Marginal Logistic Regression Models

Brady T. West

Marginal Modeling Approaches

Recall

With marginal models, there is no explicit interest in making inference about between-cluster variance in the coefficients of interest in a given model

Marginal Modeling Approaches

Recall

With marginal models, there is no explicit interest in making inference about between-cluster variance in the coefficients of interest in a given model

Last Time…

We introduced GEE as a general technique for estimating marginal regression models and accounting for the within-cluster dependency in a dependent variable, introduced by cluster sampling or repeated measurements

Marginal Modeling Approaches

Recall

With marginal models, there is no explicit interest in making inference about between-cluster variance in the coefficients of interest in a given model

Last Time…

We introduced GEE as a general technique for estimating marginal regression models and accounting for the within-cluster dependency in a dependent variable, introduced by cluster sampling or repeated measurements

In this lecture, our primary interest is in population-averaged relationships!

Marginal Models for Binary Outcomes

- GEE methods were specifically designed to **readily accommodate non-normal outcome variables measured longitudinally**

Marginal Models for Binary Outcomes

- GEE methods were specifically designed to **readily accommodate non-normal outcome variables measured longitudinally**
- Mean structure in estimating equation: Defined by a given type of generalized linear model

Marginal Models for Binary Outcomes

- GEE methods were specifically designed to **readily accommodate non-normal outcome variables measured longitudinally**
- Mean structure in estimating equation: Defined by a given type of generalized linear model
- **Example:** in logistic regression for a binary DV, the mean of the DV is the probability that the DV is equal to 1:

$$\mu_{ti} = \pi_{ti} = E(y_{ti} \mid X_{ti}) = \frac{\exp(X_{ti}\beta)}{1 + \exp(X_{ti}\beta)}$$

Marginal Logistic Regression Models

Variance of given binary observation =

Probability that the observation is 1 (the mean), multiplied by 1 minus the probability that the observation is 1 **(1 minus the mean)**

Marginal Logistic Regression Models

Variance of given binary observation =

Probability that the observation is 1 (the mean), multiplied by 1 minus the probability that the observation is 1 **(1 minus the mean)**

The means & variances of DV are **both** defined by the specified model!

Marginal Logistic Regression Models

Variance of given binary observation =

Probability that the observation is 1 (the mean), multiplied by 1 minus the probability that the observation is 1 **(1 minus the mean)**

The means & variances of DV are **both** defined by the specified model!

Given dependent data, we can specify the correlation structure that we believe holds for the binary observations

Exchangeable, first-order autoregressive, unstructured, etc.

Easy to do using GEE!

Revisiting the Smoking Example

Recall

The logistic regression model we fitted to the lifetime smoking indicator in NHANES

Revisiting the Smoking Example

Recall

The logistic regression model we fitted to the lifetime smoking indicator in NHANES

What did we find?

Evidence of variance between NHANES sampling clusters in terms of smoking prevalence (when fitting a multilevel model to the smoking data); **within-cluster dependency?**

Revisiting the Smoking Example

Recall

The logistic regression model we fitted to the lifetime smoking indicator in NHANES

What did we find?

Evidence of variance between NHANES sampling clusters in terms of smoking prevalence (when fitting a multilevel model to the smoking data); **within-cluster dependency?**

Now

Consider a marginal modeling approach using GEE, & examine whether our inferences change when fitting a population-averaged model without explicit random cluster effects!

Revisiting the Smoking Example

- Overall, we see no differences in the inferences that we would make relative to the multilevel modeling approach

* $p < 0.05$, *** $p < 0.001$

	Multilevel Approach		GEE Approach	
Predictor	**Estimate**	**SE**	**Estimate**	**SE**
Male	0.93***	0.06	0.92***	0.07
Age	0.02***	<0.01	0.02***	<0.01
Other Hispanic	0.22	0.12	0.22	0.14
White	0.66***	0.11	0.64***	0.12
Black	0.27*	0.11	0.26*	0.12
Other Race	-0.09	0.12	-0.09	0.15
BMI	<0.01	<0.01	0.01	0.01
Household Size	-0.08***	0.02	-0.08***	0.02
Family Income to Poverty Ratio	-0.19***	0.02	-0.19***	0.02

Revisiting the Smoking Example

- Overall, we see no differences in the inferences that we would make relative to the multilevel modeling approach

  * $p < 0.05$, *** $p < 0.001$

- **Remember:** the multilevel model included random cluster effects!

	Multilevel Approach		GEE Approach	
Predictor	**Estimate**	**SE**	**Estimate**	**SE**
Male	0.93***	0.06	0.92***	0.07
Age	0.02***	<0.01	0.02***	<0.01
Other Hispanic	0.22	0.12	0.22	0.14
White	0.66***	0.11	0.64***	0.12
Black	0.27*	0.11	0.26*	0.12
Other Race	-0.09	0.12	-0.09	0.15
BMI	<0.01	<0.01	0.01	0.01
Household Size	-0.08***	0.02	-0.08***	0.02
Family Income to Poverty Ratio	-0.19***	0.02	-0.19***	0.02

Model Diagnostics

- The "nuisance" estimate of the correlation was only **0.01; QIC = 6284.53**

Assuming a constant correlation within clusters

Model Diagnostics

- The "nuisance" estimate of the correlation was only **0.01; QIC = 6284.53**
- Unstructured and first-order autoregressive correlation structures don't make sense here

Assuming a constant correlation within clusters

No time ordering of the cross-sectional observations within each sampling cluster

Model Diagnostics

- The “nuisance” estimate of the correlation was only **0.01; QIC = 6284.53**
- Unstructured and first-order autoregressive correlation structures don’t make sense here
- Independence: **QIC = 6284.05**

Assuming a constant correlation within clusters

No time ordering of the cross-sectional observations within each sampling cluster

Model Diagnostics

- The "nuisance" estimate of the correlation was only **0.01; QIC = 6284.53**
- Unstructured and first-order autoregressive correlation structures don't make sense here
- Independence: **QIC = 6284.05**
- **Conclusion**: the correlation in the marginal model is fairly weak, and accounting for it is not making a difference in model fit

Assuming a constant correlation within clusters

No time ordering of the cross-sectional observations within each sampling cluster

Conclusions from the Example

1 Marginally: When looking at the overall relationships across sampling clusters, we find **essentially the same estimated fixed effects** as when we fit the multilevel model

Conclusions from the Example

1. Marginally: When looking at the overall relationships across sampling clusters, we find **essentially the same estimated fixed effects** as when we fit the multilevel model
2. Accounting for the dependency (*rather than assuming independence of observations within each cluster*) did **not** seem to improve model fit in this case

Conclusions from the Example

1. Marginally: When looking at the overall relationships across sampling clusters, we find **essentially the same estimated fixed effects** as when we fit the multilevel model
2. Accounting for the dependency (*rather than assuming independence of observations within each cluster*) did **not** seem to improve model fit in this case
3. **Remember:** We still interpret the estimated fixed effects <u>marginally</u>, across clusters *(not conditioning on a <u>given</u> cluster)*

What's Next?

Next, you will have the opportunity to practice fitting marginal models using GEE in Python, and interpreting the results!