FLORIDA STATE UNIVERSITY

Introduction to Experimental Design and Logistic Regression

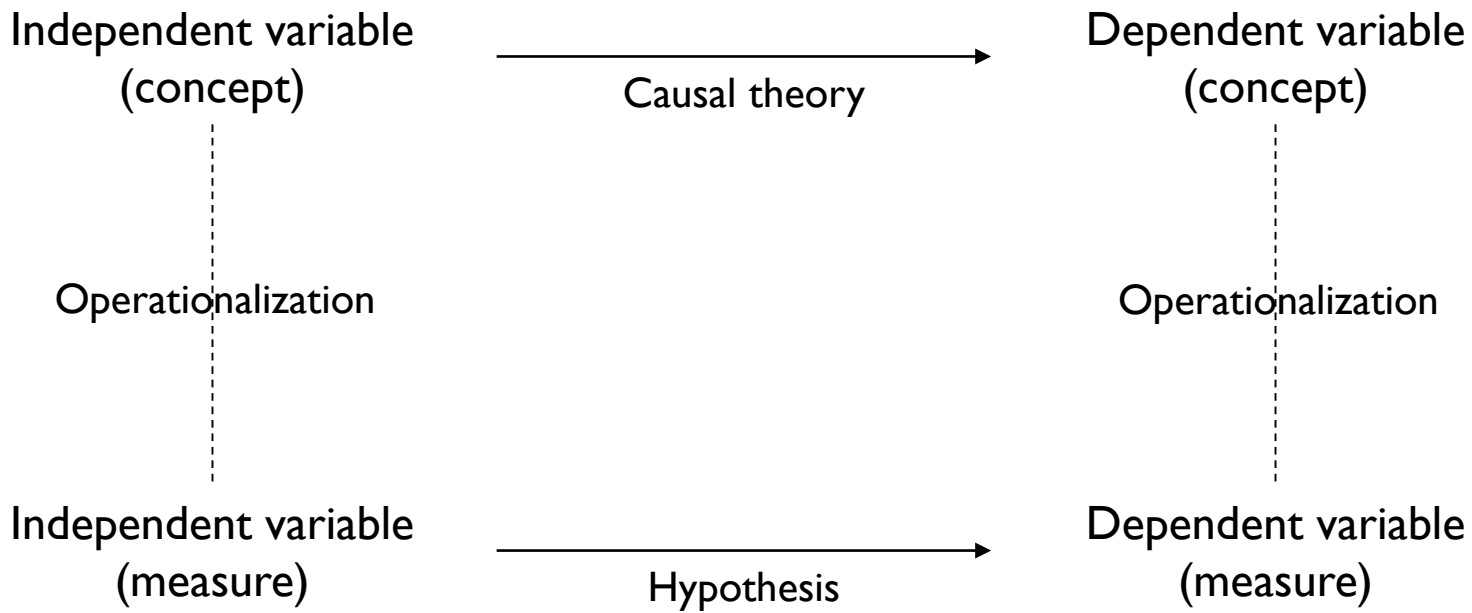
Dongfang Gaozhao | March 27, 2019 | dgaozhao@fsu.edu

The Nature of Science

- What is science?
  - Science is not a body of knowledge but a method of inquiry.
- The assumptions of science:
  - Nature is orderly and regular;
  - Phenomena have causes.

The Purposes of Social Science Research

- Description
  - Describing what, like U.S. Census does
- Explanation
  - Answering why and how
    - Theoretical level: based on reason and logic;
    - Empirical level: based on observation


Independent variable
(concept)
Causal theory
Dependent variable
(concept)
Operationalization
Operationalization
Independent variable
(measure)
Hypothesis
Dependent variable
(measure)

A Direct Effect

$$X \Rightarrow Y$$

James is happy so he smiles.

An Indirect Effect

$$X \Rightarrow Z \Rightarrow Y$$

The heat makes people thirsty and, thus, increases the sale of water.

Reverse Causation

$$X \Leftrightarrow Y$$

The U.S. and Russian military competition

A Spurious Relationship

$$X \Leftarrow Z \Rightarrow Y$$

As ice cream sales increase in a city (X), the city's murder rate increases as well (Y).

Correlated, but not causal!

Real-life Questions are Causal in Nature

We observe correlations, but ultimate goal is making causal inference! That is what we care and cause-effect relationships are a part of what public administrators do:

- Does school decentralization improve school quality?
- Does one more year of education cause higher income?
- Does conditional cash transfers cause better health outcomes in children?
- How do we improve student learning?

Causality

- Deterministic relationships
  - If some cause occurs, then the effect will occur with certainty.
- Probabilistic relationships
  - In the social sciences causation is usually understood as probabilistic;
  - An increase in x is associated with an increase (or decrease) in the probability of y occurring, so the effect is not certain.

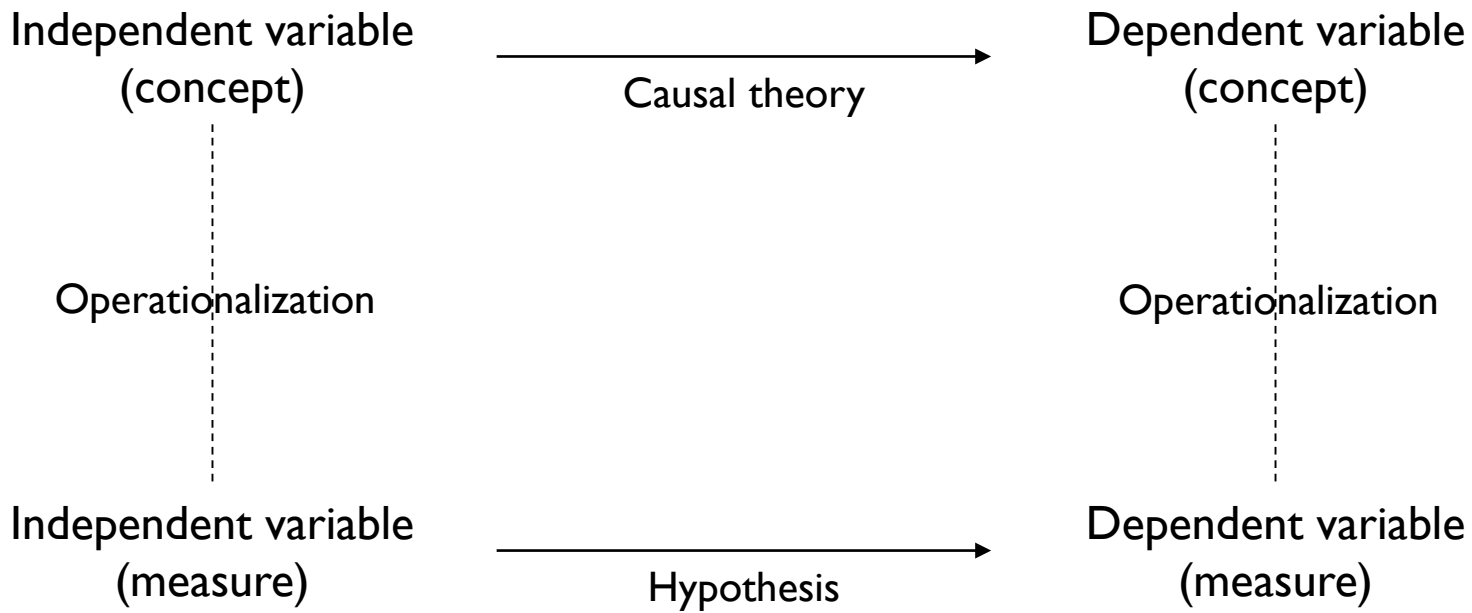
Causal Analysis

- For causal questions, we need to infer aspects of the data generation process.
- We need to be able to deduce:
  - the likelihood of events under static conditions, (as in standard statistical analysis) and also
  - the dynamics of events under changing conditions.

Experiments

Experiments are investigations in which an intervention, in all its essential elements, is under the control of the investigator.

- Two major types of control:
  - Control over assignment to treatment - this is at the heart of many field experiments;
  - Control over the treatment itself - this is at the heart of many lab experiments.
- Our focus: how does control allow researchers to make reasonable statements about causal effects?


Independent variable
(concept)
Causal theory
Dependent variable
(concept)
Operationalization
Operationalization
Independent variable
(measure)
Hypothesis
Dependent variable
(measure)

Input → Process → Output

Controlled inputs $x_1, x_2, \ldots$
Uncontrolled but observed inputs $u_1, u_2, \ldots$
Uncontrolled and unobserved inputs $v_1, v_2, \ldots$

Goals

By doing experiments, we want to understand:

- Characterization: How does inputs affect the output?
- Optimization: What input values produce the desired outputs?
- Control: How do we adjust controlled inputs to maximize control of the output?

Dealing with Inputs

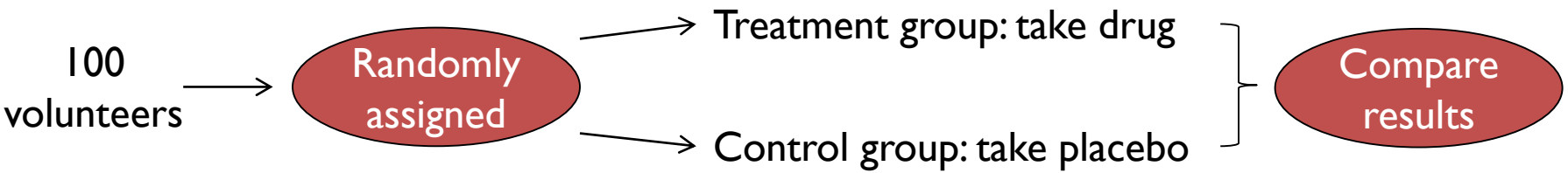
- Controlled inputs (x, i.e., education)
  - Variation + replication: Vary the inputs and repeat the experiments in a systematic way.
- Uncontrolled but observed inputs(u, i.e., gender)
  - Blocking: Group experiments into blocks, each block having a fixed value of u;
  - Analysis of covariance: Model the impact of u, then subtract it out from the model.
- Uncontrolled and unobserved inputs (v, i.e., personality)
  - Randomization: For sufficiently large sample sizes, let the impact of v average out to zero.

An Example of Experiment

- Question: A new drug causes violent behavior
- Problem: Many things other than the drug may result in violent behavior
- Solution: An experiment
  - Researchers randomly assign subjects to one of two groups: the treatment group (take drug) and the control group (placebo);
  - Researchers control the introduction of the experimental factor (the drug);
  - Measure and compare the two groups on the DV of interest (violent behavior). If there is a difference, researchers can attribute it to the experimental factor (the drug).

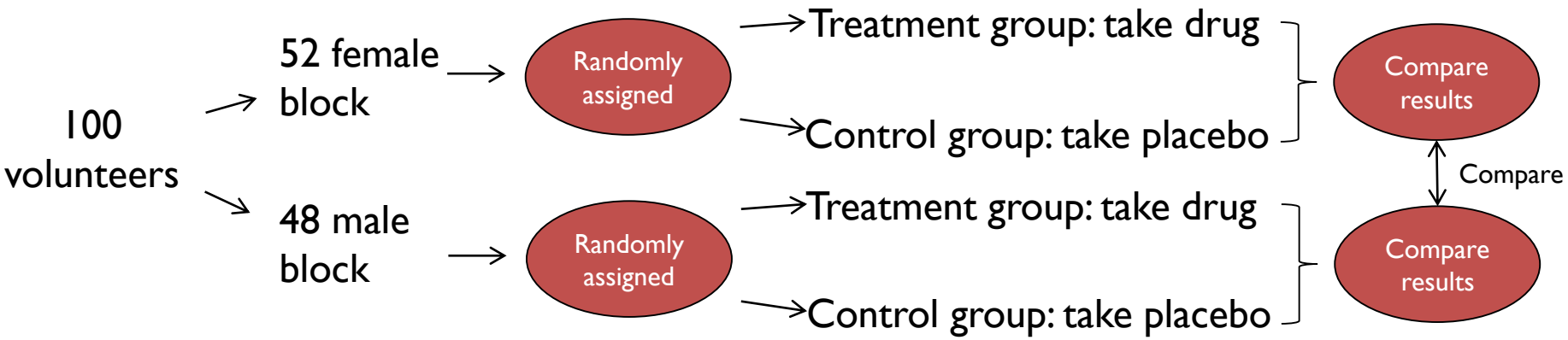
An Example of Experiment (Cont'd)

A total of 100 volunteers participate in the experiment:



An Example of Experiment (Cont'd)

If we believe that gender has an effect on the results:



Types of Experiment

- Laboratory experiment
  - Laboratory experiments take place in an artificial setting where variables and conditions are carefully controlled;
  - There must be a random assignment of participants to a condition;
  - One variable (x) is manipulated to see the effect it has on another variable (y);
  - Participants usually know they are being studied in an experiment but might not know the full purpose of the study;
  - Often high internal validity but low ecological validity.

Types of Experiment

- Field experiment
  - Similar to a lab experiment, but takes place in a more natural or ordinary setting. This type of experiment tends to be a real world setting, such as in a hospital or a school;
  - A random assignment is preferred but an opportunity sample may exist;
  - Participants often are not aware of them being in an experiment so their behavior may be more natural;
  - High ecological validity with extraneous variables and ethical issues (for instance, no informed consent);
  - Example: Bushman (1988) studies the power of uniform. A female researcher dressed either in a police-style uniform, a business executive or as a beggar stopped people in the street and told them to give change to a male researcher for an expired parking meter. The rates of obeisance are 72%, 47%, and 52%, respectively.

Types of Experiment

- Natural experiment
  - Natural experiments differ from field experiments in a way that it does not control the independent variable(s). Instead, the IV naturally occurs, like a war or a robbery;
  - Advantages of natural experiment:
    - Researchers can test in ethically sensitive areas of research;
    - High ecological validity (real life setting).
  - Disadvantage: No control over extraneous variables.

Types of Experiment

- Quasi experiment
  - No random allocation of participants to the conditions;
  - Concerns about internal validity and confounding variables;
  - But it has advantages of testing in ethically sensitive areas of research and allowing comparisons among different types of people;
  - Example: The impact of smoking on individuals' health.

Types of Experiment

- Survey experiment
  - Systematically vary one or more elements of the survey across subjects;
  - Usually do not include 'pre-test' measurement;
    - Thus, most survey experiments are not 'classic' in design;
    - "Posttest-only control group design."
  - Random assignment is critical to the design;
  - Generalizability and valid causal inference (both external and internal validity);
  - Does not have detailed information about people's reaction;
  - Useful for understanding citizens' attitudes towards certain issues.

Papers Using Survey Experiments: I

Riccucci, N. M., Van Ryzin, G. G., & Jackson, K. (2018). Representative bureaucracy, race, and policing: A survey experiment. *Journal of Public Administration Research and Theory*, *28*(4), 506-518.

- Research question: Whether varying the representation of black police officers in local agencies influences how black and white citizens judge the agency's performance, trustworthiness, and fairness in terms of civilian complaints of police misconduct.
- Results: Perceived performance, trust, and fairness increases among black citizens when the police force is composed of mostly black officers. For white citizens, the effect of greater black representation among the police is largely negative.

Papers Using Survey Experiments: I (Cont'd)

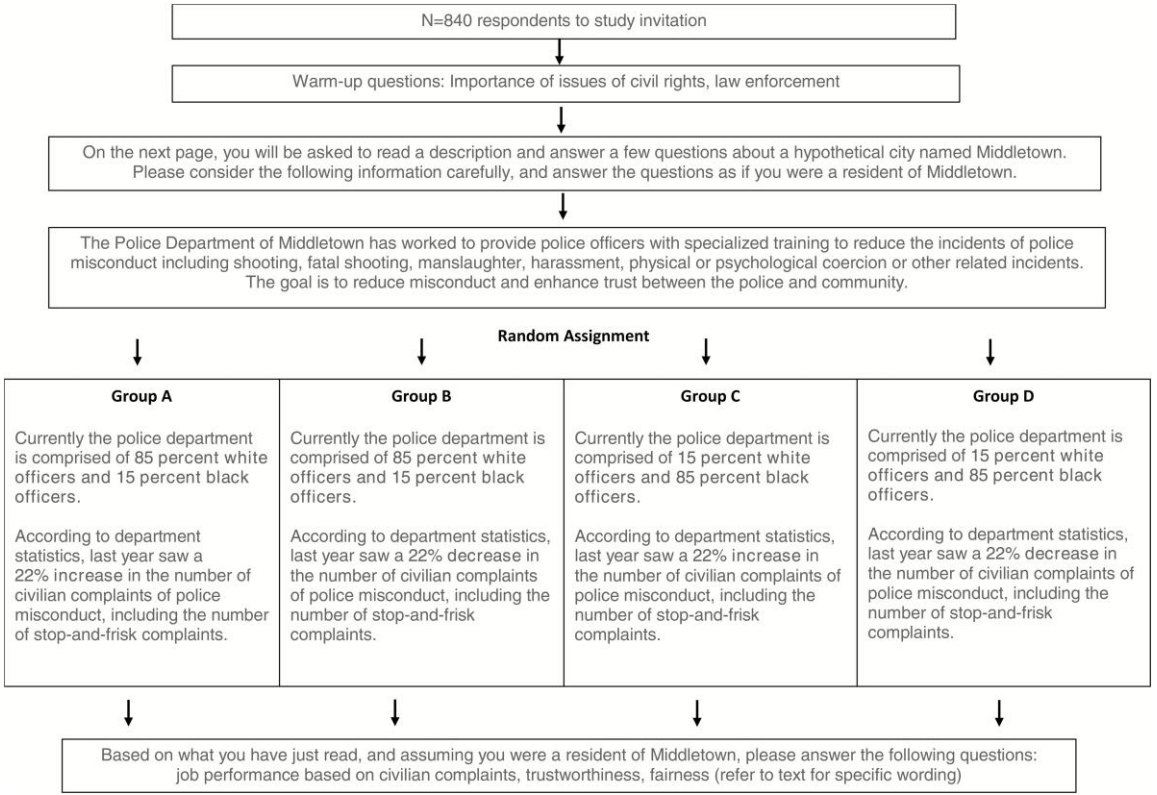


Figure 2. Experimental Design

Journal of Public Administration Research and Theory, Volume 28, Issue 4, October 2018, Pages 506–518, https://doi.org/10.1093/jopart/muy023

Papers Using Survey Experiments: 2

Bromberg, D. E., Charbonneau, É., & Smith, A. (2018). Body-worn cameras and policing: A list experiment of citizen overt and true support. *Public Administration Review*, *78*(6), 883-891.

- Research question: To what extent do citizens support for the uses of body-worn cameras?
- Results: Citizens remain skeptical about giving police officers discretion in turning on police body-worn cameras. Public polling may overestimate the level of support citizens have for police body-worn cameras. Police body-worn camera initiatives are more complex than frequently portrayed.

Papers Using Survey Experiments: 2 (Cont'd)

Control group (II) Two-fifths of the respondents, n ≈ 200	Treatment group (III) Two-fifths of the respondents, n ≈ 200
Q. How many of the following statements do you agree with (without revealing which statements):	How many of the following statements do you agree with (without revealing which statements):
Dispatch should control body cameras and activate them in appropriate situations	Dispatch should control body cameras and activate them in appropriate situations
Body cameras should be left on at all times	Body cameras should be left on at all times
Body cameras should be turned on at the point an officer is dispatched to the scene of a call	Body cameras should be turned on at the point an officer is dispatched to the scene of a call
I trust police to use their discretion in turning on their body camera	I trust police to use their discretion in turning on their body camera
Body cameras should only be used when the use of force is likely	Body cameras should only be used when the use of force is likely
Body cameras should be used to record all public encounters in relation to a call of service	Body cameras should be used to record all public encounters in relation to a call of service
Body cameras should be used to record encounters on private property in relation to a call of service	Body cameras should be used to record encounters on private property in relation to a call of service
Body cameras should primarily be used in lower-income urban centers	Body cameras should primarily be used in lower-income urban centers
All footage recorded on body cameras should be made available to the public immediately	All footage recorded on body cameras should be made available to the public immediately
I trust my local police department to disclose body camera footage when necessary	I trust my local police department to disclose body camera footage when necessary
Body camera footage should only be disclosed with a court order	Body camera footage should only be disclosed with a court order
Body camera footage should not be publicly disclosed if the suspect suffers from mental illness	Body camera footage should not be publicly disclosed if the suspect suffers from mental illness

Table 1. Design for the List Experiment (Partly)

Public Administration Review, Volume 78, Issue 6, March 2018, Pages 883-891, https://doi.org/10.1111/puar.12924

FLORIDA STATE UNIVERSITY

Let's talk about…

Logistic Regression

Continuous vs. Categorical

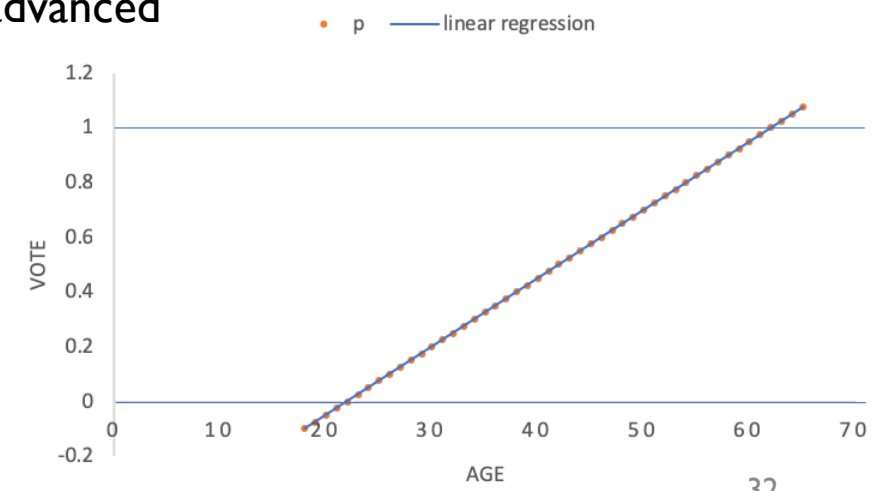
- General linear regression model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$
- Independent variables
  - Continuous: weight, income, age, and temperature
  - Categorical: gender, race, and education level
- Dependent variables
  - Continuous: spending, time, and profit
  - Categorical: yes or no

Why We Need Logistic Regression?

- Is an individual going to vote or not?
- There are two outcomes: yes ($y = 1$) and no ($y = 0$)
- We can use data to estimate the possibility of an individual voting, namely $p(vote)$, in an election and what determines hers/his decision to vote
- Here is a simplified single factor linear model: $p(vote) = \beta_0 + \beta_1 age + \varepsilon$
- By running the model, we may have: $p(vote) = -.550 + .025age$
- How to interpret the result?
- If an individual is 18, the odd of voting $= -.550 + .025 * 18 = -.1$

Issues with the Linear Model

- Probability is, by nature, bounded between 0 and 1 $(0 \leq p \leq 1)$
- The linear model, however, may generate values smaller than 0 and larger than 1
- Thus, an alternative approach needs to be advanced
- It must meet two requirements:
  - It must always be positive $(p \geq 0)$
  - It must be less than 1 $(p \leq 1)$
- We have $p = \frac{\exp(\beta_0 + \beta_1 age)}{\exp(\beta_0 + \beta_1 age) + 1}$



Inference

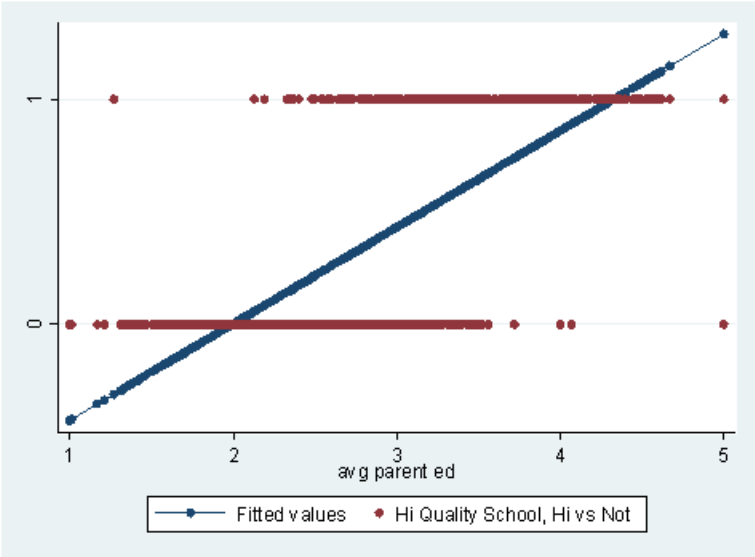
- $\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 \quad \ldots (1)$
- $p = \frac{\exp(\beta_0+\beta_1 age)}{\exp(\beta_0+\beta_1 age)+1} = \frac{e^{\beta_0+\beta_1 age}}{e^{\beta_0+\beta_1 age}+1}, \; p \in [0,1]$
- Denote RHS of (1) as t, then $p = \frac{e^t}{e^t+1}$
- $1 - p = \frac{e^t+1-e^t}{e^t+1} = \frac{1}{e^t+1}$, then $\frac{1}{1-p} = e^t + 1$
- $\ln\left(\frac{p}{1-p}\right) = \ln(\frac{e^t}{e^t+1} * (e^t + 1) = \ln(e^t) = t$

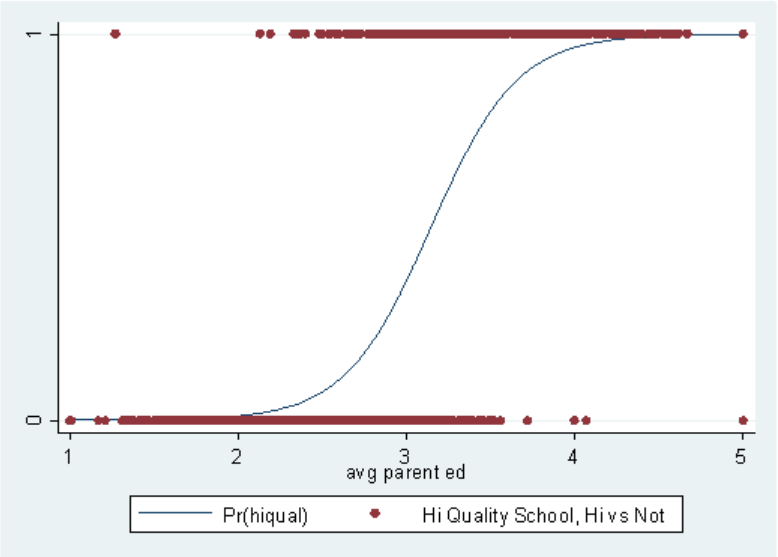
Interpreting Logistic Regression Results

- $\ln\left(\frac{p(vote)}{1-p(vote)}\right) = -.282 + .237age$
- Does it mean that each unit of increase in age will improve the possibility of voting by 23.7%?
- $p = \frac{\exp(-.282+.237age)}{\exp(-.282+.237age)+1}$
- $\Delta p = p_{t+1} - p_t$

```
. logit pvote age

Iteration 0:   log likelihood = -4.8607112
Iteration 1:   log likelihood = -4.1400003
Iteration 2:   log likelihood = -3.5297342
Iteration 3:   log likelihood = -3.4488647
Iteration 4:   log likelihood = -3.4440736
Iteration 5:   log likelihood = -3.4440668
Iteration 6:   log likelihood = -3.4440668

Logistic regression                             Number of obs     =         48
                                                LR chi2(1)        =       2.83
                                                Prob > chi2       =     0.0923
Log likelihood = -3.4440668                     Pseudo R2         =     0.2914

------------------------------------------------------------------------------
       pvote |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .2365463   .2373151     1.00   0.319    -.2285828    .7016755
       _cons |  -2.820409   5.387886    -0.52   0.601    -13.38047    7.739653
------------------------------------------------------------------------------
```

Logistic Regression in Stata: An Example





FLORIDA STATE UNIVERSITY

Summary

- Causality
- Experiments
  - Laboratory experiment
  - Field experiment
  - Natural experiment
  - Quasi experiment
  - Survey experiment
- Logistic regression

FLORIDA STATE UNIVERSITY

THANK YOU

Introduction to Experimental Design and Logistic Regression

Dongfang Gaozhao | March 27, 2019 | dgaozhao@fsu.edu