Multilevel Logistic Regression Models

Brady T. West

Model Specification

Multilevel model for **binary dependent variable *Y***,
measured on **person *i*** within **cluster *j***

$$\ln\left[\frac{P(y_{ij}=1)}{1-P(y_{ij}=1)}\right]=\mathrm{logit}[P(y_{ij}=1)]=\underbrace{\beta_0+\beta_1 x_{1ij}}_{\text{Fixed effects}}+\underbrace{u_{0j}+u_{1j}x_{1ij}}_{\text{Random effects}}$$

Model Specification

Multilevel model for **binary dependent variable *Y***,
measured on **person *i*** within **cluster *j***

$$\ln\left[\frac{P(y_{ij}=1)}{1-P(y_{ij}=1)}\right] = \text{logit}[P(y_{ij}=1)] = \beta_0 + \beta_1 x_{1ij} + u_{0j} + u_{1j}x_{1ij}$$

Fixed effects

Random effects

Could use multilevel specification if desired!

Model Specification, cont’d

Same **distributional assumptions** about random cluster effects: normally distributed, mean vector 0, unique variances and covariances

Model Specification, cont'd

Same **distributional assumptions** about random cluster effect: normally distributed, mean vector 0, unique variances and covariances

Recall: Multilevel model, because have **explicit interest** in estimating variance of random cluster effects!

Model Specification, cont'd

Same **distributional assumptions** about random cluster effect: normally distributed, mean vector 0, unique variances and covariances

Recall: Multilevel model, because have **explicit interest** in estimating variance of random cluster effects!

When we fit **generalized linear regression models** to non-normal outcomes and include random effects, **estimation is more difficult** mathematically ~ <u>**clear motivation is important!**</u>

Estimating the Model Parameters

Fitting multilevel models to non-normal outcomes

→ **difficult** to write likelihood function ~ may not be possible!

Estimating the Model Parameters

Fitting multilevel models to non-normal outcomes
→ **difficult** to write likelihood function ~ may not be possible!

1) *approximate* likelihood function
2) find parameter estimates that maximize *approximate* likelihood

Estimating the Model Parameters

Fitting multilevel models to non-normal outcomes
→ **difficult** to write likelihood function ~ may not be possible!

1) *approximate* likelihood function
2) find parameter estimates that maximize *approximate* likelihood

One approach = **adaptive Gaussian quadrature**
Simulation studies = works well in variety of scenarios
Deep dive: Reading by Kim et al. (2013)

Testing the Model Parameters

Compute confidence intervals or test hypotheses for model parameters

Test null hypotheses (e.g., fixed effect is zero, or variance component is zero – random effects don't vary!), can use **likelihood ratio testing**

Testing the Model Parameters

Compute confidence intervals or test hypotheses for model parameters

Test null hypotheses (e.g., fixed effect is zero, or variance component is zero – random effects don't vary!), can use **likelihood ratio testing**

Use same approach for multilevel logistic models, assuming large enough samples of clusters and observations per cluster

Testing the Model Parameters

Compute confidence intervals or test hypotheses for model parameters

Test null hypotheses (e.g., fixed effect is zero, or variance component is zero – random effects don't vary!), can use **likelihood ratio testing**

Use same approach for multilevel logistic models, assuming large enough samples of clusters and observations per cluster

Reading this week: Provides specific details on how to perform these types of tests for parameters in multilevel models!

Revisiting NHANES Example

- Logistic regression to model probability of ever smoking 100 cigarettes as function of selected predictors.
- Assumed all NHANES observations independent of each other … Observations came from randomly sampled clusters *(geographic areas)*!

Not True!

Revisiting NHANES Example

- Logistic regression to model probability of ever smoking 100 cigarettes as function of selected predictors.
- Assumed all NHANES observations independent of each other … Observations came from randomly sampled clusters *(geographic areas)*!

Not True!

- If smoking observations correlated within areas, *standard errors* in "naïve" logistic regression analysis *likely understated*.

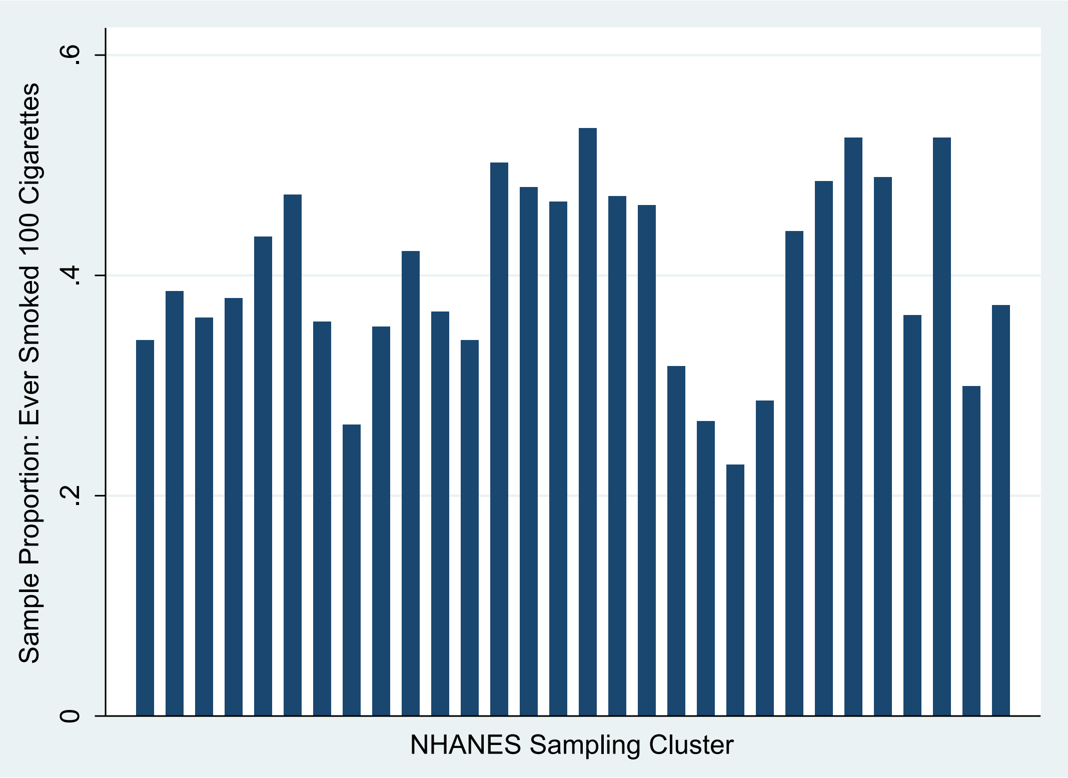
Revisiting NHANES Example

- Logistic regression to model probability of ever smoking 100 cigarettes as function of selected predictors.
- Assumed all NHANES observations independent of each other … Observations came from randomly sampled clusters *(geographic areas)*!

Not True!

- If smoking observations correlated within areas, *standard errors* in "naïve" logistic regression analysis *likely understated*.
- Plus **explicit interest** in *estimating variance* between sampling clusters in terms of probability of smoking!

Between-Cluster Variance in Smoking



Fitting Multilevel Logistic Regression Model

Logistic model including random effects of randomly sampled clusters
(allows intercepts to randomly vary across clusters; no random slopes)

Fitting Multilevel Logistic Regression Model

Logistic model including random effects of randomly sampled clusters
(allows intercepts to randomly vary across clusters; no random slopes)

- **Same inferences** regarding which predictors significant
- **Slight changes** in estimated fixed effects
- Standard errors of estimates are now **larger!**

Suppose that you examine the variability among randomly sampled higher-level clusters of observations (e.g., schools) in the proportions on a binary variable of interest, and you find visual evidence of significant variability in the proportions. You then fit a logistic regression model to these data in Python, but forget to fit a multilevel model including random effects of the clusters. **How would this affect your analysis?**

- Nothing would change; random effects do not affect our estimates of interest.
- The standard errors of the estimated fixed effects would be too high, but the fixed effects would be identical.
- The standard errors of the estimated fixed effects would be too low, but the fixed effects would be identical.
- None of the above.

Correct

Answer: d). By omitting explicit random effects in the logistic regression model, the standard errors of our estimated fixed effects would likely be too low, and the estimates of the fixed effects would likely be incorrect (because we are failing to explicitly adjust for the random effects of the higher-level clusters). Omitting random effects when they are important is the same type of model specification errors as omitting the fixed effect of an important predictor variable.

Fitting the Multilevel Logistic Regression Model

Estimated **variance** of random cluster **intercepts** = 0.046
Significant based on likelihood ratio test (p-value < 0.001)

Fitting the Multilevel Logistic Regression Model

Estimated **variance** of random cluster **intercepts** = 0.046
Significant based on likelihood ratio test (p-value < 0.001)

Even after adjusting for predictors, randomly sampled clusters still vary in terms of smoking prevalence!

Model Diagnostics

Including random cluster effects in logistic regression model **improved fit**

Model Diagnostics

Including random cluster effects in logistic regression model **improved fit**

Q: Look at distribution of **predicted** values of random interviewer effects, or EBLUPs … **outliers**?

Remember: no residuals to worry about in simple logistic regression model!

Model Diagnostics

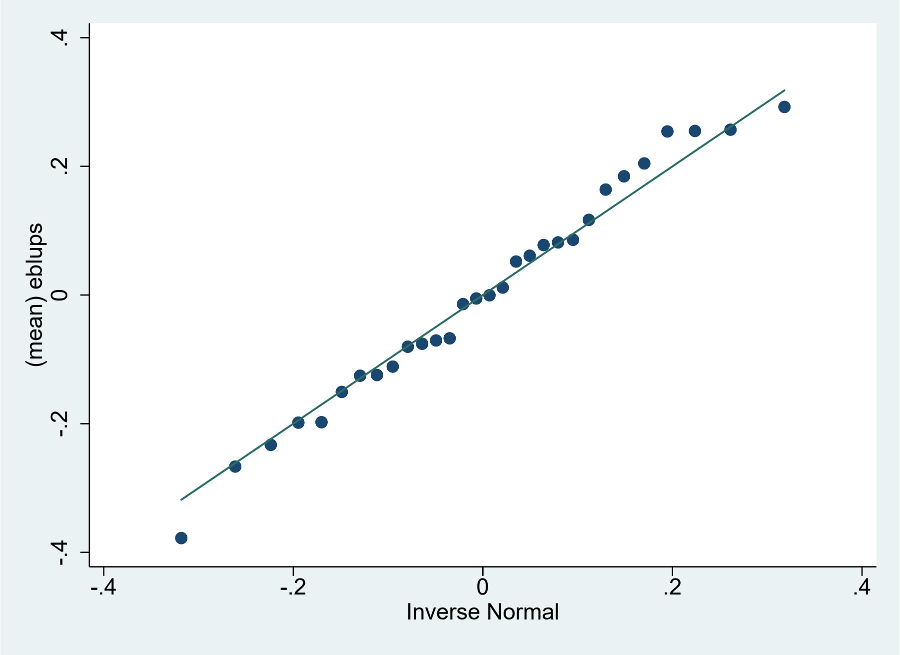
Including random cluster effects in logistic regression model **improved fit**

Q: Look at distribution of **predicted** values of random interviewer effects, or EBLUPs … **outliers**?

Remember: no residuals to worry about in simple logistic regression model!

Another Consideration:
Center continuous predictor variables so intercept is interpretable!

EBLUPs for Random Intercepts



QQ plot suggests
random effects on intercept normally distributed + no outliers!

Conclusions from the Example

- Same predictors of smoking still important!
- Given significant unexplained variance in random cluster effects → explain variance by including fixed effects of cluster-level predictors (e.g., SES)

Conclusions from the Example

- Same predictors of smoking still important!
- Given significant unexplained variance in random cluster effects → explain variance by including fixed effects of cluster-level predictors (e.g., SES)
- HOWEVER: When comparing variance components between multilevel models with different cluster-level fixed effects, **both models must include same respondent-level fixed effects**

Conclusions from the Example

- Same predictors of smoking still important!
- Given significant unexplained variance in random cluster effects → explain variance by including fixed effects of cluster-level predictors (e.g., SES)
- HOWEVER: When comparing variance components between multilevel models with different cluster-level fixed effects, **both models must include same respondent-level fixed effects**

Deeper Dive: Multilevel Analysis: Techniques and Applications, Hox et al, 3rd Edition, Section 6.5

What's Next?

- **Full example:** fitting multilevel models to longitudinal data with Python + making inference
- **Marginal models** for dependent data + alternatives for modeling **clustered** and **longitudinal** data sets