RESEARCH METHODS GROUP TRAINING

April 27, 2018

Outline

- Introduction to Research Methods (30 min)
- Introduction to OLS Regression (30 min)
- Research Idea Workshop (1 hour)
  - 40min group activity
  - 20min group activity

Research Process

- Typical Steps in Research:
  - Develop an idea
  - Formulate a testable hypothesis
  - Reviewing the literature
  - Conduct a pilot study-if collecting original data
  - Data collection
  - Data Analysis
  - Interpretation and conclusion

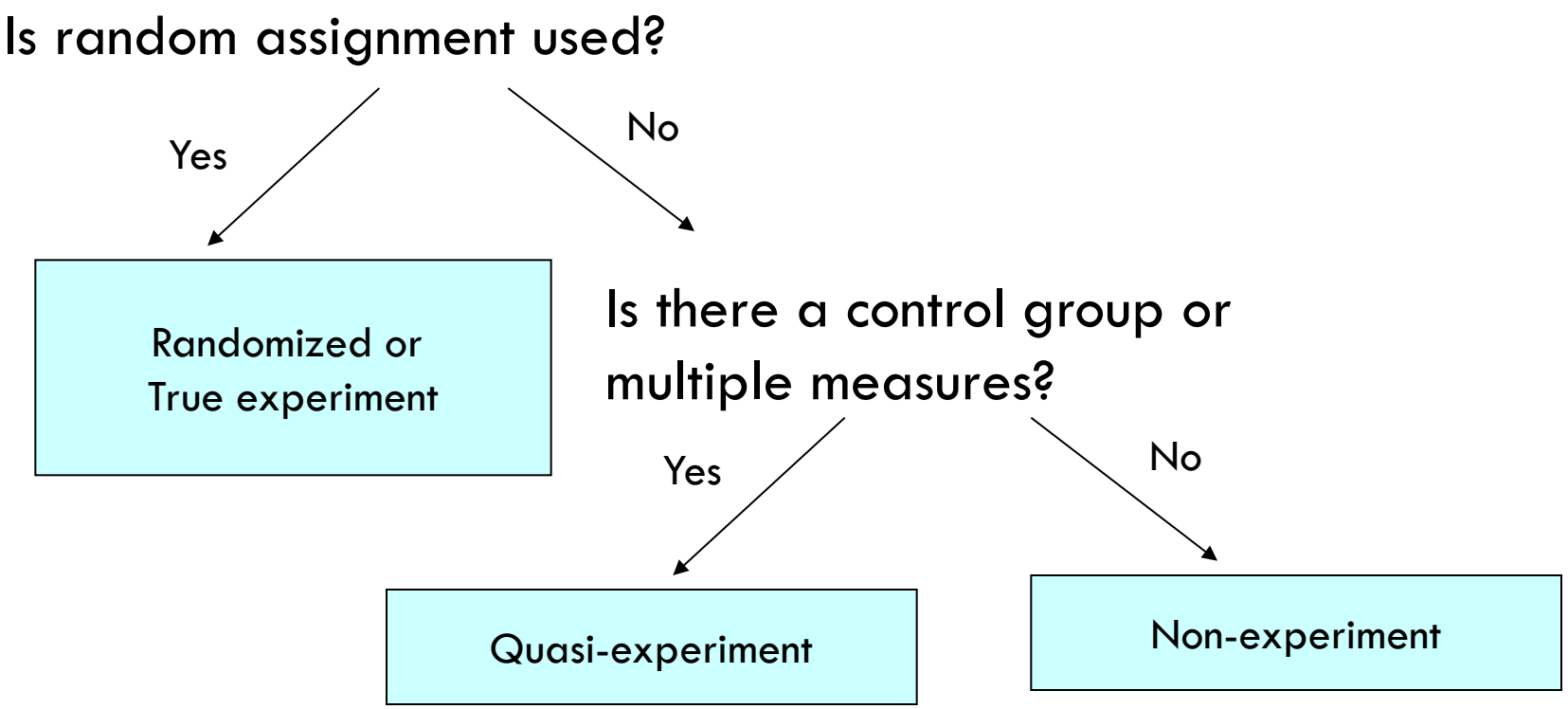
Research Methods

Method	Goal	Advantages	Disadvantages
Descriptive	Examine the current state of affairs	Shows what is occurring at a given time. Helps to give rise to research questions	Doesn't assess the strength of relationship between variables
Correlational	Examines the relationship between variables	Tests the strength of relationship between variables	Can't determine causality
Experimental	Understand causal impact of one variable on another	Determine causal relationships between variables.	In IR this type of research is often unethical

Experimental Research

- An independent variable (IV) is manipulated
- A dependent variable(s) (DV) is measured
- Subjects are randomly assigned to research groups

Types of Designs



Source: https://www.colorado.edu/geography/foote/geog5161/presentations/2009/hong_method.ppt

Quasi-Experimental Design

- Similar to the experimental design, but lacks the key ingredient, "***random assignment***"
- Easily and more frequently implemented
- Extensively used in the social sciences (Great for IR as well)
  - A useful method for measuring social variables
- Two classic quasi-experimental designs
  - Nonequivalent Design (e.g., matched pairs)
  - The Regression-Discontinuity Design

Source: https://www.colorado.edu/geography/foote/geog5161/presentations/2009/hong_method.ppt

What Does Nonequivalent Mean?

- Assignment is nonrandom.
- Researcher didn't control assignment.
- Groups *may* be different.
- Group differences *may* affect outcomes.

Source: http://webcache.googleusercontent.com/search?q=cache:G0bmP6yxWTQJ:www.cpp.edu/~smemerson/nbu-program-eval/trochimppp/Part%25204/Nonequivalent%2520Groups.ppt+&cd=1&hl=en&ct=clnk&gl=us

3 types of nonequivalent group design

(1) the differential research design.

(2) the posttest- only non-equivalent control group design.

(3) the pretest– posttest nonequivalent control group design.

Source: http://web.csulb.edu/~arezaei/EDP520/powerpoint/10-%20The%20Nonexperimental%20and%20Quasi-%20Experimental%20Strategies%20(short).pptx

The Nonequivalent Groups Design

- The most frequently used in social research
- Try to select groups that are as similar as possible to compare the treated one with the comparison one
  - e.g. two comparable classrooms or schools
  - Cannot be sure whether the groups are comparable
    - The groups may be different prior to the study
      - Any prior differences between the groups may affect the outcome of the study

Source: https://www.colorado.edu/geography/foote/geog5161/presentations/2009/hong_method.ppt

1- Differential research design

- Group differences are the primary interest not the cause an effect.
- Studies pre-existing groups
- No treatment
- This type of study often is called ex post facto research because it looks at differences " after the fact;" that is, at differences that already exist between groups.

Source http://web.csulb.edu/~arezaei/EDP520/powerpoint/10-%20The%20Nonexperimental%20and%20Quasi-%20Experimental%20Strategies%20(short).pptx

Example

- Difference between CSULB and CSULA students in their SAT math scores.

Source http://web.csulb.edu/~arezaei/EDP520/powerpoint/10-%20The%20Nonexperimental%20and%20Quasi-%20Experimental%20Strategies%20(short).pptx

2- The Posttest- Only Nonequivalent Control Group Design

- This type of study is occasionally called a static group comparison.
- Compares treatment with no-treatment group

X	O	(treatment group)
	O	(nonequivalent control group)

Source http://web.csulb.edu/~arezaei/EDP520/powerpoint/10-%20The%20Nonexperimental%20and%20Quasi-%20Experimental%20Strategies%20(short).pptx

Examples

- Example, Difference between those who take a course and those who don't.
- Comparing 2 high schools one with a pregnancy prevention program and one without
- Comparing two classes after they were taught with 2 different teaching methods.

Source http://web.csulb.edu/~arezaei/EDP520/powerpoint/10-%20The%20Nonexperimental%20and%20Quasi-%20Experimental%20Strategies%20(short).pptx

3- The Pretest– Posttest Nonequivalent Control Group Design

- A much stronger version of the nonequivalent control group design is often called a pretest– posttest nonequivalent control group design and can be represented as follows:

O	X	O	(treatment group)
O		O	(nonequivalent control group)

- The addition of the pretest measurement allows researchers to address the problem of assignment bias that exists with all nonequivalent group research.

Source http://web.csulb.edu/~arezaei/EDP520/powerpoint/10-%20The%20Nonexperimental%20and%20Quasi-%20Experimental%20Strategies%20(short).pptx

One- group pretest– posttest design

- Because the one- group pretest– posttest study precludes a cause- and- effect conclusion, this type of research is classified as non-experimental.

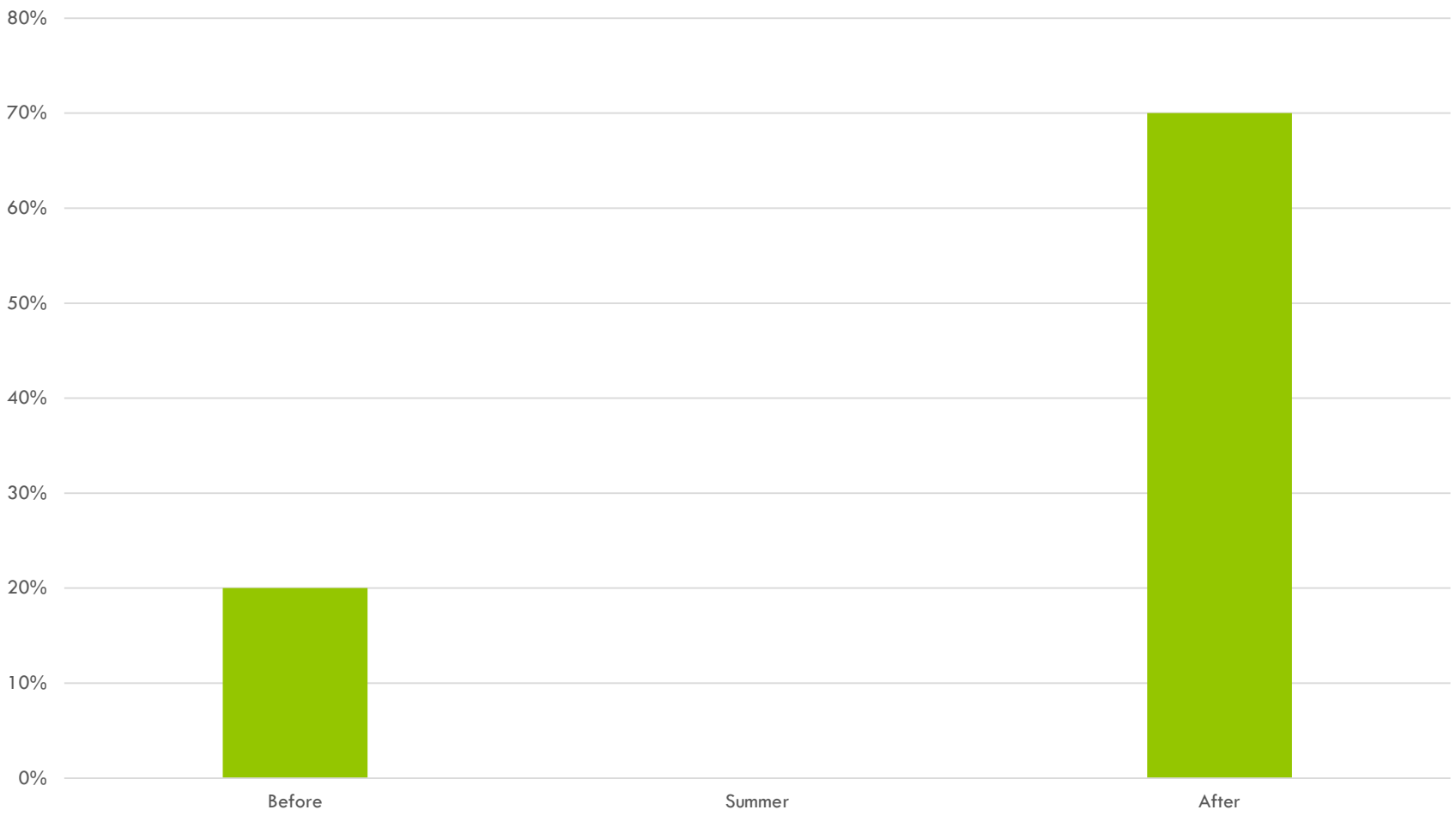
- example, political advertisement

O X O

May October

Source http://web.csulb.edu/~arezaei/EDP520/powerpoint/10-%20The%20Nonexperimental%20and%20Quasi-%20Experimental%20Strategies%20(short).pptx

A candidate's approval rate



Source http://web.csulb.edu/~arezaei/EDP520/powerpoint/10-%20The%20Nonexperimental%20and%20Quasi-%20Experimental%20Strategies%20(short).pptx

Causal Inference with Observational Data

Why is casual inference with observational data important for IR?

- It's often not possible to conduct experiments in institutional research (ethnics)
- These research methods allow us to mimic the "power" of experimentation

What is Different in These Methods?

- How the matching is done
  - Propensity Score Matching (PSM) estimates the probability that each person in both groups is (or would have been) in the treated group, based on their matching variables. That probability is the sole basis for matching.
  - Coarsened Exact Matching (CEM) matches people who match on every variable, but there is flexibility in what counts as a match. The matching rules are tight in the beginning and, if necessary, are systematically relaxed to allow for more matches, up to a pre-determined coarsening limit.
  - Regression Discontinuity Design (RDD) is a quasi-experimental pretest-posttest design that elicits the causal effects of interventions by assigning a cutoff or threshold above or below which an intervention is assigned.

Source:
https://www.researchgate.net/profile/Aran_Canes/publication/313473908_A_Scientific_Comparison_of_Coarsened_Exact_Matching_and_Propensity_Score_Matching/links/589b825192851c942ddae66a/A-Scientific-Comparison-of-Coarsened-Exact-Matching-and-Propensity-Score-Matching

What Else is Different?

- Theoretical justification
  - PSM assumes that we CAN predict the probability of assignment to either the treatment group or the untreated controls, and all other confounders are ignorable.
  - CEM assumes that the matching variables contain all the confounders, or that matching on the variables that we have will result in matches on the confounders that we don't have. Obviously, in CEM we also have to assume that any coarsening that we do results in errors that are within tolerable limits.

Source:
https://www.researchgate.net/profile/Aran_Canes/publication/313473908_A_Scientific_Comparison_of_Coarsened_Exact_Matching_and_Propensity_Score_Matching/links/589b825192851c942ddae66a/A-Scientific-Comparison-of-Coarsened-Exact-Matching-and-Propensity-Score-Matching

Coarsened Exact Matching

- Goal: Match patients so well that you could imagine that they were randomly assigned to each group
- For each patient in the treatment group, find at least one untreated patient from the comparison group who is identical or as similar as possible on all baseline characteristics
- By matching patients at the individual level, the treatment and comparison groups will be matched at the group level
- To find the R, Stata and SAS Code for CEM visit this website https://gking.harvard.edu/cem

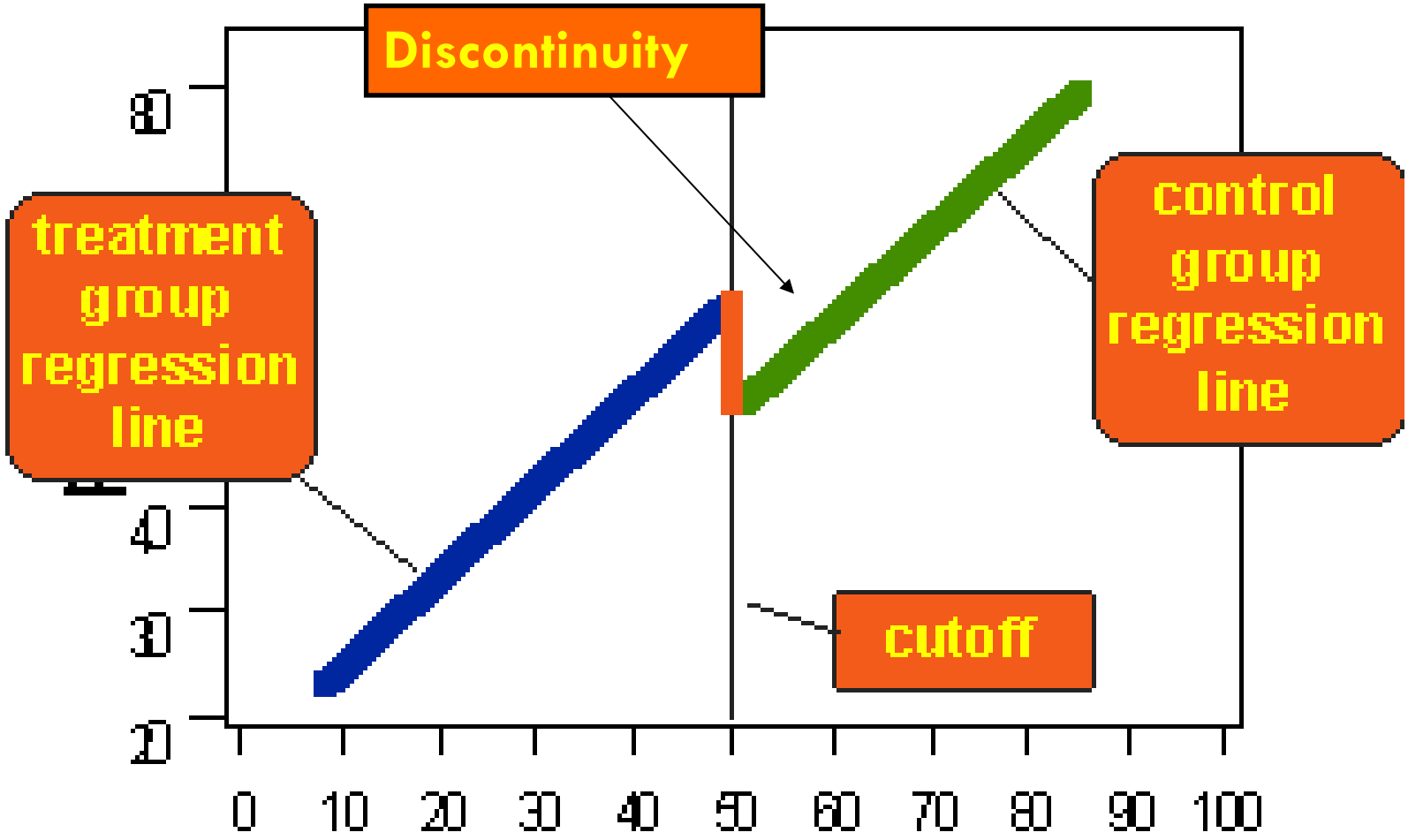
Source:
https://www.researchgate.net/profile/Aran_Canes/publication/313473908_A_Scientific_Comparison_of_Coarsened_Exact_Matching_and_Propensity_Score_Matching/links/589b825192851c942ddae66a/A-Scientific-Comparison-of-Coarsened-Exact-Matching-and-Propensity-Score-Matching

The Regression-Discontinuity Design

- A useful method for determining whether a program of treatment is effective
- Participants are assigned to program or comparison groups based on *a cutoff score* on a pretest
  - e.g. Evaluating new learning method to children who obtained low scores at the previous test.
    - Cutoff score = 50
    - The treatment group: children who obtained 0 to 50
    - The comparison group: children who obtained 51 to 100
- The program (treatment) can be given to those most in need

Source: https://www.colorado.edu/geography/foote/geog5161/presentations/2009/hong_method.ppt

The Regression-Discontinuity Design



Source: https://www.colorado.edu/geography/foote/geog5161/presentations/2009/hong_method.ppt

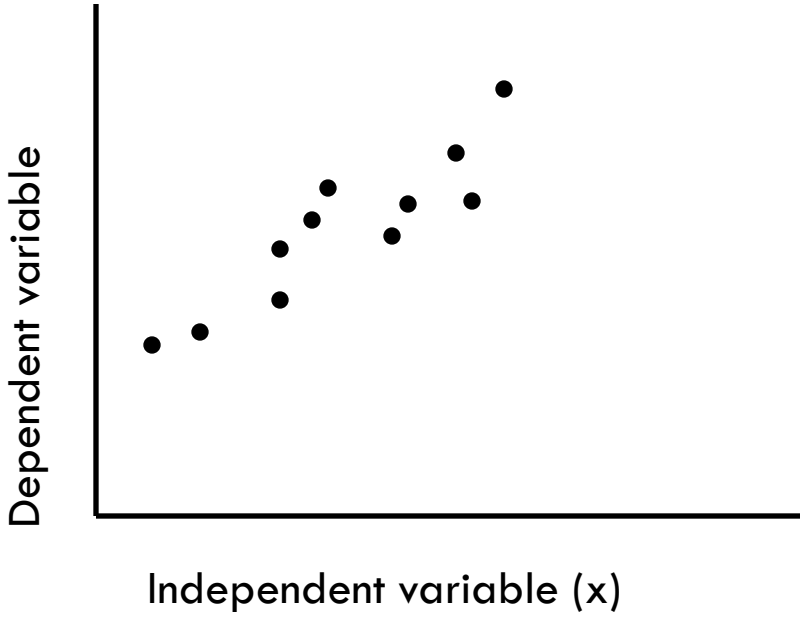
Introduction to OLS Regression

Some Types of Regression Models

Name	Description
OLS Regression or Linear Regression	the dependent variable is continuous, independent variable(s) can be continuous or discrete, and nature of regression line is linear.
Logistic Regression	Logistic regression is used to find the probability of event=Success and event=Failure. We should use logistic regression when the dependent variable is binary
Polynomial Regression	A regression equation is a polynomial regression equation if the power of independent variable is more than 1. The equation below represents a polynomial equation. In this regression technique, the best fit line is not a straight line. It is rather a curve that fits into the data points.
Stepwise Regression	This form of regression is used when we deal with multiple independent variables. In this technique, the selection of independent variables is done with the help of an automatic process, which involves *no* human intervention. Standard stepwise regression does two things. It adds and removes predictors as needed for each step. (it gives you R squared change)
Ridge Regression	Ridge Regression is a technique used when the data suffers from multicollinearity (independent variables are highly correlated)

Source: https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/

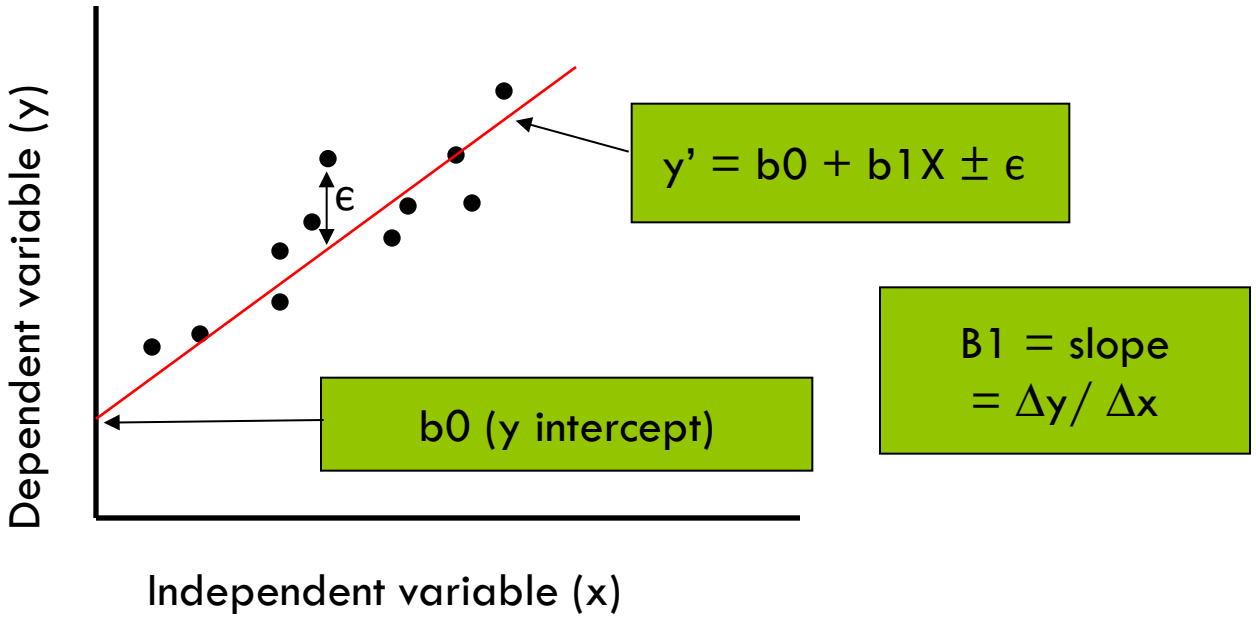
Introduction to Regression



Regression is the attempt to explain the variation in a dependent variable using the variation in independent variables.

If the independent variable(s) sufficiently explain the variation in the dependent variable, the model can be used for prediction.

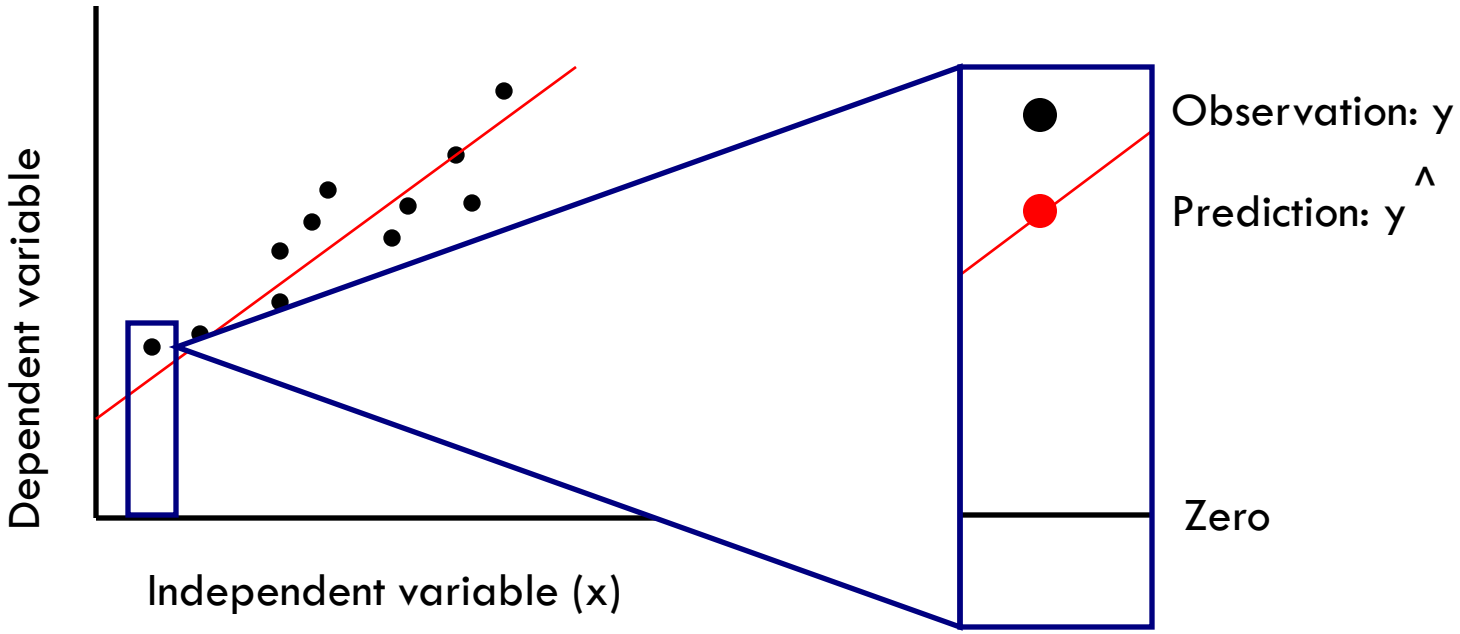
Source: http://www2.gsu.edu/~dscaas/pptdsc/regression.ppt



The output of a regression is a function that predicts the dependent variable based upon values of the independent variables.

Simple regression fits a straight line to the data.

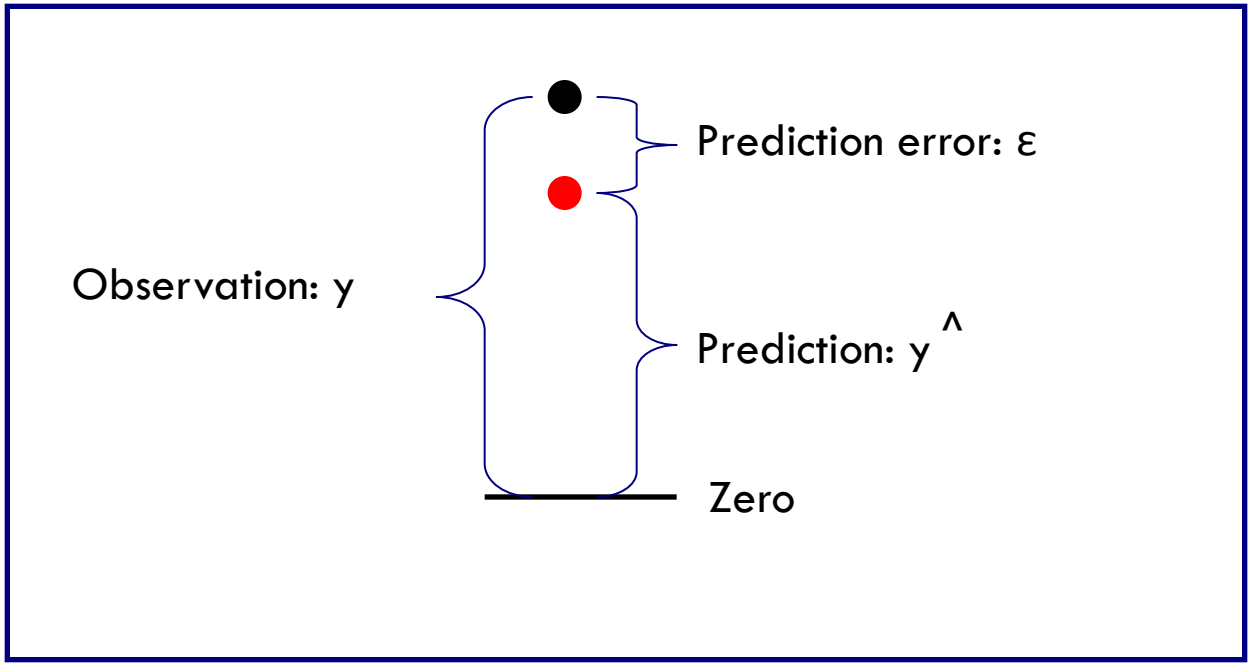
Source: http://www2.gsu.edu/~dscaas/pptdsc/regression.ppt



The function will make a prediction for each observed data point.

The observation is denoted by y and the prediction is denoted by y^.

Source: http://www2.gsu.edu/~dscaas/pptdsc/regression.ppt

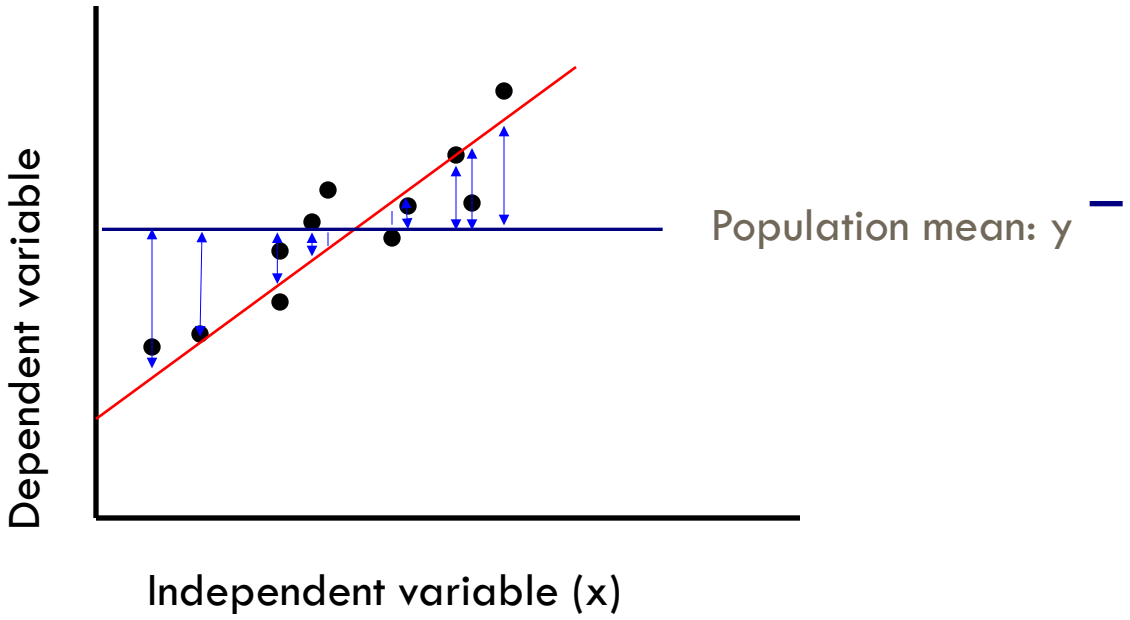


For each observation, the variation can be described as:

$$y = \hat{y} + \varepsilon$$

Actual = Explained + Error

Source: http://www2.gsu.edu/~dscaas/pptdsc/regression.ppt



The Sum of Squares Regression (SSR) is the sum of the squared differences between the prediction for each observation and the population mean.

Source: http://www2.gsu.edu/~dscaas/pptdsc/regression.ppt

OLS Regression Code

Program	Code
SAS	**proc glmselect** data=UCUESULONG; class specify your categorical predictors ; model Enter your predictor variables here /selection=none stb showpvalues; **run;**
Stata	Regress DV IV ,beta for standardized coefficients i.variable for categorical predictors
R	summary(lm(DV~ Cat.f, data = dataname)) The .f is for categorical predictors

Assumptions for OLS Regression

Assumption	Description
Linear relationship	linear regression needs the relationship between the independent and dependent variables to be linear. It is also important to check for outliers since linear regression is sensitive to outlier effects. The linearity assumption can best be tested with scatter plots
Multivariate normality	multivariate normality really ties back to all of your variables being normally distributed on a univariate level. This assumption can best be checked with a histogram or a Q-Q-Plot. When the data is not normally distributed a non-linear transformation (e.g., log-transformation) might fix this issue.
No or little multicollinearity	independent variables are too highly correlated with each other ; with VIF > 100 there is certainly multicollinearity among the variables.
No auto-correlation	Autocorrelation occurs when the residuals are not independent from each other. For instance, this typically occurs in stock prices, where the price is not independent from the previous price.
Homoscedasticity	The scatter plot is good way to check whether the data are homoscedastic (meaning the residuals are equal across the regression line).

The difference between the observed value of the dependent variable (y) and the predicted value ($\hat{y}$) is called the **residual** (e)

Sources: http://www.statisticssolutions.com/testing-multivariate-normality-in-spss/
https://www.statisticssolutions.com/assumptions-of-linear-regression/

OLS Regression Extra

- For future sessions we can discuss
  - Interaction Effects
  - Graphing main effects and interaction effects
    - Scatter plots
    - Simple slopes
      - Jeremy Dawson's Excel files http://www.jeremydawson.co.uk/slopes.htm
    - Predicted probability graphs
      - Stata and R (I have never done any graphing in SAS)
      - Tableau? Help

For Summer Meeting (Possible topics)

- Longitudinal Research Methods
  - HLM
- Regression Part Two
  - Logistic, Multinomial, Ordered Logistic, Regression for Count Data
- Predictive Modeling
- Graphing Effects (in excel and R)
  - Simple slopes
- Structural Equation Modeling
- What data's out there?
  - Use of Census data, data from ICPSR, PPIC, Pew, ANES, GSS and other sources

Idea Lab

Instructions

- Break into small groups
  - Undergraduates
  - Graduate Students
  - HR
- Go over your ideas for a research project
  - What is one (or more) outstanding "big question" in this area?
    - What is one small piece that IRAP could tackle? Is there a research question we can create?
      - What kinds of outcomes would we look for in this research? What kinds of inputs? Are these outcomes and inputs data that we have? Could we get access to them?
        - What specific research methods could be used with these data?