agricultural Economics
- *Research Topics*: Poverty Measurement and Analysis
- *Education*: Ph.D. in Agricultural Economics from Louisiana State University
- *Professional Experience*: World Bank (Poverty & Equity since 2012)

Ganesh Thapa

- *Research Areas:* Development, Food security, Nepal
- *Research Topics*: Poverty, Child Nutrition Outcomes, Agricultural Diversification, Food Safety
- *Education*: Ph.D. in Agriculture Economics from Purdue University
- *Professional Experience*: International Food Policy Research Institute (IFPRI), World Food Program (WFP), Feed the Future Nutrition Innovation Lab (FtF-NIL)

Course Materials

- Primary text: Wooldridge
- Supplemental chapters from: Deaton
- We’ll handout hardcopy of lecture notes, but will not make them available online or email them.
- Stata examples will also be in the lecture notes.

Certificate of course completion

- Participate in course lectures
- Complete all the problem sets by due date
- Certificate distributed at the end of the course

Problem Sets

- 4 problem sets will be emailed to the participants. All will be data exercises in Stata (version 12)
- Problem set 1 will be emailed this afternoon. If you do not receive this, contact Ganesh.
- The problem set is due at the beginning of the lecture on Jan 26. Please hand in hardcopy. (OPEN)
- P Sets 2-4 not yet available, but soon will be.
- TA will grade problem sets. No late problem sets accepted. (I will want to discuss in class.)

Econometrics and Statistics

- Experimental vs. Nonexperimental (obs) data
- Exp: Medical study with treatment & control
- Nonexp: regress Food Insecurity on Food Stamps (+CPS, why?); NLSS, poverty and education.

Consider comparison of means

- Many variations in between the two examples
- Usually interested in causality. Why?

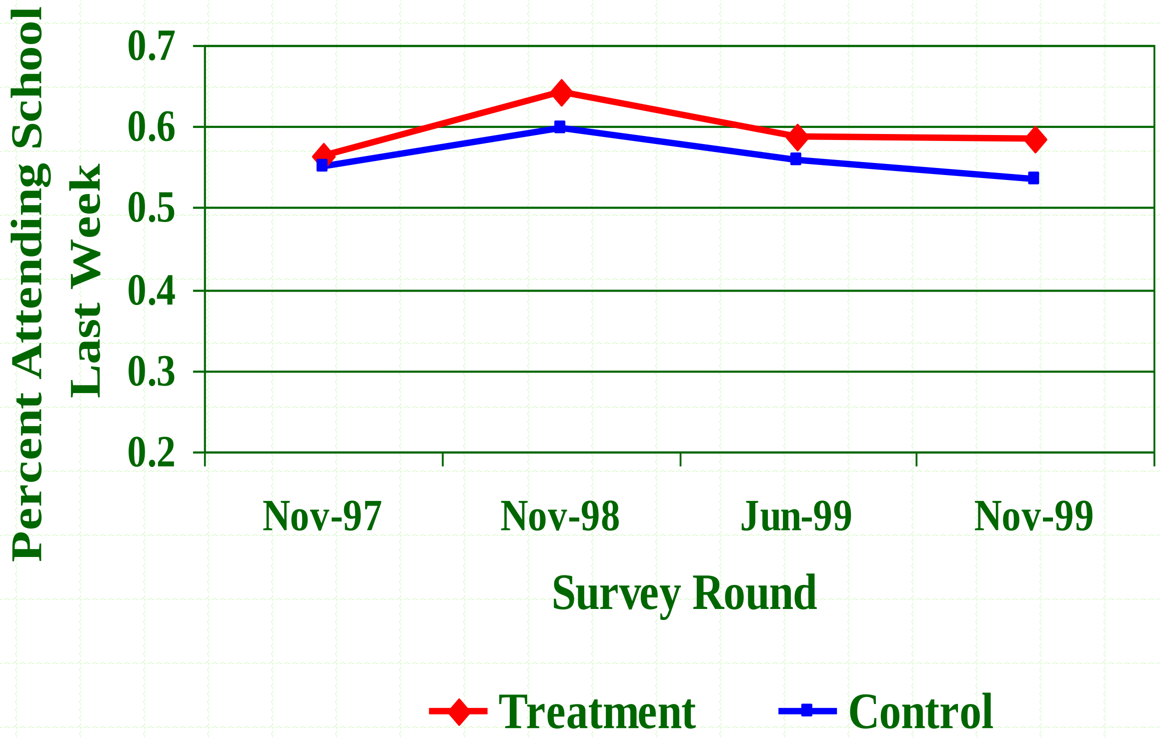
  (Can you think of exceptions to this?)

Experimental Design in the Social Sciences - Quasi Experimental

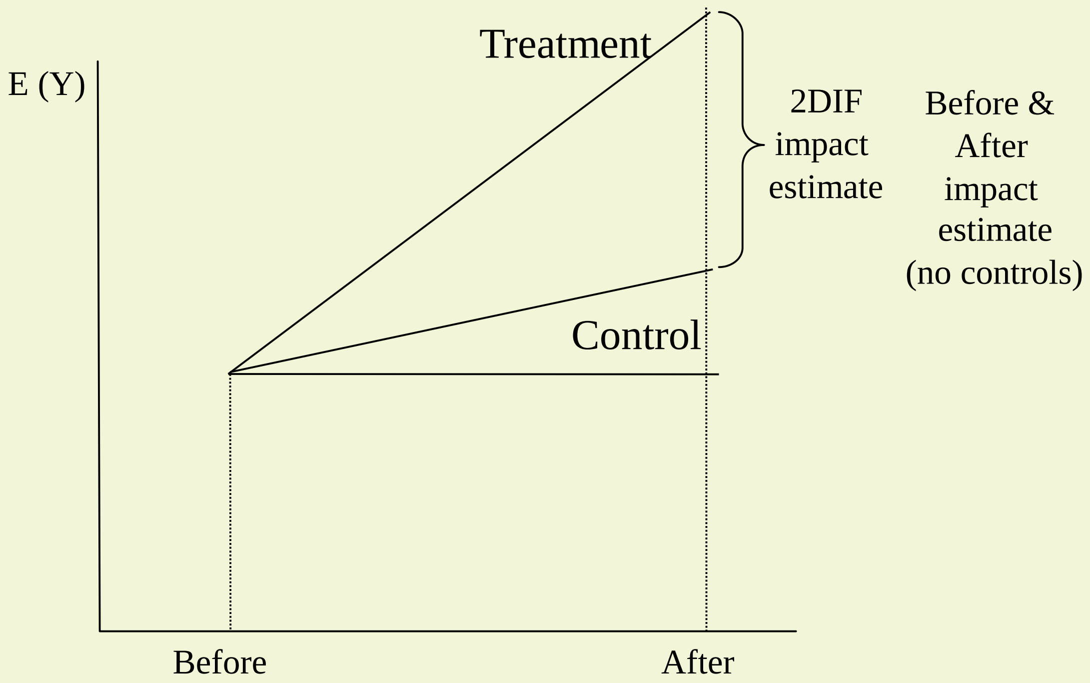
- PROGRESA / Oportunidades in Mexico
  http://www.ifpri.org/data/mexico01.asp
- Program objectives: poverty alleviation. Long term: through human capital investment (educ, health, nutrition); Short: cash transfer.
- Quasi-experimental: Program randomized at the locality level (baseline in ‘97)
- Sample of 506 localities (24,077 Hholds)
  - 186 control (no PROGRESA)
  - 320 treatment (PROGRESA)

Show Skoufias slides 29 & 30

Percent of Boys Attending School: 12-17 years old



If only observed treatment, what would we infer? How wrong would we be?


Treatment
E (Y)
2DIF
impact
estimate
Before &
After
impact
estimate
(no controls)
Control
Before
After

Experimental Design in the Social Sciences - Quasi Experimental

- World Bank Evaluation. Improving Primary Education in Kenya
- Targeting of school assistance programs typically not random (common factors: political, accessible, likelihood of success)
- 100 Schools randomly selected. In 1996, 25 rec'd textbooks; '97, 25 rec'd block grants.
- Preliminary: Impact of textbooks not as strong as some previous studies have indicated.

Non-experimental Design
Causality & Multivariate Regression

- Bivariate regression establishes correlation.
- Typically hope to say something about causality.
- Try to control for all relevant variables. Keep in mind treatment vs. control
- If we've truly controlled for enough other variables, then the estimated ceteris paribus effect can sometimes be considered to be causal

- Read short articles on smoking

The Linear Regression Model

$$y = \beta_0 + \beta_1 x_1 + u$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$$

Some Terminology

- In the simple linear regression model, where $y = b_0 + b_1x + u$, we typically refer to y as the
  - Dependent Variable, or
  - Left-Hand Side Variable, or
  - Explained Variable, or
  - Regressand
- Regress as a verb: regress y on x.

Some Terminology, cont.

- In the simple linear regression of y on x, we typically refer to x as the
  - Independent Variable, or
  - Right-Hand Side Variable, or
  - Explanatory Variable, or
  - Regressor, or
  - Covariate, or
  - Control Variables
  - Design Matrix

Simple Linear Regression (SLR) Assumptions 1-5, as in Wooldridge (JW).

- SLR1: Linear in *parameters*.
- In the population there is some linear relationship between the dependent variable and the independent variables (or transformations of these vbls).
- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$
- *Innocuous Assumption?*
- *Implicit Assumption*: We know the functional form of the population model and can measure the X's.

JW's SLR2: Random Sampling

- Population consists of N sets of (x_i,y_i)
- Sample is n draws of (x_i,y_i) from population N.

  N/n is the average expansion or raising factor.
- Simple Random Sample (*SRS*): Every element in the population has equal chance of selection.
- Complex Random Sample: every observation in the population (or universe) has some *known*, positive probability of selection.
- *Innocuous Assumption?*
- *Policy relevance?* ***(sample => population)***
- *Footnote: What is p=0? Not a focus of JW*

Examples of Violations of SLR2: Random Sampling

- The Easy Ones: Web polls, CNN, Amer. Idol
- Most common mistake: arbitrary is not random. Papers will frequently assert 'random' but provide no evidence. Be wary of assertions of 'random'.
- Somewhat subtle: Labor Force Surveys. Random draw of wage earners, but often used in empirical analysis to infer population characteristics.
- The sample frame defines the population. Weighted sample statistics provide population estimates.

CPS: www.census.gov/cps/ (who's excluded from poverty estimates?)

PSID: psidonline.isr.umich.edu/ (panel started in 1968, who's excluded?)

SPD: www.bls.census.gov/spd/ (shorter panels, but high attrition)

SLR3: Zero Conditional Mean

- $E(u|x) = 0$. (JW's critical assumption)
- If u and x are independent, then $E(u|x) = E(u)$, similarly $cov(u,x)=0$ and $corr(u,x)=0$.
- Knowing something about x does not give us any information about u, so that they are completely unrelated.

- Note: x & z independent <=> $E(xz)=E(x)E(z)$

Example of Violation of Zero Conditional Mean (SLR3)

- Consider a simple model of human capital investment. More schooling presumably leads to higher earnings. (Prior: $\beta_1 > 0$)

$$\text{Earnings} = \beta_0 + \beta_1 \text{schooling} + u$$

- What's in u?
- $E(u|\text{schooling}) = 0$?

Another Violation of SLR3, the Zero Conditional Mean Assumption

- Consider a simple model of weight loss.
- Assume we're interested in the population of overweight people and want to know whether participation in WW causes weight loss.

Weight Loss = $\alpha + \beta$WeightWatchers + ε

What's in ε? $E(\varepsilon|\text{WeightWatchers})=0$?

What's in ε

…that's not correlated with WeightWatchers?

…that is correlated with WW?

Nepal Example?

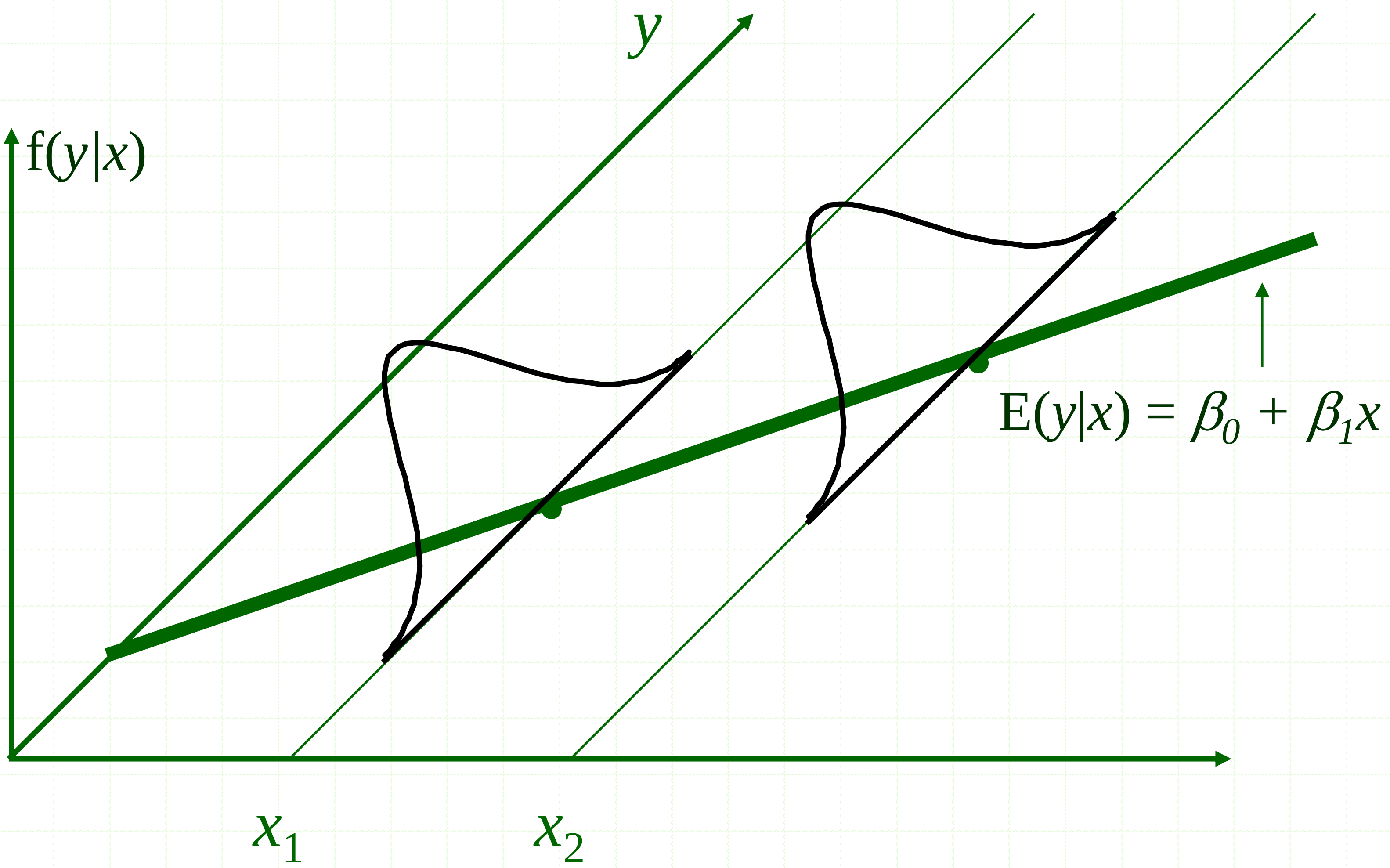
JW's SLR4: Sample variation in x

- JW asserts… never fails in interesting applications
- *Innocuous Assumption?*
- Sample vs. Population. If no pop variation in x, then the population can tell us nothing about the effect of changing x; But this is not SLR4.
- Assume that there is population variation in x and that x affects y, but assume no sample variation in x. Nothing can be learned from the sample about how changes in x affect y.
- The importance of finding the right data for the question

JW's SLR5: Homoscedasticity

- $Var(u_i|x_i) = \sigma^2$ <=> Homoscedasticity
- Independence of u and x => $Var(u_i|x_i) = \sigma^2$
- Contrast with $Var(u_i|x_i) = \sigma^2_i$ (nonconstant variance, heteroscedasticity)

Homoskedastic Case



Example of Violation of Homoscedasticity (SLR5)

- Consider a simple model of savings as a function of income. (Prior: $\beta_1 > 0$)

$$\text{Savings} = \beta_0 + \beta_1 \text{ Income} + u$$

- Do we expect $Var(u|Income) = \sigma^2$ (constant)?
- Note: min(savings)=0 & max(savings)=Income. As income increases, the range of savings possibilities increases, and so we'd expect Var(u|Income) increasing in Income.

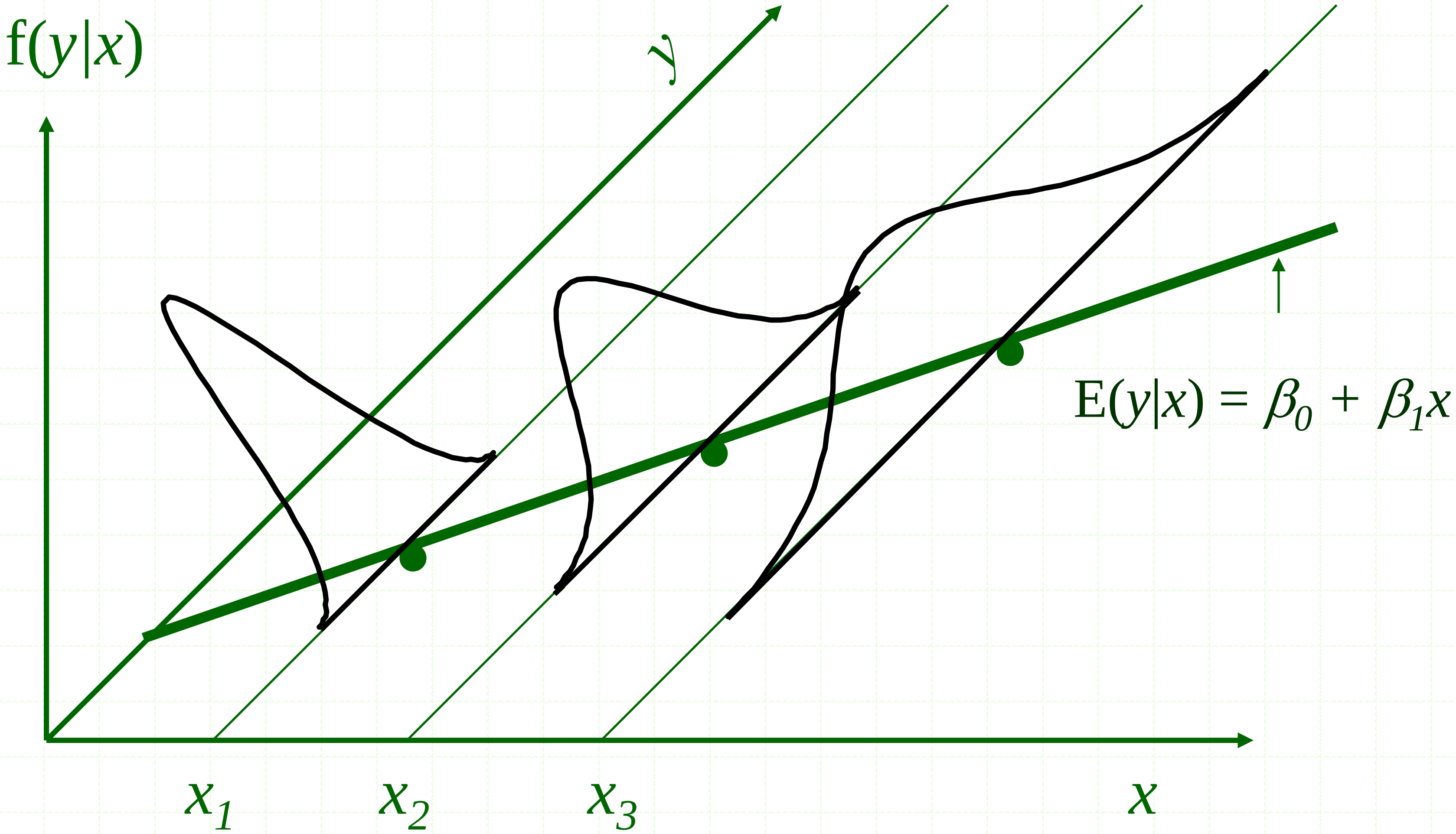
Another Candidate Case of Heteroscedasticity

- Consider a simple model of wages as a function of education and tenure with firm.

$$\text{Wages} = \beta_0 + \beta_1 \text{Education} + \beta_2 \text{Tenure} + u$$

- Do we expect $Var(u|x) = \sigma^2$ (constant)?
- Hypothesis: For a given set of characteristics, firms have narrow range of offer wages.
- As managers observe employee, they learn more about ability of employee and wage dispersion increases.

Heteroskedastic Case



SLR1-SLR5 & OLS

- SLR1: Linear in parameters
- SLR2: Random Sample
- SLR3: Zero Conditional Mean
- SLR4: Sample variation in x
- SLR1-SLR4 => OLS estimator is unbiased
- SLR5: Var(u|x) is constant
- SLR1-SLR5 => OLS is BLUE

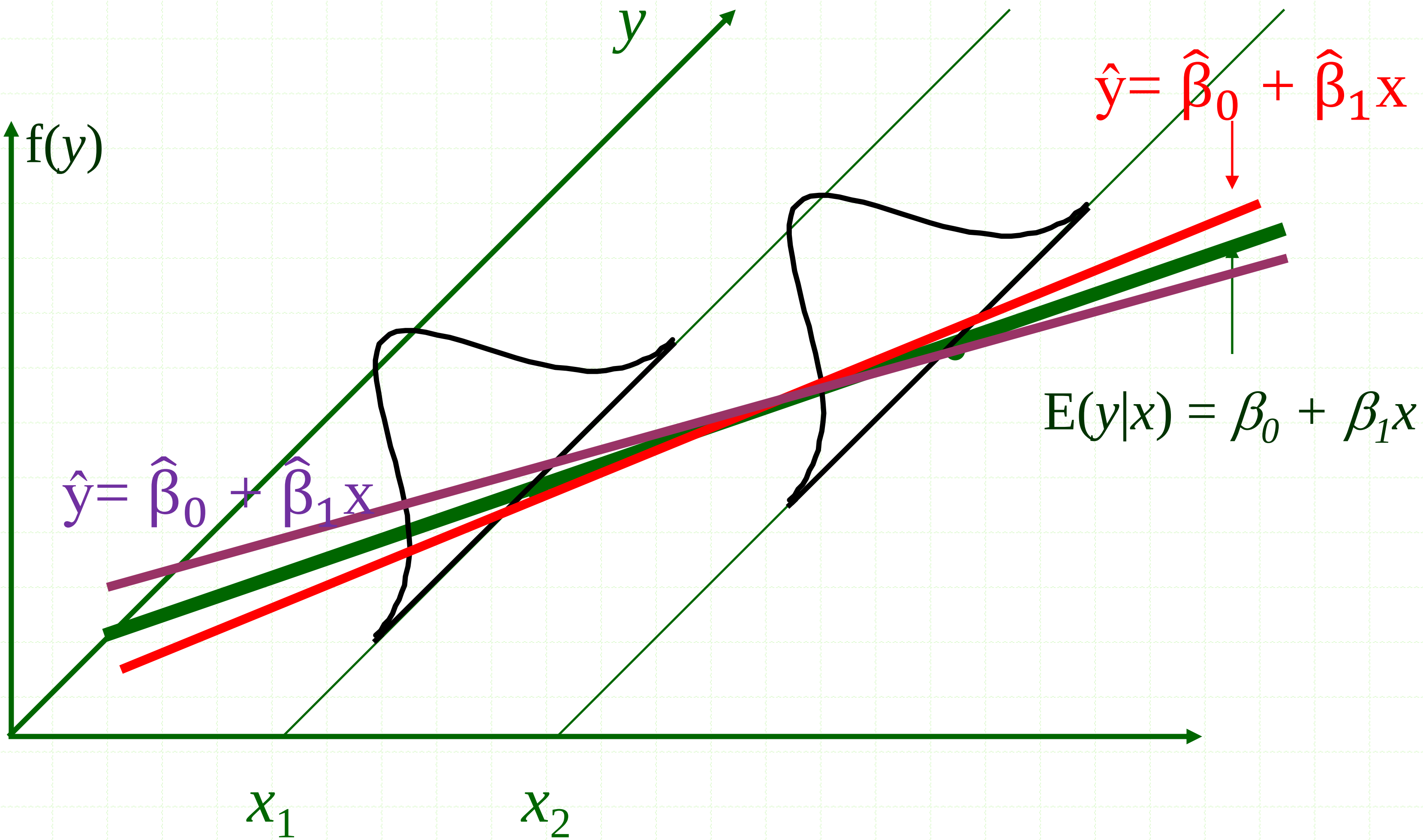
OLS estimator, bivariate model
What does unbiased mean? (BLUE)

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

assuming that $\sum_{i=1}^{n}(x_i - \bar{x})^2 > 0$ (SLR4)

Does unbiased means $\hat{\beta}_1 = \beta$?

Unbiasedness: Repeated sampling will average to population line.



OLS estimator, bivariate model What does Best mean? (**B**LUE)

- Unbiased and minimum variance
- Following slides we'll quickly derive variance for bivariate case. Show for multivariate case, and look at sample estimates of the variance.

OLS estimator, bivariate model
What does Best mean? (**B**LUE)

Unbiased and minimum variance

$$\mathrm{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}$$

Variance of OLS (cont)

- 1st derive for bivariate model, then show multivariate
- Repeat algebra results (used in next slide)

$$\sum(x_i - \bar{x})(x_i - \bar{x}) = \sum(x_i x_i) - \bar{x}\sum x_i - \bar{x}\sum x_i + n\bar{x}^2$$

$$= \sum(x_i x_i) - \bar{x}\sum x_i - n\bar{x}^2 \quad + n\bar{x}^2$$

$$= \sum(x_i - \bar{x})x_i \text{ and also}$$

$$= \sum(x_i^2) - n\bar{x}^2$$

Similarly,

$$\sum(y_i - \bar{y})(y_i - \bar{y}) = \sum(y_i - \bar{y})y_i$$

Variance of OLS (cont)

$$\hat{\beta}_1 = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \Big/ \sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\text{numerator} = \sum_{i=1}^{n}(x_i - \bar{x})\, y_i = \sum (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \mu_i\Big)$$

$$\text{numerator} = \beta_0 \underbrace{\sum (x_i - \bar{x})}_{0} + \beta_1 \sum (x_i - \bar{x}) x_i + \sum (x_i - \bar{x})\mu_i$$

$$\text{numerator} = \beta_1 \sum (x_i - \bar{x})^2) + \sum (x_i - \bar{x})\mu_i$$

$$\hat{\beta}_1 = \beta_1 + \left(\sum (x_i - \bar{x})\, \mu_i \Big/ \sum (x_i - \bar{x})^2 \right)$$

Variance of OLS (cont)

$$\mathrm{Var}(\hat{\beta}_1) = \mathrm{Var}\left\{\beta_1 + \left(\sum(x_i - \bar{x})\,\mu_i \Big/ \sum(x_i - \bar{x})^2\right)\right\}$$

If z is fixed, ε is random: $\mathrm{Var}(z_0 + z_1\varepsilon) = z_1^2\mathrm{Var}(\varepsilon)$

$$\mathrm{Var}(\hat{\beta}_1) = \left[\sum(x_i - \bar{x})^{-2}\right]^2 \sum[(x_i - \bar{x})^2]\mathrm{Var}(\mu_i)$$

SLR5: $E(\mu_i^2|x) = \sigma^2$

$$\mathrm{Var}(\hat{\beta}_1) = \sigma^2\left[\sum(x_i - \bar{x})^{-2}\right]^2 \sum[(x_i - \bar{x})^2]$$

$$\boldsymbol{\mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum(x_i - \bar{x})^2}}$$

Variance of OLS (cont)

Sampling variance of the OLS estimator differs slightly
For the multivariate model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_k x_k + u$$

$$\mathrm{Var}(\hat{\beta}_k) = \frac{\sigma^2}{\left(\sum \left(x_{i,k} - \bar{x}_k\right)^2 (1 - R_k^2\right)}$$

where R_k^2 is a goodness of fit measure R^2 from
Regressing x_k on all others x's $(x_i \ldots\ldots x_{k-1})$

Note in the bivariate case, R_k^2=0 and the
expressions for Var(β) are the same.

Error Variance Estimate (cont)

Recall $E(u_i^2) = \sigma^2$. If we observed μ_i, then an unbiased estimator of σ^2 is $\frac{1}{n}\mu_i^2$

It would seem that $\hat{\mu}_i$ could be directly substituted, but it turns out that $E\left[\sum \hat{\mu}_i^2\right] = (n - k - 1)\sigma^2$

So, an unbiased estimator of σ^2 is: $\hat{\sigma}^2 = \frac{1}{(n-k-1)}\sum \hat{\mu}_i^2$

Estimating OLS Variance (cont)

Standard error of the bivariate $\hat{\beta}_1$ is

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\mu}_i{}^2}{(n-k-1)\sum(x_i-\bar{x})^2}}$$

And for the multivariate case…

$$se(\hat{\beta}_j) = \sqrt{\frac{\hat{\mu}_i{}^2}{(n-k-1)\sum(x_{ij}-\bar{x})^2(1-R_j^2)}}$$

Interpreting Coefficients

NLSS-2011

Nepal Living Standard Survey

http://cbs.gov.np/nada/index.php/catalog/37

childnut_example.do

bmi_example.do (Command Line)