Multilevel Models for Binary Responses
Consider a 2-level hierarchical structure. Use ‘group’ as a general term for a level 2 unit (e.g. area, school).

**Notation**

- $n$ is total number of individuals (level 1 units)
- $J$ is number of groups (level 2 units)
- $n_j$ is number of individuals in group $j$
- $y_{ij}$ is binary response for individual $i$ in group $j$
- $x_{ij}$ is an individual-level predictor
Recall model for continuous $y$

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}$$

$$u_j \sim N(0, \sigma_u^2) \text{ and } e_{ij} \sim N(0, \sigma_e^2)$$
Recall model for continuous $y$

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}$$

$$u_j \sim N(0, \sigma_u^2) \text{ and } e_{ij} \sim N(0, \sigma_e^2)$$

or, expressed as model for expected value of $y_{ij}$ for given $x_{ij}$ and $u_j$:

$$E(y_{ij}) = \beta_0 + \beta_1 x_{ij} + u_j$$
Recall model for continuous $y$

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}$$

$$u_j \sim N(0, \sigma_u^2) \text{ and } e_{ij} \sim N(0, \sigma_e^2)$$

or, expressed as model for expected value of $y_{ij}$ for given $x_{ij}$ and $u_j$:

$$E(y_{ij}) = \beta_0 + \beta_1 x_{ij} + u_j$$

Model for binary $y$

For binary response $E(y_{ij}) = \pi_{ij} = \Pr(y_{ij} = 1)$, and model is

$$F^{-1}(\pi_{ij}) = \beta_0 + \beta_1 x_{ij} + u_j$$

$F^{-1}$ the link function, e.g. logit, probit clog-log
Random Intercept Logit Model: Interpretation

\[
\log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \beta_0 + \beta_1 x_{ij} + u_j
\]

\[
u_j \sim N(0, \sigma_u^2)
\]
Random Intercept Logit Model: Interpretation

\[
\log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \beta_0 + \beta_1 x_{ij} + u_j
\]

\[
u_j \sim N(0, \sigma_u^2)
\]

**Interpretation of fixed part**

- \(\beta_0\) is the log-odds that \(y = 1\) when \(x = 0\) and \(u = 0\)
Random Intercept Logit Model: Interpretation

\[
\log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \beta_0 + \beta_1 x_{ij} + u_j
\]

\[u_j \sim N(0, \sigma_u^2)\]

**Interpretation of fixed part**

- $\beta_0$ is the log-odds that $y = 1$ when $x = 0$ and $u = 0$
- $\beta_1$ is effect on log-odds of 1-unit increase in $x$ for individuals in same group (same value of $u$)
Random Intercept Logit Model: Interpretation

\[
\log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \beta_0 + \beta_1 x_{ij} + u_j
\]

\[u_j \sim N(0, \sigma_u^2)\]

**Interpretation of fixed part**

- \(\beta_0\) is the log-odds that \(y = 1\) when \(x = 0\) and \(u = 0\)
- \(\beta_1\) is effect on log-odds of 1-unit increase in \(x\) for individuals in same group (same value of \(u\))
- \(\beta_1\) is often referred to as cluster-specific or unit-specific effect of \(x\)
Random Intercept Logit Model: Interpretation

\[ \log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \beta_0 + \beta_1 x_{ij} + u_j \]

\[ u_j \sim N(0, \sigma_u^2) \]

**Interpretation of fixed part**

- \( \beta_0 \) is the log-odds that \( y = 1 \) when \( x = 0 \) and \( u = 0 \)
- \( \beta_1 \) is effect on log-odds of 1-unit increase in \( x \) for individuals in same group (same value of \( u \))
- \( \beta_1 \) is often referred to as **cluster-specific** or **unit-specific** effect of \( x \)
- \( \exp(\beta_1) \) is an odds ratio, comparing odds for individuals spaced 1-unit apart on \( x \) but in the same group
Random Intercept Logit Model: Interpretation

\[ \log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \beta_0 + \beta_1 x_{ij} + u_j \]

\[ u_j \sim N(0, \sigma_u^2) \]

**Interpretation of random part**

- \( u_j \) is the effect of being in group \( j \) on the log-odds that \( y = 1 \); also known as a level 2 residual
Random Intercept Logit Model: Interpretation

\[
\log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \beta_0 + \beta_1 x_{ij} + u_j
\]

\[u_j \sim N(0, \sigma_u^2)\]

**Interpretation of random part**

- \( u_j \) is the effect of being in group \( j \) on the log-odds that \( y = 1 \); also known as a level 2 residual.
- As for continuous \( y \), we can obtain estimates and confidence intervals for \( u_j \).
Random Intercept Logit Model: Interpretation

\[
\log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \beta_0 + \beta_1 x_{ij} + u_j
\]

\[u_j \sim N(0, \sigma_u^2)\]

**Interpretation of random part**

- \(u_j\) is the effect of being in group \(j\) on the log-odds that \(y = 1\); also known as a level 2 residual.
- As for continuous \(y\), we can obtain estimates and confidence intervals for \(u_j\).
- \(\sigma_u^2\) is the level 2 (residual) variance, or the between-group variance in the log-odds that \(y = 1\) after accounting for \(x\).
Response probability for individual $i$ in group $j$ calculated as

$$
\pi_{ij} = \frac{\exp(\beta_0 + \beta_1 x_{ij} + u_j)}{1 + \exp(\beta_0 + \beta_1 x_{ij} + u_j)}
$$
Response probability for individual $i$ in group $j$ calculated as

$$\pi_{ij} = \frac{\exp(\beta_0 + \beta_1 x_{ij} + u_j)}{1 + \exp(\beta_0 + \beta_1 x_{ij} + u_j)}$$

Substitute estimates of $\beta_0$, $\beta_1$ and $u_j$ to get predicted probability:

$$\hat{\pi}_{ij} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{ij} + \hat{u}_j)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{ij} + \hat{u}_j)}$$

We can also make predictions for 'ideal' or 'typical' individuals with particular values for $x$, but we need to decide what to substitute for $u_j$ (discussed later).
Response probability for individual $i$ in group $j$ calculated as

$$\pi_{ij} = \frac{\exp(\beta_0 + \beta_1 x_{ij} + u_j)}{1 + \exp(\beta_0 + \beta_1 x_{ij} + u_j)}$$

Substitute estimates of $\beta_0$, $\beta_1$ and $u_j$ to get predicted probability:

$$\hat{\pi}_{ij} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_{ij} + \hat{u}_j)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{ij} + \hat{u}_j)}$$

We can also make predictions for ‘ideal’ or ‘typical’ individuals with particular values for $x$, but we need to decide what to substitute for $u_j$ (discussed later).
Individuals (at level 1) within states (at level 2).
Example: US Voting Intentions

Individuals (at level 1) within states (at level 2).

Results from null logit model (no \( x \))

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 ) (intercept)</td>
<td>-0.107</td>
<td>0.049</td>
</tr>
<tr>
<td>( \sigma_u^2 ) (between-state variance)</td>
<td>0.091</td>
<td>0.023</td>
</tr>
</tbody>
</table>
Example: US Voting Intentions

Individuals (at level 1) within states (at level 2).

Results from null logit model (no $x$)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$ (intercept)</td>
<td>-0.107</td>
<td>0.049</td>
</tr>
<tr>
<td>$\sigma^2_u$ (between-state variance)</td>
<td>0.091</td>
<td>0.023</td>
</tr>
</tbody>
</table>

Questions about $\sigma^2_u$

1. Is $\sigma^2_u$ significantly different from zero?
Example: US Voting Intentions

Individuals (at level 1) within states (at level 2).

Results from null logit model (no $x$)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$ (intercept)</td>
<td>-0.107</td>
<td>0.049</td>
</tr>
<tr>
<td>$\sigma_u^2$ (between-state variance)</td>
<td>0.091</td>
<td>0.023</td>
</tr>
</tbody>
</table>

Questions about $\sigma_u^2$

1. Is $\sigma_u^2$ significantly different from zero?
2. Does $\hat{\sigma}_u^2 = 0.09$ represent a large state effect?
Testing $H_0 : \sigma_u^2 = 0$

- Likelihood ratio test. Only available if model estimated via maximum likelihood (not in MLwiN)

Example

\[
\text{Wald statistic} = \left(\hat{\sigma}_u^2 \text{se}\right)^2 = (0.0910.023)^2 = 15.65
\]

Compare with $\chi^2_1 \to$ reject $H_0$ and conclude there are state differences.

Take p-value/2 because alternative hypothesis is one-sided ($H_A : \sigma_u^2 > 0$)
Testing $H_0 : \sigma^2_u = 0$

- **Likelihood ratio test.** Only available if model estimated via maximum likelihood (not in MLwiN)
- **Wald test** (equivalent to t-test), but only approximate because variance estimates do not have normal sampling distributions

Example

Wald statistic = \left( \frac{\hat{\sigma}^2_u}{\text{SE}} \right)^2 = \left( \frac{0.091}{0.023} \right)^2 = 15.65

Compare with $\chi^2_1 \rightarrow$ reject $H_0$ and conclude there are state differences.

Take p-value/2 because alternative hypothesis is one-sided ($H_A : \sigma^2_u > 0$)
Testing $H_0 : \sigma_u^2 = 0$

- **Likelihood ratio test.** Only available if model estimated via maximum likelihood (not in MLwiN)
- **Wald test** (equivalent to t-test), but only approximate because variance estimates do not have normal sampling distributions
- **Bayesian credible intervals.** Available if model estimated using Markov chain Monte Carlo (MCMC) methods.

Example:

$$Wald statistic = \left(\hat{\sigma}_u^2 / se\right)^2 = (0.091/0.023)^2 = 15.65$$

Compare with $\chi^2_1 \rightarrow$ reject $H_0$ and conclude there are state differences.

Take p-value/2 because alternative hypothesis is one-sided ($H_A : \sigma_u^2 > 0$)
Testing $H_0 : \sigma^2_u = 0$

- **Likelihood ratio test.** Only available if model estimated via maximum likelihood (not in MLwiN)
- **Wald test** (equivalent to t-test), but only approximate because variance estimates do not have normal sampling distributions
- **Bayesian credible intervals.** Available if model estimated using Markov chain Monte Carlo (MCMC) methods.

Example

Wald statistic = \left( \frac{\hat{\sigma}_u^2}{\text{se}} \right)^2 = \left( \frac{0.091}{0.023} \right)^2 = 15.65

Compare with $\chi^2_1 \rightarrow$ reject $H_0$ and conclude there are state differences.

Take p-value/2 because alternative hypothesis is one-sided ($H_A : \sigma^2_u > 0$)
Testing $H_0 : \sigma^2_u = 0$

- Likelihood ratio test. Only available if model estimated via maximum likelihood (not in MLwiN).
- Wald test (equivalent to t-test), but only approximate because variance estimates do not have normal sampling distributions.
- Bayesian credible intervals. Available if model estimated using Markov chain Monte Carlo (MCMC) methods.

**Example**

$$\text{Wald statistic} = \left( \frac{\hat{\sigma}^2_u}{\text{se}} \right)^2 = \left( \frac{0.091}{0.023} \right)^2 = 15.65$$

Compare with $\chi^2_1 \rightarrow$ reject $H_0$ and conclude there are state differences.
Testing $H_0 : \sigma^2_u = 0$

- **Likelihood ratio test.** Only available if model estimated via maximum likelihood (not in MLwiN).
- **Wald test** (equivalent to t-test), but only approximate because variance estimates do not have normal sampling distributions.
- **Bayesian credible intervals.** Available if model estimated using Markov chain Monte Carlo (MCMC) methods.

**Example**

\[
\text{Wald statistic} = \left( \frac{\hat{\sigma}^2_u}{\text{se}} \right)^2 = \left( \frac{0.091}{0.023} \right)^2 = 15.65
\]

Compare with $\chi^2_1 \rightarrow$ reject $H_0$ and conclude there are state differences.

Take $p$-value/2 because alternative hypothesis is one-sided ($H_A : \sigma^2_u > 0$)
Calculate $\hat{\pi}$ for ‘average’ states ($u = 0$) and for states that are 2 standard deviations above and below the average ($u = \pm 2\hat{\sigma}_u$).

$\hat{\sigma}_u = \sqrt{0.091} = 0.3017$
State Effects on Probability of Voting Bush

Calculate $\hat{\pi}$ for ‘average’ states ($u = 0$) and for states that are 2 standard deviations above and below the average ($u = \pm 2\hat{\sigma}_u$).

$\hat{\sigma}_u = \sqrt{0.091} = 0.3017$

$u = -2\hat{\sigma}_u = -0.603 \rightarrow \hat{\pi} = 0.33$

$u = 0 \rightarrow \hat{\pi} = 0.47$

$u = +2\hat{\sigma}_u = +0.603 \rightarrow \hat{\pi} = 0.62$

Under a normal distribution assumption, expect 95% of states to have $\hat{\pi}$ within ($0.33, 0.62$).
State Effects on Probability of Voting Bush

Calculate $\hat{\pi}$ for ‘average’ states ($u = 0$) and for states that are 2 standard deviations above and below the average ($u = \pm 2\hat{\sigma}_u$).

$\hat{\sigma}_u = \sqrt{0.091} = 0.3017$

$u = -2\hat{\sigma}_u = -0.603 \rightarrow \hat{\pi} = 0.33$

$u = 0 \rightarrow \hat{\pi} = 0.47$

$u = +2\hat{\sigma}_u = +0.603 \rightarrow \hat{\pi} = 0.62$

Under a normal distribution assumption, expect 95% of states to have $\hat{\pi}$ within (0.33, 0.62).
$\hat{u}_j$ with 95% Confidence Intervals for $u_j$
Adding Income as a Predictor

$x_{ij}$ is household annual income (grouped into 9 categories), centred at sample mean of 5.23

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$ (constant)</td>
<td>$-0.099$</td>
<td>0.056</td>
</tr>
<tr>
<td>$\beta_1$ (income, centered)</td>
<td>0.140</td>
<td>0.008</td>
</tr>
<tr>
<td>$\sigma_u^2$ (between-state variance)</td>
<td>0.125</td>
<td>0.030</td>
</tr>
</tbody>
</table>

$-0.099$ is the log-odds of voting Bush for household of mean income living in an 'average' state.

0.140 is the effect on the log-odds of a 1-category increase in income.

Expected odds of voting Bush to be $\exp(8 \times 0.14) = 3.1$ times higher for an individual in the highest income band than for an individual in the same state but in the lowest income band.
**Adding Income as a Predictor**

$x_{ij}$ is household annual income (grouped into 9 categories), centred at sample mean of 5.23

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$ (constant)</td>
<td>−0.099</td>
<td>0.056</td>
</tr>
<tr>
<td>$\beta_1$ (income, centered)</td>
<td>0.140</td>
<td>0.008</td>
</tr>
<tr>
<td>$\sigma^2_u$ (between-state variance)</td>
<td>0.125</td>
<td>0.030</td>
</tr>
</tbody>
</table>

−0.099 is the log-odds of voting Bush for household of mean income living in an ‘average’ state
Adding Income as a Predictor

$x_{ij}$ is household annual income (grouped into 9 categories), centred at sample mean of 5.23

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$ (constant)</td>
<td>$-0.099$</td>
<td>$0.056$</td>
</tr>
<tr>
<td>$\beta_1$ (income, centered)</td>
<td>$0.140$</td>
<td>$0.008$</td>
</tr>
<tr>
<td>$\sigma^2_u$ (between-state variance)</td>
<td>$0.125$</td>
<td>$0.030$</td>
</tr>
</tbody>
</table>

- $-0.099$ is the log-odds of voting Bush for household of mean income living in an ‘average’ state
- $0.140$ is the effect on the log-odds of a 1-category increase in income
Adding Income as a Predictor

$x_{ij}$ is household annual income (grouped into 9 categories), centred at sample mean of 5.23

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Standard error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$ (constant)</td>
<td>−0.099</td>
<td>0.056</td>
</tr>
<tr>
<td>$\beta_1$ (income, centered)</td>
<td>0.140</td>
<td>0.008</td>
</tr>
<tr>
<td>$\sigma_u^2$ (between-state variance)</td>
<td>0.125</td>
<td>0.030</td>
</tr>
</tbody>
</table>

−0.099 is the log-odds of voting Bush for household of mean income living in an ‘average’ state

0.140 is the effect on the log-odds of a 1-category increase in income

expect odds of voting Bush to be $\exp(8 \times 0.14) = 3.1$ times higher for an individual in the highest income band than for an individual in the same state but in the lowest income band
Prediction Lines by State: Random Intercepts
As in the single-level case, consider a latent continuous variable $y^*$ that underlines observed binary $y$ such that:

$$y_{ij} = \begin{cases} 
1 & \text{if } y_{ij}^* \geq 0 \\
0 & \text{if } y_{ij}^* < 0 
\end{cases}$$
As in the single-level case, consider a latent continuous variable $y^*$ that underlines observed binary $y$ such that:

$$y_{ij} = \begin{cases} 
1 & \text{if } y_{ij}^* \geq 0 \\
0 & \text{if } y_{ij}^* < 0 
\end{cases}$$

Threshold model

$$y_{ij}^* = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}^*$$
As in the single-level case, consider a latent continuous variable $y^*$ that underlies observed binary $y$ such that:

$$y_{ij} = \begin{cases} 
1 & \text{if } y_{ij}^* \geq 0 \\
0 & \text{if } y_{ij}^* < 0
\end{cases}$$

**Threshold model**

$$y_{ij}^* = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}^*$$

As in a single-level model:
- $e_{ij}^* \sim N(0, 1) \rightarrow \text{probit model}$
- $e_{ij}^* \sim \text{standard logistic (with variance } \approx 3.29) \rightarrow \text{logit model}$
As in the single-level case, consider a latent continuous variable $y^*$ that underlines observed binary $y$ such that:

$$y_{ij} = \begin{cases} 
1 & \text{if } y_{ij}^* \geq 0 \\
0 & \text{if } y_{ij}^* < 0 
\end{cases}$$

**Threshold model**

$$y_{ij}^* = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}^*$$

As in a single-level model:

- $e_{ij}^* \sim N(0, 1) \rightarrow \text{probit model}$
- $e_{ij}^* \sim \text{standard logistic (with variance } \approx 3.29) \rightarrow \text{logit model}$

So the level 1 residual variance, $\text{var}(e_{ij}^*)$, is fixed.
Recall single-level logit model expressed as a threshold model:

$$y_i^* = \beta_0 + \beta_1 x_i + e_i^*$$

Adding random effects has increased the residual variance → scale of $y_i^*$ stretched out → $\beta_0$ and $\beta_1$ increase in absolute value.
Recall single-level logit model expressed as a threshold model:

\[ y_i^* = \beta_0 + \beta_1 x_i + e_i^* \]

\[ \text{var}(y_i^* | x_i) = \text{var}(e_i^*) = 3.29 \]
Recall single-level logit model expressed as a threshold model:

\[ y_i^* = \beta_0 + \beta_1 x_i + e_i^* \]

\[ \text{var}(y_i^* | x_i) = \text{var}(e_i^*) = 3.29 \]

Now add random effects:

\[ y_{ij}^* = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}^* \]
Impact of Adding $u_j$ on Coefficients

Recall single-level logit model expressed as a threshold model:

$$y_i^* = \beta_0 + \beta_1 x_i + e_i^*$$

$$\text{var}(y_i^*|x_i) = \text{var}(e_i^*) = 3.29$$

Now add random effects:

$$y_{ij}^* = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}^*$$

$$\text{var}(y_{ij}^*|x_{ij}, u_j) = \text{var}(u_j) + \text{var}(e_{ij}^*) = \sigma_u^2 + 3.29$$

Adding random effects has increased the residual variance → scale of $y$ stretched out → $\beta_0$ and $\beta_1$ increase in absolute value.
Recall single-level logit model expressed as a threshold model:

\[ y_i^* = \beta_0 + \beta_1 x_i + e_i^* \]

\[ \text{var}(y_i^*|x_i) = \text{var}(e_i^*) = 3.29 \]

Now add random effects:

\[ y_{ij}^* = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij}^* \]

\[ \text{var}(y_{ij}^*|x_{ij}, u_j) = \text{var}(u_j) + \text{var}(e_i^*) = \sigma_u^2 + 3.29 \]

Adding random effects has increased the residual variance → scale of \( y^* \) stretched out → \( \beta_0 \) and \( \beta_1 \) increase in absolute value.
\( \beta^{RI} \) coefficient from a random intercept model

\( \beta^{SL} \) coefficient from the corresponding single-level model
Single-level vs Random Intercept Coefficients

$\beta^{RI}$ coefficient from a random intercept model

$\beta^{SL}$ coefficient from the corresponding single-level model

For a logit model

$$\beta^{RI} = \beta^{SL} \sqrt{\frac{\sigma_u^2 + 3.29}{3.29}}$$

Replace 3.29 by 1 to get expression for relationship between probit coefficients.
Single-level vs Random Intercept Coefficients

$\beta^{RI}$ coefficient from a random intercept model

$\beta^{SL}$ coefficient from the corresponding single-level model

For a logit model

$$\beta^{RI} = \beta^{SL} \sqrt{\frac{\sigma^2_u + 3.29}{3.29}}$$

Replace 3.29 by 1 to get expression for relationship between probit coefficients.

NOTE: Adding random effects to a continuous response model does not ‘scale up’ coefficients because the level 1 variance is not fixed and so: $\text{var}(e_i) \approx \text{var}(u_j) + \text{var}(e_{ij})$
Simulated data where distribution of $x_1$ and $x_2$ same in each level 2 unit.

$\hat{\sigma}_u^2 = 1.018$ so expected RI:SL ratio is $\sqrt{1.018 + 3.29} = 1.14$

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\beta_{SL}$</th>
<th>$\beta_{RI}$</th>
<th>$\beta_{RI}/\beta_{SL}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.221</td>
<td>0.257</td>
<td>1.163</td>
</tr>
<tr>
<td>$x_1$</td>
<td>0.430</td>
<td>0.519</td>
<td>1.207</td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.498</td>
<td>0.613</td>
<td>1.231</td>
</tr>
</tbody>
</table>

In practice, RI:SL ratio for a given $x$ may be quite different from that expected if distribution of $x$ differs across level 2 units.
Simulated data where distribution of $x_1$ and $x_2$ same in each level 2 unit.

$\hat{\sigma}_u^2 = 1.018$ so expected RI:SL ratio is $\sqrt{\frac{1.018+3.29}{3.29}} = 1.14$
Simulated data where distribution of $x_1$ and $x_2$ same in each level 2 unit.

$\hat{\sigma}_u^2 = 1.018$ so expected RI:SL ratio is $\sqrt{\frac{1.018+3.29}{3.29}} = 1.14$

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\beta^{SL}$</th>
<th>$\beta^{RI}$</th>
<th>$\beta^{RI} / \beta^{SL}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.221</td>
<td>0.257</td>
<td>1.163</td>
</tr>
<tr>
<td>$x_1$</td>
<td>0.430</td>
<td>0.519</td>
<td>1.207</td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.498</td>
<td>0.613</td>
<td>1.231</td>
</tr>
</tbody>
</table>
Simulated data where distribution of $x_1$ and $x_2$ same in each level 2 unit.

$\hat{\sigma}_u^2 = 1.018$ so expected RI:SL ratio is $\sqrt{\frac{1.018 + 3.29}{3.29}} = 1.14$

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\beta^{SL}$</th>
<th>$\beta^{RI}$</th>
<th>$\beta^{RI}/\beta^{SL}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>0.221</td>
<td>0.257</td>
<td>1.163</td>
</tr>
<tr>
<td>$x_1$</td>
<td>0.430</td>
<td>0.519</td>
<td>1.207</td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.498</td>
<td>0.613</td>
<td>1.231</td>
</tr>
</tbody>
</table>

In practice, RI:SL ratio for a given $x$ may be quite different from that expected if distribution of $x$ differs across level 2 units.
In random effects model for continuous $y$

- Reduction in level 1 residual variance $\sigma_e^2$
Impact of Adding level 1 \(x\)

In random effects model for **continuous** \(y\)

- Reduction in level 1 residual variance \(\sigma_e^2\)
- Reduction in total residual variance \(\sigma_u^2 + \sigma_e^2\)
Impact of Adding level 1 \( x \)

In random effects model for **continuous** \( y \)

- Reduction in level 1 residual variance \( \sigma_e^2 \)
- Reduction in total residual variance \( \sigma_u^2 + \sigma_e^2 \)
- Coefficients of variables correlated with \( x \) will change (increase or decrease)
In random effects model for continuous $y$

- Reduction in level 1 residual variance $\sigma^2_e$
- Reduction in total residual variance $\sigma^2_u + \sigma^2_e$
- Coefficients of variables correlated with $x$ will change (increase or decrease)
In random effects model for continuous $y$

- Reduction in level 1 residual variance $\sigma^2_e$
- Reduction in total residual variance $\sigma^2_u + \sigma^2_e$
- Coefficients of variables correlated with $x$ will change (increase or decrease)

In random effects model for binary $y$

- Level 1 residual variance $\sigma^2_{e^*}$ cannot change; fixed at 3.29 (logit) or 1 (probit)
In random effects model for **continuous** \( y \)
- Reduction in level 1 residual variance \( \sigma_e^2 \)
- Reduction in total residual variance \( \sigma_u^2 + \sigma_e^2 \)
- Coefficients of variables correlated with \( x \) will change (increase or decrease)

In random effects model for **binary** \( y \)
- Level 1 residual variance \( \sigma_e^2 \) cannot change; fixed at 3.29 (logit) or 1 (probit)
- So addition of \( x \) will tend to increase proportion of variance that is at level 2, i.e. ratio of \( \sigma_u^2 \) to \( \sigma_e^2 \)
Impact of Adding level 1 $x$

In random effects model for continuous $y$

- Reduction in level 1 residual variance $\sigma_e^2$
- Reduction in total residual variance $\sigma_u^2 + \sigma_e^2$
- Coefficients of variables correlated with $x$ will change (increase or decrease)

In random effects model for binary $y$

- Level 1 residual variance $\sigma_{e*}^2$ cannot change; fixed at 3.29 (logit) or 1 (probit)
- So addition of $x$ will tend to increase proportion of variance that is at level 2, i.e. ratio of $\sigma_u^2$ to $\sigma_{e*}^2$
- $\rightarrow$ increase in level 2 residual variance $\rightarrow$ stretches scale of $y^*$
Impact of Adding level 1 $x$

In random effects model for continuous $y$
- Reduction in level 1 residual variance $\sigma^2_e$
- Reduction in total residual variance $\sigma^2_u + \sigma^2_e$
- Coefficients of variables correlated with $x$ will change (increase or decrease)

In random effects model for binary $y$
- Level 1 residual variance $\sigma^2_{e*}$ cannot change; fixed at 3.29 (logit) or 1 (probit)
- So addition of $x$ will tend to increase proportion of variance that is at level 2, i.e. ratio of $\sigma^2_u$ to $\sigma^2_{e*}$
- → increase in level 2 residual variance → stretches scale of $y^*$
- → increase in absolute value of coefficients of other variables
Variance Partition Coefficient for Binary $y$

Usual formula is:

$$VPC = \frac{\text{level 2 residual variance}}{\text{level 1 residual variance} + \text{level 2 residual variance}}$$

From threshold model for latent $y^*$, we obtain

$$VPC = \sigma^2_u \sigma^2_e^* + \sigma^2_u$$

where $\sigma^2_e^* = 1$ for probit model and 3.29 for logit model.

In voting intentions example, $\hat{\sigma}^2_u = 0.125$, so $VPC = 0.037$. Adjusting for income, 4% of the remaining variance in the propensity to vote Bush is attributable to between-state variation.
Variance Partition Coefficient for Binary $y$

Usual formula is:

$$VPC = \frac{\text{level 2 residual variance}}{\text{level 1 residual variance} + \text{level 2 residual variance}}$$

From threshold model for latent $y^*$, we obtain

$$VPC = \frac{\sigma_u^2}{\sigma_{e^*}^2 + \sigma_u^2}$$

where $\sigma_{e^*}^2 = 1$ for probit model and 3.29 for logit model.
Usual formula is:

\[
VPC = \frac{\text{level 2 residual variance}}{\text{level 1 residual variance} + \text{level 2 residual variance}}
\]

From threshold model for latent \( y^* \), we obtain

\[
VPC = \frac{\sigma_u^2}{\sigma_{e^*}^2 + \sigma_u^2}
\]

where \( \sigma_{e^*}^2 = 1 \) for probit model and 3.29 for logit model.

In voting intentions example, \( \hat{\sigma}_u^2 = 0.125 \), so VPC = 0.037. Adjusting for income, 4% of the remaining variance in the propensity to vote Bush is attributable to between-state variation.
Marginal Model for Clustered $y$

When $y$ are clustered, an alternative to a random effects model is a **marginal model**.
When $y$ are clustered, an alternative to a random effects model is a **marginal model**.

A marginal model has two components

1. Generalised linear model specifying relationship between $\pi_{ij}$ and $x_{ij}$
Marginal Model for Clustered $y$

When $y$ are clustered, an alternative to a random effects model is a **marginal model**.

A marginal model has two components

1. Generalised linear model specifying relationship between $\pi_{ij}$ and $x_{ij}$
2. Specification of structure of correlations between pairs of individuals in the same group
Marginal Model for Clustered $y$

When $y$ are clustered, an alternative to a random effects model is a marginal model.

A marginal model has two components

1. Generalised linear model specifying relationship between $\pi_{ij}$ and $x_{ij}$
2. Specification of structure of correlations between pairs of individuals in the same group
   - Exchangeable - equal correlation between every pair (as in random intercept model)
When \( y \) are clustered, an alternative to a random effects model is a **marginal model**.

A marginal model has two components:

1. Generalised linear model specifying relationship between \( \pi_{ij} \) and \( x_{ij} \)
2. Specification of structure of correlations between pairs of individuals in the same group
   - **Exchangeable** - equal correlation between every pair (as in random intercept model)
   - **Autocorrelation** - used for longitudinal data where correlation a function of time between measures

Estimated using Generalised Estimating Equations (GEE)
Marginal Model for Clustered $y$

When $y$ are clustered, an alternative to a random effects model is a **marginal model**.

A marginal model has two components

1. Generalised linear model specifying relationship between $\pi_{ij}$ and $x_{ij}$
2. Specification of structure of correlations between pairs of individuals in the same group
   - **Exchangeable** - equal correlation between every pair (as in random intercept model)
   - **Autocorrelation** - used for longitudinal data where correlation a function of time between measures
   - **Unstructured** - all pairwise correlations estimated

Estimated using Generalised Estimating Equations (GEE)
When \( y \) are clustered, an alternative to a random effects model is a marginal model.

A marginal model has two components

1. Generalised linear model specifying relationship between \( \pi_{ij} \) and \( x_{ij} \)
2. Specification of structure of correlations between pairs of individuals in the same group
   - Exchangeable - equal correlation between every pair (as in random intercept model)
   - Autocorrelation - used for longitudinal data where correlation a function of time between measures
   - Unstructured - all pairwise correlations estimated

Estimated using Generalised Estimating Equations (GEE)
Marginal Model for Clustered $y$

When $y$ are clustered, an alternative to a random effects model is a **marginal model**.

A marginal model has two components

1. Generalised linear model specifying relationship between $\pi_{ij}$ and $x_{ij}$
2. Specification of structure of correlations between pairs of individuals in the same group
   - **Exchangeable** - equal correlation between every pair (as in random intercept model)
   - **Autocorrelation** - used for longitudinal data where correlation a function of time between measures
   - **Unstructured** - all pairwise correlations estimated

Estimated using Generalised Estimating Equations (GEE)
Marginal vs Random Effects Approaches

**Marginal**
- Accounts for clustering and adjusts standard errors
Marginal
- Accounts for clustering and adjusts standard errors
- Clustering regarded as a nuisance
**Marginal vs Random Effects Approaches**

**Marginal**
- Accounts for clustering and adjusts standard errors
- Clustering regarded as a nuisance
- No parameter representing between-group variance
Marginal vs Random Effects Approaches

Marginal

- Accounts for clustering and adjusts standard errors
- Clustering regarded as a nuisance
- No parameter representing between-group variance
- No distributional assumptions about group effects, but no estimates of group effects either
Marginal vs Random Effects Approaches

**Marginal**
- Accounts for clustering and adjusts standard errors
- Clustering regarded as a nuisance
- No parameter representing between-group variance
- No distributional assumptions about group effects, but no estimates of group effects either

**Random effects**
- Accounts for clustering and adjusts standard errors
- Clustering of substantive interest
- Estimate between-group variance $\sigma_u^2$
- Level 2 residuals $\hat{u}_j$ interpreted as group effects
- Can allow between-group variance to depend on $x$
Marginal vs Random Effects Approaches

**Marginal**
- Accounts for clustering and adjusts standard errors
- Clustering regarded as a nuisance
- No parameter representing between-group variance
- No distributional assumptions about group effects, but no estimates of group effects either

**Random effects**
- Accounts for clustering and adjusts standard errors
Marginal vs Random Effects Approaches

Marginal
- Accounts for clustering and adjusts standard errors
- Clustering regarded as a nuisance
- No parameter representing between-group variance
- No distributional assumptions about group effects, but no estimates of group effects either

Random effects
- Accounts for clustering and adjusts standard errors
- Clustering of substantive interest
Marginal vs Random Effects Approaches

**Marginal**
- Accounts for clustering and adjusts standard errors
- Clustering regarded as a nuisance
- No parameter representing between-group variance
- No distributional assumptions about group effects, but no estimates of group effects either

**Random effects**
- Accounts for clustering and adjusts standard errors
- Clustering of substantive interest
- Estimate between-group variance $\sigma_u^2$
Marginal vs Random Effects Approaches

**Marginal**
- Accounts for clustering and adjusts standard errors
- Clustering regarded as a nuisance
- No parameter representing between-group variance
- No distributional assumptions about group effects, but no estimates of group effects either

**Random effects**
- Accounts for clustering and adjusts standard errors
- Clustering of substantive interest
- Estimate between-group variance $\sigma^2_u$
- Level 2 residuals $\hat{u}_j$ interpreted as group effects
Marginal vs Random Effects Approaches

**Marginal**
- Accounts for clustering and adjusts standard errors
- Clustering regarded as a nuisance
- No parameter representing between-group variance
- No distributional assumptions about group effects, but no estimates of group effects either

**Random effects**
- Accounts for clustering and adjusts standard errors
- Clustering of substantive interest
- Estimate between-group variance $\sigma_u^2$
- Level 2 residuals $\hat{u}_j$ interpreted as group effects
- Can allow between-group variance to depend on $x$
Marginal and Random Effects Models

- Marginal $\beta$ have a population-averaged (PA) interpretation
- Random effects $\beta$ have a cluster-specific (CS) interpretation
Marginal and Random Effects Models

- Marginal $\beta$ have a population-averaged (PA) interpretation
- Random effects $\beta$ have a cluster-specific (CS) interpretation

Random intercept logit model

$$\text{logit}(\pi_{ij}) = \beta_{0}^{CS} + \beta_{1}^{CS} x_{ij} + u_{j}$$

where $u_{j} \sim N(0, \sigma_{u}^{2})$
Marginal and Random Effects Models

- Marginal $\beta$ have a population-averaged (PA) interpretation
- Random effects $\beta$ have a cluster-specific (CS) interpretation

Random intercept logit model

$$\text{logit}(\pi_{ij}) = \beta_0^{CS} + \beta_1^{CS} x_{ij} + u_j$$

where $u_j \sim N(0, \sigma_u^2)$

Marginal logit model

$$\text{logit}(\pi_{ij}) = \beta_0^{PA} + \beta_1^{PA} x_{ij}$$

Plus specification of structure of within-cluster covariance matrix
Interpretation of CS and PA Effects

Cluster-specific

\[ \beta_{1}^{CS} \] is the effect of a 1-unit change in \( x \) on the log-odds that \( y = 1 \) for a given cluster, i.e. holding constant (or conditioning on) cluster-specific unobservables.
Interpretation of CS and PA Effects

Cluster-specific

- $\beta_{1}^{CS}$ is the effect of a 1-unit change in $x$ on the log-odds that $y = 1$ for a given cluster, i.e. holding constant (or conditioning on) cluster-specific unobservables.
- $\beta_{1}^{CS}$ contrasts two individuals in the same cluster with $x$-values 1 unit apart.
Cluster-specific

\( \beta_{1}^{CS} \) is the effect of a 1-unit change in \( x \) on the log-odds that \( y = 1 \) for a given cluster, i.e. holding constant (or conditioning on) cluster-specific unobservables.

\( \beta_{1}^{CS} \) contrasts two individuals in the same cluster with \( x \)-values 1 unit apart.
Interpretation of CS and PA Effects

Cluster-specific

- $\beta_{1}^{CS}$ is the effect of a 1-unit change in $x$ on the log-odds that $y = 1$ for a given cluster, i.e. holding constant (or conditioning on) cluster-specific unobservables.
- $\beta_{1}^{CS}$ contrasts two individuals in the same cluster with $x$-values 1 unit apart.

Population-averaged

- $\beta_{1}^{PA}$ is the effect of a 1-unit change in $x$ on the log-odds that $y = 1$ in the study population, i.e. averaging over cluster-specific unobservables.
Consider a longitudinal study designed to assess cancer patients’ tolerance to different doses of chemotherapy.

$y_{ij}$ indicates whether patient $j$ has an adverse reaction at occasion $i$ to (time-varying) dose $x_{ij}$. 
Consider a longitudinal study designed to assess cancer patients’ tolerance to different doses of chemotherapy.

\( y_{ij} \) indicates whether patient \( j \) has an adverse reaction at occasion \( i \) to (time-varying) dose \( x_{ij} \).

\( \beta_{CS} \) is effect of 1-unit increase in dose, holding constant time-invariant unobserved individual characteristics represented by \( u_j \). If patients are assigned to different doses at random, could be interpreted as a causal effect.
Consider a longitudinal study designed to assess cancer patients’ tolerance to different doses of chemotherapy.

\( y_{ij} \) indicates whether patient \( j \) has an adverse reaction at occasion \( i \) to (time-varying) dose \( x_{ij} \).

\( \beta_1^{CS} \) is effect of 1-unit increase in dose, holding constant time-invariant unobserved individual characteristics represented by \( u_j \). If patients are assigned to different doses at random, could be interpreted as a causal effect.

\( \beta_1^{PA} \) compares individuals whose dosage \( x_{ij} \) differs by 1 unit, averaging over between-individual differences in tolerance.
Suppose we add a level 2 variable, gender ($x_{2j}$), with coefficient $\beta_2$. 
Example: PA vs. CS Interpretation (2)

Suppose we add a level 2 variable, gender \((x_{2j})\), with coefficient \(\beta_2\).

- Because \(x_{2j}\) is fixed over time, we cannot interpret \(\beta_2^{CS}\) as a within-person effect. Instead \(\beta_2^{CS}\) compares men and women with the same value of \(x_{ij}\) and \(u_j\), i.e. the same dose and the same combination of unobserved time-invariant characteristics.

\[ \beta_2^{PA} \] compares men and women receiving the same dose \(x_{ij}\), averaging over individual unobservables.

\[ \beta_2^{PA} \] may be of more interest.
Example: PA vs. CS Interpretation (2)

Suppose we add a level 2 variable, gender ($x_{2j}$), with coefficient $\beta_2$.

- Because $x_{2j}$ is fixed over time, we cannot interpret $\beta_2^{CS}$ as a within-person effect. Instead $\beta_2^{CS}$ compares men and women with the same value of $x_{ij}$ and $u_j$, i.e. the same dose and the same combination of unobserved time-invariant characteristics.

- $\beta_2^{PA}$ compares men and women receiving the same dose $x_{ij}$, averaging over individual unobservables.
Suppose we add a level 2 variable, gender \((x_{2j})\), with coefficient \(\beta_2\).

- Because \(x_{2j}\) is fixed over time, we cannot interpret \(\beta_{CS}^2\) as a within-person effect. Instead \(\beta_{CS}^2\) compares men and women with the same value of \(x_{ij}\) and \(u_j\), i.e. the same dose and the same combination of unobserved time-invariant characteristics.

- \(\beta_{PA}^2\) compares men and women receiving the same dose \(x_{ij}\), averaging over individual unobservables.

For a level 2 variable, \(\beta_{PA}^2\) may be of more interest.
In general $|\hat{\beta}^{CS}| > |\hat{\beta}^{PA}|$
In general $|\hat{\beta}^{CS}| > |\hat{\beta}^{PA}|$

The relationship between the CS and PA logit coefficients for a variable $x$ is approximately:

$$\beta^{CS} = \sqrt{\frac{\sigma_u^2 + 3.29}{3.29}} \beta^{PA}$$
In general $|\hat{\beta}^{CS}| > |\hat{\beta}^{PA}|$

The relationship between the CS and PA logit coefficients for a variable $x$ is approximately:

$$\beta^{CS} = \sqrt{\frac{\sigma_u^2 + 3.29}{3.29}} \beta^{PA}$$

When there is no clustering, $\sigma_u^2 = 0$ and $\beta^{CS} = \beta^{PA}$. Coefficients move further apart as $\sigma_u^2$ increases.
Comparison of PA and CS Coefficients

- In general $|\hat{\beta}^{CS}| > |\hat{\beta}^{PA}|$
- The relationship between the CS and PA logit coefficients for a variable $x$ is approximately:

$$\beta^{CS} = \sqrt{\frac{\sigma_u^2 + 3.29}{3.29}} \beta^{PA}$$

- When there is no clustering, $\sigma_u^2 = 0$ and $\beta^{CS} = \beta^{PA}$. Coefficients move further apart as $\sigma_u^2$ increases
- Note that marginal models can also be specified for continuous $y$, but in that case CS and PA coefficients are equal
Response probability for individual $i$ in group $j$ calculated as

$$
\pi_{ij} = \frac{\exp(\beta_0 + \beta_1 x_{ij} + u_j)}{1 + \exp(\beta_0 + \beta_1 x_{ij} + u_j)}
$$

where we substitute estimates of $\beta_0$, $\beta_1$ and $u_j$ to get predicted probabilities.
Response probability for individual $i$ in group $j$ calculated as

$$
\pi_{ij} = \frac{\exp(\beta_0 + \beta_1 x_{ij} + u_j)}{1 + \exp(\beta_0 + \beta_1 x_{ij} + u_j)}
$$

where we substitute estimates of $\beta_0$, $\beta_1$ and $u_j$ to get predicted probabilities.

Rather than calculating probabilities for each individual, however, we often want predictions for specific values of $x$. But what do we substitute for $u_j$?
Suppose we want predictions for $x = x^*$. What do we do about $u$?

1. Substitute the mean $u_j = 0$. But predictions are not the mean response probabilities for $x = x^*$ because $\pi$ is a nonlinear function of $u_j$. Value of $\pi$ at mean of $u_j \neq$ mean of $\pi$.

2. Integrate out $u_j$ to obtain an expression for mean $\pi$ that does not involve $u$. Leads to probabilities that have a PA interpretation, but requires some approximation.

3. Average over simulated values of $u_j$. Also gives PA probabilities, but easier to implement. Now available in MLwiN.
Suppose we want predictions for $x = x^*$. What do we do about $u$?

1. **Substitute the mean $u_j = 0$.** But predictions are not the mean response probabilities for $x = x^*$ because $\pi$ is a nonlinear function of $u_j$. Value of $\pi$ at mean of $u_j \neq$ mean of $\pi$. 
Suppose we want predictions for $x = x^*$. What do we do about $u$?

1. **Substitute the mean $u_j = 0$.** But predictions are not the mean response probabilities for $x = x^*$ because $\pi$ is a nonlinear function of $u_j$. Value of $\pi$ at mean of $u_j \neq$ mean of $\pi$.

2. **Integrate out $u_j$** to obtain an expression for mean $\pi$ that does not involve $u$. Leads to probabilities that have a PA interpretation, but requires some approximation.
Suppose we want predictions for $x = x^*$. What do we do about $u$?

1. **Substitute the mean $u_j = 0$.** But predictions are not the mean response probabilities for $x = x^*$ because $\pi$ is a nonlinear function of $u_j$. Value of $\pi$ at mean of $u_j \neq$ mean of $\pi$.

2. **Integrate out $u_j$** to obtain an expression for mean $\pi$ that does not involve $u$. Leads to probabilities that have a PA interpretation, but requires some approximation.

3. **Average over simulated values of $u_j$.** Also gives PA probabilities, but easier to implement. Now available in MLwiN.
1. Generate $M$ values for random effect $u$ from $N(0, \hat{\sigma}_u^2)$, and denote generated values by $u^{(1)}, u^{(2)}, \ldots, u^{(M)}$.
1. Generate $M$ values for random effect $u$ from $N(0, \hat{\sigma}^2_u)$, and denote generated values by $u^{(1)}, u^{(2)}, \ldots, u^{(M)}$.

2. For each simulated value ($m = 1, \ldots, M$) compute, for given $x$,

$$\pi^{(m)} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x + u^{(m)})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x + u^{(m)})}$$
Predictions via Simulation

1. Generate $M$ values for random effect $u$ from $N(0, \sigma_u^2)$, and denote generated values by $u^{(1)}, u^{(2)}, \ldots, u^{(M)}$

2. For each simulated value ($m = 1, \ldots, M$) compute, for given $x$,

$$
\pi^{(m)} = \frac{\exp(\beta_0 + \beta_1 x + u^{(m)})}{1 + \exp(\beta_0 + \beta_1 x + u^{(m)})}
$$

3. Calculate the mean of $\pi^{(m)}$:

$$
\pi = \frac{1}{M} \sum_{m=1}^{M} \pi^{(m)}
$$
Predictions via Simulation

1. Generate $M$ values for random effect $u$ from $N(0, \hat{\sigma}_u^2)$, and denote generated values by $u^{(1)}, u^{(2)}, \ldots, u^{(M)}$

2. For each simulated value ($m = 1, \ldots, M$) compute, for given $x$,

$$\pi^{(m)} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x + u^{(m)})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x + u^{(m)})}$$

3. Calculate the mean of $\pi^{(m)}$:

$$\pi = \frac{1}{M} \sum_{m=1}^{M} \pi^{(m)}$$

4. Repeat 1-3 for different value of $x$
## Predicted Probabilities for Voting Bush

<table>
<thead>
<tr>
<th></th>
<th>Random intercept model</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Method 1</td>
<td>Method 3</td>
<td>Marginal</td>
<td></td>
</tr>
<tr>
<td><strong>Household income</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>0.374</td>
<td>0.378</td>
<td>0.377</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>0.444</td>
<td>0.446</td>
<td>0.445</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>0.564</td>
<td>0.564</td>
<td>0.562</td>
<td></td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>0.510</td>
<td>0.510</td>
<td>0.510</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>0.442</td>
<td>0.444</td>
<td>0.444</td>
<td></td>
</tr>
</tbody>
</table>

In this case, \( \hat{\pi} \) from Methods 1 and 3 are very similar. This is because (i) predictions are all close to 0.5, and (ii) \( \hat{\sigma}^2_u \) is small, so that \( \hat{\beta}_{CS} \) is close to \( \hat{\beta}_{PA} \). In longitudinal applications, where \( \hat{\sigma}^2_u \) can be large, there will be bigger differences between Methods 1 and 3.
<table>
<thead>
<tr>
<th></th>
<th>Random intercept model</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Method 1</td>
<td>Method 3</td>
<td>Marginal model</td>
</tr>
<tr>
<td>Household income</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>0.374</td>
<td>0.378</td>
<td>0.377</td>
</tr>
<tr>
<td>Medium</td>
<td>0.444</td>
<td>0.446</td>
<td>0.445</td>
</tr>
<tr>
<td>High</td>
<td>0.564</td>
<td>0.564</td>
<td>0.562</td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>0.510</td>
<td>0.510</td>
<td>0.510</td>
</tr>
<tr>
<td>Female</td>
<td>0.442</td>
<td>0.444</td>
<td>0.444</td>
</tr>
</tbody>
</table>

In this case, $\hat{\pi}$ from Methods 1 and 3 are very similar. This is because (i) predictions are all close to 0.5, and (ii) $\hat{\sigma}_u^2$ is small, so that $\beta^{CS}$ is close to $\beta^{PA}$. 
### Predicted Probabilities for Voting Bush

<table>
<thead>
<tr>
<th></th>
<th>Random intercept model</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Method 1</td>
<td>Method 3</td>
<td>Marginal model</td>
<td></td>
</tr>
<tr>
<td>Household income</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>0.374</td>
<td>0.378</td>
<td>0.377</td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>0.444</td>
<td>0.446</td>
<td>0.445</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>0.564</td>
<td>0.564</td>
<td>0.562</td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>0.510</td>
<td>0.510</td>
<td>0.510</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>0.442</td>
<td>0.444</td>
<td>0.444</td>
<td></td>
</tr>
</tbody>
</table>

- In this case, $\hat{\pi}$ from Methods 1 and 3 are very similar. This is because (i) predictions are all close to 0.5, and (ii) $\hat{\sigma}_u^2$ is small, so that $\beta^{CS}$ is close to $\beta^{PA}$
- In longitudinal applications, where $\hat{\sigma}_u^2$ can be large, there will be bigger differences between Methods 1 and 3
So far we have allowed \( \pi_{ij} \) to vary from group to group by including a group-level random component in the intercept: \( \beta_{0j} = \beta_0 + u_{0j} \).

BUT we have assumed the effect of any predictor \( x \) is the same in each group. We now consider a random slope model in which the slope of \( x \) (\( \beta_1 \)) is replaced by \( \beta_{1j} = \beta_1 + u_{1j} \).
So far we have allowed $\pi_{ij}$ to vary from group to group by including a group-level random component in the intercept: $\beta_{0j} = \beta_0 + u_{0j}$.

BUT we have assumed the effect of any predictor $x$ is the same in each group. We now consider a random slope model in which the slope of $x$ ($\beta_1$) is replaced by $\beta_{1j} = \beta_1 + u_{1j}$.

$$\log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \beta_0 + \beta_1 x_{ij} + u_{0j} + u_{1j} x_{ij}$$

where $(u_{0j}, u_{1j})$ follow a bivariate normal distribution:

$$u_{0j} \sim N(0, \sigma_{u0}^2), \quad u_{1j} \sim N(0, \sigma_{u1}^2), \quad \text{cov}(u_{0j}, u_{1j}) = \sigma_{u01}$$
Example: Random Slope for Income

Extend random intercept logit model for relationship between probability of voting Bush and household income to allow income effect to vary across states.
Example: Random Slope for Income

Extend random intercept logit model for relationship between probability of voting Bush and household income to allow income effect to vary across states.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Random int.</th>
<th></th>
<th>Random slope</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est.</td>
<td>se</td>
<td>Est.</td>
<td>se</td>
</tr>
<tr>
<td>$\beta_0$ (constant)</td>
<td>$-0.099$</td>
<td>$0.056$</td>
<td>$-0.087$</td>
<td>$0.057$</td>
</tr>
<tr>
<td>$\beta_1$ (Income, centred)</td>
<td>$0.140$</td>
<td>$0.008$</td>
<td>$0.145$</td>
<td>$0.013$</td>
</tr>
</tbody>
</table>

**State-level random part**

<table>
<thead>
<tr>
<th></th>
<th>Random int.</th>
<th></th>
<th>Random slope</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est.</td>
<td>se</td>
<td>Est.</td>
<td>se</td>
</tr>
<tr>
<td>$\sigma^2_{u_0}$ (intercept variance)</td>
<td>$0.125$</td>
<td>$0.031$</td>
<td>$0.132$</td>
<td>$0.032$</td>
</tr>
<tr>
<td>$\sigma^2_{u_1}$ (slope variance)</td>
<td>-</td>
<td>-</td>
<td>$0.003$</td>
<td>$0.001$</td>
</tr>
<tr>
<td>$\sigma_{u01}$ (intercept-slope covariance)</td>
<td>-</td>
<td>-</td>
<td>$0.018$</td>
<td>$0.006$</td>
</tr>
</tbody>
</table>
Allowing $x$ to have a random slope introduces 2 new parameters: $\sigma^2_{u1}$ and $\sigma_{u01}$. 

Testing $H_0$:

Test $H_0: \sigma^2_{u1} = \sigma_{u01} = 0$ using a likelihood ratio test or (approximate) Wald test on 2 d.f.

For the income example, Wald = 9.72. Comparing with $\chi^2_2$ gives a two-sided $p$-value of 0.0008 $\Rightarrow$ income effect does vary across states.
Allowing $x$ to have a random slope introduces 2 new parameters: $\sigma_{u1}^2$ and $\sigma_{u01}$.

Test $H_0: \sigma_{u1}^2 = \sigma_{u01} = 0$ using a likelihood ratio test or (approximate) Wald test on 2 d.f.
Allowing $x$ to have a random slope introduces 2 new parameters: $\sigma_{u1}^2$ and $\sigma_{u01}$.

Test $H_0 : \sigma_{u1}^2 = \sigma_{u01} = 0$ using a likelihood ratio test or (approximate) Wald test on 2 d.f.

For the income example, Wald = 9.72. Comparing with $\chi^2_2$ gives a two-sided p-value of 0.0008

$\implies$ income effect does vary across states.
Prediction Lines by State: Random Slopes
Intercept vs. Income Slope Residuals

Bottom left: Washington DC
Top right: Montana and Utah
In a random slope model, the between-group variance is a function of the variable(s) with a random coefficient $x$:

$$\text{var}(u_{0j} + u_{1j}x_{ij}) = \text{var}(u_{0j}) + 2x_{ij}\text{cov}(u_{0j}, u_{1j}) + x_{ij}^2\text{var}(u_{1j})$$

$$= \sigma_{u0}^2 + 2\sigma_{u01}x_{ij} + \sigma_{u1}^2x_{ij}^2$$
In a random slope model, the between-group variance is a function of the variable(s) with a random coefficient \( x \):

\[
\text{var}(u_{0j} + u_{1j}x_{ij}) = \text{var}(u_{0j}) + 2x_{ij}\text{cov}(u_{0j}, u_{1j}) + x_{ij}^2\text{var}(u_{1j})
\]

\[
= \sigma_{u0}^2 + 2\sigma_{u01}x_{ij} + \sigma_{u1}^2x_{ij}^2
\]

Between-state variance in log-odds of voting Bush

\[0.132 + 0.036 \textbf{Income} + 0.003 \textbf{Income}^2\]
A major advantage of the multilevel approach is the ability to explore effects of group-level (level 2) predictors, while accounting for the effects of unobserved group characteristics.
A major advantage of the multilevel approach is the ability to explore effects of group-level (level 2) predictors, while accounting for the effects of unobserved group characteristics.

A random intercept logit model with a level 1 variable $x_{1ij}$ and a level 2 variable $x_{2j}$ is:

$$\log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} + u_j$$

$\beta_2$ is the contextual effect of $x_{2j}$. 
A major advantage of the multilevel approach is the ability to explore effects of group-level (level 2) predictors, while accounting for the effects of unobserved group characteristics.

A random intercept logit model with a level 1 variable $x_{1ij}$ and a level 2 variable $x_{2j}$ is:

$$\log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} + u_j$$

$\beta_2$ is the contextual effect of $x_{2j}$.

Especially important to use a multilevel model if interested in contextual effects as $\text{se}(\hat{\beta}_2)$ may be severely estimated if a single-level model is used.
Individual religiosity measured by dummy variable for frequency of attendance at religious services (1=weekly or more, 0=other)

State religiosity is proportion of respondents in state who attend a service weekly or more.
Individual and Contextual Effects of Religiosity

Individual religiosity measured by dummy variable for frequency of attendance at religious services (1 = weekly or more, 0 = other)

State religiosity is proportion of respondents in state who attend a service weekly or more.

<table>
<thead>
<tr>
<th>Variable</th>
<th>No contextual effect</th>
<th>Contextual effect</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est.</td>
<td>se</td>
</tr>
<tr>
<td>Individual religiosity</td>
<td>0.556</td>
<td>0.037</td>
</tr>
<tr>
<td>State religiosity</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Between-state variance</td>
<td>0.083</td>
<td>0.022</td>
</tr>
</tbody>
</table>

(Model also includes age, sex, income and marital status.)
Suppose we believe that the effect of an individual characteristic on $\pi_{ij}$ depends on the value of a group characteristic.
Suppose we believe that the effect of an individual characteristic on $\pi_{ij}$ depends on the value of a group characteristic.

We can extend the contextual effects model to allow the effect of $x_{1ij}$ to depend on $x_{2j}$ by including a cross-level interaction:

$$\log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} + \beta_3 x_{1ij} x_{2j} + u_j$$
Suppose we believe that the effect of an individual characteristic on $\pi_{ij}$ depends on the value of a group characteristic.

We can extend the contextual effects model to allow the effect of $x_{1ij}$ to depend on $x_{2j}$ by including a cross-level interaction:

$$\log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2j} + \beta_3 x_{1ij} x_{2j} + u_j$$

The null hypothesis for a test of a cross-level interaction is $H_0 : \beta_3 = 0$. 
Example of Cross-Level Interaction

Suppose we believe that the effect of individual age on the probability of voting Bush might depend on the conservatism of their state of residence, which we measure by state religiosity.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Est.</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.012</td>
<td>0.005</td>
</tr>
<tr>
<td>State prop. attending religious services weekly</td>
<td>4.206</td>
<td>0.716</td>
</tr>
<tr>
<td>Age × State religiosity</td>
<td>-0.043</td>
<td>0.013</td>
</tr>
</tbody>
</table>

Z-ratio for interaction coefficient is $-0.043/0.013 = 3.31$ which is highly significant $\Rightarrow$ effect of age depends on state religiosity.
Suppose we believe that the effect of individual age on the probability of voting Bush might depend on the conservatism of their state of residence, which we measure by state religiosity.

---

**Selected coefficients from interaction model**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Est.</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.012</td>
<td>0.005</td>
</tr>
<tr>
<td>State prop. attending religious services weekly</td>
<td>4.206</td>
<td>0.716</td>
</tr>
<tr>
<td>Age × State religiosity</td>
<td>−0.043</td>
<td>0.013</td>
</tr>
</tbody>
</table>
Example of Cross-Level Interaction

Suppose we believe that the effect of individual age on the probability of voting Bush might depend on the conservatism of their state of residence, which we measure by state religiosity.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Est.</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>0.012</td>
<td>0.005</td>
</tr>
<tr>
<td>State prop. attending religious services weekly</td>
<td>4.206</td>
<td>0.716</td>
</tr>
<tr>
<td>Age $\times$ State religiosity</td>
<td>$-0.043$</td>
<td>0.013</td>
</tr>
</tbody>
</table>

Z-ratio for interaction coefficient is $0.043/0.013 = 3.31$ which is highly significant $\Rightarrow$ effect of age depends on state religiosity.
Age effects on log-odds of voting Bush

<table>
<thead>
<tr>
<th>Proportion attending services weekly</th>
<th>Age Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.16</td>
<td>$0.012 - (0.043 \times 0.16) = 0.005$</td>
</tr>
<tr>
<td>0.30</td>
<td>$0.012 - (0.043 \times 0.30) = -0.0009$</td>
</tr>
<tr>
<td>0.64</td>
<td>$0.012 - (0.043 \times 0.64) = -0.016$</td>
</tr>
</tbody>
</table>

So age effect is weakly positive for the least religious states, and becomes less strongly positive and then more strongly negative as state-level religiosity increases.

Difference between young and old respondents in voting intentions is greatest in most religious states.
Effect of Age by State Religiosity

Age effects on log-odds of voting Bush

<table>
<thead>
<tr>
<th>Proportion attending services weekly</th>
<th>Age Effect</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.16</td>
<td>$0.012 - (0.043 \times 0.16) = 0.005$</td>
</tr>
<tr>
<td>0.30</td>
<td>$0.012 - (0.043 \times 0.30) = -0.0009$</td>
</tr>
<tr>
<td>0.64</td>
<td>$0.012 - (0.043 \times 0.64) = -0.016$</td>
</tr>
</tbody>
</table>

So age effect is weakly positive for the least religious states, and becomes less strongly positive and then more strongly negative as state-level religiosity increases.
So age effect is weakly positive for the least religious states, and becomes less strongly positive and then more strongly negative as state-level religiosity increases.

⇒ Difference between young and old respondents in voting intentions is greatest in most religious states.
Multilevel models for continuous responses are usually estimated via maximum likelihood (ML)
Multilevel models for continuous responses are usually estimated via maximum likelihood (ML).

For binary (and other discrete) responses, there is a range of options:
A Brief Overview of Estimation Procedures

- Multilevel models for continuous responses are usually estimated via maximum likelihood (ML).
- For binary (and other discrete) responses, there is a range of options:
  - Direct ML via numerical quadrature (software includes SAS, Stata, MIXOR, aML).
Multilevel models for continuous responses are usually estimated via maximum likelihood (ML).

For binary (and other discrete) responses, there is a range of options:

- Direct ML via **numerical quadrature** (software includes SAS, Stata, MIXOR, aML)
- Quasi-likelihood (MLwiN, HLM)
Multilevel models for continuous responses are usually estimated via maximum likelihood (ML). For binary (and other discrete) responses, there is a range of options:

- Direct ML via numerical quadrature (software includes SAS, Stata, MIXOR, aML)
- Quasi-likelihood (MLwiN, HLM)
- Markov chain Monte Carlo (MCMC) methods (WinBUGS, MLwiN)
A Brief Overview of Estimation Procedures

- Multilevel models for continuous responses are usually estimated via maximum likelihood (ML).
- For binary (and other discrete) responses, there is a range of options:
  - Direct ML via numerical quadrature (software includes SAS, Stata, MIXOR, aML)
  - Quasi-likelihood (MLwiN, HLM)
  - Markov chain Monte Carlo (MCMC) methods (WinBUGS, MLwiN)
- In some situations, different procedures can lead to quite different results.
Rodríguez and Goldman (2001, *J. Roy. Stat. Soc.*) simulated a 3-level data structure with 2449 births (level 1) from 1558 mothers (level 2) in 161 communities (level 3), and one predictor at each level.
Rodríguez and Goldman (2001, *J. Roy. Stat. Soc.*.) simulated a 3-level data structure with 2449 births (level 1) from 1558 mothers (level 2) in 161 communities (level 3), and one predictor at each level.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>True value</th>
<th>MQL1</th>
<th>MQL2</th>
<th>PQL2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Child-level $x$</td>
<td>1</td>
<td>0.74</td>
<td>0.85</td>
<td>0.96</td>
</tr>
<tr>
<td>Family-level $x$</td>
<td>1</td>
<td>0.74</td>
<td>0.86</td>
<td>0.96</td>
</tr>
<tr>
<td>Community-level $x$</td>
<td>1</td>
<td>0.77</td>
<td>0.91</td>
<td>0.96</td>
</tr>
<tr>
<td><strong>Random effect st. dev.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Family</td>
<td>1</td>
<td>0.10</td>
<td>0.28</td>
<td>0.73</td>
</tr>
<tr>
<td>Community</td>
<td>1</td>
<td>0.73</td>
<td>0.76</td>
<td>0.93</td>
</tr>
</tbody>
</table>
Rodríguez and Goldman (2001) also analysed real data on child immunisation in Guatemala.
Comparison of Estimation Procedures

Rodríguez and Goldman (2001) also analysed real data on child immunisation in Guatemala.

<table>
<thead>
<tr>
<th></th>
<th>PQL2</th>
<th>PQL1-B</th>
<th>ML</th>
<th>MCMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Family</td>
<td>1.75</td>
<td>2.69</td>
<td>2.32</td>
<td>2.60</td>
</tr>
<tr>
<td>Community</td>
<td>0.84</td>
<td>1.06</td>
<td>1.02</td>
<td>1.13</td>
</tr>
</tbody>
</table>

PQL-B is PQL with a bias correction; ML is maximum likelihood; MCMC is Markov chain Monte Carlo (Gibbs sampling)
Guidelines on Choice of Estimation Procedure

- ML via numerical quadrature preferred for simple models, but estimation times can be lengthy when there are several correlated random effects.

Quasi-likelihood methods are quick and useful for model screening, but biased (especially for small cluster sizes).

MCMC methods are flexible and becoming increasingly computationally feasible; the recommended method in MLwiN.
Guidelines on Choice of Estimation Procedure

- **ML via numerical quadrature** preferred for simple models, but estimation times can be lengthy when there are several correlated random effects
- **Quasi-likelihood methods** quick and useful for model screening, but biased (especially for small cluster sizes)
Guidelines on Choice of Estimation Procedure

- **ML via numerical quadrature** preferred for simple models, but estimation times can be lengthy when there are several correlated random effects.
- **Quasi-likelihood methods** quick and useful for model screening, but biased (especially for small cluster sizes).
- **MCMC methods** are flexible and becoming increasingly computationally feasible; the recommended method in MLwiN.