Generalized Linear Models for Proportions and Categorical Outcomes

- Today's Class:
  - **Review of 3 parts of a generalized model**
  - Models for proportion and percent correct outcomes
  - Models for categorical outcomes

3 Parts of Generalized (Multilevel) Models

1. Non-Normal Conditional Distribution of y_{ti} ← **2. Link Function** = **3. Linear Predictor of Fixed and Random Effects**

1. **Non-normal conditional distribution of y_{ti}:**
   - General MLM uses a *normal* conditional distribution to describe the y_{ti} variance remaining after fixed + random effects → we called this the level-1 residual variance, which is estimated separately and usually assumed constant across observations (unless modeled otherwise)
   - Other distributions will be more plausible for bounded/skewed y_{ti}, so the ML function maximizes the likelihood using those instead
   - **Why?** To get the most correct **standard errors** for fixed effects
   - Although you can still think of this as *model for the variance*, not all conditional distributions will actually have a separately estimated residual variance (e.g., binary → Bernoulli, count → Poisson)

3 Parts of Generalized (Multilevel) Models

1. Non-Normal Conditional Distribution of y_{ti} ⇐ **2. Link Function** = **3. Linear Predictor of Fixed and Random Effects**

2. **Link Function = $g(\cdot)$:** How the conditional mean to be predicted is transformed so that the model predicts an **unbounded** outcome instead
   - **Inverse link** $g^{-1}(\cdot)$= how to go back to conditional mean in y_{ti} scale
   - Predicted outcomes (found via inverse link) will then stay within bounds
   - e.g., binary outcome: conditional mean to be predicted is probability of a 1, so the model predicts a linked version (when inverse-linked, the predicted outcome will stay between a probability of 0 and 1)
   - e.g., count outcome: conditional mean is expected count, so the log of the expected count is predicted so that the expected count stays > 0
   - e.g., for normal outcome: an "identity" link function (y_{ti} * 1) is used given that the conditional mean to be predicted is already unbounded…

3 Parts of Generalized (Multilevel) Models

1. Non-Normal Conditional Distribution of y_{ti} ⇐ **2. Link Function** = **3. Linear Predictor of Fixed and Random Effects**

3. **Linear Predictor:** How the fixed and random effects of predictors combine additively to predict a link-transformed conditional mean
   - This works the same as usual, except the linear predictor model **directly predicts the link-transformed conditional mean**, which we then convert (via inverse link) back into the original conditional mean
   - That way we can still use the familiar "one-unit change" language to describe effects of model predictors (on the linked conditional mean)
   - You can think of this as "model for the means" still, but it also includes the level-2 random effects for dependency of level-1 observations
   - Fixed effects are no longer determined: they now have to be found through the ML algorithm, the same as the variance parameters

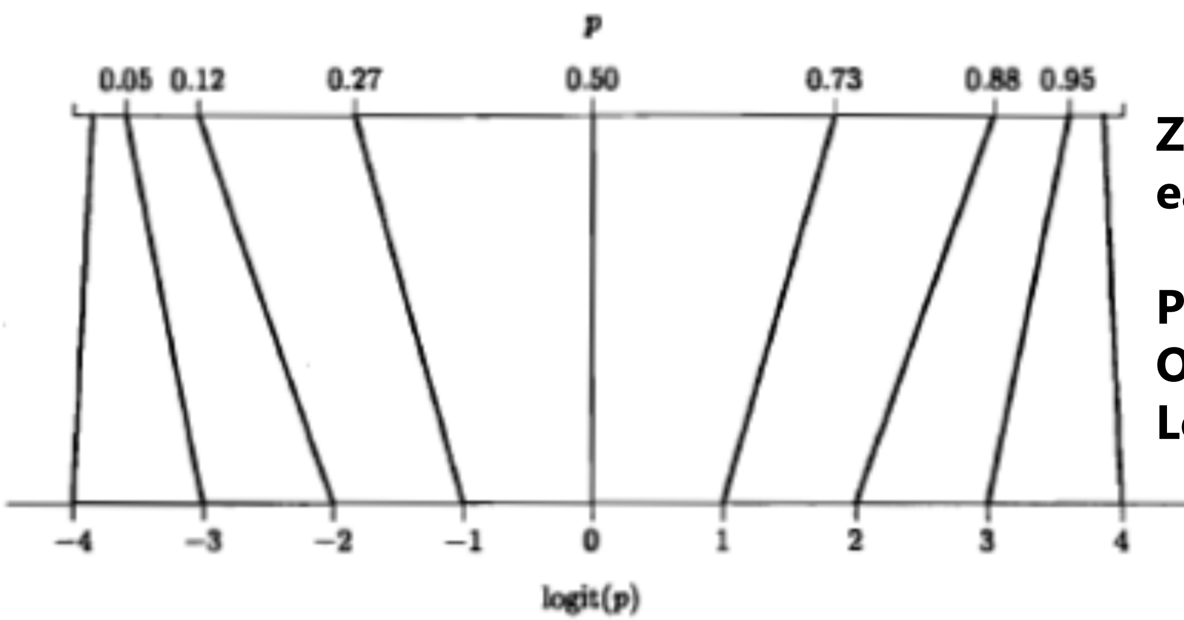
Probability, Odds, and Logits

- **A Logit link is a nonlinear transformation of probability:**
  - Equal intervals in logits are NOT equal intervals of probability
  - The logit goes from ±∞ and is symmetric about prob = .5 (logit = 0)
  - Now we can use a linear model → the model will be **linear with respect to the predicted logit**, which translates into a nonlinear prediction with respect to probability → **the conditional mean outcome shuts off at 0 or 1 as needed**

Probability: $p(y_i = 1)$

Odds: $\left[\frac{p}{1-p}\right]$

Logit (log odds): $\text{Log}\left[\frac{p}{1-p}\right]$



Zero-point on each scale:

Prob = .5
Odds = 1
Logit = 0

Too Logit to Quit: Predicting Proportions

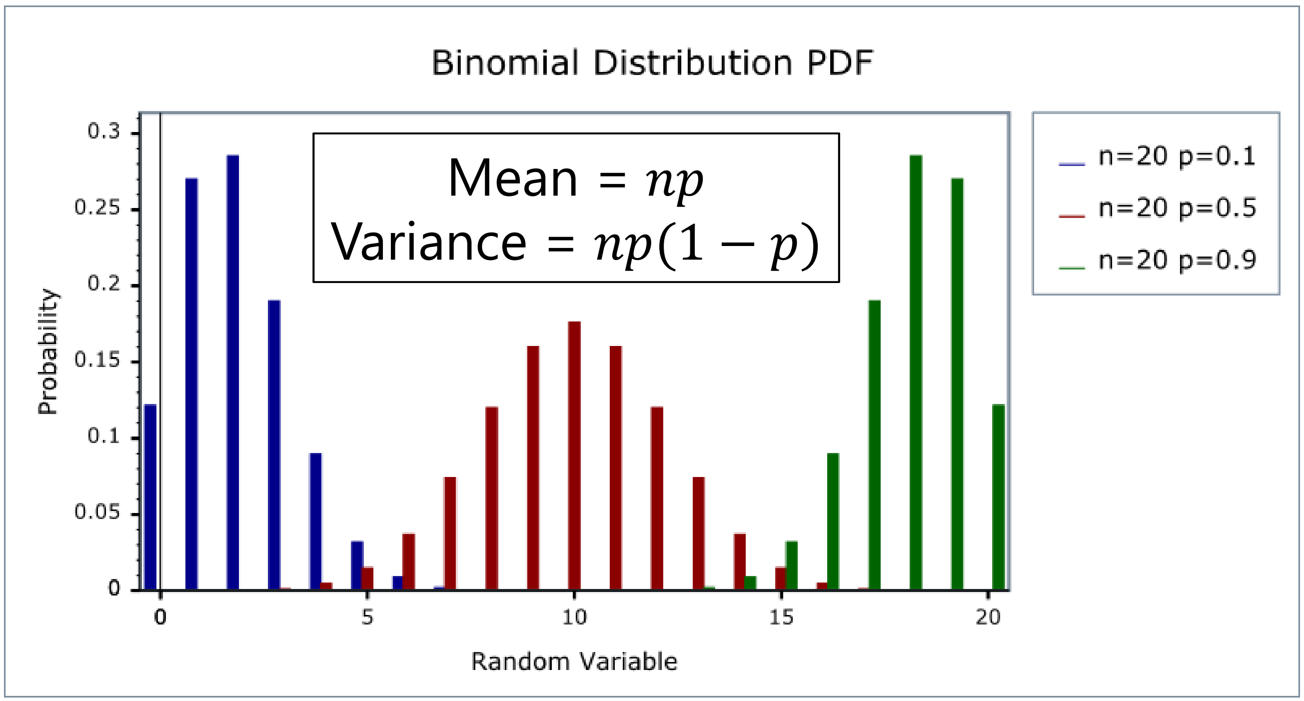
- The logit link can also be useful in predicting proportions:
  - Range between 0 and 1, so model needs to "shut off" predictions for conditional mean as they approach those ends, just as in binary data
  - Data to model: → μ in logits $= \text{Log}\left(\frac{p}{1-p}\right)$ ← $g(\cdot)$ **Link**
  - Model to data → $p = \frac{\exp(\mu)}{1+\exp(\mu)}$ ← $g^{-1}(\cdot)$ **Inverse-Link**

- However, because the outcome values aren't just 0 or 1, a Bernoulli conditional distribution won't work for proportions
- Two distributions: **Binomial** (discrete) vs. **Beta** (continuous)
  - Binomial: Less flexible (just one hump), but can include 0 and 1 values
  - Beta: Way more flexible (????), but cannot directly include 0 or 1 values
    - There are "zero-inflated" and/or "one-inflated" versions for these cases

Binomial Distribution for Proportions

- The discrete **binomial** distribution can be used to predict $\boldsymbol{c}$ correct responses given $\boldsymbol{n}$ trials
  - Bernoulli for binary = special case of binomial when n=1
  - $Prob(y = c) = \frac{n!}{c!(n-c)!} p^c (1-p)^{n-c}$

p = probability of 1



As p gets closer to .5 and n gets larger, the binomial pdf will look more like a normal distribution.

But if many people show floor/ceiling effects, a normal distribution is not likely to work well… so use a binomial!

Binomial Distribution for Proportions

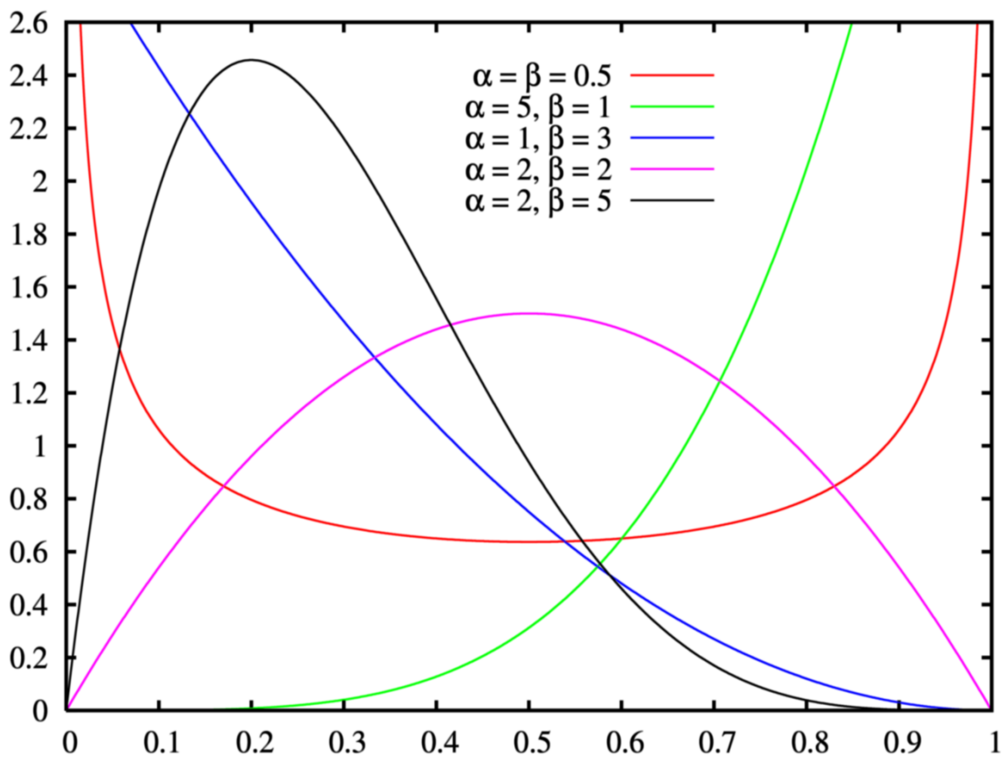
- SAS PROC GLIMMIX allows the outcome variable to be defined as ***#events/#trials*** on MODEL statement
  - LINK=LOGIT so that the conditional mean stays bounded between 0 and 1 as needed (or alternatively, CLOGLOG/LOGLOG)
  - DIST=BINOMIAL so variance (and SEs) are determined by that mean, as they should be assuming independent events
- STATA MELOGIT does the same with this option after ||:
  - Binomial(*VarforNtrials*); outcome then has number of events

- Be careful of **overdispersion**
  - Overdispersion = more variability than the mean would predict (cannot happen in binary outcomes, but it can for binomial)
  - Indicated by Pearson $\chi^2/\mathrm{df} > 1$ in SAS GLIMMIX output

Beta Distribution for Proportions

- The continuous **beta** distribution (SAS GLIMMIX LINK=LOGIT, DIST=BETA) can predict <u>percent</u> correct $\boldsymbol{p}$ (must be $0 < \boldsymbol{p} < 1$)
  - $F(y|\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1}$

α and β are "shape" parameters (> 0)



Mean = $\mu = \frac{\alpha}{\alpha+\beta}$

"Scale" = $\phi = \alpha + \beta$

Variance = $\frac{\mu(1-\mu)}{1+\phi}$

SAS GLIMMIX will provide a fixed intercept as logit(μ) and the "scale" ϕ

Beta Distribution for Proportions

- STATA appears to do beta regression models via a "betabin" add-on installed separately
- Does not appear to have a mixed effects version…?

- The beta distribution is extremely flexible (i.e., can take on many shapes), but outcomes must be $0 < y < 1$
  - If have 0's in outcome, need to add "zero-inflation" factor:
    → predicts logit of 0, then beta after 0 via two simultaneous models
  - If have 1's in outcome, need to add "one-inflation" factor:
    → predicts beta, then logit of 1 via two simultaneous models
  - Need both inflation factors if your outcome has 0s and 1s (3 models)
  - Can be used with outcomes that have other ranges of possible values if they are rescaled into 0 to 1

Too Logit to Quit…http://www.youtube.com/watch?v=Cdk1gwWH-Cg

- The **logit** is the basis for many other generalized models for categorical (ordinal or nominal; polytomous) outcomes

- Next we'll see how C possible response categories can be predicted using $C - 1$ binary "submodels" that involve carving up the categories in different ways, in which each binary submodel uses a logit link to predict its outcome

- Types of categorical outcomes:
  - Definitely ordered categories: "**cumulative logit**"
  - Maybe ordered categories: "**adjacent category logit**" (not used much)
  - Definitely NOT ordered categories: "**generalized logit**"

Logit-Based Models for C Ordinal Categories

- Known as "**cumulative logit**" or "**proportional odds**" model in generalized models; known as "graded response model" in IRT
  - LINK=CLOGIT, (DIST=MULT) in SAS GLIMMIX; MEOLOGIT or MEGLM in STATA
- Models the probability of **lower vs. higher** cumulative categories via $C-1$ submodels (e.g., if $C=4$ possible responses of $c=0{,}1{,}2{,}3$):

$\underbrace{\textbf{0} \text{ vs. } \textbf{1, 2,3}}_{\text{Submodel}_1}$ $\underbrace{\textbf{0,1} \text{ vs. } \textbf{2,3}}_{\text{Submodel}_2}$ $\underbrace{\textbf{0,1,2} \text{ vs. } \textbf{3}}_{\text{Submodel}_3}$

I've named these submodels based on what they predict, but program output will name them their own way…

- What the binary submodels predict depends on whether the model is predicting **DOWN ($\mathbf{y_i = 0}$)** or **UP ($\mathbf{y_i = 1}$)** **cumulatively**
- **Example predicting UP in an empty model (subscripts=parm,submodel)**
- Submodel 1: $\text{Logit}[p(\text{y}_\text{i} > 0)] = \beta_{01}$ $\rightarrow p(\text{y}_\text{i} > 0) = \exp(\beta_{01})/[1 + \exp(\beta_{01})]$
- Submodel 2: $\text{Logit}[p(\text{y}_\text{i} > 1)] = \beta_{02}$ $\rightarrow p(\text{y}_\text{i} > 1) = \exp(\beta_{02})/[1 + \exp(\beta_{02})]$
- Submodel 3: $\text{Logit}[p(\text{y}_\text{i} > 2)] = \beta_{03}$ $\rightarrow p(\text{y}_\text{i} > 2) = \exp(\beta_{03})/[1 + \exp(\beta_{03})]$

Logit-Based Models for C Ordinal Categories

- Models the probability of **lower vs. higher** cumulative categories via $C-1$ submodels (e.g., if $C=4$ possible responses of $c=0{,}1{,}2{,}3$):

$$\underbrace{\text{0 vs. 1,2,3}}_{\text{Submodel}_1 \rightarrow \text{Prob}_1} \qquad \underbrace{\text{0,1 vs. 2,3}}_{\text{Submodel}_2 \rightarrow \text{Prob}_2} \qquad \underbrace{\text{0,1,2 vs. 3}}_{\text{Submodel}_3 \rightarrow \text{Prob}_3}$$

$$\text{Logit}[p(\mathrm{y_i} > 2)] = \beta_{03}$$

$$\rightarrow p(\mathrm{y_i} > 2) = \frac{\exp(\beta_{03})}{1+\exp(\beta_{03})}$$

- What the binary submodels predict depends on whether the model is predicting **DOWN $(\mathbf{y_i = 0})$** or **UP $(\mathbf{y_i = 1})$** **cumulatively**
  - Either way, the model predicts the middle category responses ***indirectly***

- **Example if predicting UP with an empty model:**
  - Probability of 0 = 1 – Prob_1
    Probability of 1 = Prob_1– Prob_2
    Probability of 2 = Prob_2– Prob_3
    Probability of 3 = Prob_3– 0

The cumulative submodels that create these probabilities are each estimated using **all the data** (good, especially for categories not chosen often), but **assume order in doing so** (may be bad or ok, depending on your response format).

Logit-Based Models for C Ordinal Categories

- Ordinal models usually use a logit link transformation, but they can also use cumulative log-log or cumulative complementary log-log links
  - LINK= CUMLOGLOG or CUMCLL in SAS GLIMMIX; CLOGLOG link in MEGLM in STATA

- Almost always assume **proportional odds**, that effects of predictors are the same across binary submodels—for example (subscripts = parm, submodel)
  - Submodel 1: $\mathrm{Logit}[p(\mathrm{y_i} > 0)] = \boldsymbol{\beta_{01}} + \beta_1 \mathrm{X_i} + \beta_2 \mathrm{Z_i} + \beta_3 \mathrm{X_i Z_i}$
  - Submodel 2: $\mathrm{Logit}[p(\mathrm{y_i} > 1)] = \boldsymbol{\beta_{02}} + \beta_1 \mathrm{X_i} + \beta_2 \mathrm{Z_i} + \beta_3 \mathrm{X_i Z_i}$
  - Submodel 3: $\mathrm{Logit}[p(\mathrm{y_i} > 2)] = \boldsymbol{\beta_{03}} + \beta_1 \mathrm{X_i} + \beta_2 \mathrm{Z_i} + \beta_3 \mathrm{X_i Z_i}$

- Proportional odds essentially means no interaction between submodel and predictor effects, which greatly reduces the number of estimated parameters
  - Despite the importance of this assumption, there appears to be no way to test it directly in most software packages for mixed effects models (except SAS NLMIXED)
  - If the proportional odds assumption fails, you can use a nominal model instead (dummy-coding to create separate outcomes can approximate a nominal model)

Logit-Based Models for C Categories

- Uses multinomial distribution, whose PDF for $C = 4$ categories of $c = 0,1,2,3$, an observed $y_i = c$, and indicators I if $c = y_i$

$$f(\mathrm{y_i} = \mathrm{c}) = p_{\mathrm{i0}}^{\mathrm{I[y_i=0]}} p_{\mathrm{i1}}^{\mathrm{I[y_i=1]}} p_{\mathrm{i2}}^{\mathrm{I[y_i=2]}} p_{\mathrm{i3}}^{\mathrm{I[y_i=3]}}$$

Only p_{ic} for the response $y_i = c$ gets used

  - Maximum likelihood is then used to find the most likely parameters in the model to predict the probability of each response through the (usually logit) link function; probabilities sum to 1: $\sum_{\mathrm{c=1}}^{\mathrm{C}} p_{\mathrm{ic}} = 1$

- Other models for categorical data that use the multinomial:
  - Adjacent category logit (partial credit): Models the probability of **each next highest** category via $C - 1$ submodels (e.g., if $C = 4$):

    0 vs. **1**   **1** vs. **2**   **2** vs. **3**

  - Baseline category logit (**nominal**): Models the probability of **reference vs. other** category via $C - 1$ submodels (e.g., if $C = 4$ and $0 = \mathrm{ref}$):

    0 vs. **1**   **0** vs. **2**   **0** vs. **3**

In **nominal** models, all parameters are estimated **separately** per submodel

One More Idea…

- Ordinal data can sometimes also be approximated with a logit link and binomial distribution instead
  - Example: Likert scale from $0-4$ → $\#\,\mathrm{trials} = 4, \#\,\mathrm{correct} = \mathrm{y_i}$
  - Model predicts p of binomial distribution, $p * \#trials\ =\ mean$
  - $p(\mathrm{y_i})$ = proportion of sample expected in that y_i response category
- Advantages:
  - Only estimates one parameter that creates a conditional mean for each response category, instead of $C-1$ cumulative intercepts or thresholds
  - Can be used even if there is sparse data in some categories
  - Results may be easier to explain than if using cumulative sub-models
- Disadvantages:
  - # persons in each category will not be predicted perfectly to begin with, so it may not fit the data as well without the extra intercept parameters

Generalized MLM: Summary

- Statistical models come from probability distributions
  - Conditional outcomes are assumed to have some distribution
  - The normal distribution is one choice, but there are lots of others: so far we've seen Bernoulli, binomial, beta, and multinomial
  - ML estimation tries to maximize the height of the data using that distribution along with the model parameters

- Generalized models have three parts:
  1. Non-normal conditional outcome distribution
  2. Link function: how bounded conditional mean of y_{ti} gets transformed into something unbounded we can predict linearly
     - So far we've seen identity, logit, probit, log-log, and cumulative log-log
  3. Linear predictor: how we predict that linked conditional mean