

Econometrics with Observational Data

Introduction and Identification

Todd Wagner

Goals for Course

- To enable researchers to conduct careful analyses with existing VA (and non-VA) datasets.
 - We will
 - Describe econometric tools and their strengths and limitations
 - Use examples to reinforce learning
-

Goals of Today's Class

- Understanding causation with observational data
 - Describe elements of an equation
 - Example of an equation
 - Assumptions of the classic linear model
-

Terminology

- Confusing terminology is a major barrier to interdisciplinary research
 - Multivariable or multivariate
 - Endogeneity or confounding
 - Interaction or Moderation
 - Right or Wrong
 - Maciejewski ML, Weaver ML and Hebert PL. (2011) *Med Care Res Rev* 68 (2): 156-176
-

Polls

- What is your background with analyzing observational data?
 - 1. Beginner. Understand averages, medians and variance, but don't run regression
 - 2.
 - 3. Modest experience. Familiar with linear or logistic regression
 - 4.
 - 5. Reasonably advanced. Have used statistical methods to control for unobserved heterogeneity or endogeneity.
-

What is your formal educational training?

- BA
 - Masters
 - Clinical training (MD, RN, Pharm D, clinical psych)
 - Clinical training with health services fellowship
 - PhD in a social science with no clinical training
-

Years since last degree

- 1
 - 2-3
 - 3-4
 - 5-7
 - 8+
-

Understanding Causation: Randomized Clinical Trial

- RCTs are the gold-standard research design for assessing causality
 - What is unique about a randomized trial?
 - The treatment / exposure is randomly assigned
 - Benefits of randomization:
 - Causal inferences
-

Randomization

- Random assignment distinguishes experimental and non-experimental design
 - Random assignment should not be confused with random selection
 - Selection can be important for generalizability (e.g., randomly-selected survey participants)
 - Random assignment is required for understanding causation
-

Limitations of RCTs

- Generalizability to real life may be low
 - Exclusion criteria may result in a select sample
 - Hawthorne effect (both arms)
 - RCTs are expensive and slow
 - Can be unethical to randomize people to certain treatments or conditions
 - Quasi-experimental design can fill an important role
-

Can Secondary Data Help us understand Causation?

Study: coffee may make you lazy *Coffee not linked to psoriasis*

Coffee, exercise may decrease risk of skin cancer
Coffee: An effective weight loss tool
Coffee poses no threat to hearts, may reduce diabetes risk: EPA
Coffee may make high achievers slack off

Observational Data

- Widely available (especially in VA)
 - Permit quick data analysis at a low cost
 - May be realistic/ generalizable

 - Key independent variable may not be exogenous – it may be endogenous
-

Endogeneity

- A variable is said to be **endogenous** when it is correlated with the error term (assumption 4 in the classic linear model)
 - If there exists a loop of causality between the independent and dependent variables of a model leads, then there is endogeneity
-

Endogeneity

- Endogeneity can come from:
 - Measurement error
 - Autoregression with autocorrelated errors
 - Simultaneity
 - Omitted variables
 - Sample selection
-

Example of Endogeneity

- Qx: does greater use of PET screening decrease lung cancer mortality
 - You observe that some facilities do a lot of PET screening while others do little
 - You compare patients across facilities and find a negative correlation between PET screening intensity and mortality
 - Why is PET screening intensity endogenous?
-

Econometrics v Statistics

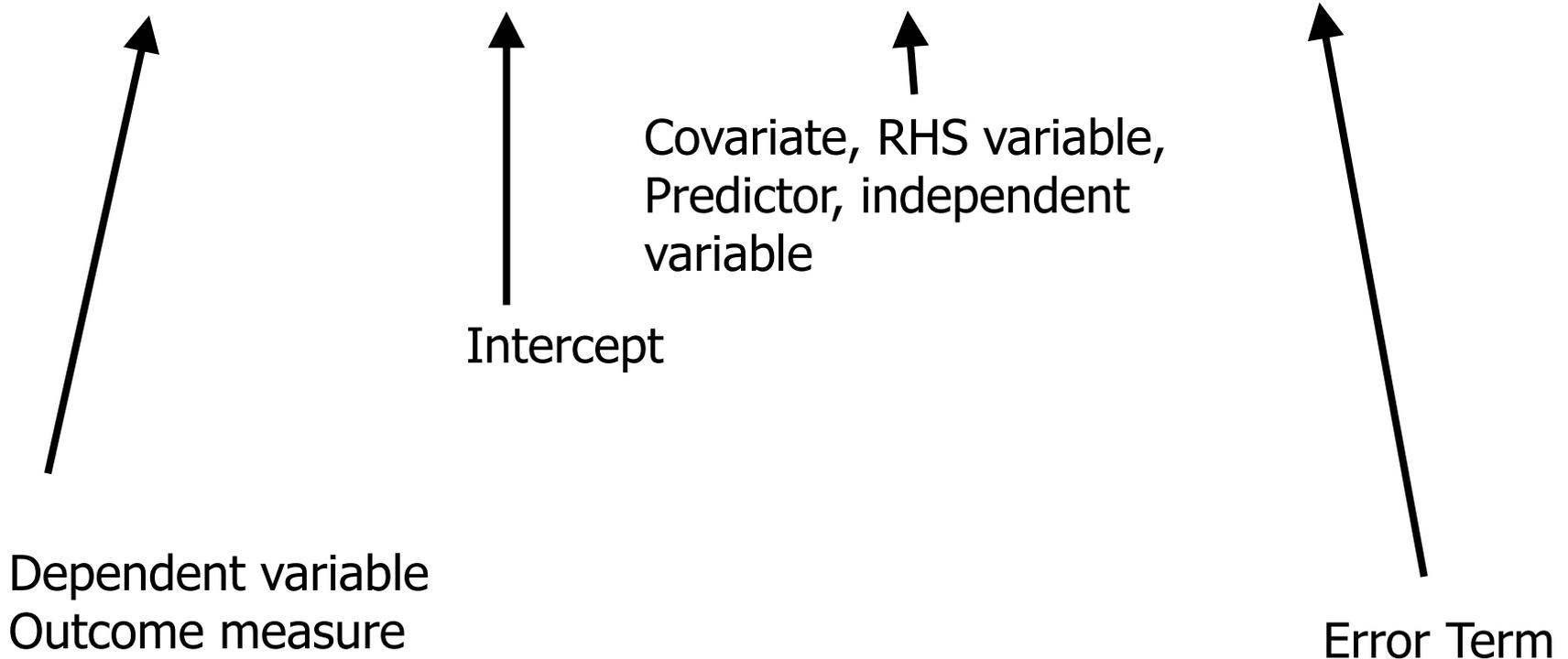
- Often use different terms
 - Cultural norms—if it seems endogenous, it probably is
 - Underlying data generating model is economic. Rational actors concerned with
 - Profit maximization
 - Quantity maximization
 - Time minimization
-

Elements of an Equation

Terms

- Univariate— the statistical expression of one variable
 - Bivariate— the expression of two variables
 - Multivariate— the expression of more than one variable (can be dependent or independent variables)
-

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



Note the similarity to the equation of a line ($y=mx+B$)

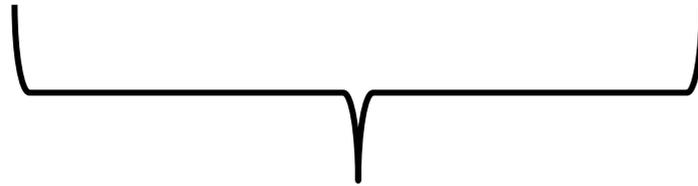
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

“i” is an index.

If we are analyzing people, then this typically refers to the person

There may be other indexes

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \varepsilon_i$$



Two covariates



DV



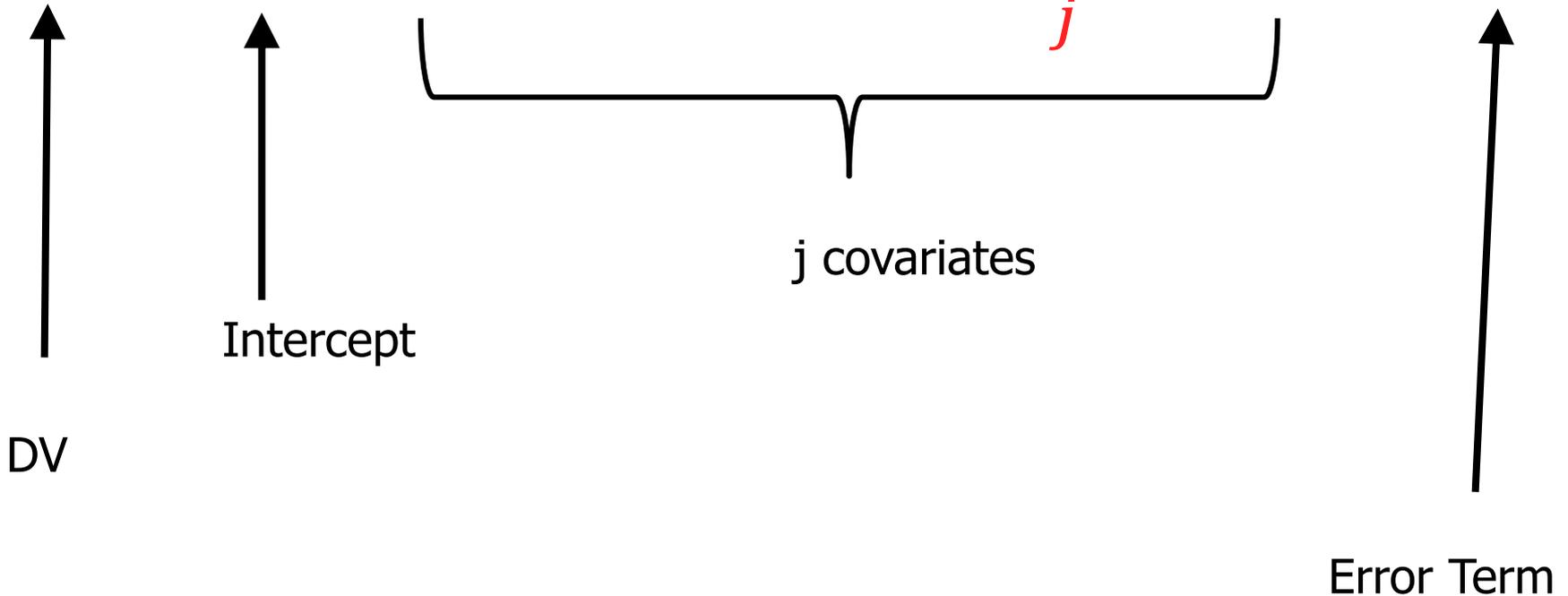
Intercept



Error Term

Different notation

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \underbrace{B_{ij} X_{ij}}_j + \varepsilon_i$$



Error term

- Error exists because
 1. Other important variables might be omitted
 2. Measurement error
 3. Human indeterminacy
 - Understand error structure and minimize error
 - Error can be additive or multiplicative
-

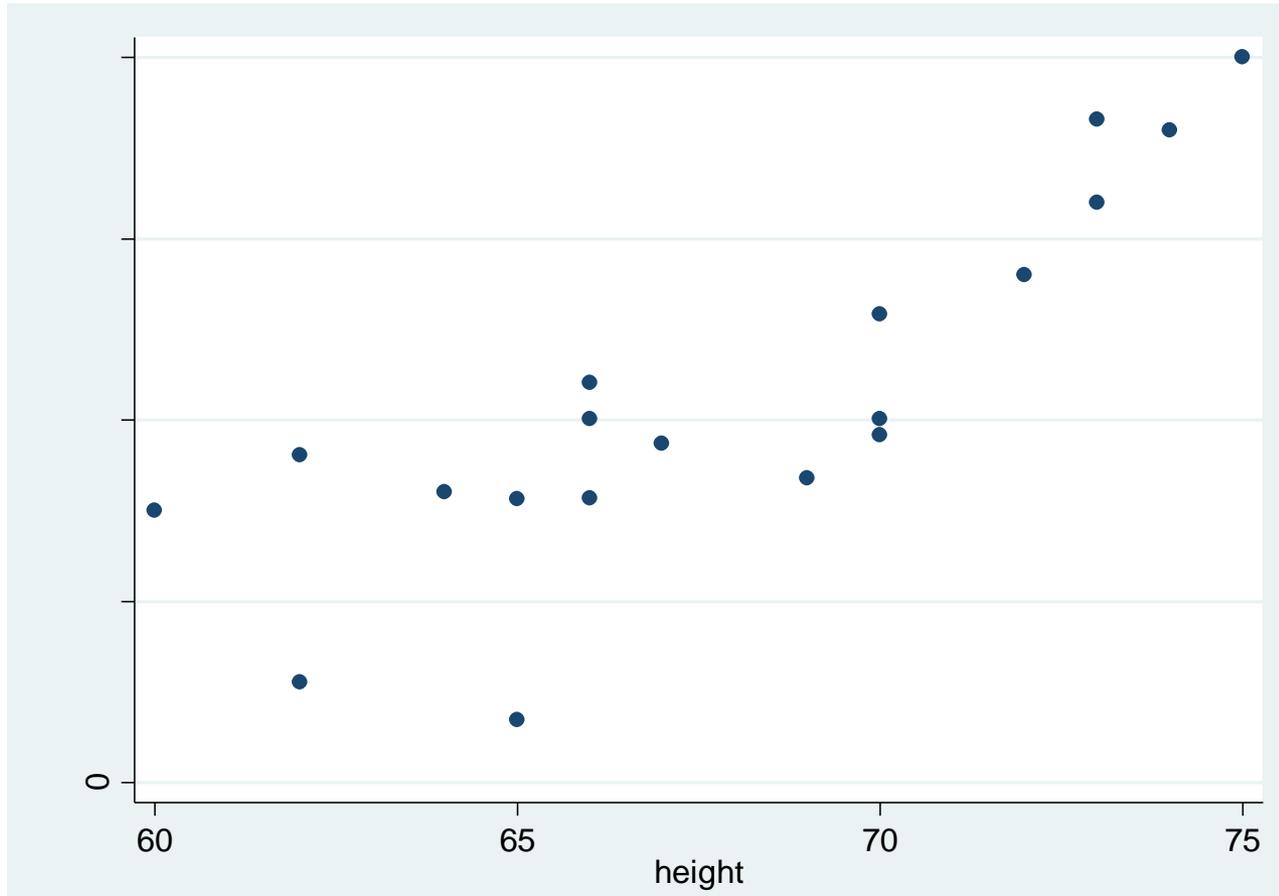
**Example: is height associated with
income?**



$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

- Y =income; X =height
 - Hypothesis: Height is not related to income ($\beta_1=0$)
 - If $\beta_1=0$, then what is β_0 ?
-

Height and Income

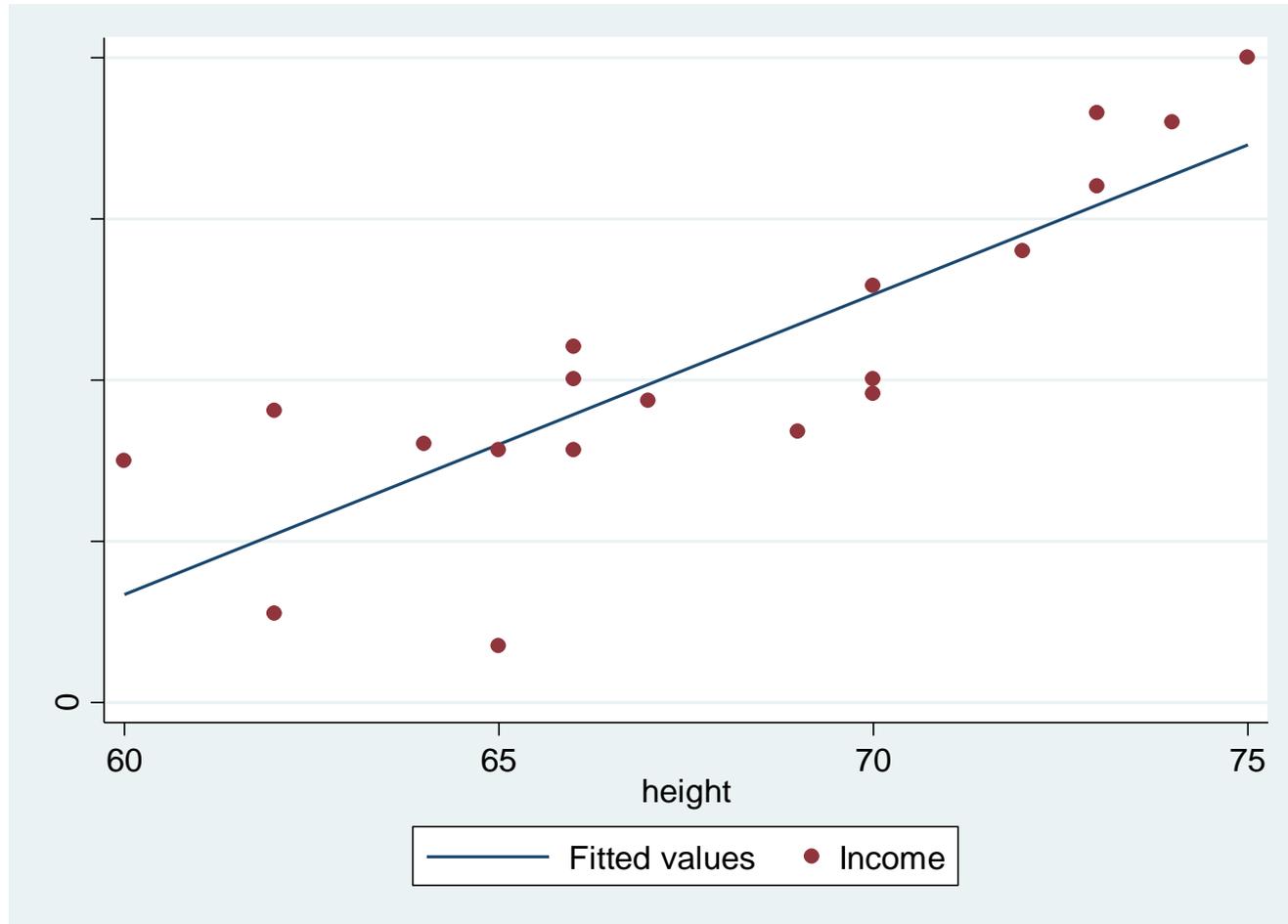


How do we want to describe the data?

Estimator

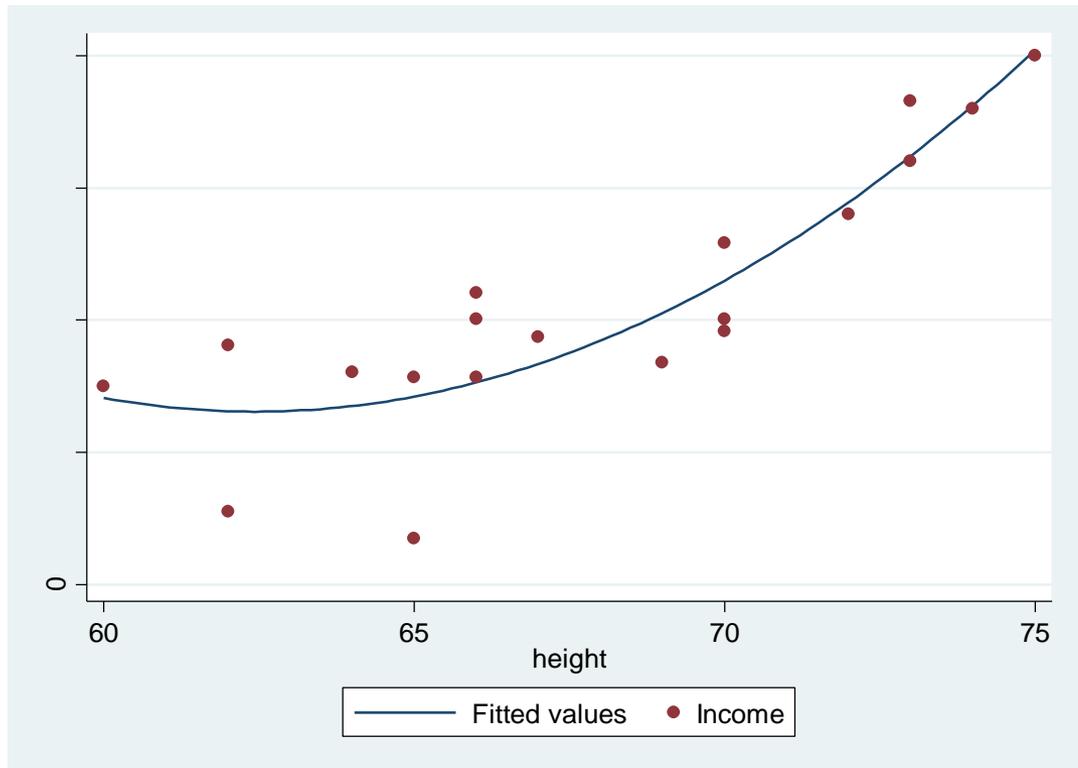
- A statistic that provides information on the parameter of interest (e.g., height)
 - Generated by applying a function to the data
 - Many common estimators
 - Mean and median (univariate estimators)
 - Ordinary least squares (OLS) (multivariate estimator)
-

Ordinary Least Squares (OLS)



Other estimators

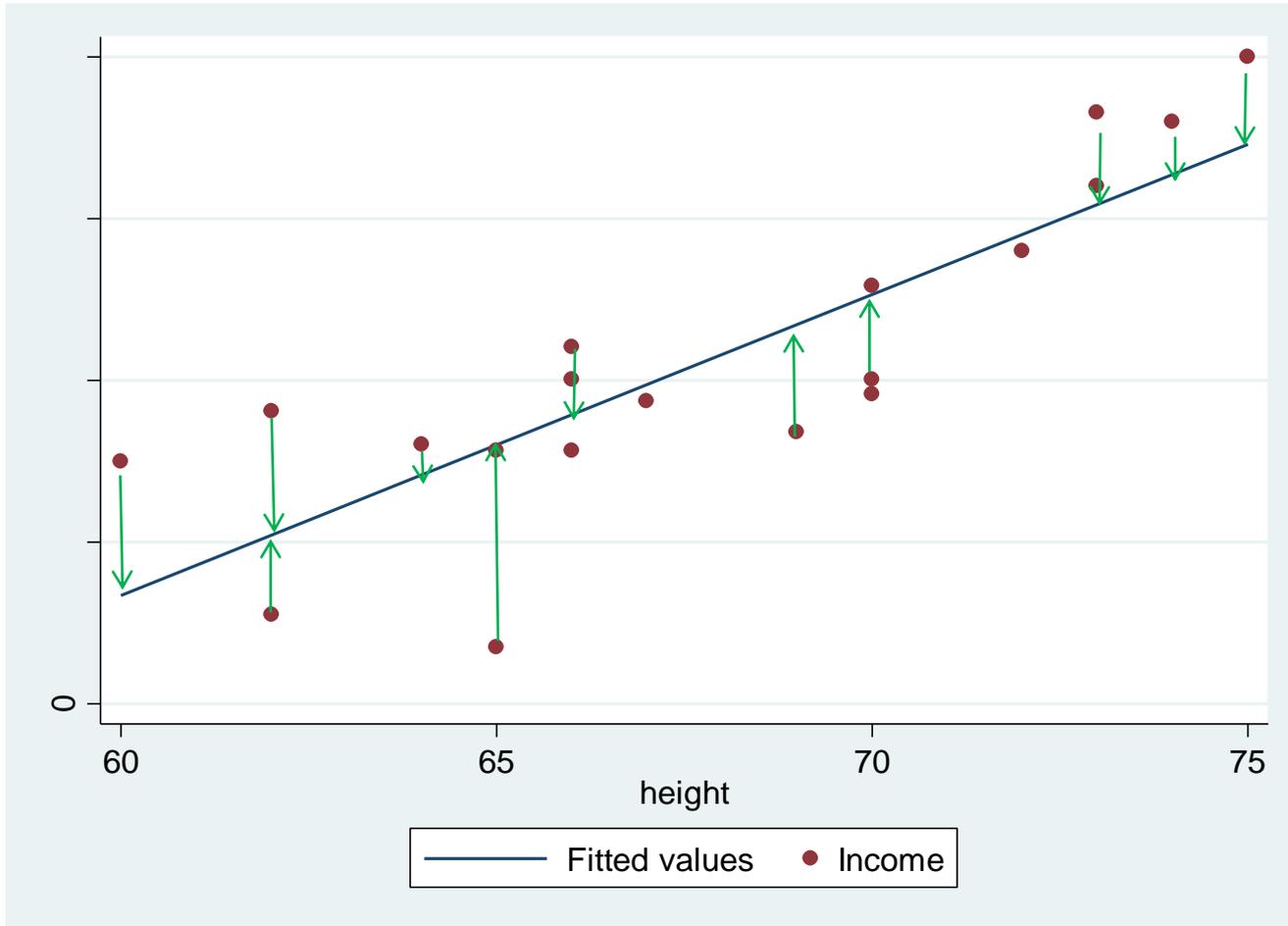
- Least absolute deviations
- Maximum likelihood



Choosing an Estimator

- Least squares
 - Unbiasedness
 - Efficiency (minimum variance)
 - Asymptotic properties
 - Maximum likelihood
 - Goodness of fit
-
- We'll talk more about identifying the “right” estimator throughout this course.
-

How is the OLS fit?



What about gender?

- How could gender affect the relationship between height and income?
 - Gender-specific intercept
 - Interaction
-

Gender Indicator Variable

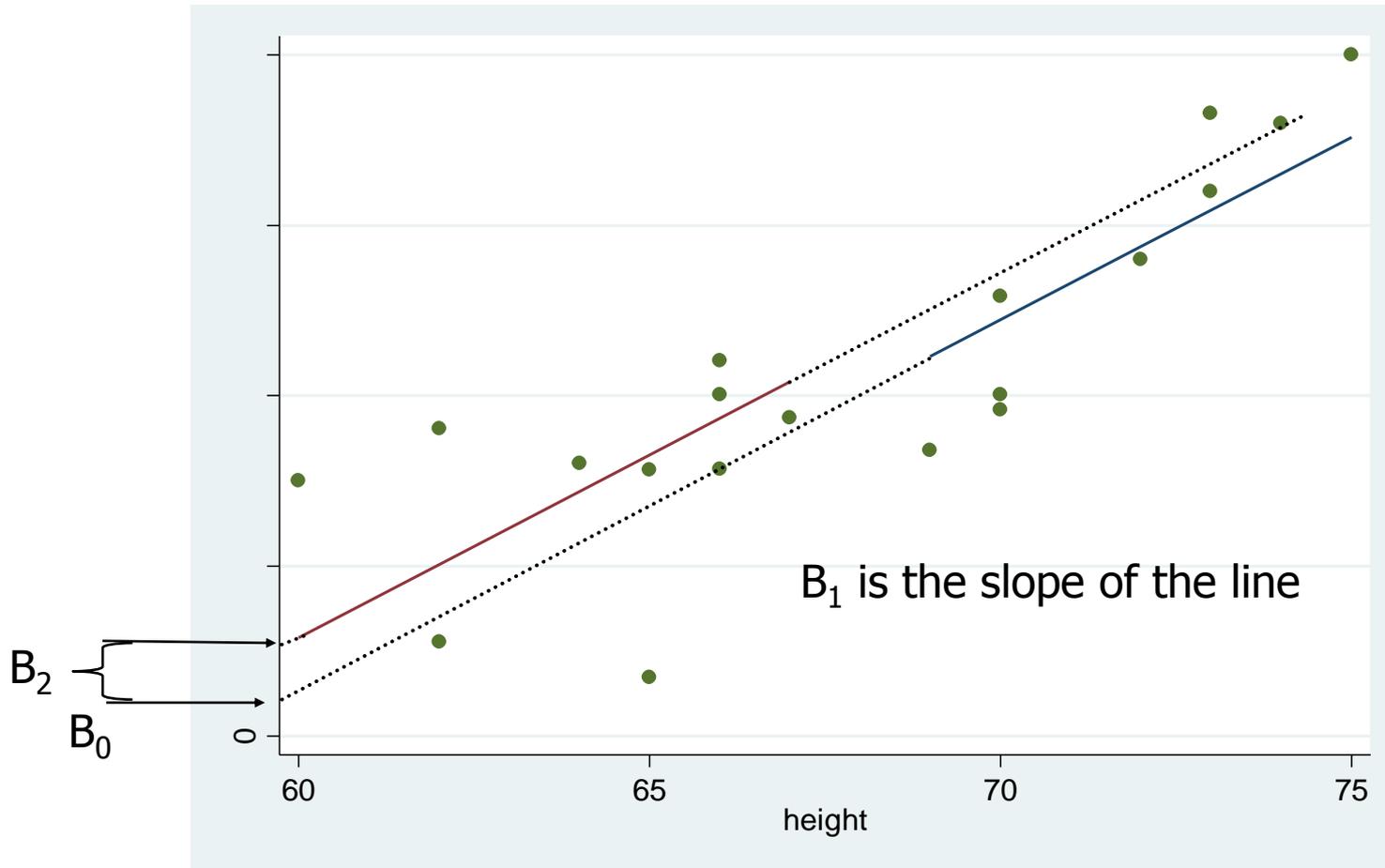
height

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \varepsilon_i$$



Gender Intercept

Gender-specific Indicator



Interaction

height

gender

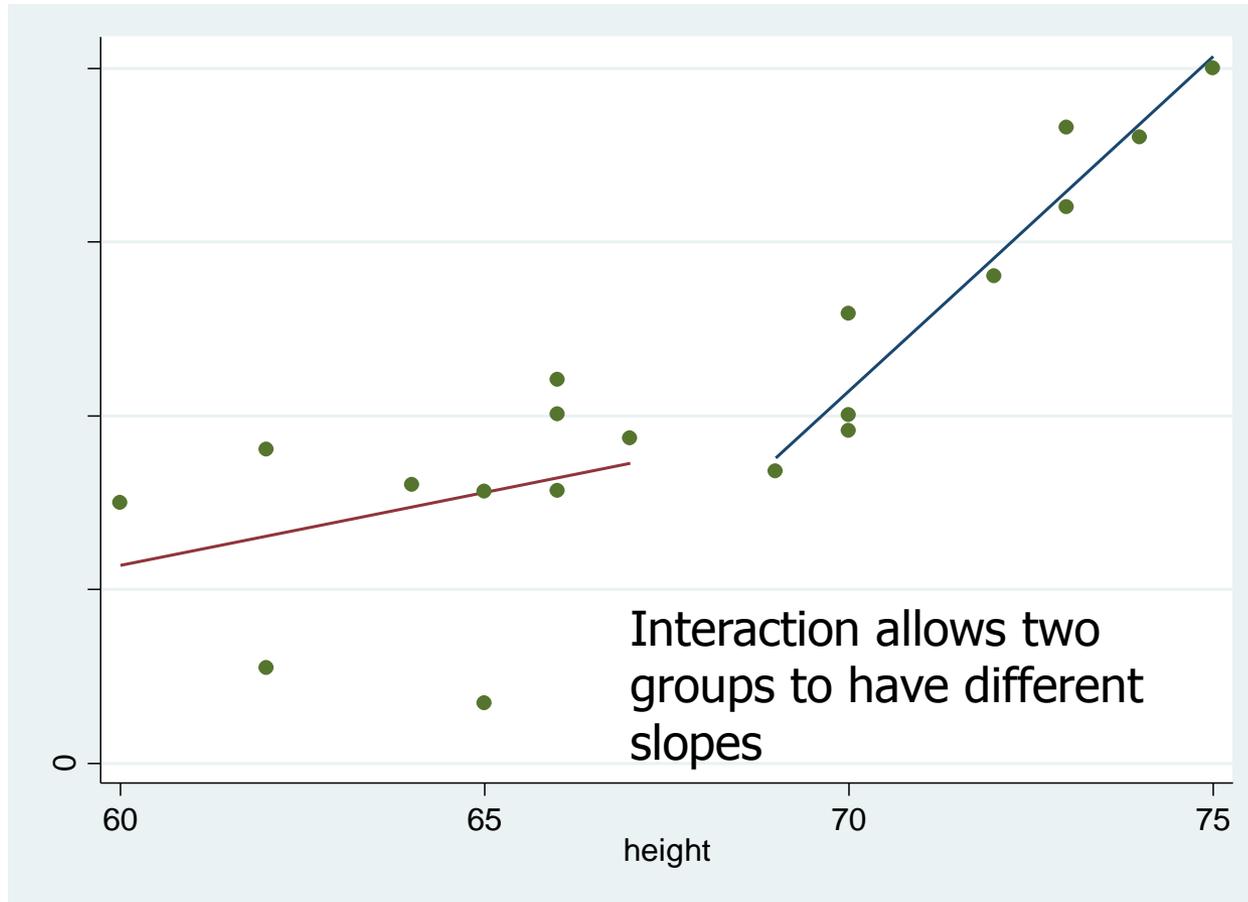
$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + \varepsilon_i$$



Interaction Term,
Effect modification,
Modifier

Note: the gender “main effect”
variable is still in the model

Gender Interaction



Classic Linear Regression (CLR)

Assumptions

Classic Linear Regression

- No “superestimator”
 - CLR models are often used as the starting point for analyses
 - 5 assumptions for the CLR
 - Variations in these assumption will guide your choice of estimator (and happiness of your reviewers)
-

Assumption 1

- The dependent variable can be calculated as a linear function of a specific set of independent variables, plus an error term
- For example,

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \beta_3 X_i Z_i + \varepsilon_i$$

Violations to Assumption 1

- Omitted variables
 - Non-linearities
 - Note: by transforming independent variables, a nonlinear function can be made from a linear function
-

Testing Assumption 1

- Theory-based transformations (e.g., Cobb-Douglas production)
- Empirically-based transformations
- Common sense
- Ramsey RESET test
- Pregibon Link test

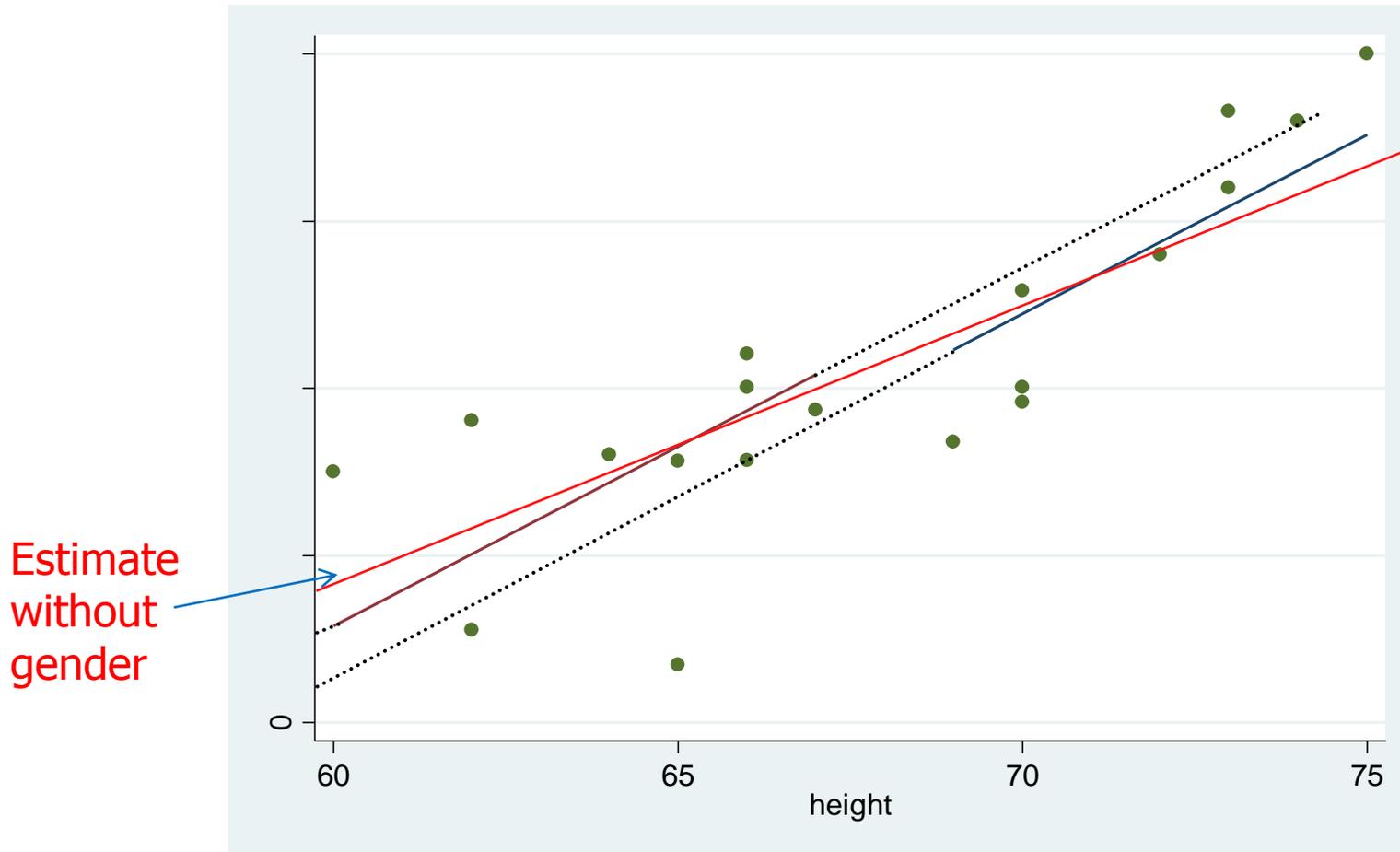
Ramsey J. Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society*. 1969;Series B(31):350-371.

Pregibon D. Logistic regression diagnostics. *Annals of Statistics*. 1981;9(4):705-724.

Assumption 1 and Stepwise

- Statistical software allows for creating models in a “stepwise” fashion
 - Be careful when using it.
 - Little penalty for adding a nuisance variable
 - BIG penalty for missing an important covariate
-

Bias if Gender is Ignored



Assumption 2

- Expected value of the error term is 0

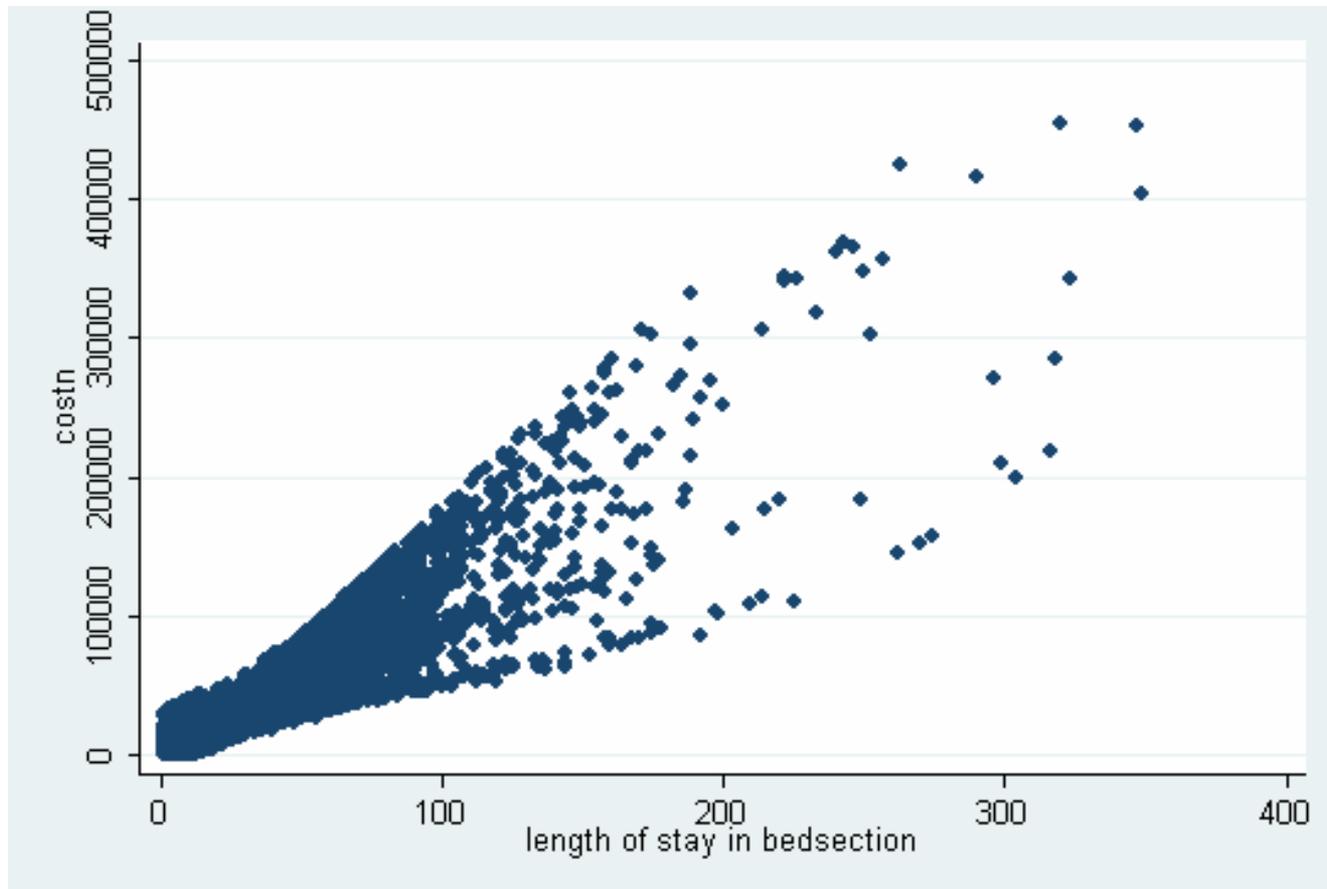
$$E(u_i)=0$$

- Violations lead to biased intercept
 - A concern when analyzing cost data
(Smearing estimator when working with logged costs)
-

Assumption 3

- IID– Independent and identically distributed error terms
 - Autocorrelation: Errors are uncorrelated with each other
 - Homoskedasticity: Errors are identically distributed
-

Heteroskedasticity



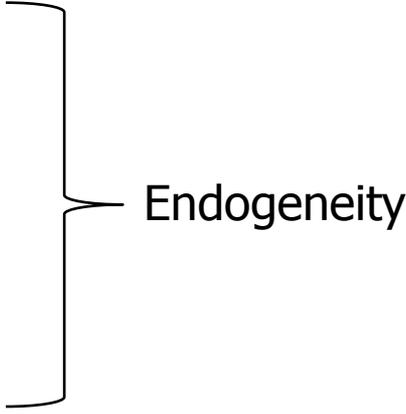
Violating Assumption 3

- Effects
 - OLS coefficients are unbiased
 - OLS is inefficient
 - Standard errors are biased
 - Plotting is often very helpful
 - Different statistical tests for heteroskedasticity
 - GWHet--but statistical tests have limited power
-

Fixes for Assumption 3

- Transforming dependent variable may eliminate it
 - Robust standard errors (Huber White or sandwich estimators)
-

Assumption 4

- Observations on independent variables are considered fixed in repeated samples
 - $E(x_i u_i | x) = 0$
 - Violations
 - Errors in variables
 - Autoregression
 - Simultaneity
- Endogeneity
- 

Assumption 4: Errors in Variables

- Measurement error of dependent variable (DV) is maintained in error term.
 - OLS assumes that covariates are measured without error.
 - Error in measuring covariates can be problematic
-

Common Violations

- Including a lagged dependent variable(s) as a covariate
 - Contemporaneous correlation
 - Hausman test (but very weak in small samples)
 - Instrumental variables offer a potential solution
-

Assumption 5

- Observations $>$ covariates
 - No multicollinearity
 - Solutions
 - Remove perfectly collinear variables
 - Increase sample size
-

Any Questions?



Statistical Software

- I frequently use SAS for data management
 - I use Stata for my analyses
 - Stattransfer
-

Regression References

- Kennedy A Guide to Econometrics
 - Greene. Econometric Analysis.
 - Wooldridge. Econometric Analysis of Cross Section and Panel Data.
 - Winship and Morgan (1999) The Estimation of Causal Effects from Observational Data *Annual Review of Sociology*, pp. 659-706.
-