Big Data and Security

Jeffrey Borowitz, PhD

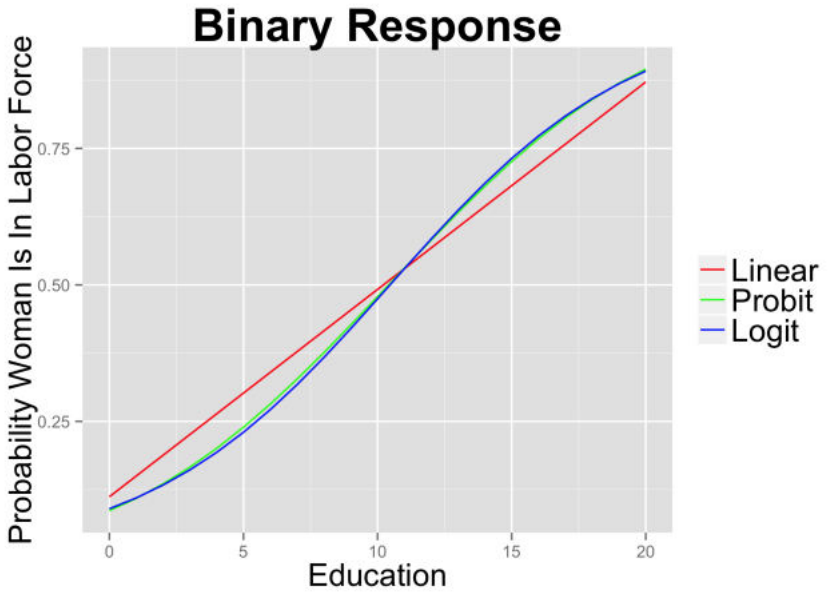
Lecturer

Sam Nunn School of International Affairs

Generalizations of Parametric Models, Part 1

Logistic Regression vs. Linear Regression

- Logistic regression is used often in practice, but often many models give similar outputs



Georgia Tech

Example: Predicting Whether a Woman Works for Pay

- Whether a woman works is binary: it's either yes or no
- What things might affect whether a woman works for pay?
  - If she has young children, that might make her less likely to work
  - If she's more educated, she might want to work (didn't spend all that time in school if she didn't want to use it)

Practicalities of Different Models

- Linear regression in R:
  - lm(inlf ~ educ, data = wages)
- Logistic regression in R:
  - glm(inlf ~ educ, data = wages, family='binomial')
- Probit regression in R:
  - glm(inlf ~ educ, data = wages, binomial(link = 'probit'))

Georgia Tech

Multinomial Logistic Regression

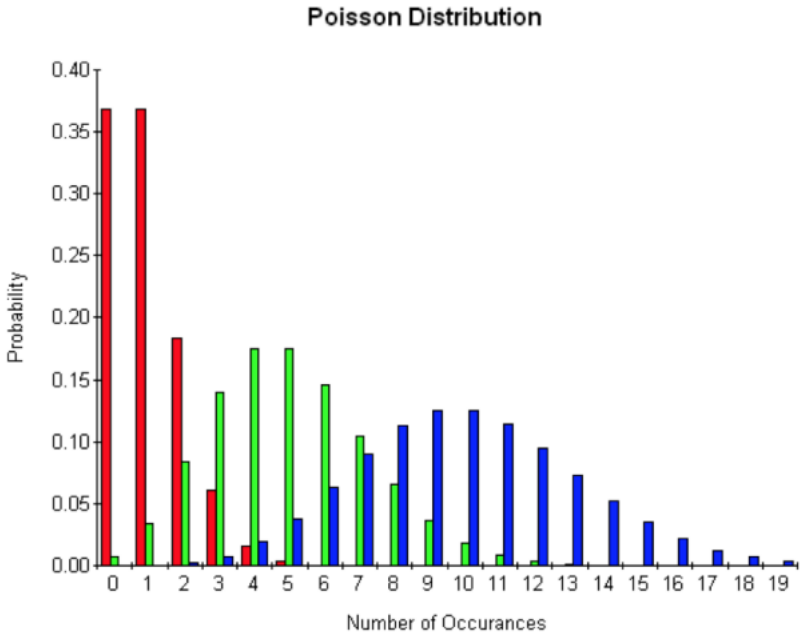
- What if you have outcome variables which could take on many different values, which might not be obviously related to each other?
  - For example, will a given resident of Atlanta commute to work by car, bus, walking, or Marta?
- Basically instead of y being just 0 or 1, it now takes a countable number of different values
  - Create a new variable z_1 which is 1 if y is 1 and 0 otherwise
  - Create a new variable z_2 which is 1 if y is 2 and 0 otherwise, etc
- So basically you run a series of single logistic regressions for whether each data point takes each potential value of y
- This still fits in the same framework! You are just picking the best set of parameters to maximize your likelihood!
- This is a really useful tool a lot of places: often we want to decide what category a particular thing belongs to, e.g. Google wants to decide if you're doing a news search or not

Georgia Tech

And More Extensions, E.g. Poisson for Count Data

- If we were modeling e.g. how many people walk into a store on a given day, this data has a different distribution
  - It's not just 0 or 1
  - It can be any positive whole number
  - We call it "count" data
- Turns out we have specific models for how these processes behave.
  - The intuition is that the number of people is a random variable with a particular class of distributions (called "Poisson")
  - Our X factors shift the mean of the distribution
  - We choose the best parameters for each X so that we have the smallest error in our predictions, just like before

Poisson Distribution



Georgia Tech

Instrumental Variables and Causality

- How can we make a claim about causality?
- Sometimes we might be willing to say that the relationship between X and Y , as channeled through a third variable Z, is actually causal
  - Example: for education and wage, variables such as season of birth and closeness to college have been used.
- We replace the assumption that X is uncorrelated with ε from regression with the assumption that Z is uncorrelated with ε
- In R, this is implemented in the ivreg command of the AER package.

Lesson Summary

- A probit regression uses specific functions for G and H related to the shape of the normal distribution
- Multinomial logistic regressions are used when variables could be one of many output values
- Instrumental variables provides a mechanism to make an alternative causal assumption