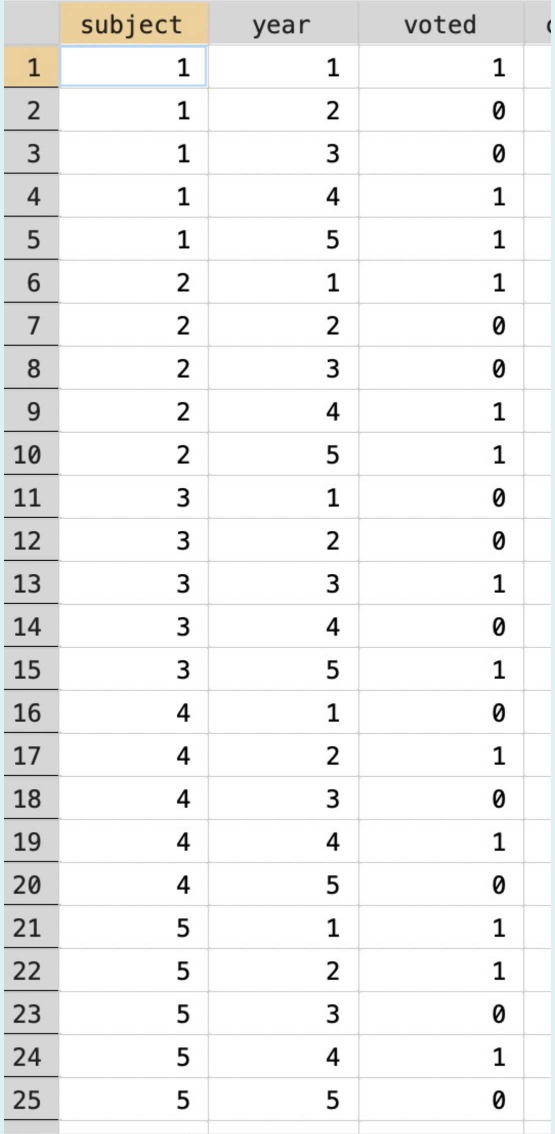
Unit 3
Longitudinal and Multilevel Models
3-1: Models for Dichotomous Outcomes

PS2730-2021

Week 10

Professor Steven Finkel

- Can extend the models we've considered in the course so far to handle *panel* and *multilevel* data, where observations are gathered on the same units over time (panel data), or on units that are "nested" at multiple levels of a data hierarchy (multilevel data), such as individuals "nested" within countries, students nested within classrooms, etc.
- Panel data itself is also a kind of multilevel data, where the waves of observation are "nested" within individual units; this means that the methods we'll consider next will be applicable to all kinds of multilevel structures, including individual (unit)-level panel data.



	subject	year	voted
1	1	1	1
2	1	2	0
3	1	3	0
4	1	4	1
5	1	5	1
6	2	1	1
7	2	2	0
8	2	3	0
9	2	4	1
10	2	5	1
11	3	1	0
12	3	2	0
13	3	3	1
14	3	4	0
15	3	5	1
16	4	1	0
17	4	2	1
18	4	3	0
19	4	4	1
20	4	5	0
21	5	1	1
22	5	2	1
23	5	3	0
24	5	4	1
25	5	5	0

- See how there are 5 years or "waves" of observations for each individual?
- Waves are "nested" within individuals
- Vote/No Vote varies for each individual at each point in time
- Other variables to explain voting: can be time-varying if they change over time (i.e., contacted by parties) or time-invariant (e.g., demographic characteristics)

Multilevel Models are Ubiquitous in the Social Sciences!

- **Panel data**, waves of observation or measurement (Level 1) are nested within individuals or units (Level 2)
- **Education research**, children (Level 1) are nested within classrooms (Level 2), nested within schools (Level 3).
- **Survey research**: respondents (Level 1) are (hypothetically) nested within blocks (Level 2) which are nested within Primary Sampling Units (Level 3) which are nested within, for example, U.S. states (Level 4)
- **Cross-national survey research**: respondents (Level 1) are nested within countries (Level 2). Example: World Values, LAPOP Surveys from one wave
- **Repeated Surveys on the Same Cross-Sectional Unit (Trend Analysis)**: respondents (Level 1) are nested within Time (Level 2), since in this kind of design *different individuals* from the same aggregate unit (country, state) are interviewed at multiple points in time. This is the *reverse* of panel data, where time of measurement (Level 1) is nested within individuals (Level 2). Example: Pooled NES Election data, 1948-2012.
- **Repeated Surveys on Multiple Cross-Sectional Units:** Individuals (Level 1) nested within country-years (Level 2) nested within country and year (Level 3). Example: World Values, LAPOP Surveys from multiple waves

- With this kind of data, it is imperative to take the multilevel structure of the data into account, at minimum to arrive at correct standard errors of parameter estimates. Since observations are not independent from other observations at the same level, there is *clustering* of the data that needs to be taken into account
- But panel/multilevel data have important benefits for modeling causal effects, so exploiting this kind of data goes beyond simply taking care of the statistical "nuisance" of clustered observations
- Panel data in particular allows the direct incorporation of individual (unit)-level *heterogeneity* into the analysis, and thus allows the analyst to model, and, under some circumstances, to "control for" the stable unobserved unit-level factors that may also both the independent and dependent variables and thus confound estimates of causal effects
- The corresponding heterogeneity in a non-longitudinal multilevel data structure is at "Level 2" or above, e.g., countries, neighborhoods, dyads, classrooms, schools, all of which may have stable characteristics which affect all of the "Level 1" units nested within it, aside from the Level-1 varying independent variables of interest
- So with panel/multilevel models, we have the ability to make more rigorous causal inferences than is possible with cross-sectional or single-level data!

- PS2701 (Longitudinal Analysis) covers these issues primarily for continuous outcomes; here we will cover basic models for handling categorical and count outcomes
- Many other uses of longitudinal data: modeling temporal dynamics, unit-level trajectories, measurement error, reciprocal effects. Take PS2701!

Dichotomous Dependent Variables

- Begin with familiar logit model:

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

- We construct the odds P(Y=1)/P(Y=0) as:

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

- And taking the log-odds of this expression is the "logit link" from the non-linear P(Y=1) to the linearized η in the GLM framework:

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

- 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. ML finds the parameters that maximize the joint likelihood of observing the outcomes for the given sample

Longitudinal and Multilevel Extensions

- The situation becomes more complex with longitudinal observations on Y_{it}, so that we observe Y_i at time points t=1, 2, 3…*t*. 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, as noted earlier, 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 multilevel case, the clustering could be by country, classroom, etc., and there may be unit effects at these levels – stable, unmeasured factors – that push the unit higher or lower on an outcome, regardless of the characteristics of the individuals/students/etc. at the lower "within cluster" level
- It may 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. 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

- Empirical example: In one high school, we randomly assign 140 of the seniors to participate in a special semester-long civic education program after school hours, with the remaining 77 being non-participants. We can ask: to what extent does the civic education "treatment" impact the students' likelihood of voting compared with traditional factors influencing turnout, such as feelings about candidates and mobilization by political parties?
- The marginal distributions for the treatment and control groups over time is:

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 from time-varying covariates

- Need to take into account explicitly individual-level heterogeneity, or unobservables that cause outcomes for individual *i* to be similar over *t*
- This is done through a unit-level, time-invariant heterogeneity term (ζ_i) 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_i term.
- Either assume that ζ_i 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_i and treat them as "fixed effects", leading to **Fixed Effects Logit** (there is no Fixed Effect Probit).
- There are analogues for these procedures that we will apply later to ordinal, multinomial, and count outcomes

Random Effects Logit

- Adding the ζ_i unit-level heterogeneity in the logit model leads to the expression for the probability that Y=1:

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

- 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
- Note: the "Random Effects" model we are considering here is also referred to as the "Random Intercept" model, since each unit has its own intercept, based on the magnitude of ζ_i

- The crucial assumptions of this version of the Random Effects (RE) model is that $E(X_{it}\ \zeta_i)=0$, i.e. that there is no covariation between the independent variables and the unit-specific effect. **This assumption restricts the ability of the RE model to estimate causal effects, since allowing for a correlation between X_{it} and ζ_i is what "controls" for the unobservables in a causal sense**
- A simple example: individuals who are contacted by political parties were already more interested in politics, so they have a higher ζ_i than individuals who were not contacted. And since interest → turnout, ζ_i is also related to the outcome Y_{it}. This is the classic omitted variables problem: ζ_i confounds the relationship between X_{it} and Y_{it}.
- This limits the applicability of the RE model for causal inference but:
  - the inclusion of ζ_i can handle clustering as a "nuisance" term
  - Further, in most analyses there is an effort to *predict* the level of ζ_i with other unit-level variables in attempting to understand why some units are higher or lower at all time points; this is the beginnings of stronger causal analysis
- Let's start with this model and then elaborate as needed!

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

$$\ln\frac{P(Y=1|X_{it})}{P(Y=0|X_{it})} = X_{it}B + \zeta_i = \beta_0 + \beta_1 X_{it} + \beta_2 X_{it} + \ldots \beta_k X_{kt} + \zeta_i$$

- 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_i), 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:

$$P(Y=1|X_{it}) = \Phi(X_{it}B + \zeta_i)$$
$$\Phi^{-1} = X_{it}B + \zeta_i = \beta_0 + \beta_1 X_{it} + \beta_2 X_{it} + \ldots \beta_k X_{kt} + \zeta_i$$

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

- Important note: In nearly all longitudinal models we need to deal with the effects of time such that *every* unit has a higher or lower probability of "1" at some times more than others. For example, voting is higher in presidential election years than non-presidential election years
- This is usually taken care of with dummy variables for time *(*t_1, t_2, etc.*)*

$$\ln \frac{P(Y=1|X_{it})}{P(Y=0|X_{it})} = X_{it}B + \zeta_i = \beta_0 + \beta_1 X_{it} + \beta_2 X_{it} + \ldots \beta_k X_{kt} + \ldots t_t + \zeta_i$$

or with a time trend/time counter variable. In our case, substantively speaking, we might have such a time trend "year" variable because we want to see how voting rates develop over time for the civic education and non-civic education groups.

- I won't put the t_t or time trend in every equation in the longitudinal section of this unit just to keep things simple in the presentation, but they are (almost always) there!

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:

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

$$L = \prod \left(\frac{e^{\beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \ldots \beta_\mathrm{k} \mathrm{X}_\mathrm{k} + \zeta_i}}{1 + e^{\beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \ldots \beta_\mathrm{k} \mathrm{X}_\mathrm{k} + \zeta_i}}\right)^{Y_i} \left(\frac{1}{1 + e^{\beta_0 + \beta_1 \mathrm{X}_1 + \beta_2 \mathrm{X}_2 + \ldots \beta_\mathrm{k} \mathrm{X}_\mathrm{k} + \zeta_i}}\right)^{1-Y_i}$$

- 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 XTLOGIT or MELOGIT procedure as robustness check -- the default in Stata is 7

Interpretation of RE Logit Coefficients

- The RE model:

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

- 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:

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

and when X=1 as:

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

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 .

Subject-Specific Versus Population-Average Effects

- Important: the effects in RE Logit are calculated at a *fixed* level of the random effect ζ_i, so it will NOT result in the same effect on P(Y=1) for all individuals!!!! This means that RE Logit provides what are called ***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_i. When ζ_i is very large or very small, a unit-change in X will not produce very large changes in the probability, while values of ζ_i that put the individual's 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:

$$\ln(\frac{P(Y=1|X)}{P(Y=0|X)}) = (X\text{B} + \zeta_i) = .5\text{X} + \zeta_i$$

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_i, 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_i, 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_i.

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_i. 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:**

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

- 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).
- Pooled logit gives you a population averaged estimate, as does a class of estimators for longitudinal data called Generalized Estimating Equations (GEE), which we won't have time to cover

The Multilevel Modeling Framework

- We can also arrive at the longitudinal RE Logit model through multilevel modeling
- Level 1 is the model for the outcome of each unit at the different waves or points in time: $\ln \frac{P(Y=1|X_{it})}{P(Y=0|X_{it})} = X_{it}B = \beta_{0i} + \beta_1 X_{it} + \beta_2 X_{it} + \ldots \beta_k X_{kt}$

  with β_{oj} as the unit-level intercept that **does not vary** over time.
- We then model β_{oi} at Level 2 with a "grand mean" population intercept and a unit-level random intercept that deviates from the grand mean

  Level 2: $\beta_{oi} = \beta_{00} + \zeta_{0i}$

  Putting this together yields the so-called "mixed" formulation:

  Mixed: $\ln \frac{P(Y=1|X_{it})}{P(Y=0|X_{it})} = X_{it}B = \beta_{00} + \beta_1 X_{it} + \beta_2 X_{it} + \ldots \beta_k X_{kt} + \zeta_{0i}$

which is identical to the RE model developed earlier except for the added notation in the grand mean intercept and random effect terms. In the multilevel world, the β are (confusingly) called "**fixed effects**", and ζ is a "**random effect"**.

- This is one kind of "**multilevel generalized mixed effects model**", which extends the GLM framework we've already considered to a multilevel framework that allows effects of IVs on non-continuous DVs to be estimated with multilevel or longitudinal data, so long as the model can be "linearized" via the "link" function to a linear specification.
- Mixed effects models
  - Mixed Effects Logit 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).
  - Mixed Effects Probit uses the "probit link" to move to P(Y=1) using the cumulative normal distribution
  - Mixed Effects Poisson Regression uses the "poisson" link to move to the "rate" from an exponentiated linear function
- In Stata, these models for longitudinal or multilevel data 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"

- Multilevel/longitudinal modeling is usually extended to include explanatory models at Level 2, where the variables specified at that level are used as independent variables for the Level 1 parameters
- Level 1 again is the model for the outcome of each unit at the different waves or points in time:

$$\ln \frac{P(Y=1|X_{it})}{P(Y=0|X_{it})} = X_{it}B = \beta_{0i} + \beta_1 X_{it} + \beta_2 X_{it} + \ldots \beta_k X_{kt}$$

with β_{oj} as the unit-level intercept.

- We then model β_{oi} at Level 2 with the "grand mean" population intercept, a unit-level random intercept, and **other explanatory variables at the unit-level**. For example, more highly educated individuals may generally vote at higher rates, men at higher rates than women, etc.

- Level 2: $\beta_{oi} = \beta_{oi} + \gamma_1 Z_i + \gamma_2 Z_i + \ldots \gamma_m Z_{mi} + \zeta_{0i}$
- The Z_{mi} are the *m* Level 2 explanatory variables that are constant for each unit (so they do not have a "t" subscript)

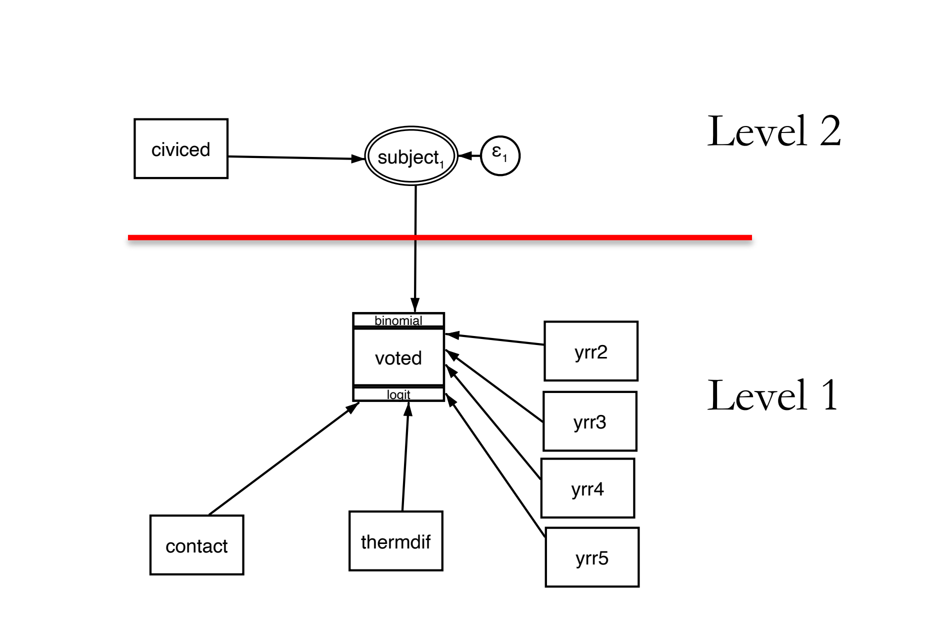
- Putting this together again yields the “mixed” formulation:

$$ln\frac{P(Y=1|X_{it})}{P(Y=0|X_{it})} = \beta_{00} + \beta_1 X_{it} + \beta_2 X_{it} + \ldots \beta_k X_{kt} + \gamma_1 Z_i + \gamma_2 Z_i + \ldots \gamma_m Z_{mi} + \zeta_{0i}$$

- So we are predicting the time-specific (log-odds or) probabilities of a “1” with a series of *k* time-varying X_{it} factors and a series of *m* time-constant Z_i factors, along with a unit-specific ζ_{0i} that pushes the probability of a “1” higher or lower, over and above all of the Level 1 factors and the Level 2 factors in the model
- Again, in the multilevel world, the β_k and γ_m are the “fixed” parts of the model and the ζ_{0i} is the “random” part – that’s what makes it “mixed”!

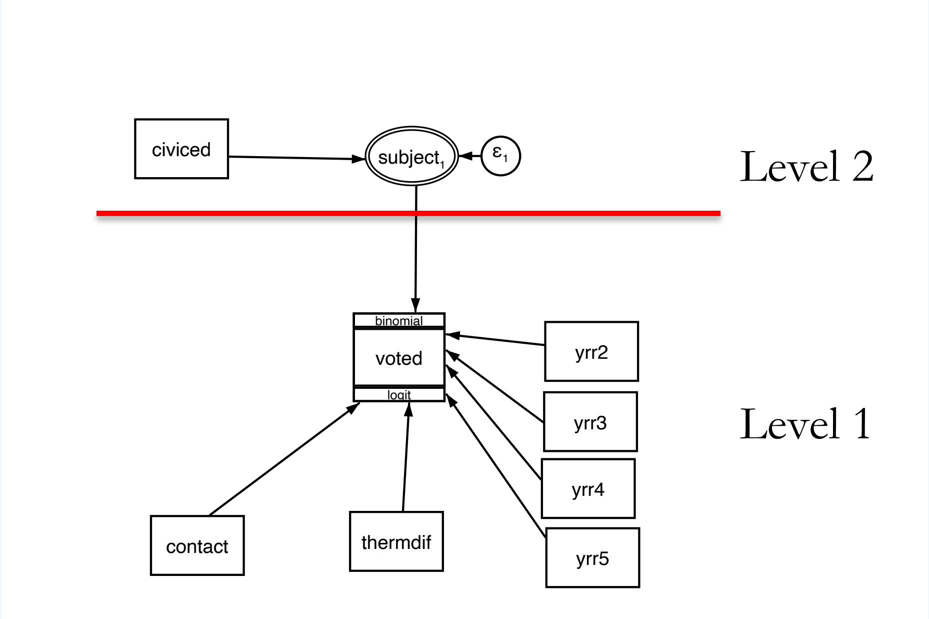
- There are other additions that can be included in multilevel models, e.g. another random effect for the β_k -- the slopes of the X_i variables -- so that the effect of, say, civic education, may differ for different kinds of individuals as well as the slope having its own random or unexplained component ζ_{1i} (see slides 29-30).
- We can include additional nesting levels as well. We won't have time to cover these additions but they can be very useful – see Rabe-Hesketh and Skrondal (or PS2701) for more!
- Finally: these models are applicable to *all* kinds of multilevel analyses, not only longitudinal/panel applications. If you have individuals nested within countries in the World Values Survey for 2005, just substitute "individuals" for Level 1 in the models we've considered, and "countries" for Level 2. Each country would have a random intercept, and you can explain the size of the random intercept with different country-level factors. You can also explain individual-level outcomes with individual characteristics and/or individual-country variable interactions, and/or add random slopes as mentioned above

- It can also be useful to depict multilevel models visually
- For example, here is the random intercept longitudinal model from our Stata do file



- This is from Stata’s GSEM (Generalized Structural Equation Modeling) module, an alternative way to estimate generalized multilevel models
- You can see that the bottom portion of the figure is the “Level 1” portion, with Level 2 in the top portion. (The levels are separated by the red line for emphasis)

- The model includes both level 1 explanatory variables (X_{it}) and a Level 2 explanatory variable (Z_i)

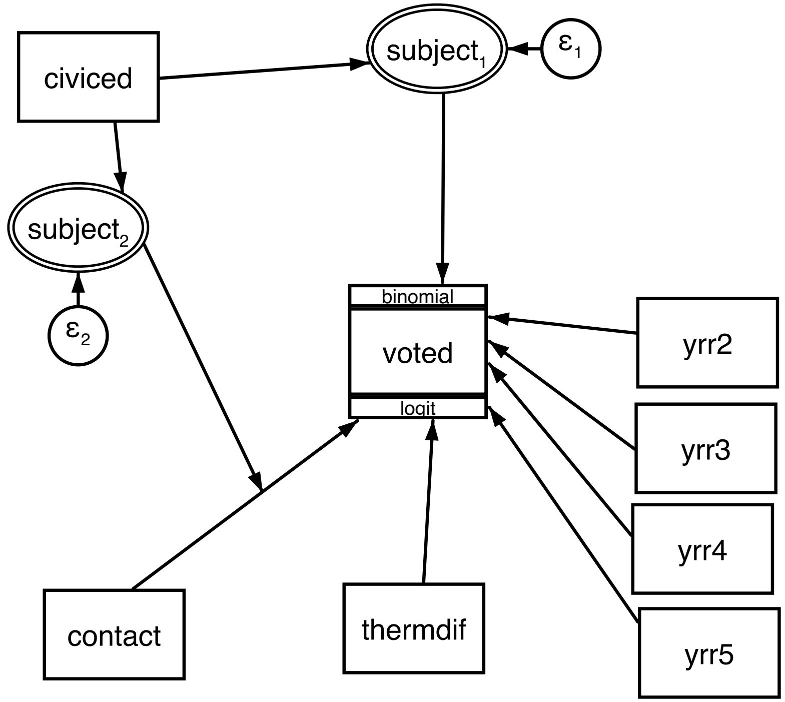


- At Level 1, we have time-varying factors that affect the vote – whether or not the person was contacted by a political party, and the difference between the feeling thermometer scores for each candidate.
- At Level 2, we explain the unit-level intercept with one time-invariant variable, whether or not the person received civic education in high school, as well as a random component

Advantages of MLM/GSEM (Briefly)

- Why use multilevel modeling instead of "straight up" random effects/random intercept modeling?
- Or, in Stata terminology, why use the "ME" suite or GSEM instead of XTLOGIT?
- Two major advantages of ME/GSEM: ability to include additional nests in the data hierarchy, each with their own random intercept (e.g., time nested within individuals nested within neighborhoods), and the ability to include random effects for other parameters in the model. In our example, it may be that the **effect** (slope) of contact on voter turnout varies randomly across individuals; further, this variation may be partially accounted for by Level 2 factors such as whether the individual was exposed to civic education
- We won't focus too much on these models here, but an example is on the next slide in GSEM diagram form.
- GSEM has advantages of its own: multiple equation estimation, mediation analysis, measurement error, and others we won't have time to cover

Random Intercept and Random Slope Model



Notice how there are now 2 random effects: the intercept at the subject level which represents the random variation in the general level of the dependent variable, and a random slope of the effect of contact on turnout (represented by the double circle "subject$_2$"). Both random effects are modeled as caused by high-school level exposure to civic education along with an error term which is the unexplained variation in the random effect, or the random effect that remains after including civic education as a predictor.

Fixed Effects and Hybrid RE Logit

- But what about the assumption in RE Logit that the unit-specific term ζ_i is unrelated to the Xs? As noted earlier that is ***the* big problem** with RE models generally.
- Recall the example mentioned earlier: people who are contacted by parties during campaigns are also those that may have the highest latent inclination to vote based on unobserved factors (ζ_i) such as their political interest or their normative integration into society. So it is *not* party contact that matters, rather the factors that led them to be contacted are the same (unobserved) ones that lead them to vote. These unobservables are folded into ζ_i, leading to biased estimation of the effects of contact (X_{it}) unless corrective steps are taken.
- Solution(s): 1) Fixed Effects Logit; and 2) Hybrid Random Effects, a multilevel model with unit-level means added to Level 2 equation

- In the linear case, we deal with the possible X_{it}-ζ_i correlation in a number of different ways -- all of them result in an estimation equation that **eliminates the ζ_i altogether** so the estimated effects of the X_{it} are "purged" of their possible correlation with ζ_i .
- For example: subtracting out the individual-level means on the X_{it} results in the **"Fixed Effects"** model (not to be confused with the "fixed effects" in mixed models with represent the β_k – arguably the most confusing terminology in all of statistical analysis!)
- $Y_{it} = \beta_0 + \beta_k X_{ikt} + \zeta_i + \varepsilon_{it}$
- $\bar{Y}_i = \beta_0 + \beta_k \bar{X}_{ik} + \bar{\zeta}_i + \bar{\varepsilon}_i$
- $(Y_{it} - \bar{Y}_i) = \alpha + \beta_k \, (X_{ikt} - \bar{X}_{ik}) \; + \varepsilon_i^*$
- See how the ζ_i are eliminated from consideration while estimating the β_k? This is a huge step towards controlling for unobservables and strengthening causal inference in non-experimental research!

- Can achieve same thing by including dummy variables for each case (Least Squared Dummy Variables) or by "First Differences", subtracting the equation for time period *t-1* from the equation for time period *t*
- **Unfortunately, none of these options are available for almost all models for non-continuous DVs!**
- For dichotomous outcomes, if we add a dummy variable for each case, we run up against the "incidental parameters" problem in ML estimation; 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 increases directly with sample size
- And neither de-meaning nor first-differencing works to subtract out the ζ, either, since the logit/log-odds model is intrinsically non-linear
- (Actually this statement is true only for multi-wave data, i.e., more than two waves. See Allison pp. 28-32 for the two-wave first difference case).
- So what to do? "Fixed Effects Logit" or "Hybrid Random Effects Logit"

- **The Fixed Effects Logit Model**
- This turns out to be the exact same model we used for conditional logit in the cross-sectional case a few weeks ago!
- Remember: we were trying then to model why an individual would choose alternative A, B or C based on attributes of the alternatives. Each alternative had its own row of data with a set of attributes which varied over alternatives (rows).
- We modeled the log-odds of choosing alternative (row) A versus B or C with

$$P(y = m|X) = \frac{exp^{X_m\beta}}{\sum_{j=1}^{J} exp^{X_j\beta}}$$

- The model was developed by the economist Gary Chamberlain in the early 1980s, based on "Conditional Maximum Likelihood" methods (which is why the procedure is called "conditional logit" in the first place)

- This has direct analogies to the longitudinal case. Think of the multiple alternatives ABC, each with its own row of data, as the waves of observation in a panel. Each wave has its own row, with the same variables down the columns registering differing values from wave to wave – in the same way as the different attributes have different values across alternatives in the cross-sectional conditional logit model
- We try to model why one or more rows has a “1” on the outcome while other rows have “0” based on the differing values of X_{it}
- The difference is that in the cross-sectional case, one and only one row had a “1” – the row with the alternative that was chosen. With longitudinal data there can be multiple rows with 1.
- **As long as at least 1 row but not all rows have a 1**, the case can figure into the likelihood function (cases with all 1s or all 0s over time on the dependent variable drop out of consideration, as we will see).

- Here is the basic likelihood for the longitudinal logit model:

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

$$L = \prod \left(\frac{e^{\beta_0+\beta_1 \mathrm{X}_1+\beta_2 \mathrm{X}_2+\ldots\beta_\mathrm{k} \mathrm{X}_\mathrm{k}+\zeta_i}}{1+e^{\beta_0+\beta_1 \mathrm{X}_1+\beta_2 \mathrm{X}_2+\ldots\beta_\mathrm{k} \mathrm{X}_\mathrm{k}+\zeta_i}}\right)^{Y_i} \left(\frac{1}{1+e^{\beta_0+\beta_1 \mathrm{X}_1+\beta_2 \mathrm{X}_2+\ldots\beta_\mathrm{k} \mathrm{X}_\mathrm{k}+\zeta_i}}\right)^{1-Y_i}$$

- 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*** $\boldsymbol{Y_i}$***, 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, as noted, 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.

- Let’s use the 2 wave case as an example:
- We want to maximize this quantity:

$$L = \prod_i \prod_t \left(\frac{e^{\beta_0+\beta_1\mathrm{X}_1+\beta_2\mathrm{X}_2+\ldots\beta_\mathrm{k}\mathrm{X}_\mathrm{k}+\zeta_i}}{1+e^{\beta_0+\beta_1\mathrm{X}_1+\beta_2\mathrm{X}_2+\ldots\beta_\mathrm{k}\mathrm{X}_\mathrm{k}+\zeta_i}}\right)^{Y_i}\left(\frac{1}{1+e^{\beta_0+\beta_1\mathrm{X}_1+\beta_2\mathrm{X}_2+\ldots\beta_\mathrm{k}\mathrm{X}_\mathrm{k}+\zeta_i}}\right)^{1-Y_i} \mid \Sigma(Y_i)$$

- 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 as:

$$L = \left(\frac{1}{1+e^{\beta_0+\beta_1\mathrm{X}_1+\beta_2\mathrm{X}_2+\ldots\beta_\mathrm{k}\mathrm{X}_\mathrm{k}+\zeta_i}}\right)\left(\frac{1}{1+e^{\beta_0+\beta_1\mathrm{X}_1+\beta_2\mathrm{X}_2+\ldots\beta_\mathrm{k}\mathrm{X}_\mathrm{k}+\zeta_i}}\right) \mid 0$$

- 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 maximizes 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 time 1 and "1" is observed at time 2, we maximize:

$$L = \frac{\left(\frac{e^{\beta_0+\beta_1 \mathrm{X}_1(2)+\beta_2 \mathrm{X}_2(2)+\ldots\beta_k \mathrm{X}_k(2)}}{1+e^{\beta_0+\beta_1 \mathrm{X}_1(2)+\beta_2 \mathrm{X}_2(2)+\ldots\beta_k \mathrm{X}_k(2)}}\right)}{\left(\frac{e^{\beta_0+\beta_1 \mathrm{X}_1(1)+\beta_2 \mathrm{X}_2(1)+\ldots\beta_k \mathrm{X}_k(1)}}{1+e^{\beta_0+\beta_1 \mathrm{X}_1(1)+\beta_2 \mathrm{X}_2(1)+\ldots\beta_k \mathrm{X}_k(1)}}\right)+\left(\frac{e^{\beta_0+\beta_1 \mathrm{X}_1(2)+\beta_2 \mathrm{X}_2(2)+\ldots\beta_k \mathrm{X}_k(2)}}{1+e^{\beta_0+\beta_1 \mathrm{X}_1(2)+\beta_2 \mathrm{X}_2(2)+\ldots\beta_k \mathrm{X}_k(2)}}\right)}$$

- 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:

$$L = \frac{\left(\frac{e^{\beta_0+\beta_1 X_1(1)+\beta_2 X_2(1)+\ldots\beta_k X_k(1)}}{1+e^{\beta_0+\beta_1 X_1(1)+\beta_2 X_2(1)+\ldots\beta_k X_k(1)}}\right)}{\left(\frac{e^{\beta_0+\beta_1 X_1(1)+\beta_2 X_2(1)+\ldots\beta_k X_k(1)}}{1+e^{\beta_0+\beta_1 X_1(1)+\beta_2 X_2(1)+\ldots\beta_k X_k(1)}}\right)+\left(\frac{e^{\beta_0+\beta_1 X_1(2)+\beta_2 X_2(2)+\ldots\beta_k X_k(2)}}{1+e^{\beta_0+\beta_1 X_1(2)+\beta_2 X_2(2)+\ldots\beta_k X_k(2)}}\right)}$$

- So the denominator has both ways of getting a 1 (at either wave 1 or at wave 2); the numerator has the particular wave that a 1 was observed

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_i . Or: they provide us with no information in the CL procedure (if X(1)=X(2) then the predicted P is the same for all points in time and the procedure breaks down).
- 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 waves 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 "FIXED EFFECT" (ζ) instead of trying to attribute the pattern of responses to something that we do know.

The Hybrid Random Effects Logit Model

- Alternative: apply what can be called a "hybrid" RE model to the dichotomous DV case, which adds the unit-level means of the time-varying Xs to predicting the unit-level (subject-specific) intercept
- In multilevel terms:

Level 1: $\ln(\frac{P(Y=1|X)}{P(Y=0|X)}) = \beta_{0i} + \beta_1 \mathrm{X}_{1it} + \beta_2 \mathrm{X}_{2it} + \ldots \beta_k \mathrm{X}_{kit}$

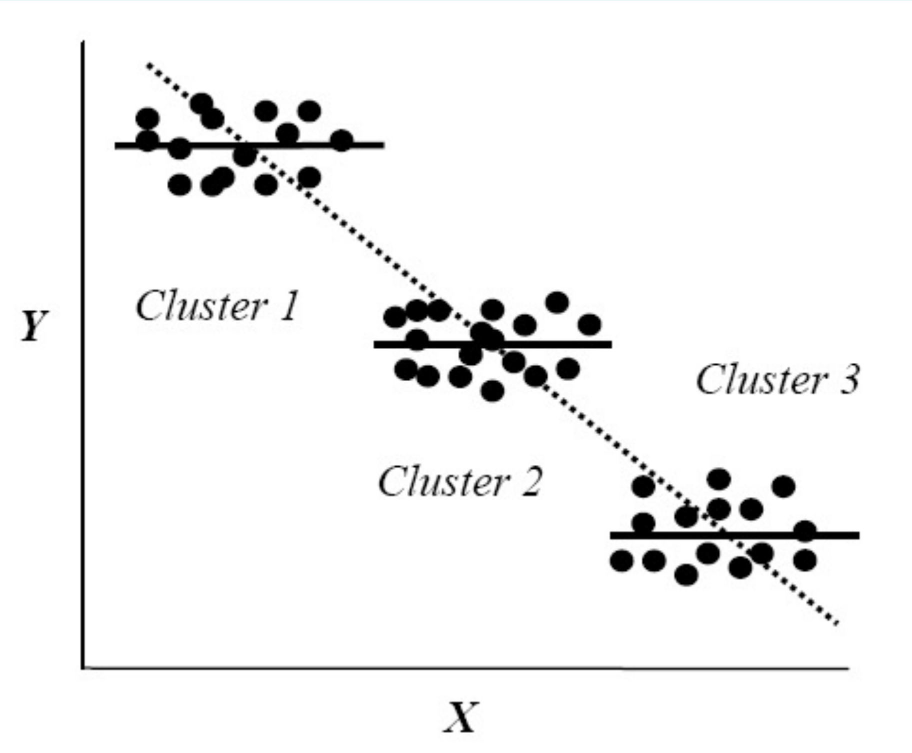
Level 2: $\beta_{0i} = \beta_{00} + \beta_{01} \bar{X}_{1i} + \zeta_{0i}$

and

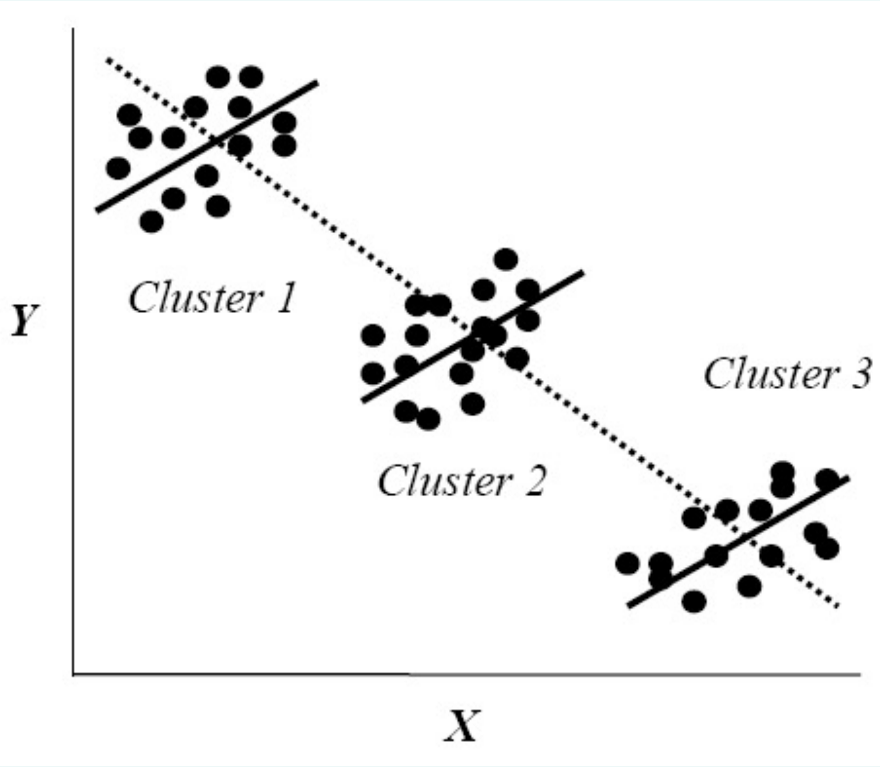
Mixed $\ln(\frac{P(Y=1|X)}{P(Y=0|X)}) = \beta_{00} + \beta_1 \mathrm{X}_{1it} + \beta_{01} \bar{X}_{1i} + \beta_2 \mathrm{X}_{2it} + \ldots \beta_k \mathrm{X}_{kit} + \zeta_{0i}$

- All this uses information from *all* cases, not only those changing on the DV! This is a huge potential improvement in the model!
- Another advantage: as in other RE models, you can include Z_i time-invariant variables at Level 2. So you get information on the factors that explain the level of the subject-specific intercept at Level 1

What does including X-bar do? It picks up the X_{it}-ζ_i correlation! For example, here units low on X(bar) are high on ζ_i , units high on X(bar) are low on ζ_i. Controlling for X-bar allows estimation of the "within-unit" effect regardless of where ζ_i generally puts the unit! The effect of X-bar on Y is called the "between effect" so in the hybrid model you estimate both "within" and "between" effects simultaneously, along with a random intercept. Great model!



Null "within-unit" effect, negative "between-unit" and negative overall effect



Positive "within-unit" effect, negative "between-unit" and negative overall effect

- Two Versions of the Hybrid Model: “Raw” and “Mean-Deviated”
- See slide 42: first version is “raw”

$$\ln(\frac{P(Y=1|X)}{P(Y=0|X)}) = \beta_{00} + \beta_1 \mathrm{X}_{1\mathrm{it}} + \beta_{01}\bar{X}_{1i} + \beta_2 \mathrm{X}_{2\mathrm{it}} + \ldots \beta_{\mathrm{k}} \mathrm{X}_{\mathrm{kit}} + \zeta_{0i}$$

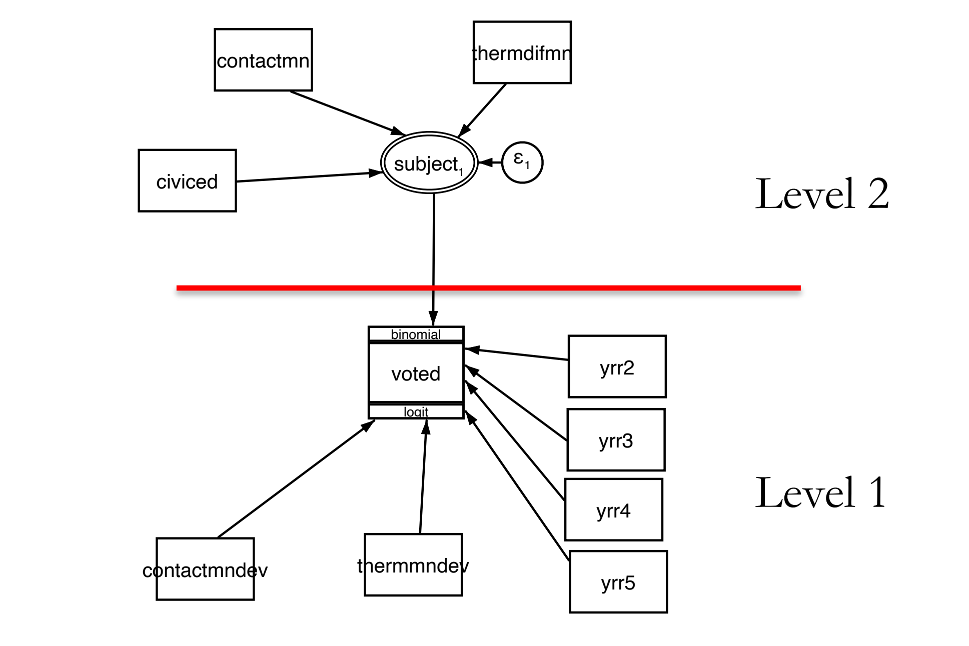
- Can also estimate this models with **mean-deviated X** and **mean X** as IVs:

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

- (This is done by subtracting and adding $\beta_1 \bar{X}$ and gathering terms)
- In this form of the model, β_1 and β_k give you the “within-unit” effects – controlling for the ζ_i -- and $\beta*_3$ and β^* give you the sum of the “within” and “between” effects. Testing whether β_1 equals $\beta*_{3,}$ or whether β_k equals β^* tests for whether the effect of the unit mean is zero. If they are equal, traditional RE logit is fine. If they are not equal, then the hybrid model is better.
- 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!

- Allison (blog post, statisticalhorizons.com) recommends the "raw" version of the hybrid model as opposed to the mean-deviated model, as this allows the incorporation of non-linear effects of the unit mean X-bar, which you can incorporate by adding X-bar-squared, cubed, etc. This is especially relevant with the multilevel (as opposed to longitudinal) version of the model
- Other advantages of the hybrid model:
  - Information on time-invariant IVs
  - The possibility of specifying more complex random coefficient models where the *effect* or slope of given covariates may vary randomly at higher levels (i.e., not only the intercept)
  - 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, in a Multilevel Structural Equation Modeling framework (ML-SEM) for estimating the "between" effects (see PS2701!)

- Here is the RE-Hybrid Logit Model in GSEM diagram form



- Level 1: mean-deviated contact and mean-deviated feeling thermometer differences, time dummies predicting votes; Level 2 mean contact, mean feeling thermometers, civic education predicting the individual-level random effect
- Test equality of coefficients between mean and mean-deviated variables; if equal then can go back to simple RE model