R Lecture #4

November 24th, 2017

Hyeokkoo Eric Kwon (KAIST)

hkkwon7@business.kaist.ac.kr

Econometrics: Paradigm

Econometrics

- Econometrics is the application of statistical techniques and analyses to the study of problems and issues in economics.
- Economics suggests important relationships, often with policy implications, but virtually never suggests quantitative magnitudes of causal effects.

1. *What is the quantitative effect of reducing class size on student achievement?*
2. *How does another year of education change earnings?*
3. *What is the price elasticity of cigarettes?*
4. *What is the effect on output growth of a 1 percentage point increase in interest rates by the Fed?*
5. *What is the effect on housing prices of environmental improvements?*

Steps in Empirical Economic Analysis

1. Specify an economic model.
2. Specify an econometric model.
3. Gather data.
4. Analyze data according to econometric model.
5. Draw conclusions about your economic model.

Step 1. Economic Model of Education

- What is the effect of education on wages?
  1. *wage=f(educ,exper,tenure)*
  2. *educ=years of education*
  3. *exper=years of workforce experience*
  4. *tenure=years at current job*

Step 2: Specify an Econometric Model

- In the wage example, we can’t reasonably observe all of the varia bles. For example, what matters?
- We need to specify an econometric model based on observable factors.
  1. *Wage=f(educ, exper, tenure)* + ε

Step 3: Gathering Data

Types of Data:

- Cross-Sectional Data
- Time Series Data
- Panel/Longitudinal Data

Cross-Sectional Data

- A sample of individuals, households, firms, cities, states, or other units, taken at a given point in time
- Random Sampling
- Mostly used in applied microeconomics

- Examples
  1. *General Social Survey*
  2. *US Census*
  3. *Most other surveys*

Cross-Sectional Data (Cont'd)

Obs	wage	educ	exper	female	married
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	6.00	11	3	0	1
…	…	…	…	…	…
525	3.50	16	4	0	0
526	4.25	14	5	1	0

Time Series Data

- Observations on a variable or several variables over time
- E.g. stock prices, money supply, CPI, GDP, annual homicide rates, etc.
- Because past events can influence future events, and lags in behavior are common in economics, time is an important dimension of time-series
- More difficult to analyze than cross-sectional data
- Observations across time are not independent
- May also have to control for seasonality

Time Series Data (Cont'd)

Obs	year	avgmin	avgcov	unemp	gnp
1	1950	0.20	20.1	15.4	878.7
2	1951	0.21	20.7	16.0	925.0
3	1952	0.23	22.6	14.8	1015.9
…	…	…	…	…	…
37	1986	3.35	58.1	18.9	4281.6
38	1987	3.35	58.2	16.8	4496.7

Panel/Longitudinal Data

- A panel data set consists of a time series for each cross-sectional member
- E.g. select a random sample of 500 people, and follow each for 10 years.

obs	personid	year	wage	dinout
1	1	1990	5.50	2
2	1	1992	6.50	4
3	1	1994	6.75	4
4	2	1990	10.50	6
5	2	1992	10.50	5
6	2	1994	11.25	2
7	3	1990	7.75	5
…	…	…	…	…
900	300	1994	15.00	2

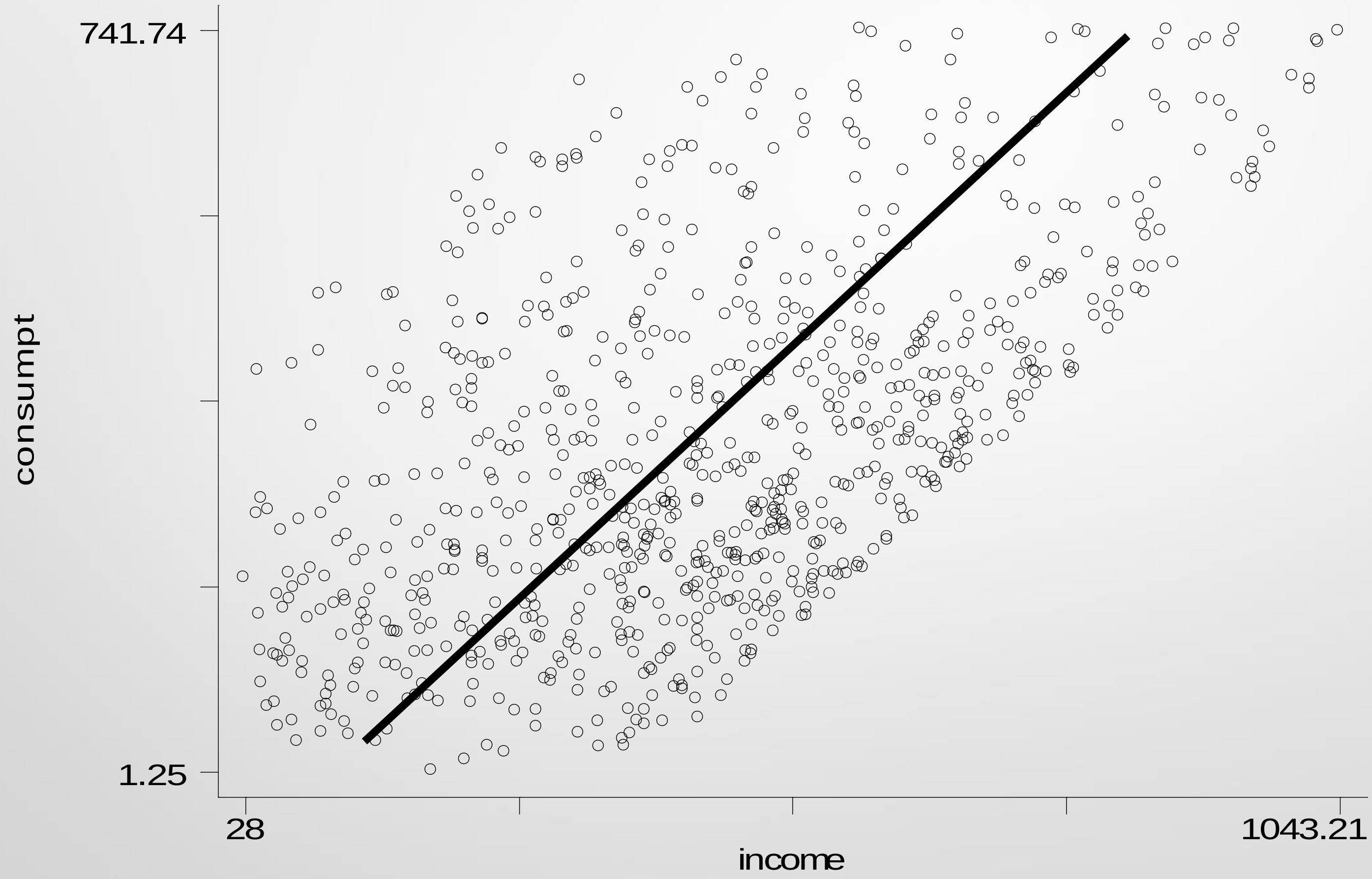
Steps 4 & 5:
Analyzing Data and Draw Conclusion

- Analyze the data based on the econometric estimation, validate the econometric findings, and draw conclusions.

- Why Use The Econometric Framework?
  1. *Understanding covariation*
  2. *Prediction of the outcome of interest*
  3. *The search for "causal" effects*

Model Estimation

Income & Consumption



Where OLS comes from

- Think of fitting a line to the data. This will never pass through every point
- Let *ui* be the deviation associated with the *i*th observation ("residual")

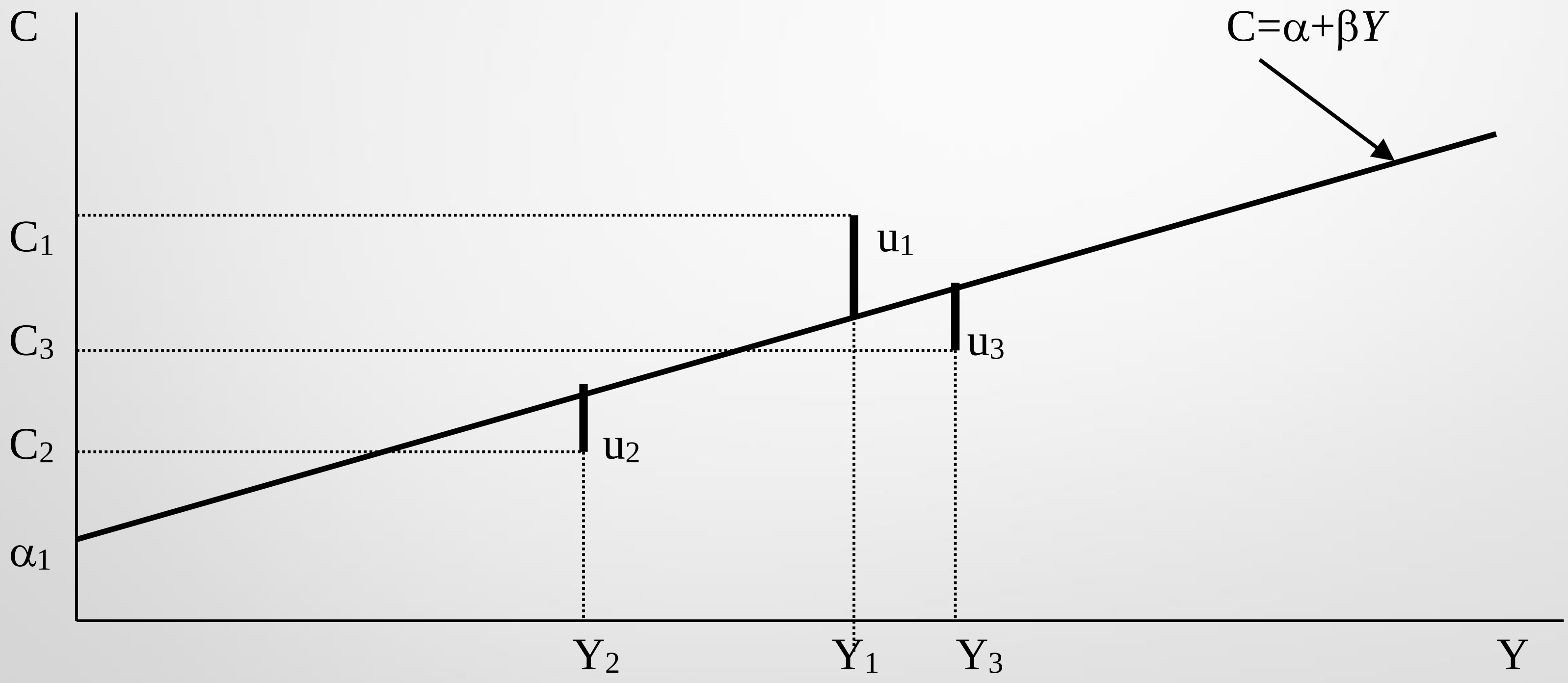
$$C_i = \alpha + \beta Y_i + u_i$$

$$\hat{C}_i = \alpha + \beta Y_i$$

$$C_i = \hat{C}_i + u_i$$

- Every choice of *a* and *b* will generate a new set of *ui*
- OLS chooses *a* and *b* to minimize sum of squared *ui*
- "Best fit" : *R2*

Where OLS comes from (Cont’d)



OLS Assumptions

- Model is linear in parameters.
- The data are a random sample of the population (independent).
- The expected value of the errors is always zero.
- The residuals have constant variance.
- The errors are normally distributed.

Simple OLS in R

lm(formula, data)

- **formula** : a symbolic description of the model to be fitted
- **data** : a data frame, list or environment containing the variables in the model

1. *install.packages("car")*
2. *library(car)*
3. *Prestige*

education : Average education of occupational incumbents, years, in 1971.

income : Average income of incumbents, dollars, in 1971.

women : Percentage of incumbents who are women.

prestige : Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s.

census : Canadian Census occupational code.

*type : Type of occupation. A factor with levels (**bc**, Blue Collar; **prof**, Professional, Managerial, and Technical; **wc**, White Collar)*

Simple OLS in R (Cont'd)

- **Linear Regression**
  1. *reg1 <- lm(prestige ~ education + income + women, data = Prestige)*
  2. *summary(reg1)*

Dependent Variable or *response variable, explained variable, outcome variable*

Independent Variable or *regressor, explanatory variable, predictor variable*

Simple OLS in R (Cont'd)

- **Linear Regression**
  1. *reg1 <- lm(prestige ~ education + income + women, data = Prestige)*
  2. *summary(reg1)*

```
Call:
lm(formula = prestige ~ education + income + women, data = Prestige)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.8246  -5.3332  -0.1364   5.1587  17.5045 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.7943342  3.2390886  -2.098   0.0385 *  
education    4.1866373  0.3887013  10.771  < 2e-16 ***
income       0.0013136  0.0002778   4.729 7.58e-06 ***
women       -0.0089052  0.0304071  -0.293   0.7702    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.846 on 98 degrees of freedom
Multiple R-squared:  0.7982,    Adjusted R-squared:  0.792 
F-statistic: 129.2 on 3 and 98 DF,  p-value: < 2.2e-16
```

a% risk of concluding that a relationship exists when there is no actual relationship

Simple OLS in R (Cont'd)

- **Linear Regression**
  1. *reg1 <- lm(prestige ~ education + income + women, data = Prestige)*
  2. *summary(reg1)*

```
Call:
lm(formula = prestige ~ education + income + women, data = Prestige)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.8246  -5.3332  -0.1364   5.1587  17.5045 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.7943342  3.2390886  -2.098   0.0385 *  
education    4.1866373  0.3887013  10.771  < 2e-16 ***
income       0.0013136  0.0002778   4.729 7.58e-06 ***
women       -0.0089052  0.0304071  -0.293   0.7702    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.846 on 98 degrees of freedom
Multiple R-squared:  0.7982,	Adjusted R-squared:  0.792 
F-statistic: 129.2 on 3 and 98 DF,  p-value: < 2.2e-16
```

About 80 percent of variance can be explained by the variables

Simple OLS in R (Cont'd)

- **Linear Regression**
  1. *reg1 <- lm(prestige ~ education + income + women, data = Prestige)*
  2. *summary(reg1)*

```
Call:
lm(formula = prestige ~ education + income + women, data = Prestige)

Residuals:
     Min       1Q   Median       3Q      Max 
-19.8246  -5.3332  -0.1364   5.1587  17.5045 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -6.7943342  3.2390886  -2.098   0.0385 *  
education    4.1866373  0.3887013  10.771  < 2e-16 ***
income       0.0013136  0.0002778   4.729 7.58e-06 ***
women       -0.0089052  0.0304071  -0.293   0.7702    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.846 on 98 degrees of freedom
Multiple R-squared:  0.7982,	Adjusted R-squared:  0.792 
F-statistic: 129.2 on 3 and 98 DF,  p-value: < 2.2e-16
```

Prestige Score increases as the education and income level increase

Simple OLS in R (Cont'd)

- **Factor variable regression with no interactions**
  1. *reg2 <- lm(prestige ~ education + income + type, data = Prestige)*

```
Call:
lm(formula = prestige ~ education + income + type, data = Prestige)

Residuals:
     Min       1Q   Median       3Q      Max 
-14.9529  -4.4486   0.1678   5.0566  18.6320 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.6229292  5.2275255  -0.119    0.905    
education    3.6731661  0.6405016   5.735 1.21e-07 ***
income       0.0010132  0.0002209   4.586 1.40e-05 ***
typeprof     6.0389707  3.8668551   1.562    0.122    
typewc      -2.7372307  2.5139324  -1.089    0.279    

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.095 on
  (4 observations deleted due to 
Multiple R-squared:  0.8349,    A
F-statistic: 117.5 on 4 and 93 DF
```

Ceteris paribus, Prof. and Wc. have higher and lower Prestige Score than Bc., Respectively.

	Bc	Prof	Wc
Intercept	**-0.62**	-0.62 + **6.04** **=5.42**	-0.62 **- 2.74** **= -3.36**
Education	3.67	3.67	3.67
Income	0.001	0.001	0.001

Simple OLS in R (Cont'd)

- **Factor variable regression with interactions**
  1. *reg3 <- lm(prestige ~ income + type*education, data = Prestige)*

```
Call:
lm(formula = prestige ~ income + type * education, data = Prestige)

Residuals:
     Min       1Q   Median       3Q      Max 
-15.1168  -4.1751   0.4384   5.1625  15.2362 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)       -2.331e+00  7.783e+00  -0.299    0.765    
income             1.052e-03  2.201e-04   4.782 6.66e-06 ***
typeprof           2.209e+01  1.520e+01   1.454    0.149    
typewc            -2.822e+01  1.959e+01  -1.440    0.153    
education          3.852e+00  9.406e-01   4.096 9.12e-05 ***
typeprof:education -1.227e+00  1.304e+00  -0.941    0.349    
typewc:education    2.270e+00  1.872e+00   1.213    0.228    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.036 on 91 degrees of freedom
  (4 observations deleted due
Multiple R-squared:  0.8411,
F-statistic: 80.27 on 6 and 9
```

The positive effect of Education Level on Prestige Score is greater for Prof. but smaller for Wc. compared to Bc.

	Bc	**Prof**	**Wc**
Intercept	-2.33	-2.33+**2.21** **=-0.12**	-2.33**-2.82** **=-5.15**
Education	3.85	3.85**-1.23** **=2.62**	3.85+**2.27** **=6.12**
Income	1.05	1.05	1.05

Logistic Regression for Binary Dependent Variable

glm(formula, data, family = "binomial")

- **formula** : a symbolic description of the model to be fitted
- **data** : a data frame, list or environment containing the variables in the model

1. *logitex <- read.csv("https://stats.idre.ucla.edu/stat/data/binary.csv")*
2. *logitreg1<- glm(admit ~ gre + gpa + rank, logitex , family = "binomial")*
3. *summary(logitreg1)*

Poisson Regression for Count Dependent Variable

glm(formula, data, family = "poisson")

- **formula** : a symbolic description of the model to be fitted
- **data** : a data frame, list or environment containing the variables in the model

1. *poissonex <- read.csv("https://stats.idre.ucla.edu/stat/data/poisson_sim.csv")*
2. *poissonreg1<- glm(num_awards ~ prog + math, poissonex, family = "poisson")*
3. *summary(poissonreg1)*

Fixed Effect Regression for Panel Data

plm(formula, data, index, model="within")

- **formula** : a symbolic description of the model to be fitted
- **data** : a data frame, list or environment containing the variables in the model
- **index** : the individual and time indexes

1. *install.packages("plm")*
2. *library(plm)*
3. *panel_data <-read.csv("panel data.csv", sep=",")*
4. *FEregression<-plm(invest~mvalue+kstock, panel_data, index=c("company","year"), model="within")*
5. *summary(FEregression)*

Experiments

Difficulties Arising from Observational to Estimate Causal Effects

- Ideally, we would like an experiment
- But almost always we only have observational (nonexperimental) data
- Challenges include
  1. *confounding effects (omitted factors)*
  2. *simultaneous causality*
  3. *"correlation does not imply causation"*

Causality & Ceteris Paribus

- What we really want to know is: does the independent variable have a causal effect on the dependent variable
- But: <u>Correlation does not imply causation</u>
- Suppose we want to know if higher education leads to higher worker wage

Causality & Ceteris Paribus (Cont'd)

- If we find a relationship between education and wages, we don't know much
- Why? What if highly educated people have higher IQs, and it's really high IQ that leads to higher wages?
- If you give a random person more education, will they get higher wages?

Causality & Ceteris Paribus (Cont'd)

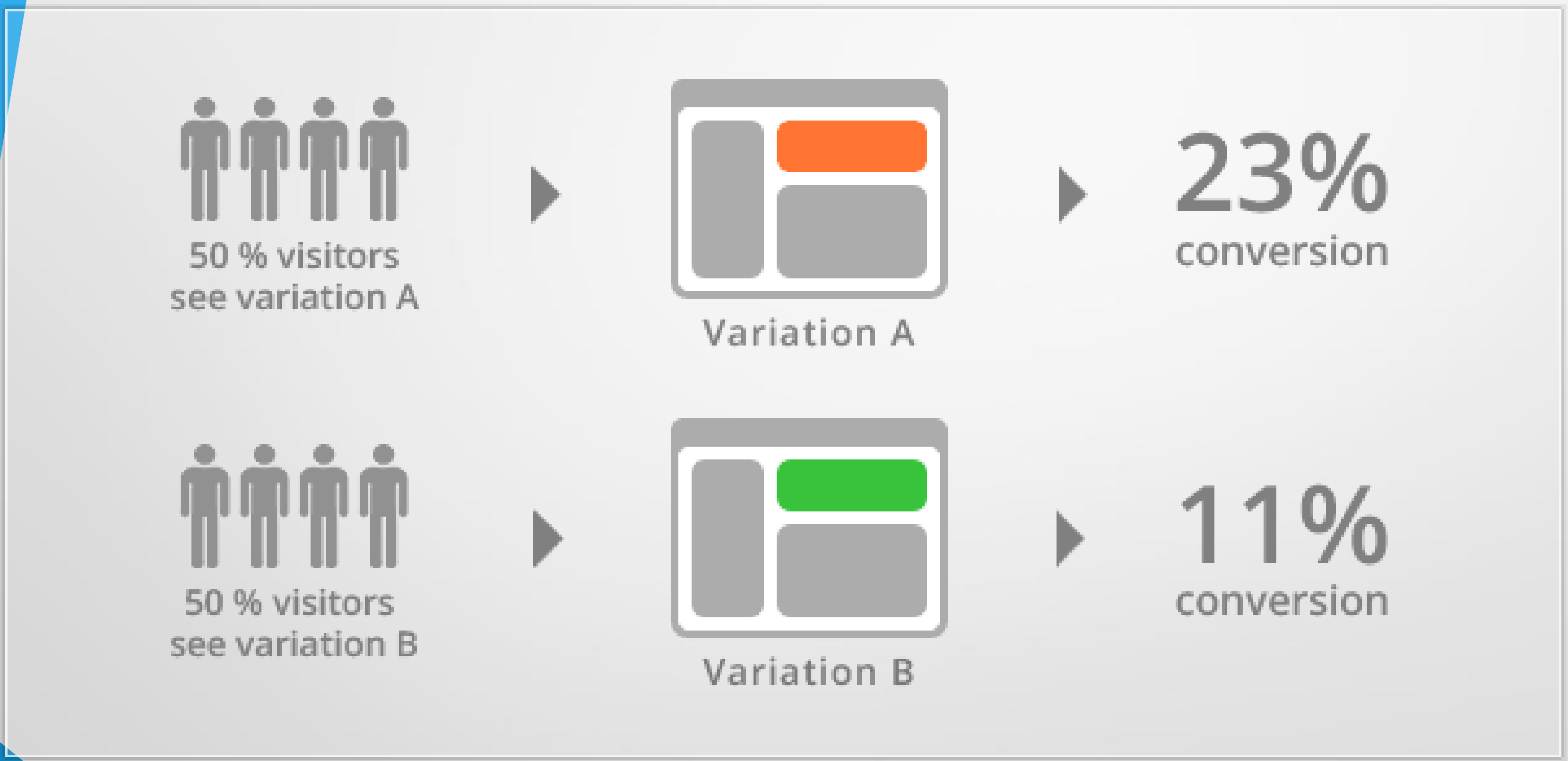
- What we want to know is… Does higher education lead to higher wages ceteris paribus… <u>holding all else constant</u>
- We have to control for IQ, experience, gender, job training, etc.
- But we can't control for everything!

What is an Experiment?

- Research method in which
  1. *conditions (extraneous) are controlled*
  2. *so that 1 or more independent variables can be manipulated to test a hypothesis about a dependent variable.*

- Allows
  1. *evaluation of causal relationships among variables*
  2. *while all other variables are eliminated or controlled.*

A/B Test



50 % visitors
see variation A
Variation A
23%
conversion
50 % visitors
see variation B
Variation B
11%
conversion

Definitions

Experimental Treatments

- Alternative manipulations of the independent variable being investigated

Experimental Group

- Group of subjects exposed to the experimental treatment

Control Group

- Group of subjects exposed to the control condition
- Not exposed to the experimental treatment

Randomization

- Assignment of subjects and treatments to groups is based on chance

Pretest-Posttest Control Group Design

- A.K.A., **Before-After with Control**
- True experimental design
- Experimental group tested before and after treatment exposure
- Control group tested at same two times without exposure to experimental treatment
- Includes random assignment to groups
- Effect of all extraneous variables assumed to be the same on both groups
- Do run the risk of a testing effect

Pretest-Posttest Control Group Design (Cont'd)

X = exposure of a group to an experimental treatment
O = observation or measurement of the dependent variable
$\boxed{\textbf{R}}$ = random assignment of test units;
individuals selected as subjects for the experiment are randomly assigned to the experimental groups

- **Diagrammed as**

	Before		After
1. Treatment Group:	O_1	X	O_2
2. Control Group:	O_3		O_4

- **Effect of the experimental treatment equals**

$$(O_2 - O_1) \text{ -- } (O_4 - O_3)$$

Difference-in-Difference

- **Diagrammed as**

	Before		After
1. Treatment Group:	O_1	X	O_2
2. Control Group:	O_3		O_4

- **Effect of the experimental treatment equals**

$$(O_2 - O_1) -- (O_4 - O_3)$$

- $y_{it} = \beta_0 + \beta_1 \text{treat}_i + \beta_2 \text{after}_t + \beta_3 \text{treat}_i{*}\text{after}_t$

		$\text{after}_t = 0$	$\text{after}_t = 1$
		Before	**After**
$\text{treat}_i = 1$	Treatment	β_1	$\beta_1+\beta_2+\beta_3$
$\text{treat}_i = 0$	Control	0	β_2

- **Effect of the experimental treatment equals**

$$(O_2 - O_1) -- (O_4 - O_3) = \boldsymbol{\beta_3}$$

Difference-in-Difference (Ex)

What is the effect of increasing the minimum wage on employment at fast food restaurants?

Confounding factor: national recession

Treatment group = NJ (New Jerwery)
Control group = PA (Pennsylvania)

Before = Feb 92
After = Nov 92

1. *did <- read.csv("did.csv")*
2. *didreg = lm(fte ~ treated + t + treated*t, did)*
3. *summary(didreg)*

id	Store ID
t	Feb. 1992 = 0; Nov. 1992 = 1
treated	New Jersey = 1; Pennsylvania = 0
fte	Output: Full Time Employment
bk	Burger King == 1
kfc	Kentuky Fried Chiken == 1
roys	Roy Rogers == 1
wendys	Wendy's == 1