

Econometrics with Observational Data

Introduction and Identification

Todd Wagner

January 2023



U.S. Department of Veterans Affairs
Veterans Health Administration
Health Services Research & Development Service

One Perspective:

FULL TEXT ARTICLE 

Real-world studies no substitute for RCTs in establishing efficacy  

Hertzel C Gerstein, John McMurray and Rury R Holman

Lancet, The, 2019-01-19, Volume 393, Issue 10168, Pages 210-211, Copyright © 2019 Elsevier Ltd

“In the absence of randomisation, analyses of most observational data from the real world, regardless of their sophistication, can only be viewed as hypothesis generating.”

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

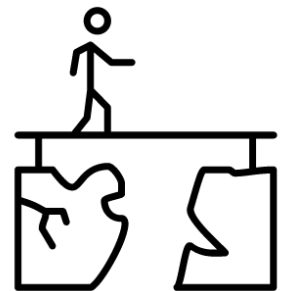
Can secondary data be used to understand causation?

Coffee, exercise may decrease risk of skin cancer
Coffee: An effective weight loss tool

Coffee poses no threat to hearts, may reduce diabetes risk: EPIC data
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 covariate may not be exogenous – it may be endogenous



Endogeneity

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

Example of Endogeneity: Testosterone Injections



- Research has correlated bone density and testosterone in men.
- Men generate different levels of testosterone.
 - This is endogenous testosterone.
 - There may be many reasons why a man's internal testosterone is low/high
- Giving men exogenous testosterone (an injection) may lead to very different effects from those studies that examine endogenous testosterone.

Testosterone

- Endogeneity isn't necessarily a problem if you observe everything and can control for it.
- Different approaches
 - Control for observables as best we can (propensity scores)
 - Focus on variation that is exogenous (instrumental variables, regression discontinuity)
- But, there is no way to control for everything

Endogeneity

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

Econometrics vs Statistics

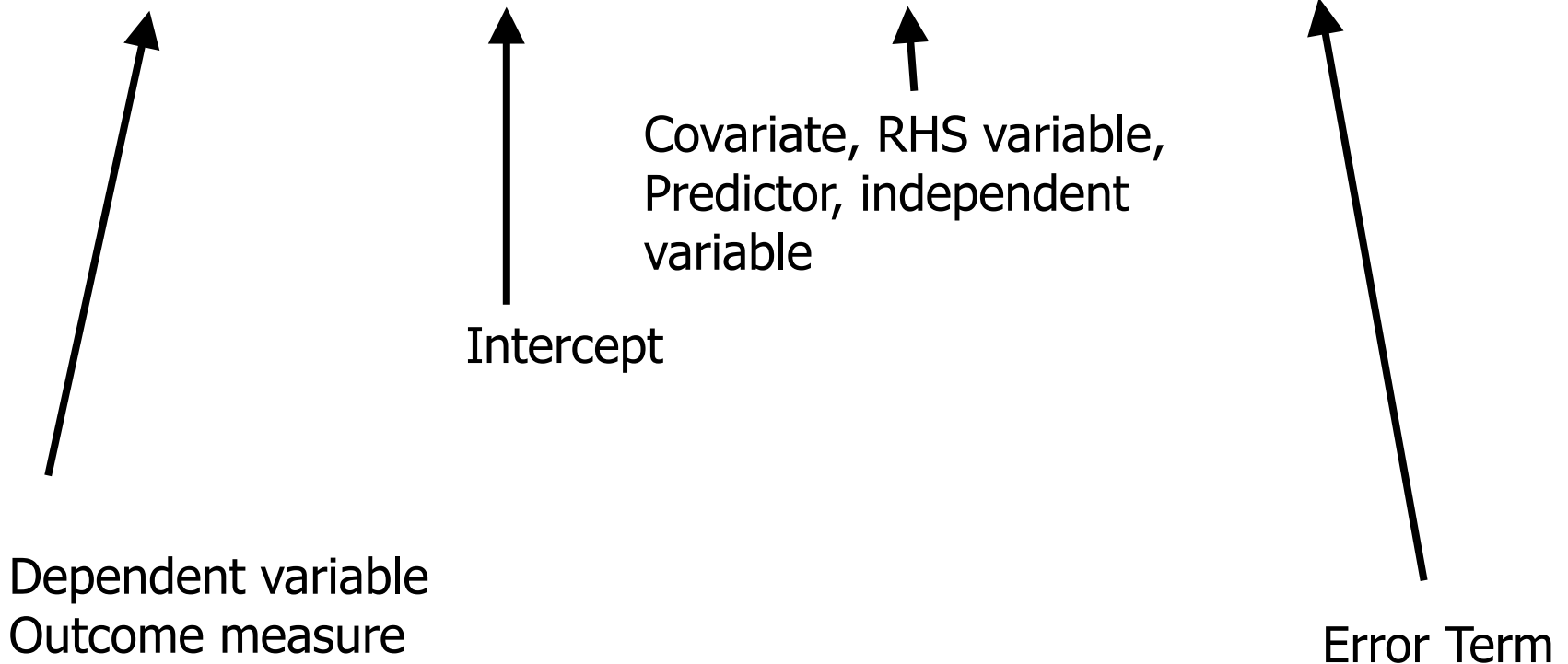
- Often use different terms
- Cultural norms
 - In health economics if it seems endogenous, it probably is
 - Underlying data generating model is economic. Rational actors concerned with
 - Profit maximization
 - Quantity maximization
 - Time minimization
 - Random and fixed effects
 - Propensity scores

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

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

DV




Intercept



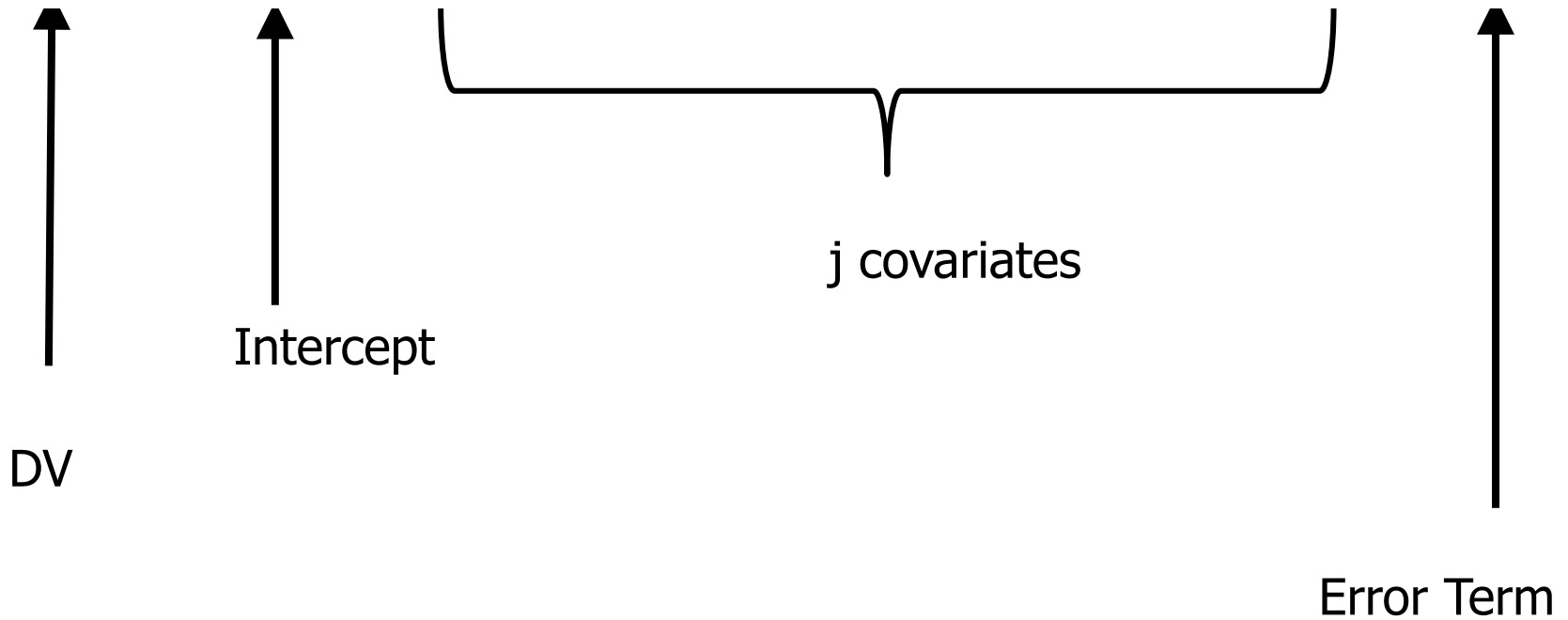
Two covariates



Error Term



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



Error term

- Error exists because
 1. Other important variables might be omitted
 2. Measurement error
 3. Human indeterminacy

- Your goal
 - Understand error structure
 - minimize error

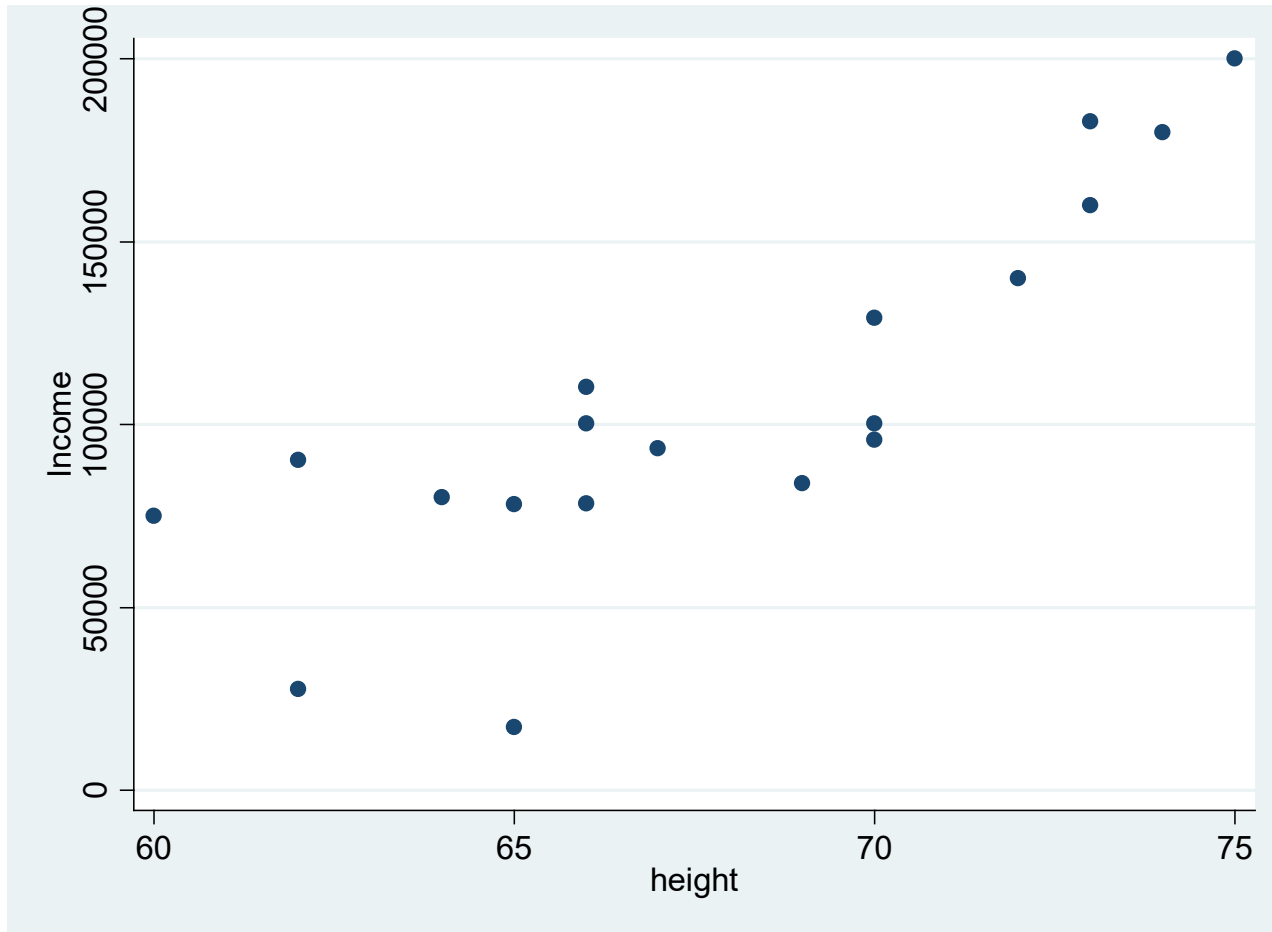
See Kennedy, P. [A Guide to Econometrics](#)

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

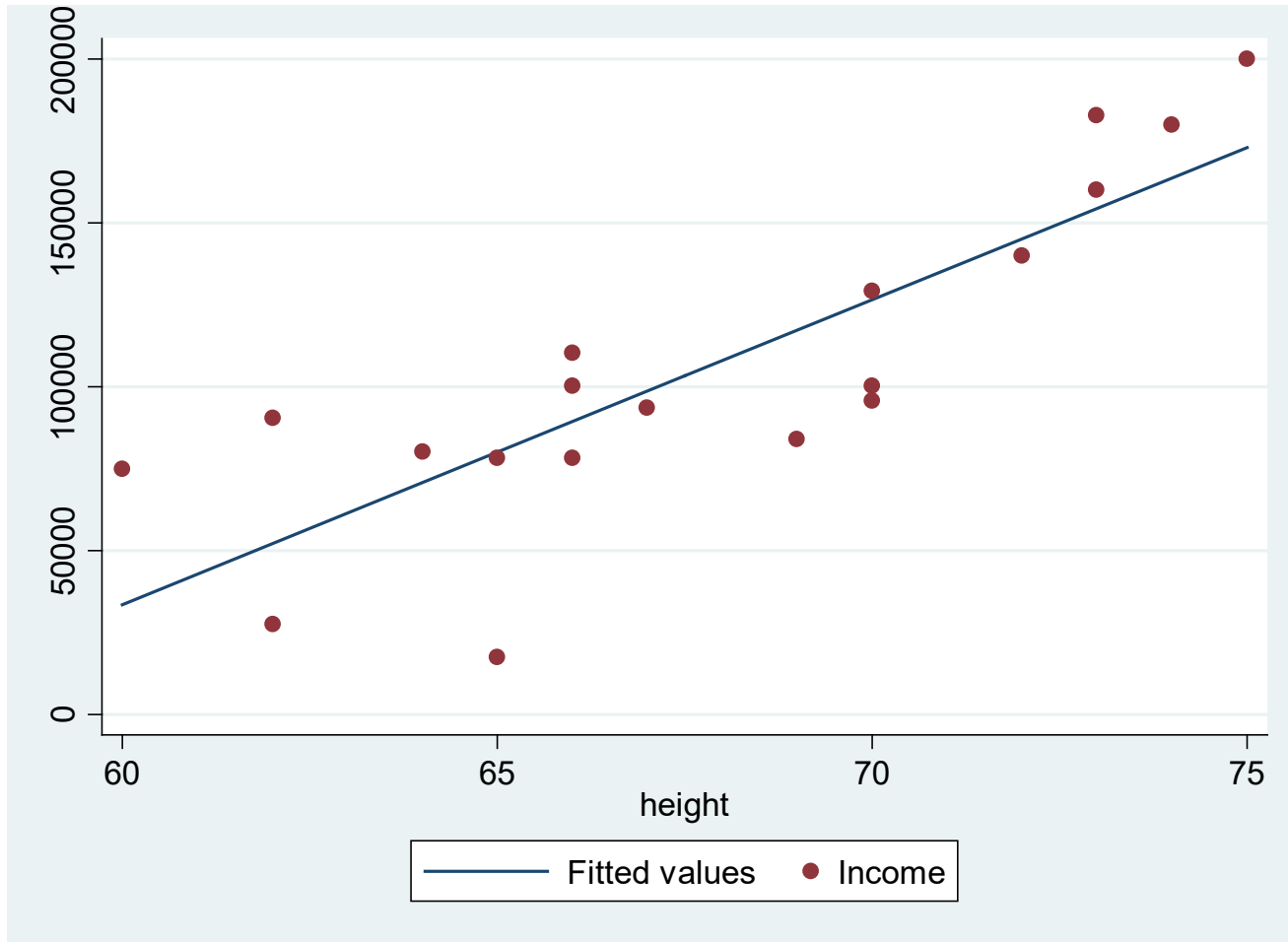


How do we want to describe the data?

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 / median of income (univariate)
 - Mean of income and by height (bivariate)
 - Mean of and by height controlling for other variable (multivariate)

Ordinary Least Squares (OLS)

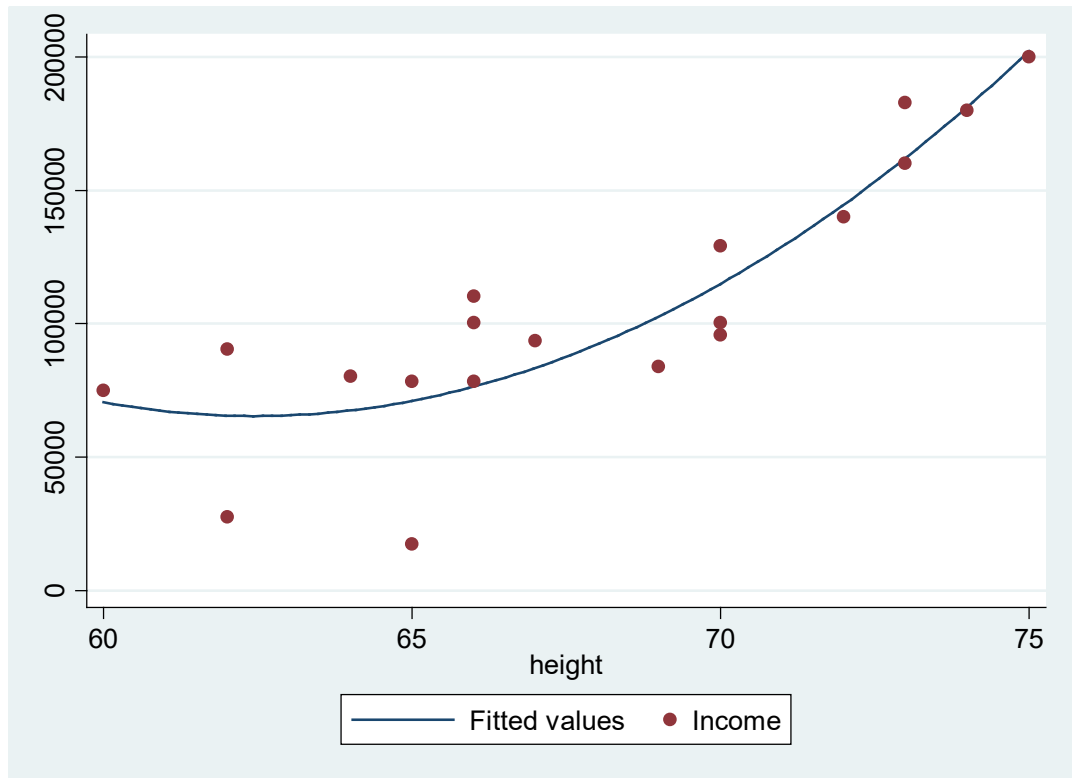


We are using this line to represent a relationship between height and income

Is this linear relationship correct?

Other estimators

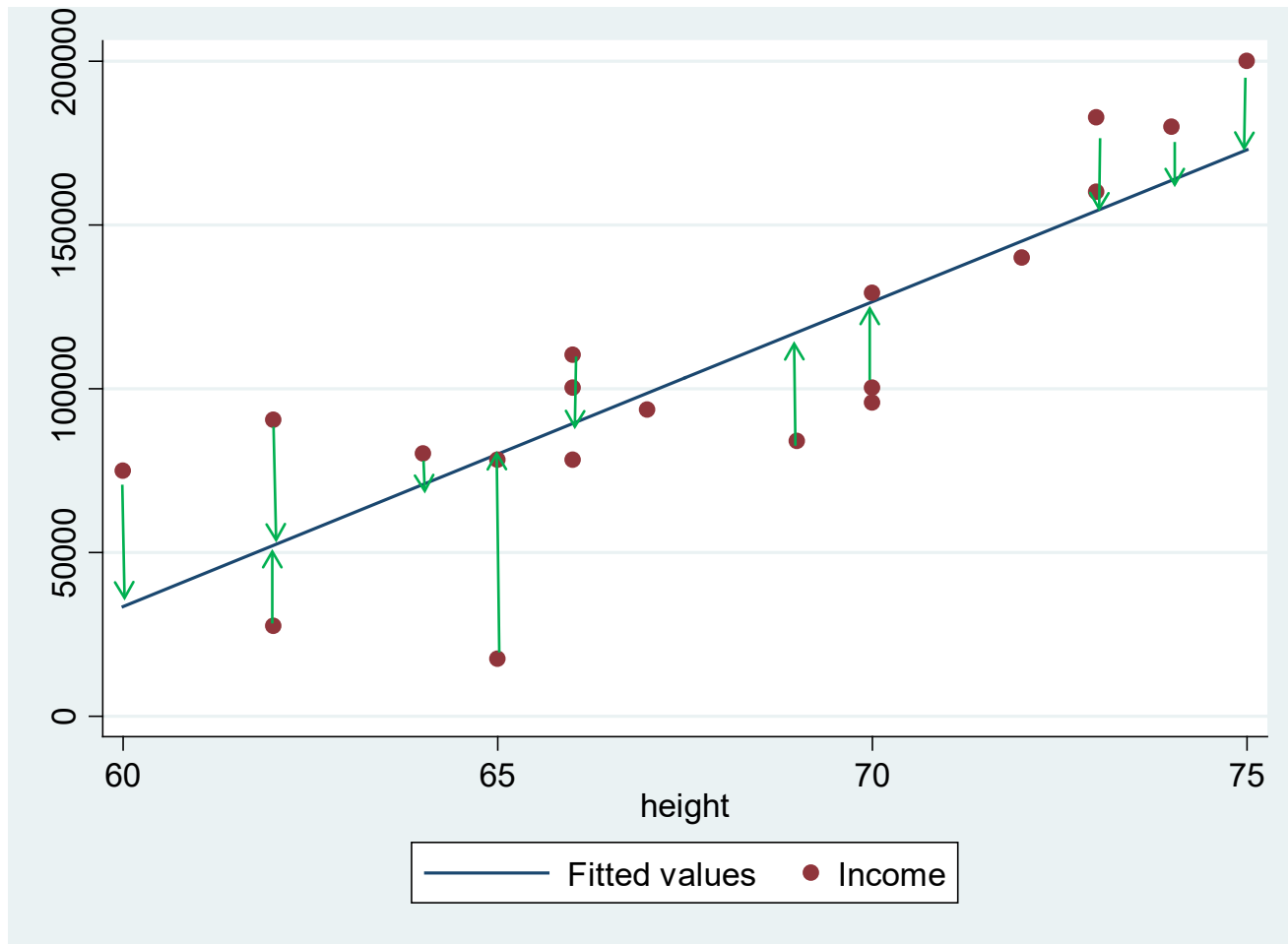
- Least absolute deviations
- Maximum likelihood
- Non-linear



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

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

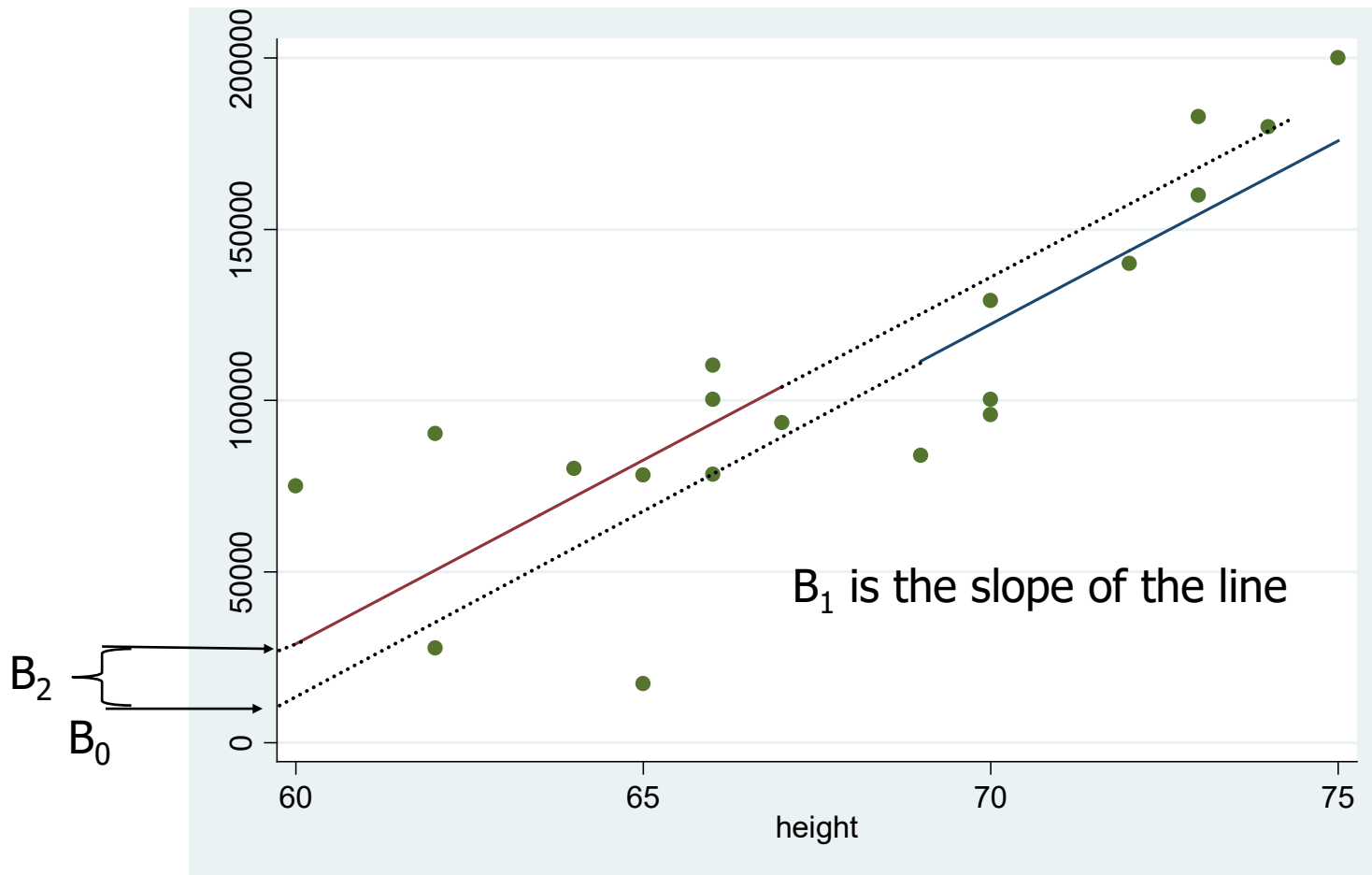
income

height



Gender Intercept

Gender-specific Indicator



Interaction

height

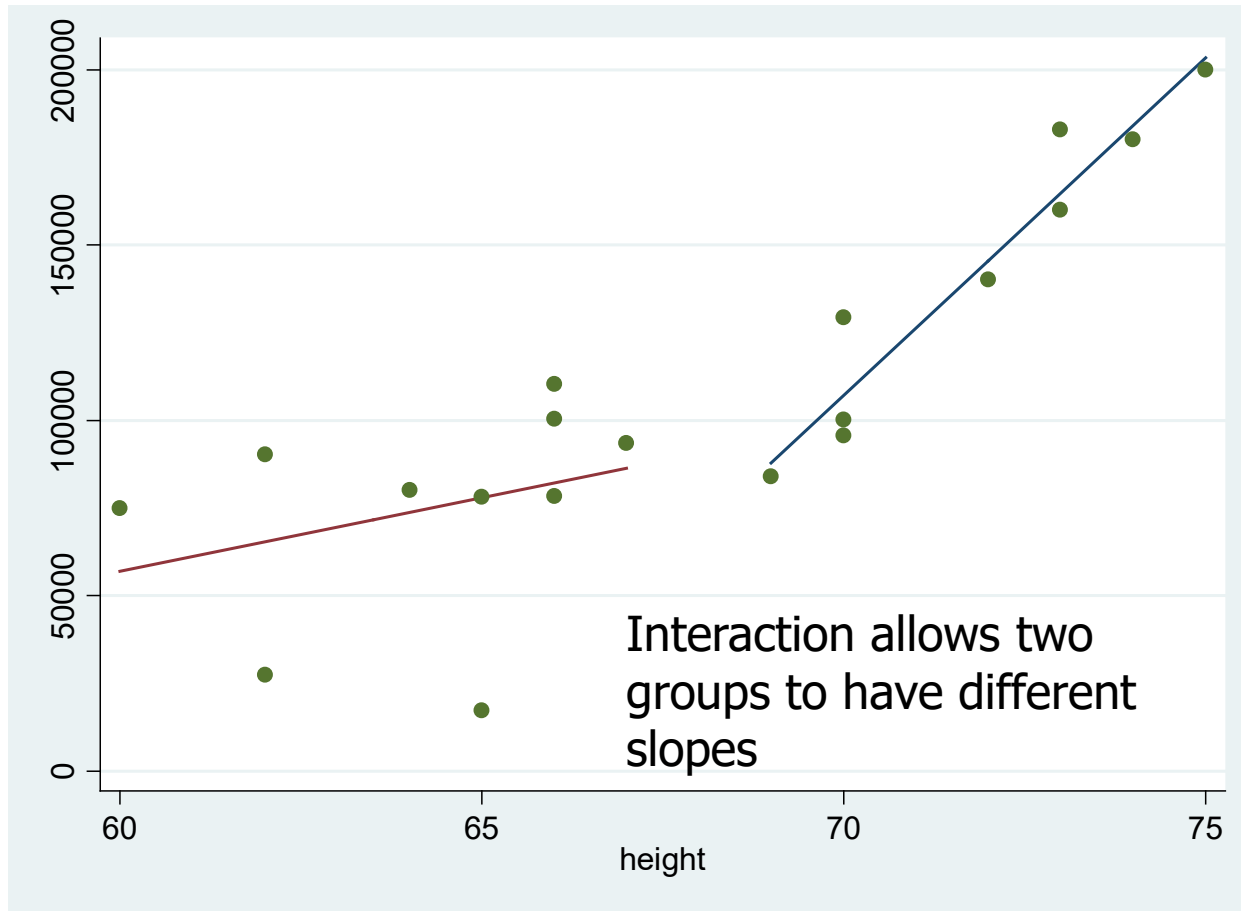
gender

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

Interaction Term,
Effect modification,
Modifier

Note: the gender "main effect"
variable is still in the model

Gender Interaction



Identification

- Is an association meaningful?
- Should we change behavior or make policy based on associations?
- For many people, associations are insufficient evidence, and we need to identify the causal relationship
- Identification requires that we meet all 5 assumptions in the classic linear model

Questionable science can lead to questionable policy

- Example: Bicycle helmet laws
- In laboratory experiments, helmets protect the head
- This may not translate to the real road
 - Do bikers behave differently when wearing a helmet?
 - Do drivers behave differently around bikers with/without helmets?
 - Do helmet laws have unintended consequences? (low uptake of bike share)

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
- Variation in these assumptions 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

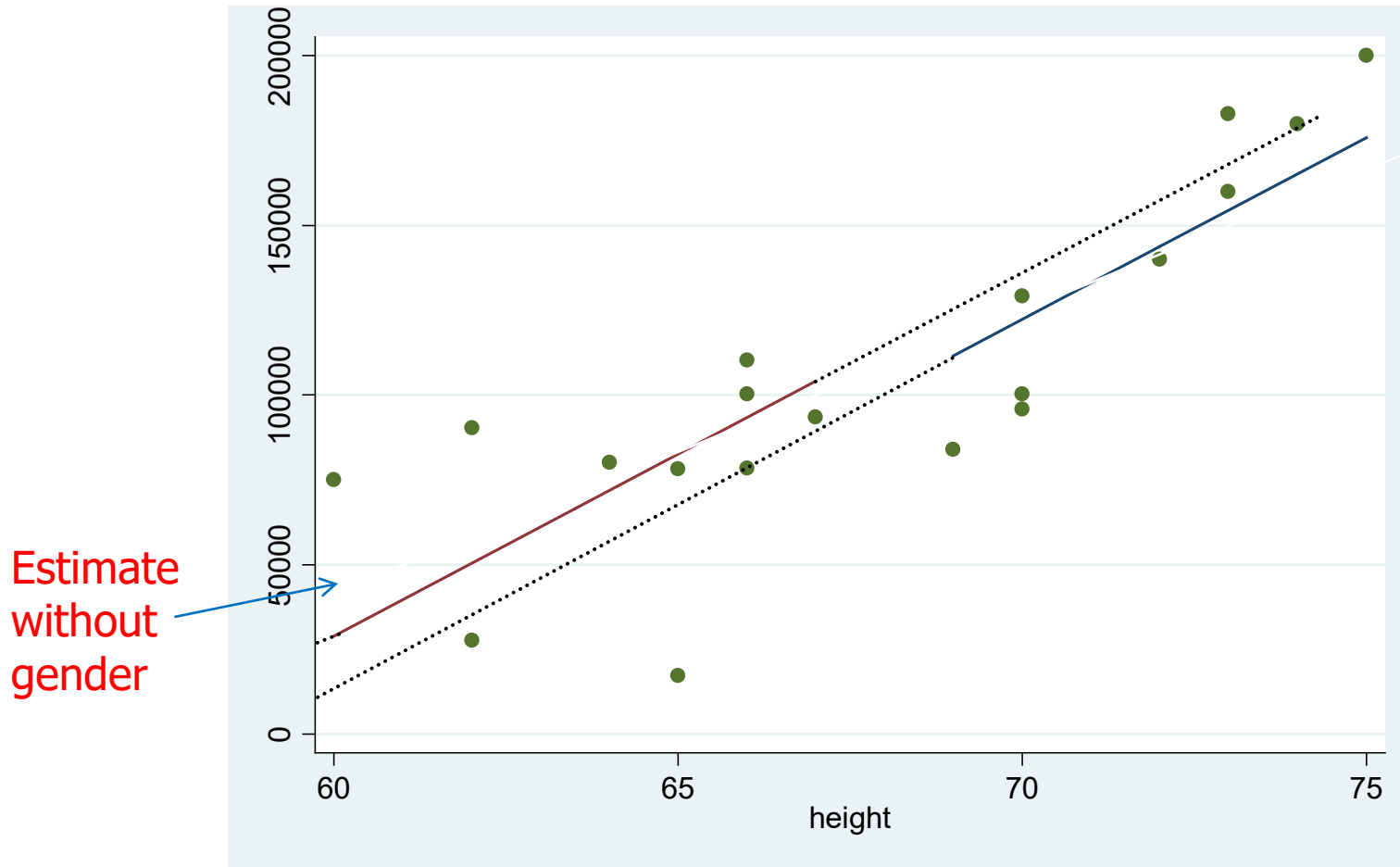
Ramsey J. Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society*. 1969;Series B(31):350-371.

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
- Don't use it
 - Little penalty for adding a nuisance variable
 - BIG penalty for missing an important covariate
- There are better methods for model building

Bias if Gender is Ignored



Assumption 2

- Expected value of the error term is 0

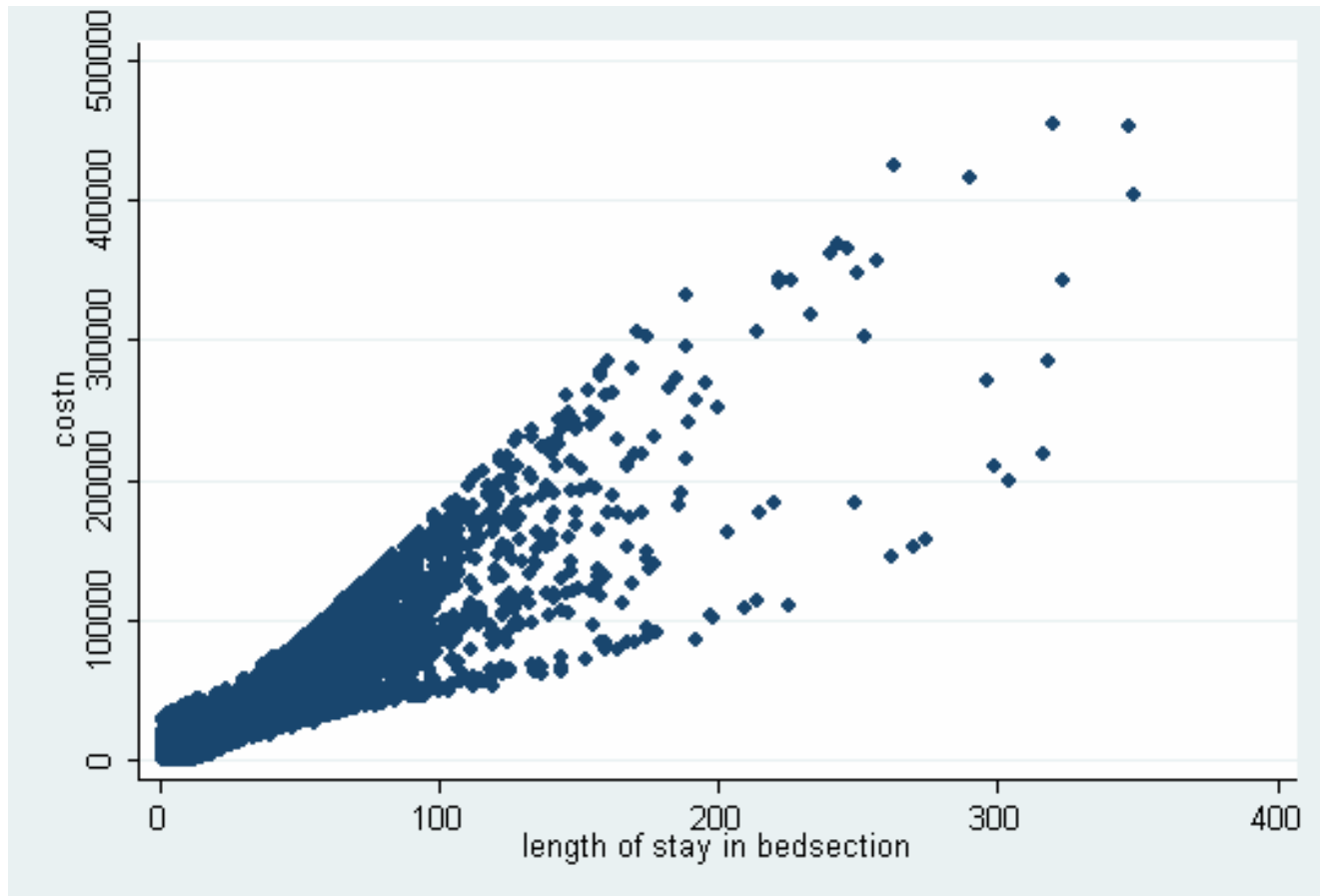
$$E(u_i)=0$$

- 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



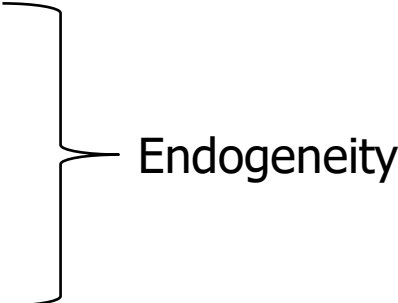
Violating Assumption 3

- 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)
- Potential solutions: instrumental variables, regression discontinuity (discussed in future classes)

Assumption 5

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

Regression References

- Kennedy A Guide to Econometrics
- Greene. Econometric Analysis.
- Wooldridge. Econometric Analysis of Cross Section and Panel Data.

Any Questions?

herc@va.gov
Todd.wagner@va.gov
twagner@Stanford.edu



@herc_va
@toddwagner