Maximum Likelihood Estimation: Categorical and Limited Dependent Variable Models Introduction to the Course

PS2730-2021

Week 1

Professor Steven Finkel

- Course is designed to cover appropriate methods for the analysis of **categorical or limited** dependent variables
- PS2030 Political Research and Analysis focused for most part – as did most of the discipline until the 1980s – on the analysis of *continuous* outcomes, where DV can take on any value from negative to positive infinity
- Estimation within the "least squares" framework
  - Linear additive regression models estimated via OLS if core assumptions met
  - Variations if core assumptions not met
    - WLS/GLS for heteroskedastic or autocorrelated errors
    - Polynomial or exponential terms to deal with non-linearities
    - Interaction terms to deal with non-additivity
    - Instrumental variables/2SLS to deal with endogeneity (omitted variables, reciprocal causality, measurement error)

- These methods--irrespective of all the technical advances that have been made within the least squares framework– are not adequate to handle many of the dependent variables of interest to us as political scientists
- Many DVs can only take on only a *limited* number of values, or are continuous within *limited* bounds, or are *limited* in the conditions under which they are observed
- These variables **fail to satisfy** the distributional assumptions underlying OLS and many other LS estimation procedures

Examples of Categorical and Limited Dependent Variables

- Dichotomous Variables: two categories (e.g., vote/abstain)
- Ordinal Variables: more than two ranked categories without necessarily equal distance between the categories (e.g. "low" "middle" or "upper" class)
- Multinomial Variables: more than two unranked categories (e.g., post-PhD career choice – academics, policy, government, other)
- Count Variables: more than two non-negative integer categories (e.g., political participation, terrorist attacks per year)
- Censored Variables: continuous up to (or down to) a threshold (e.g., demand for an undergraduate lecture class)
- Sample-Selected Variables: continuous but observed only when another variable is at specific values (e.g. survey responses filtered by an initial income question to set a threshold for inclusion; wages observed only if individuals are employed)

- Modeling all of these kinds DVs is done instead via *Maximum Likelihood Estimation* methods and procedures
- MLE models we will consider include:
  - Logit/probit models for dichotomous outcomes
  - Ordinal logit/probit models for ordered outcomes
  - Multinomial and conditional logit models for unordered outcomes
  - Poisson and negative binomial models for counts
  - Tobit models for censored outcomes
  - Longitudinal/multilevel versions of many of these models that take clustering, non-independence of the observations and unobserved heterogeneity into account
  - Maximum likelihood treatment effects, bivariate probit and other models taking *endogeneity* and *sample-selection* biases into account in order to make more rigorous causal inferences

Course Goals

- Expand your statistical toolkit to include all of these models which you can use as appropriate in future research. If dependent variable you are interested in modeling is categorical or limited, generally expected to use the appropriate procedure and not simply plug in an OLS estimate (though occasionally we'll see some arguments in favor of linear specifications for these models)
- Develop understanding of Maximum Likelihood Estimation as a general framework for estimating parameters from sample data. Not "limited" (no pun intended) to categorical limited dependent variable models, but can be also applied to continuous outcomes, multiple systems of equations, missing data models, Bayesian analysis, and others. We'll cover features, advantages and limitations of MLE so that you can apply these methods with confidence more broadly

- Expand ability to absorb and critique articles modeling categorical and limited dependent variables and/or utilizing MLE methods in published articles and books
- Expand statistical computing skills, primarily via STATA applications of these models
  - logit, probit, ologit, mlogit, asmprobit, clogit, poisson, nbreg, tobit, heckreg, ivregress, ivprobit
  - Interpretation, presentation, graphing of effects using "margins" and Long and Freese's SPOST suite of commands
  - glm for "generalized linear models"
  - Mixed effects ("me" suite) for multilevel and longitudinal models
  - "extended regression", "treatment effects" and GSM ("generalized structural equation") models for causal inference with endogenous and sample-selected variables
- Not averse to R but: may not have corresponding commands in all instances, and I can give you some, but not authoritative, guidance

Requirements etc.

- Requirements: homework exercises and seminar paper
- In-class exercises
- Stata 17/Spost 13
- Books and readings

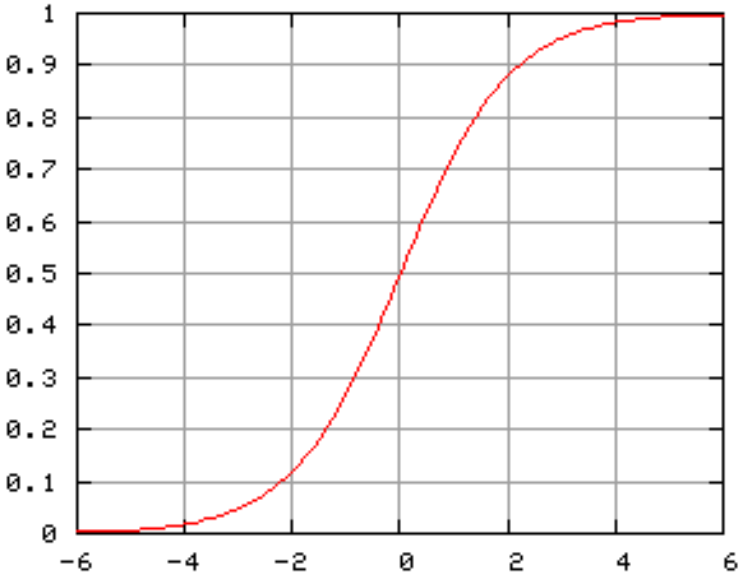
Your Professor

- Academic Positions
  - Daniel Wallace Professor of Political Science, University of Pittsburgh, 2005—present. Department Chair, 2011-2018,
  - Professor of Quantitative and Qualitative Methods, Hertie School of Governance, Berlin, Germany, 2005 – 2008
  - Assistant, Associate, Full Professor, University of Virginia, 1984-2005
  - Guest Professor: Arizona State University (1994); University of Bern (2006); Aarhus University (2013); Gothenburg University (2015); University of Copenhagen (2017)
  - PhD Stony Brook University 1984
- Research Interests
  - Political Behavior, Participation, Democratization
  - Evaluation of Civic Education, Democracy Promotion and Countering Violent Extremism Programs of USAID and International Donors
  - Statistical Methods for Longitudinal and Panel Data
- Home Page: www.pitt.edu/~finkel
- Google Scholar: https://scholar.google.com/citations?hl=en&user=nJTbi8oAAAAJ

Outline of Course Topics

Unit 1: Logit/Probit Models for Dichotomous Outcomes (Weeks 1-4)

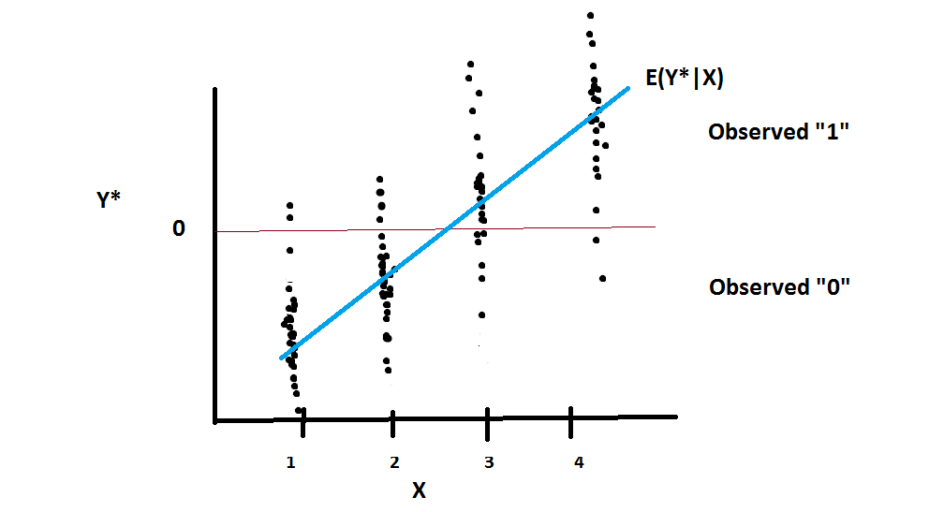
- Backbone of many of the models in the class
- Two derivations
  - Non-linear probability model: P(Y=1), the probability that Y equals 1, as a non-linear function of the independent variables



$$P(Y=1|X)=\frac{\exp(X\mathrm{B})}{1+\exp(X\mathrm{B})}$$

Non-linear in the probabilities, linear in the "log-odds" or logits

- Alternative Derivation: Latent Variable Threshold Model
- Underlying variable (Y*) is *unobserved and continuous*; e.g., individuals have an underlying unobserved *propensity to vote* which is unbounded; or they derived unbounded unobserved *utility* from voting, either positive or negative
- We observe Y as 1 if Y* crosses a threshold (arbitrarily set to 0), and we observe Y as 0 if Y* fails to cross the threshold



Assuming a normal distribution for the error terms of Y*, we arrive at the *probit* model for observed Y:

$$P(Y=1|X)=\Phi(X\mathrm{B})$$

which is also non-linear in the P(Y=1)

- Foregoing suggests a more general formulation: start with a non-linear probability for some dependent variable, and convert it to a linear model through some "link function", or vice versa from a linear formulation via the "inverse link" function to the non-linear probability specification

Probability→ linear specification

$$\ln \frac{P(Y=1|X)}{1-P(Y=1)|x} = XB \qquad \text{Logit}$$

$$\Phi^{-1}\big(P(Y=1|X)\big) = XB \qquad \text{Probit}$$

Linear specification→ probability

$$P(Y=1\,|\,X) = \Phi(X\mathrm{B}) \qquad \text{Probit}$$

$$P(Y=1\,|\,X) = \frac{\exp(X\mathrm{B})}{1+\exp(X\mathrm{B})} \qquad \text{Logit}$$

- All subsumed under the "Generalized Linear Model" (GLM) framework which we will apply to variety of DVs in the course

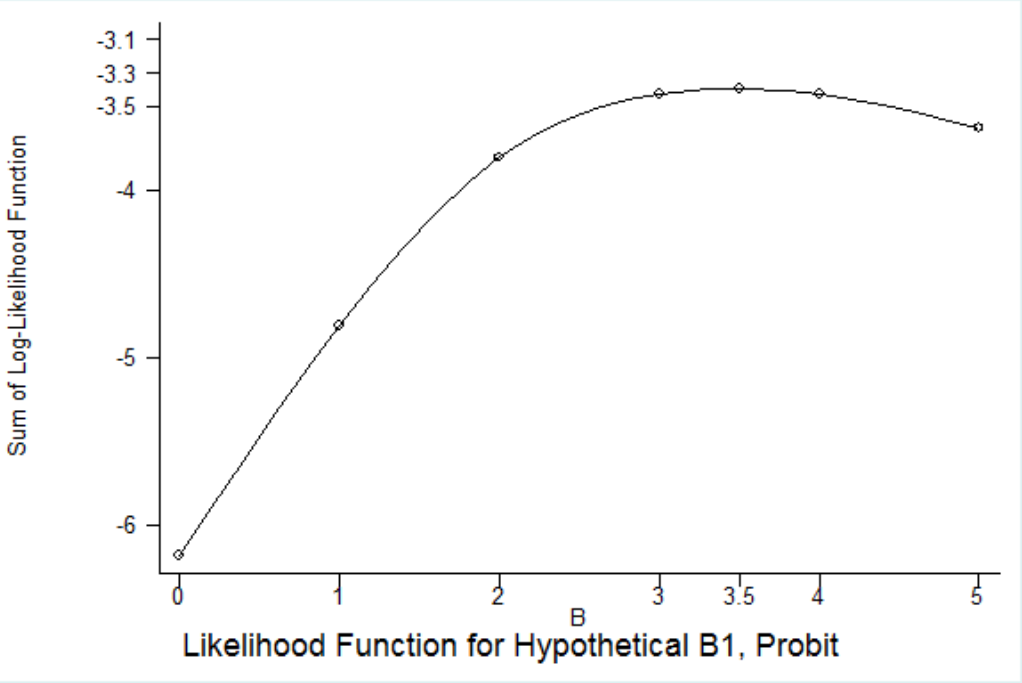
- Next: how to estimate these models? OLS inappropriate (non-normal distribution of Y; unobserved P(Y=1) or Y*)
- *Maximum Likelihood Estimation* instead: **find the model parameters which maximize the overall probability of having observed the sample data, given an assumed probability distribution for each of the Y**
- For logit/probit, this means finding the regression coefficients for the independent variables that maximize the joint probability of having observed the sample of 1s and 0s that we did observe
- In other words, we want to find the β, which generate the *largest* predicted P(Y=1) for the 1s, and the *smallest* predicted P(Y=1) for the 0s, according to the probability specification of

$$\text{logit} \quad P(Y=1|X) = \frac{\exp(X\mathrm{B})}{1+\exp(X\mathrm{B})}$$

or

$$\text{probit} \quad P(Y=1|X) = \Phi(X\mathrm{B})$$

- MLE is a general framework for estimating parameters from sample data, so we'll use logit/probit as a window into MLE procedures, properties, and interpretations
- Steps in any MLE
  - Define a probability model for each data point (like the logit or probit specification)
  - Define the "likelihood" of observing the sample data; this will be equivalent to the joint probability of observing the individual units in the sample, given the probability model and a given population parameter; assuming independent observations, this will be the product of the individual probabilities for each data point
  - Search over all possible parameter values for the ones which maximize the joint (log-) likelihood for the observed data
  - Sometimes the MLE estimates can be derived analytically via differential calculus; other times through iterative optimization procedures



Example: can see how different possible parameter values produce different overall likelihoods of observing the "1s" and "0s" for a given sample of data. The ML estimate is the one where this curve is at a **maximum**.

Shape of the curve (or surface in multivariate analyses) will also be important in calculating standard errors for parameter estimates

We'll cover estimation, significance testing, goodness of fit statistics, model comparisons and model diagnostics (weeks 2-3)

- Unit 1, weeks 3-4: consider some important issues/problems in estimation and interpretation in logit/probit models
- Given non-linearity in P(Y=1), marginal effects of X will differ depending on the unit's location on the underlying cumulative logistic or normal function, i.e., based on the intercept and values of other IVs. How should one best interpret the coefficients in terms of their effect on the dependent variable? And how should one most effectively present those results?

  Several options:

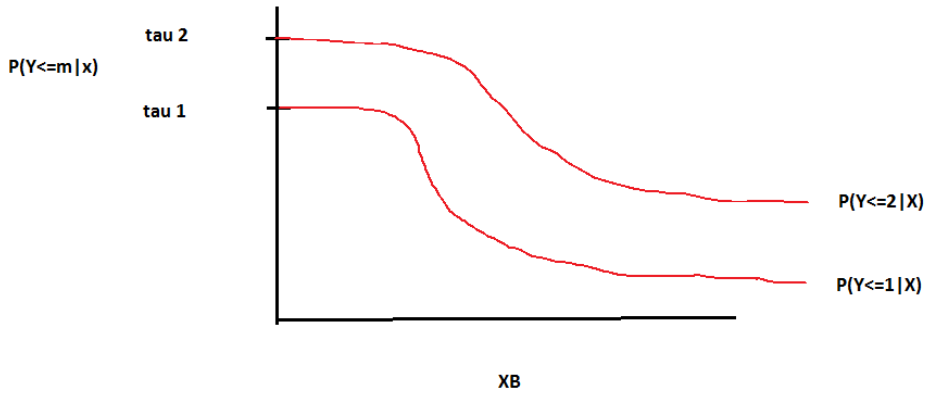
  – In terms of the linear effect of X on the probit z-score, or on the logit log-odds or odds
  – In terms of the non-linear effect of X on the P(Y=1). Here there are issues in terms of selecting appropriate values of the other IVs for assessing the effects of the others. Means? Observed values?
  – In terms of the effect of X on the unobserved latent Y* in probit models

Other issues:

- In latent variable models with Y*, can we distinguish "true" causal effects from heteroskedasticity, where some values of X may have larger error variance and hence push greater proportions of individual units over the threshold for "non-causal" reasons? [There are other error term issues in logit/probit models we'll discuss as well]
- Given non-linearities in the probabilities, are interaction terms interpretable since there is already (self) interaction in the model?
- How to handle the problem of "separation", whereby a given category of X may predict Y perfectly, e.g., **all** PhDs in a given sample are voters, which leads to breakdown in ML estimation and loss of data. We'll discuss some possible solutions, notably "penalized ML" or "Firth regression"

Unit 2: Extensions to Other Categorical and Limited Dependent Variables (Weeks 4-8)

- All involve extensions of either (or both) the non-linear specification or the latent variable framework for modeling dichotomous dependent variables via logit and probit
- As with logit/probit, for all these models we'll discuss: MLE estimation, interpretation of effects, goodness of model fit, model comparisons and the relative "importance" of variables
- Start with ordered outcomes, e.g., individuals trust government "not at all", "some", or "a lot"
- Ordered logit: models the probability of *being at or below a given category*
- Ordered probit: extends the Y* threshold model to include multiple thresholds

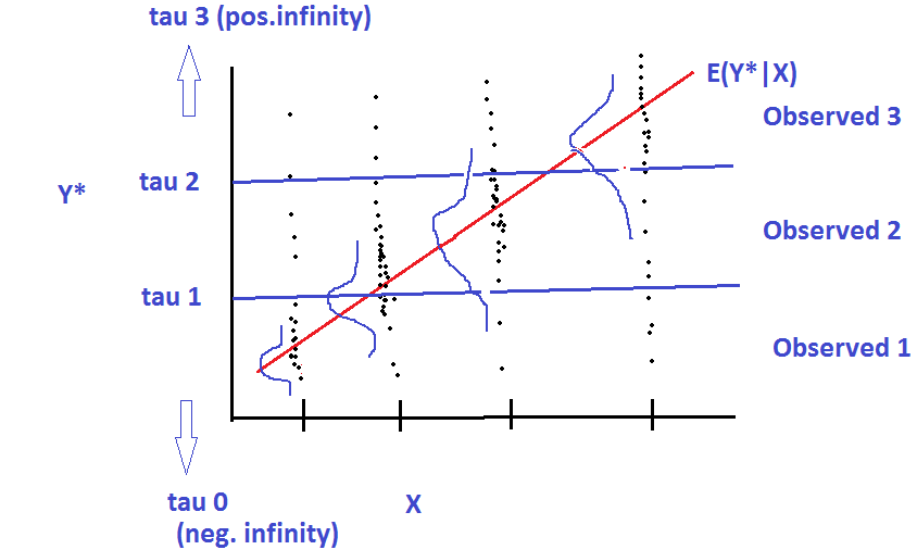


Ordered Logit

$$P(Y \le m \mid X) = \frac{\exp^{\tau_m - XB}}{1 + \exp^{\tau_m - XB}}$$

Ordered Probit

$$P(Y=1|X)=\Phi(\tau_1 - XB) - \Phi(\tau_0 - XB)$$



Multinomial outcomes

- voting behavior in multiparty democracies, or US when third parties run strongly – here DV has three or more categories, unordered with no intrinsic ranking
- Multinomial Logit: set one outcome to be 'baseline" and model the probability of each other outcome relative to the baseline
- Will arrive (for three categories) at two different sets of regression coefficients for each IV, which can then be manipulated to arrive at the predicted probability of each outcome, given X

$$P(y = m) = \frac{e^{XB_{m1|b}}}{e^{XB_{m1|b}} + e^{XB_{m2|b}} + 1}$$

- Extensions to multinomial regression
  - MNL predicts one outcome relative to another based on variables that unit-specific, i.e., age, education, Party ID etc.
  - But some (most?) choice situations are affected by "alternative" or choice-specific as well. For example, you might choose a post-PhD career based on your age at graduation (unit-specific) *and* the expected lifetime earnings you'll receive for each career in your choice set (choice-specific). Traditional MNL cannot handle choice-specific variables, but we can extend this in what is called **"conditional logit"** analysis
  - This involves reshaping the data into "long form" (one row per alternative for each respondent), coding "1" on a choice variable in the row corresponding to the alternative that was chosen, and 0 otherwise. (Similar to "long form" panel data.)

- Traditional MNL also proceeds under the fairly restrictive assumption of the "Independence of Irrelevant Alternatives" (IIA). This means that the odds of choosing alternative A over alternative B *do not change* if we expand the choice set to include alternative C. So if I am twice as likely to choose A over B, I should still be twice as likely to choose A over B if I also have C as an option (I might choose C sometimes but not in a way that would affect my A:B ratio).
- If IIA is violated, then MNL produces biased results
- Political Science: Are odds of choosing Biden over Trump the same in a two-candidate race as they would be in a three-candidate race with Trump, Biden, and Bernie (a relatively far-left Democrat?). Almost certainly not, so traditional MNL would not work

- When IIA is violated, it is usually the case that some alternatives are "closer" to each other than to other options, and, depending on different assumptions, lead to different models of choice:
  - **"nested' or "sequential" logits**, where individuals first choose A or (B or C), and then choose between the two remaining alternatives (e.g., Biden v. Bernie first, and then Trump v. the winner next)
  - when no natural order or sequence of choices but there are still choices that cluster together, can model via **alternative-specific multinomial probit**, which allows correlations between the error terms of related choices

Count Models

- Many variables of interest to political scientists are expressed as "counts" of events, that is, number of times something has happened in a given period of time.
  - Ex: Number of vetoes in a presidential term, number of conflicts a country might get involved in, number of acts of political participation by individual in a given year.
- Distinguishing features: 1) Cannot be less than zero, and 2) must take on integer values. So DV in these cases goes from 0, 1, 2, 3, …..up to positive infinity. Not continuous --- limited in the values that it can take on, so OLS or variants no longer appropriate.
- Appropriate method: Poisson Regression or some variants we will discuss, all of which force expected values to be non-negative

- Poisson model

$$P(y \mid u) = \frac{e^{(-\mu)} u^{y}}{y!}$$

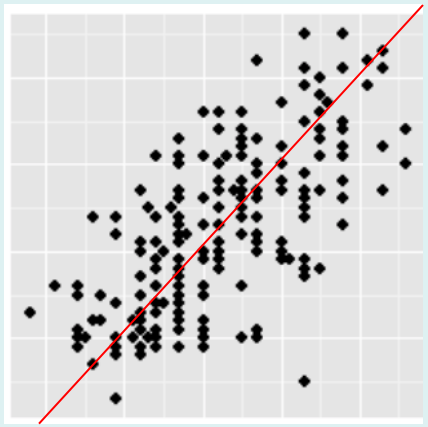
- Probability of observing a count of Y, given a "rate parameter" μ, which we model as an exponentiated linear function of the X independent variables

$$\mu = e^{XB}$$

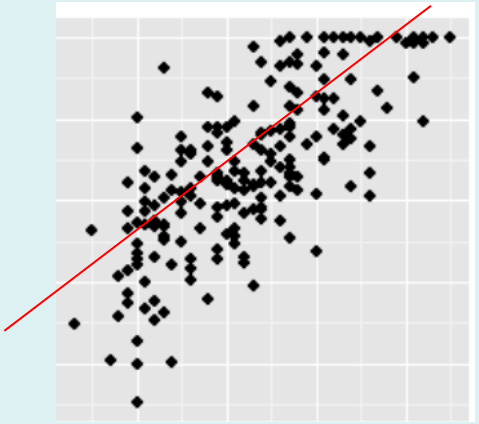
- Other variants for count models we'll consider:
  - **Negative binomial model**: incorporates possibility of individual heterogeneity and contagion of outcomes
  - **Zero Inflated Poisson/Negative binomial models**: splits sample and the analysis into two groups – those units that are (and always will be) 0, and another whose counts can be positive

Limited Dependent Variables

- Sometimes we may have limited information on DV. It could theoretically take on infinite range of values but our measurement scheme only goes up to a certain point or down to a certain point.
  - For example, thermometer scores in attitude research go from 0 degrees to 100 degrees, presumably very warm toward particular stimuli. What about person who is at 200 degrees? 120? All pushed to 100, the upper bound. Call that *right-censored* variable. Here a person who feels –50, -100, is also pushed to 0, so thermometer scores are *left-censore*d as well.
- Bounding the data in this way leads to bias in OLS estimation.
- Gives rise to the Tobit model, after economist James Tobit who developed it in 1950s. Idea of model is that it provides estimates of probability of NOT being censored, and then, expected values of Y, conditional on not being censored



True relationship between restrictive immigration attitudes and support for Trump



OLS relationship between immigration attitudes and support for Trump with thermometer score censored at 100

Unit 3: Longitudinal and Multilevel Models (Weeks 8-11)

- We'll extend the models considered so far to encompass multilevel and longitudinal data
- Both kinds of data structures result in observations that are non-independent, hence simple pooling approaches will yield inefficient and potentially biased results
  - In multilevel structures, units are nested within higher-level units, e.g., students within classrooms, individuals within neighborhoods, legislators within parties
  - In longitudinal structures, time of observation is nested within units
- Explanatory variables at each level of the data structure may be important

- Some of the clustering in longitudinal/multilevel data structures problem may be ***unobserved unit-level heterogeneity***, such that each unit has an individual "unit effect" due to stable unmeasured variables that lead the unit to be higher or lower than the overall population average, regardless of the values of the other independent variables
- This is typically handled by adding a unit-level, time-invariant (or, in multilevel models, a cluster-level) heterogeneity term (ζ) into the model, i.e., a subject (cluster)-specific intercept.
- **Random Effects Logit**: adds an additional *random* heterogeneity term to the logit model, which accounts for all stable unobserved factors which make the individual more or less likely, for example, to vote at each point in time, regardless of the values on any other factor

$$P(Y=1|X)=\frac{\exp(X\mathrm{B}+\zeta_i)}{1+\exp(X\mathrm{B}+\zeta_i)}$$

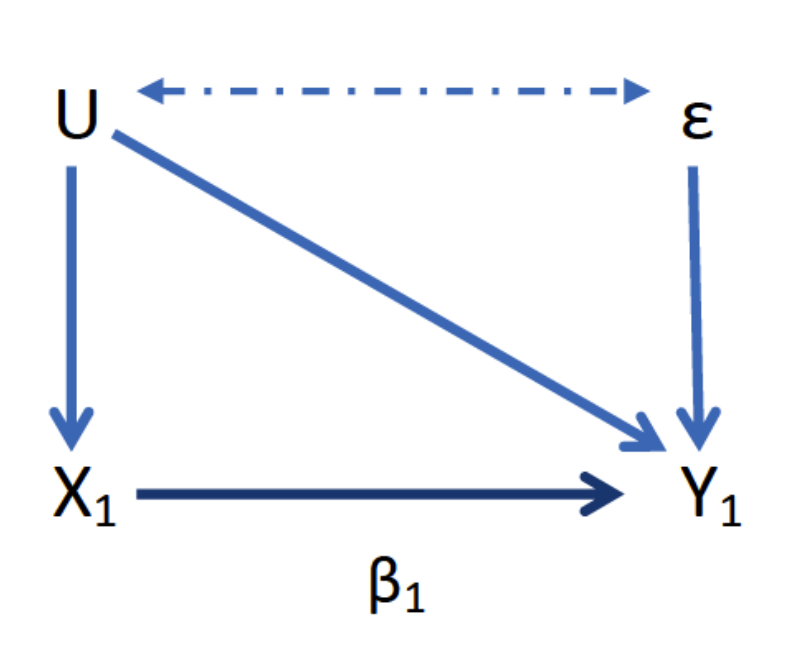
- There is a corresponding **Random Effects Probit** model as well

- Big problem: if the unit effect (ζ) is **correlated** with the observed variables in the model, then the estimates of the β in the pooled model will be biased. This would be an example of an omitted variable or *endogeneity problem* that is common in observational research, and for which panel data represents one possible solution
- **Fixed Effects Logit** allows the X to correlate with ζ, thus providing stronger causal inferences by controlling for stable unobservable variables that could confound the effect of the Xs on outcomes
- We'll consider these kinds of "random" and "fixed" effects models in longitudinal/multilevel structures for ordered, multinomial, and count outcomes.
- All of these models are examples of **Generalized Multilevel Mixed Models"** – a combination of GLM with random or fixed effects for higher level clustering in the data. In Stata, they are subsumed within the "ME" suite (melogit, meologit, mepoisson, metobit, meglm for "mixed effects glm")

Unit 4: Endogeneity and Sample-Selection Biases

- Final unit will expand our coverage of models that deal more generally with the problem of *endogeneity* biases, in the context of both cross-sectional and longitudinal/multilevel models
- Endogeneity occurs whenever independent variables are related to the error term (disturbance) of the outcome they are assumed to predict, i.e., $E(Xu) \neq 0$
- Occurs if:
  - Omitted variables are related to both X and Y
  - There is reciprocal causality between X and Y
  - There is random measurement error in X
- In any of these instances we say that X is an *endogenous regressor*, and OLS estimates of the effect of X on Y will be *biased* as a result
- This is *the* fundamental problem in causal inference to overcome

- This is the identical problem of estimating the causal effects of **treatments**, where the same factors that affect whether a unit will receive a given treatment may be related to the outcome of interest (e.g., via self-selection or omitted variables)



- Solutions usually involve bringing specific kinds of outside *observed* variables into the process so that they can serve as proxies or *instruments* for the endogenous regressors or treatments.
- We'll consider many kinds of instrumental variable models: for example, for continuous outcomes and dichotomous treatments (ivregress in Stata), for dichotomous outcomes and continuous treatments (ivprobit), dichotomous outcomes and dichotomous treatments (biprobit), and extensions to other kinds of non-continuous variables
- Can extend these models to the longitudinal/multilevel case, where the data structures allows more options for instrumental variable estimation (another advantage of panel/multilevel data for causal inference)
- Alternative endogeneity models can also be estimated in the SEM (structural equation modeling) tradition via ML methods. We *may* have time to consider them.

- Finally, a closely related problem arises if there is *sample selection,* such that we only observe the dependent variable for a unit IF the unit passes a threshold or must have a specific value on another variable(s).
- Example: survey data on political participation, but participation (and all other variables) are observed only if the individual agrees to participate in the survey.
- If agreement to participate in the survey is random and unrelated to political participation, no problems
- But almost certainly the unobserved factors that lead individuals to participate in a survey will be positively correlated with political participation. So there will be correlation between the errors of a *selection equation* and an *outcome equation*, which will bias OLS estimation. This is another kind of endogeneity in the process.
- We'll consider the **Heckman regression model** for sample-selected outcomes, with links to more general **endogenous regressor, treatment effect (and possibly SEM) models**