Unit 4
V. Models for Non-Continuous Dependent Variables

PS2701-2019

Longitudinal Analysis

Weeks 12-13

Professor Steven Finkel

Non-Continuous DVs

- Numerous longitudinal models exist for dependent variables that are non-continuous, such as:
  1) Dichotomous variables (vote/no vote, war/no war)
  2) Ordered variables (attitudinal scales with 1-5 responses, subjective social class)
  3) Multinomial variables with unordered outcomes (vote in multiparty elections, regime type)
  4) Count Data (number of acts of political participation, number of treaty violations)
- Same goals as before in class: estimate causal effects from IVs to DV, controlling for clustered (multilevel) responses, and taking into account unit-level heterogeneity, endogeneity concerns and other possible sources of temporal dependence in longitudinal data
- Methods will differ due to nature of the DV, which will produce differences from linear estimation procedures

Longitudinal Models for Non-Continuous DVs

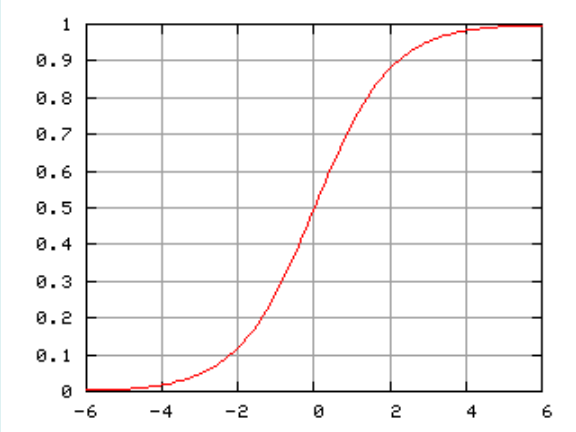
- **Generalized Estimating Equations (GEE)**, a general procedure for taking into account clustered or correlated longitudinal data that produces "population averaged" estimates, as opposed to the "subject-specific" estimates from random or fixed effects models that include U_i or ζ_{0i} in the specification; and
- **Random Effects, or "Generalized Multilevel Mixed Models"**, i.e., adding random unit-level effect(s) to a "linearized" version of a particular model, e.g. the logit model predicting the log-odds of an outcome occurring
- **Fixed Effects Models**, an alternative model for taking into account unit-level heterogeneity in logit and other non-continuous DVs, one that allows for correlation between the U_i or ζ_{0i} and the included X variables
- **Transition Models**, where the value of Y_{it} is modeled as a function of its value at Y_{it-1} or other Y_{it-s} time periods
- We'll focus on RE and FE models due to time constraints

Modeling Dichotomous Variables

- The logit model:

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

- Model that has a **non-linear functional form** of the effects of X on the P(Y=1), where the P(Y=1) are bounded by 0 and 1 and the (unbounded) Xs have more impact at middle levels of the distribution than at the tails
- As XB goes to ∞, P(Y=1) goes to 1 but never gets there; as XB goes to -∞, P(Y=1) goes to 0 but never gets there; when XB is 0, P(Y=1)=.5.



- The probability of Y being "0" or (1-P(Y=1)) = $\frac{1}{1+\exp(X\mathrm{B})}$
- We construct the quantity P(Y=1)/P(Y=0) --- what we call the "odds" of Y being 1 – as:

$$\frac{P(Y=1)}{P(Y=0)} = \frac{\frac{\exp(X\mathrm{B})}{1+\exp(X\mathrm{B})}}{\frac{1}{1+\exp(X\mathrm{B})}} = \exp(X\mathrm{B})$$

- This means that, if we exponentiate the estimated β, we get the effect of the Xs on the "odds-ratio" (OR), or the change in the odds of changing X by 1 unit.
- And taking the natural logarithm of both sides (to the base "e") gives

$$(2) \qquad \ln\frac{P(Y=1)}{P(Y=0)} = (X\mathrm{B}) = \beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \ldots \beta_k \mathrm{X}_k$$

- We call the log of the odds that P(Y=1) the "logit" of Y, and so we can say that the logit model is *linear in the logits*, an increase of a unit in X produces a constant change in the *logits*, but is non-linear in the probabilities (and odds). The effects of X on P(Y=1) differ depending on where the unit is otherwise on the logistic function

- We estimate the logit model through ML methods, finding the β parameters that yield the highest joint likelihood that P(Y=1) for all the Y=1 observations and P(Y=0) for all the Y=0 observations. Thus ML finds the parameters that maximize the likelihood of observing the outcomes for the given sample that were, in fact, observed.
- We maximize the Likelihood Function with respect to the β as:

(3) $$L = \prod P(Y=1)^{Y_i} P(Y=0)^{1-Y_i}$$

$$L = \prod \left(\frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\beta_k X_k}}\right)^{Y_i} \left(\frac{1}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\beta_k X_k}}\right)^{1-Y_i}$$

- Taking the logs of this expression and then the derivative with respect to β yields a solution for β:

$$\sum_{i=1}^{N} X_i \left(Y_i - \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\beta_k X_k}}\right) = 0$$

- This equation does not have an algebraic solution, so the estimates are produced through an iterative procedure that has been implemented in Stata and all other standard software packages.

Longitudinal Extensions

- The situation becomes more complex with longitudinal observations on Y_{it}. As we have discussed in the course so far, in these situations the Y_{it} are not independent observations, but rather are **clustered by unit**. So we cannot simply run "pooled logit" across all the cases and all time periods and arrive at correct estimates of the β and their associated standard errors.
- Moreover, the source of the clustering problem may be unobserved unit-level heterogeneity, such that each unit has some unit effect due to stable unmeasured variables that makes the unit higher or lower than the overall population average, regardless of the values of the other independent variables in the model.
- It may even be the case that this unit effect is correlated with the observed variables in the model, such that the estimates of the β in the pooled model would be inconsistent as well as inefficient. Finally, it may be that there is other temporal dependence among the Y_i such that prior values of Y_i *determine* in some way the current value, either in addition to, or instead of, the unit effect or other sources of unit-level over-time clustering.

- Empirical example: We randomly assign 140 high school students to a special semester-long civic education program after school hours, with the remaining 77 forming the control group. We can ask: to what extent does the civic education treatment impact the students' initial likelihood of voting, and to what extent do the treatment and other independent variables impact the rate of growth in their voting likelihoods over subsequent elections?
- We can see the following marginal distribution for the treatment and control group over time:

Election	**0**	**1**	**2**	**3**	**4**
Control	**.48**	**.42**	**.51**	**.52**	**.58**
Treatment	**.80**	**.78**	**.83**	**.86**	**.86**
OVERALL	**.69**	**.65**	**.71**	**.74**	**.76**

- Treatment group is more likely to vote at every election, but the control group increases over time at a faster rate.
- We want to model each individual's election-specific probability of voting from treatment/control and time-(in)varying covariates

- Need to take into account individual-level heterogeneity, or unobservables that cause outcomes for individual *i* to be similar over *t*
- This is done through the a unit-level, time-invariant heterogeneity term (U_i or ζ_{0i}) into the model, i.e., a subject-specific intercept.
- Controlling for civic education exposure and all other covariates, some individuals are generally predisposed at all time periods to have greater P(Y=1) than others, depending on the size of the ζ_{0i} term.
- Either assume that ζ_{0i} is a random variable with a known distribution (e.g., normal) – leading to **Random Effects Logit or Probit**, or we can make no assumption about the distribution of ζ_{0i} and treat them as "fixed effects", leading to **Fixed Effects Logit** (there is no Fixed Effect Probit).
- There are analogues for most of these procedures that can be applied to ordinal, multinomial, and count outcomes as well

Random Effects Logit

- One way to incorporate heterogeneity is by adding the ζ_i to the expression for the probability that Y=1 in the dichotomous logistic regression model

$$(4) \qquad P(Y=1|X)=\frac{\exp(X\mathrm{B}+\zeta_i)}{1+\exp(X\mathrm{B}+\zeta_i)}$$

- where ζ_i is the unit effect that pushes the probability (intercept) up or down for a given case at all time periods, regardless of the levels of the X covariates.
- If we assume that ζ_i is a normally distributed random variable (with variance of, say, ψ^2 , we arrive at the ***Random Effects Logit*** model
- Of course, we have all the other important assumptions of RE as well, namely that E(Xζ)=0, i.e. that there is no covariation between the independent variables and the unit-specific effect.

- We can also express this model in its "linearized" form, as:

$$(5) \quad \ln\left(\frac{P(Y=1|X)}{P(Y=0|X)}\right) = (X\mathrm{B}+\zeta_i) = \beta_0+\beta_1\mathrm{X}_1+\beta_2\mathrm{X}_2+\ldots\beta_\mathrm{k}\mathrm{X}_\mathrm{k}+\zeta_i$$

- We could call this the "logistic-normal" model, because the response variable Y is assumed to result from the logit link from this linear function, and the linear function contains a normally-distributed error term (ζ). If we apply the probit link to the response, given the normally-distributed error term (ζ), we end up with the ***Random Effects Probit*** Model, or the "normal-normal" model as the probit link is based on the cumulative normal distribution of z-scores:

$$(6) \quad P(Y=1|X) = \Phi(X\mathrm{B}+\zeta_i)$$

$$\Phi^{-1}(X\mathrm{B}+\zeta_i) = \beta_0+\beta_1\mathrm{X}_1+\beta_2\mathrm{X}_2+\ldots\beta_\mathrm{k}\mathrm{X}_\mathrm{k}+\zeta_i$$

- The "logistic-normal" model is implemented in STATA as XTLOGIT and the "normal-normal" probit model as XTPROBIT

- Finally, we can arrive at the RE Logit model through multi-level modeling as well:
- Level 1:

$$(7)\quad \ln\left(\frac{P(Y=1|X)}{P(Y=0|X)}\right)=\beta_{0i}+\beta_1 X_{1i}+\beta_2 X_{2i}+\ldots\beta_k X_{ki}$$

- Level 2:

$$(8)\quad \beta_{0i}=\beta_{00}+\zeta_{0i}$$

and

- Mixed:

$$(9)\quad \ln\left(\frac{P(Y=1|X)}{P(Y=0|X)}\right)=\beta_{00}+\beta_1 X_1+\beta_2 X_2+\ldots\beta_k X_k+\zeta_{0i}$$

- which is identical to (5) except for the added notation in the intercept and random effect terms. We see that the β are the **fixed effects** in the model, and the ζ the **random effect**.
- This is one kind of "**multilevel generalized mixed effects model**", which refers to a "generalized" framework that allows effects on IVs on non-continuous DVs to be estimated, so long as the model can be "linearized" via what is called a "link" function to a linear specification.

- For example, the mixed logistic regression model uses the "logit link" function to move from the "linear" logit specification to the non-linear probability of observing (Y=1) for a dichotomous variable (and vice versa). The "probit link" moves to P(Y=1) using the cumulative normal distribution; the "poisson" link moves to predicting a non-negative integer count variable from an exponentiated linear function, etc.
- In Stata, these models can be estimated with the "ME" family of commands (for "Mixed Effects"). So **MELOGIT** and **MEPROBIT** for logit and probit, **MEOLOGIT** and **MEOPROBIT** for ordered logit/probit, **MEPOISSON** for Mixed Effects count models.
- **MEGLM** is the most general command, standing for "Mixed Effects Generalized Linear Model"
- With "ME" family, can also take advantage of all other aspects of the multilevel modeling framework

- For example, possible in MELOGIT/MEPROBIT to build more complex multilevel models that have random coefficient terms for the X covariates at Level 2. For example, the slope for X_{1i}, β_1, may vary across individuals, be dependent on other Level 2 factors and/or an additional random effect ζ_{1i}, for example. Everything you can do with linear mixed models, you can more or less do with multilevel generalized linear mixed models. The caveat is that the estimation procedures are not as exact and they become **very slow** as you add more and more random effects, compared to the "simple" linear case.

Estimation of the RE Logit Model

- Maximum likelihood method: we seek to maximize the likelihood of observing the pattern of 1s and 0s that we observed in the data, given values of the covariates, the β fixed effects and the variance of the ζ term ψ^2.
- This gives us: $$L = \prod P(Y=1)^{Y_i} P(Y=0)^{1-Y_i}$$
$$L = \prod \left(\frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\beta_k X_k+\zeta_i}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\beta_k X_k+\zeta_i}}\right)^{Y_i}\left(\frac{1}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\beta_k X_k+\zeta_i}}\right)^{1-Y_i}$$
- Solution via integral calculus, by integrating out this expression with respect to the ζ_i, so that we arrive at a marginal likelihood for all of the clusters (units) that depends only on the β and the value of ψ^2 and not on the individual ζ_i. We then search for the values of the β and ψ^2 that maximize the joint marginal likelihood.
- No closed form solution, so estimation is done via such technically complex procedures such as "Gauss-Hermite quadrature," "adaptive quadrature," or other alternatives. *Very* slow when you have more than one random effect.
- You can test different "intpoints" (quadrature points) in XTMELOGIT procedure as robustness check -- the default in Stata is 7

Interpretation of RE Logit Coefficients

- The RE model:

$$P(Y=1|X)=\frac{\exp(X\mathrm{B}+\zeta_i)}{1+\exp(X\mathrm{B}+\zeta_i)} \qquad \ln\left(\frac{P(Y=1|X)}{P(Y=0|X)}\right)=(X\mathrm{B}+\zeta_i)=\beta_0+\beta_1\mathrm{X}_1+\zeta_i$$

- Take an individual with a given ζ_i, that is, a given unit effect that predisposes him/her to respond on 1 at all points in time. For such an individual, we may express the log-odds that Y=1 when X=0 as:

$$\ln\left(\frac{P(Y=1\|\zeta)}{P(Y=0\|\zeta)}\right)=\beta_0+\zeta_i$$

and when X=1 as:

$$\ln\left(\frac{P(Y=1\|\zeta)}{P(Y=0\|\zeta)}\right)=\beta_0+\beta_1+\zeta_i$$

So a unit change in X produces a $(\beta_0+\beta_1+\zeta_i)-(\beta_0+\zeta_i)$ change in the log-odds that Y=1, or simply β_1 .

- But note: these effects are calculated at a *fixed* level of the random effect ζ, so it will NOT result in the same effect on P(Y=1) for all individuals!!!! This means that RE Logit provides what are referred to as ***subject-specific*** effects
- Subject-specific effects, in probability terms, are dependent on the subject's latent predisposition to be on 1, or the level of the random effect ζ. When ζ is very large or very small, a unit-change in X will not produce very large changes in the probability, while values of ζ that put the individuals prior probability nearer to .5 will see greater effects on the probability from the same unit change in X.
- These differences *only* exist because of the non-linearities of the logit/probit model; in a linear model the effect of X on Y produced the same change on Y no matter what the value of ζ happens to be

- Can see this on the table with hypothetical data below with individuals with different ζ. Assume the logit model is:

$$\ln\left(\frac{P(Y=1\mid X)}{P(Y=0\mid X)}\right)=(X\mathrm{B}+\zeta_i)=.5\mathrm{X}+\zeta_i$$

Individual	ζ	P(Y=1)|X=0	P(Y=1)|X=1	Change in P(Y=1)	Log-Odds Difference
A	1.50	0.82	0.88	0.06	**.5**
B	1.00	0.73	0.82	0.09	**.5**
C	0.00	0.50	0.62	0.12	**.5**
D	-1.00	0.27	0.38	0.11	**.5**
E	-1.50	0.18	0.27	0.09	**.5**
Population Average	0	.50	.59	.09	**.38**

- Individual A has a high probability of .82 of Y=1 when X=0, changes to .88 when X=1. B goes from .73 to .82, and so on until individual E, who has a very negative ζ, starts at .18 and increases to .27.
- But if you calculate the difference in the log-odds of these probabilities *for each subject*, it turns out to be the same: **.5.** So the above equation is the *subject-specific* model for the log-odds that Y=1

$$\ln\left(\frac{P(Y=1 \,||\, X)}{P(Y=0 \,||\, X)}\right) = .5X_i + \zeta_i$$

Note that, depending on the individual's ζ, adding the β_1 of .5 to the logit produces anywhere from a .06 to .09 to .12 change in the probability that Y=1. So we can say that, for **all** levels of ζ, a change in X produces a change of β_1 to the logit, but this will produce a different change in the probability that Y=1 for every value of ζ.

Note that these subject-specific changes in probabilities will differ from one another more, to the extent that there is more variance in the ζ. If all the ζ are the same or closely bunched together, there will be no difference in the changes in P(Y=1) from case to case.

- Now, look at what the table says about the population marginals. When X=0, the average P(Y=1) is .5. And when X=1, the average P(Y=1) is .59. This translates into a logit difference of .38, which means that, **had we estimated a pooled logit with these data, we would have obtained a β estimate of .38:**

$$\ln\left(\frac{P(Y=1\,||\,X)}{P(Y=0\,||\,X)}\right)=.38X_i$$

- exp(.38)/(1+exp(.38))=.59, versus exp(0)/(1+exp(0))=.5, so difference is the observed average difference in P(Y=1) of .09.
- So, for a given distribution of Y, population-averaged estimates will be **attenuated** from RE logit estimates, and this attenuation will be greater to the extent that there is more variance in the random effects (i.e. greater ψ^2).
- Side note: "GEE" models are another kind of "population average" or marginal models that take clustering into account in more complex ways (similar to robust standard error approaches in regular regression)

Fixed Effects Logit

- But what about the assumption in RE Logit that the unit-specific term is unrelated to the Xs? Isn't that the big problem with RE models generally speaking, and what do we do about it in the case of non-continuous dependent variables?
- Answer: **Yes, it is still a big problem!**
- Solution(s): **1) Fixed Effects Logit**; and **2) Hybrid RE or multilevel model with cluster level means (or a latent mean as in ML-SEM) added to Level 2 equation**
- Example: People who are contacted by parties during campaigns are also those that have the highest latent inclination to vote based on unobserved factors (ζ_i) such as their normative integration into society. So it is not that party contact matters for voter turnout, it is rather the factors that led them to be contacted are the same (unobserved) ones that lead them to vote

- In the linear case, we were able to specify the heterogeneity model with ζ_i (or U_i) and then "sweep it out" of the equation through "de-meaning" or first differencing, or simply by adding a dummy variable to signify each case in the LSDV procedure. This allowed us to estimate the β effects of the X variables without the need to assume that the ζ_i and X were unrelated, and without the need to assume anything at all about the distribution of the ζ_i in terms of normality or other forms. So the FE model, based on "within-unit" variation, has some attractive features.
- But, unfortunately, "sweeping out" the heterogeneity is not that easy in the non-continuous DV case!
- Why not?

- If we add a dummy variable for each case, we run up against the "incidental parameters" problem that we mentioned briefly in the linear case, as this comes into play here with the maximum likelihood estimation procedures for non-continuous DVs. As the number of dummy (or nuisance) variables increases, there is inconsistency in the estimation of the β
- Technically, ML assumptions are violated, since the number of parameters to be estimated increased directly with sample size
- If you take the logit model, there is no way to subtract out the ζ, either from this version or from the logit/log-odds version because of the intrinsic non-linearities of the model.

- So what to do?

- A **Fixed Effects Logit** model was developed by the economist Gary Chamberlain in the early 1980s, based on what is known as "Conditional Maximum Likelihood" methods (which is why the procedure is sometimes called "conditional logit" –there is a cross-sectional analog which will be covered in MLE)
- Recall equation (3) above which expressed the idea of "regular" MLE for the logit model. Find the parameters β that, given the Xs, maximizes the following expression.

$$L = \prod P(Y=1)^{Y_i} P(Y=0)^{1-Y_i}$$

$$L = \prod \left(\frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\beta_k X_k+\zeta_i}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\beta_k X_k+\zeta_i}}\right)^{Y_i} \left(\frac{1}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\beta_k X_k+\zeta_i}}\right)^{1-Y_i}$$

- Chamberlain's method maximizes *not* this overall likelihood (or what could be called the "unconditional likelihood" of observing the sample values Y_i), but rather maximizes the *conditional likelihood of observing the sample values Y_i, given the SUM of the 1s and 0s for a given case*. This is an ingenious method because it turns out that this eliminates the ζ from consideration in the likelihood function altogether, and one can estimate the β free from the potential contaminating effect of the ζ-X correlation. The drawback is that *only* cases that show some change on the dependent variable over time provide any information whatsoever in the estimation procedure, so you lose all the cases that are always 1 or always 0.

- This is a **serious** potential problem, but TANSTAAFL !!!!
- Let's illustrate this method, assuming T=2 for simplicity.

- We want to maximize this quantity:

$$(11) \quad L = \prod_i \prod_t \left(\frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\beta_k X_k+\zeta_i}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\beta_k X_k+\zeta_i}}\right)^{Y_i}\left(\frac{1}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\beta_k X_k+\zeta_i}}\right)^{1-Y_i} | \Sigma(Y_i)$$

- So for two time periods, we have the following possible sequences for Y_i:

  Sequence 1: $Y_1=0$, $Y_2=0$ SUM=0
  Sequence 2: $Y_1=1$, $Y_2=1$ SUM=2
  Sequence 3: $Y_1=0$, $Y_2=1$ SUM=1
  Sequence 4: $Y_1=1$, $Y_2=0$ SUM=1

- Taking the first sequence, we have their contribution to the overall conditional likelihood in (11) as:

$$(12) \quad L = \left(\frac{1}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\beta_k X_k+\zeta_i}}\right)\left(\frac{1}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots\beta_k X_k+\zeta_i}}\right) | 0$$

- What is the probability of getting the sequence 0-0 if the sum of the Ys is 0? That's right, the probability is 1. And this is the case regardless of what value we would estimate for the β, so we know the answer before we start the estimation procedure. These case thus contribute *nothing* to the conditional likelihood. They are lost to FE logit!!
- Another way to looks at it: you can see that, logically, for every case that has all 0s over time, all you would need to do in regular MLE is posit an arbitrarily small ζ_i term (i.e. a very large negative value), and you would generate a predicted P(Y=0) of 1 at all times, no matter what the values on the Xs or the βs. So the cases that are all 0s give you no information whatsoever on the effects of the Xs.
- The same thing happens for Sequence 2, when both Ys are 1 and the sum of the Ys is 2.

- Therefore, in FE logit, we consider **only** those cases which change on the DV over time
- In those cases, the task becomes: given that one (or more, but fewer than T) outcome is "1", find the β which maximize the likelihood of observing the particular "1" outcome(s) of those that were observed, relative to the "0" outcome(s) that were observed
- For example, for the cases where "0" is observed at the time 1 and "1" is observed at time 2, we maximize:

$$(13) \qquad L = \frac{\left(\frac{e^{\beta_0+\beta_1X_1(2)+\beta_2X_2(2)+\ldots\beta_kX_k(2)}}{1+e^{\beta_0+\beta_1X_1(2)+\beta_2X_2(2)+\ldots\beta_kX_k(2)}}\right)}{\left(\frac{e^{\beta_0+\beta_1X_1(1)+\beta_2X_2(1)+\ldots\beta_kX_k(1)}}{1+e^{\beta_0+\beta_1X_1(1)+\beta_2X_2(1)+\ldots\beta_kX_k(1)}}\right)+\left(\frac{e^{\beta_0+\beta_1X_1(2)+\beta_2X_2(2)+\ldots\beta_kX_k(2)}}{1+e^{\beta_0+\beta_1X_1(2)+\beta_2X_2(2)+\ldots\beta_kX_k(2)}}\right)}$$

- All of this is done *without any consideration of the* ζ– and hence provides estimates of the β regardless of whatever correlation may or may not exist between ζ and X.

- And for cases where "1" is observed at t1 and "0" at t2, we maximize:

$$(14) \quad L = \frac{\left(\frac{e^{\beta_0+\beta_1X_1(1)+\beta_2X_2(1)+\ldots\beta_kX_k(1)}}{1+e^{\beta_0+\beta_1X_1(1)+\beta_2X_2(1)+\ldots\beta_kX_k(1)}}\right)}{\left(\frac{e^{\beta_0+\beta_1X_1(1)+\beta_2X_2(1)+\ldots\beta_kX_k(1)}}{1+e^{\beta_0+\beta_1X_1(1)+\beta_2X_2(1)+\ldots\beta_kX_k(1)}}\right)+\left(\frac{e^{\beta_0+\beta_1X_1(2)+\beta_2X_2(2)+\ldots\beta_kX_k(2)}}{1+e^{\beta_0+\beta_1X_1(2)+\beta_2X_2(2)+\ldots\beta_kX_k(2)}}\right)}$$

[NOTE: for those of you who will take MLE, you will see how these equations are the same as Conditional Logit models in predicting whether a person would choose an outcome 1, 2 or 3, given the attributes of the choices on variable X for choice 1, 2, or 3. In that model, we don't lose any cases because everybody chooses 1 and only 1 of the outcomes, so there is always variation in the outcome at the individual level, i.e. no one is 1 on all choices or 0 on all choices– everyone is 1 on only 1 choice and 0 on all the others, and CL finds the β that maximize the probability of choosing the outcome that the individual did in fact choose, relative to the overall probability of choosing all the other outcomes].

Two-Wave FE Logit

- Actually, with two waves it is straightforward to show that the FE logit procedure just outlined is exactly the same as a "difference model", along the lines of our FD linear model
- Among units that change on the DV between time 1 and time 2, it can be shown that:

$$(15) \quad \ln \frac{\mathrm{P}((Y_2 - Y_1) = 1)}{1 - ((\mathrm{P}(Y_2 - Y_1) = 1))} = \beta_0 + \beta_1 (X_{12} - X_{11}) + \beta_2 (X_{22} - X_{21})$$

- which in turn is the same as predicting the probability that Y=1 at time 2 from the differenced independent variables

$$(16) \quad \ln \frac{\mathrm{P}(Y_2 = 1)}{1 - (\mathrm{P}(Y_2 = 1))} = \beta_0 + \beta_1 (X_{12} - X_{11}) + \beta_2 (X_{22} - X_{21})$$

- Since if Y_2-Y_1 =1, then Y_2 must equal 1 among those cases that changed over time, and if Y_2-Y_1=-1, then Y_2 must equal 0 among those cases that changed over time
- So we have the same logic as "linear" FD in "sweeping out" the ζ (U_i)

Problems with FE Logit

- We can potentially lose a *lot* of cases, so it is a *much* more inefficient method than RE, and if the ζ are not related to X, then FE is really a poor choice. Inferences can be shaky.
- As with FE in the continuous case, all time-invariant variables drop out because they are perfectly correlated with the ζ . Or rather, they provide us with no information in the CL procedure (see equation (15) or (16) for the logic of this, and assume that X(1)=X(2)).
- FE logit asks you to believe that the reasons for cases being all 1 or all 0 are not interesting, and the only thing that matters is *which* of the mixed cases pops up as 1. This is not so ideal from a theoretical point of view, because we are in effect throwing up our hands for the "all 1" or "all 0" cases and saying "UNIT EFFECT" instead of trying to attribute the pattern of responses to something that we do know.
- NOTE: we can, however, interact TIVs with time in the FE context, just as we did in the linear case. Plus TIV*TV interactions too.

- Alternative: extend the hybrid model to the dichotomous DV case, and add the cluster-level means of the time-varying Xs to predicting the cluster-level (subject-specific) intercept
- In multilevel terms:
- Level 1: (17) $\ln(\frac{P(Y=1|X)}{P(Y=0|X)}) = \beta_{0i} + \beta_1 X_{1it} + \beta_2 X_{2it} + \ldots \beta_k X_{kit}$
- Level 2: (18) $\beta_{0i} = \beta_{00} + \beta_{01}\bar{X}_{1i} + \zeta_{0i}$

  and
- Mixed (19) $\ln(\frac{P(Y=1|X)}{P(Y=0|X)}) = \beta_{00} + \beta_1 X_{1it} + \beta_{01}\bar{X}_{1i} + \beta_2 X_{2it} + \ldots \beta_k X_{kit} + \zeta_{0i}$

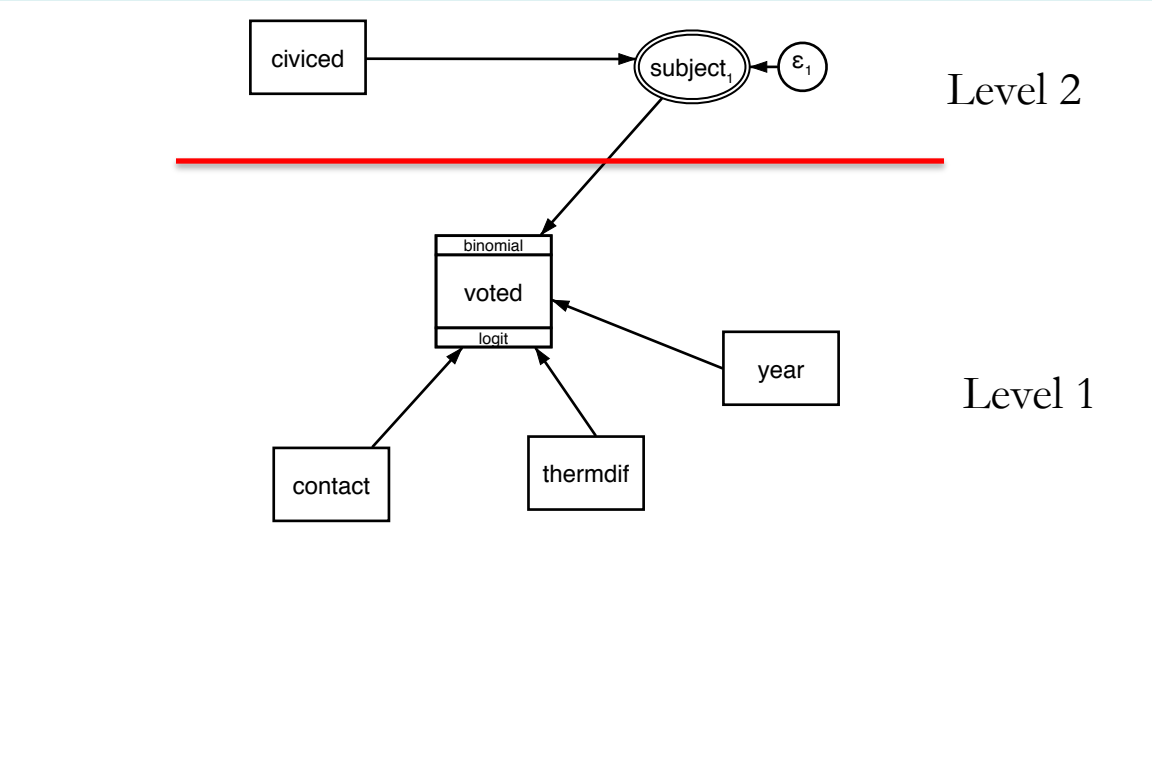
And, as we saw with the linear hybrid model, we also can estimate this with mean-deviated X and mean X as IVs:

(20) $\ln(\frac{P(Y=1|X)}{P(Y=0|X)}) = \beta_{00} + \beta_1 (X_{1it} - \bar{X}_{1i}) + \beta_3 \bar{X}_{1i} + \ldots . \beta_k (X_{kit} - \bar{X}_{ki}) + \beta * \bar{X}_{ki} + \zeta_i$

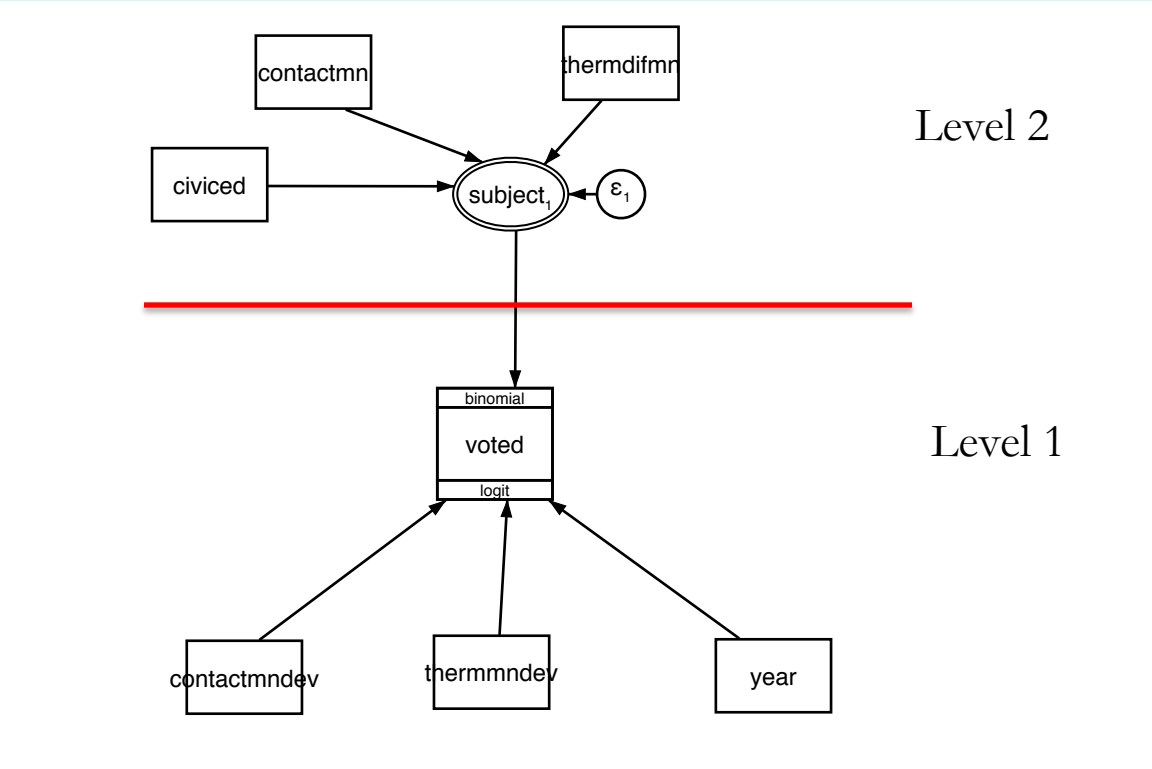
All this, and uses information from *all* cases, not only those changing on the DV! This is a huge potential improvement in the model!

- β_1 and β_k give you the "FE" within effect and β_3 and β^* give you the sum of the within and "between" effects. Testing whether β_1 equals β_3, or whether β_k equals β^* is the version of the Hausman test for the equality of FE and RE effects. If the coefficients are equal, then this means that β_{01} (or β_{0k}) are zero in equation (18), and hence traditional RE is fine. If the test shows they are not equal, then the hybrid model is better, with separate within and between effects
- Note that the hybrid model will not *exactly* reproduce the FE Logit because it doesn't lose all the case that FE logit does, so the results will only be approximate. But the general logic is fine!
- Using this formulation also gives:
  - Information on time-invariant IVs (in the context of a model controlling for endogeneity)
  - The possibility of specifying more complex random coefficient models
  - The possibility of adding more levels and random effects to the data hierarchy
  - The possibility of using a latent cluster mean as opposed to the sample mean value for estimating the "between" effect

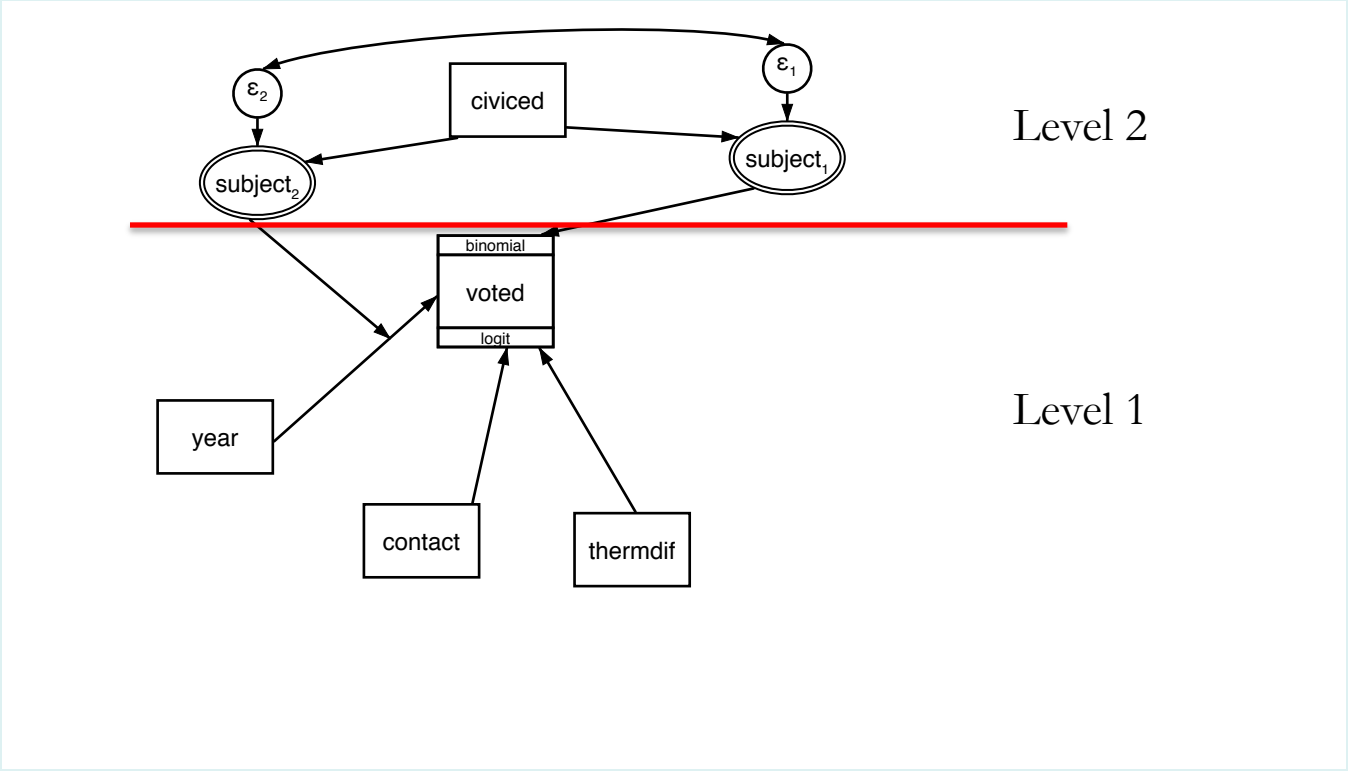
GSEM RE Logit



GSEM RE Hybrid Logit



GSEM RE Logit with Random Time Trend and Level 2 IV Predicting Time Trend



Full ML-SEM Model: Contact on Vote with three Random Effects