BIO 226: APPLIED LONGITUDINAL ANALYSIS

LECTURE 18

Contrasting Marginal and Mixed Effects Models

Contrasting Marginal and Mixed Effects Models for Longitudinal Data

So far, we have discussed two main extensions of generalized linear models:

1. Marginal Models

2. Generalized Linear Mixed Models

There are important distinctions between these two broad classes of models that go beyond simple differences in approaches for accounting for the within-subject association.

These two classes of models have somewhat different targets of inference and address subtly different questions regarding longitudinal change in the response.

A **marginal model** for the mean response is given by

$$g(\mu_{ij}) = g[E(Y_{ij}|X_{ij})] = X'_{ij}\beta = \beta_1 X_{ij1} + \cdots + \beta_p X_{ijp},$$

where $g(\cdot)$ is an appropriate non-linear link function (e.g., logit or log).

In marginal models, β's have interpretation in terms of changes in the transformed mean response in the study population, and their relation to covariates.

The population means can be expressed in terms of the inverse link function, say $h(\cdot) = g^{-1}(\cdot)$,

$$h[g(\mu_{ij})] = \mu_{ij} = E(Y_{ij}|X_{ij}) = h(\beta_1 X_{ij1} + \cdots + \beta_p X_{ijp}).$$

Next, consider the **generalized linear mixed model**

$$g[E(Y_{ij}|X_{ij}, b_i)] = X'_{ij}\beta^* + Z'_{ij}b_i,$$

where the random effects b_i have a distribution with mean zero and covariance matrix G.

The regression coefficients β^* have subject-specific interpretations in terms of changes in the transformed mean response for a specific individual.

β^* do not describe changes in the transformed mean response in the study population.

Recall: Linear Mixed Effects Models (Lecture 11)

In the linear mixed effects model,

$$Y_{ij} = X'_{ij}\beta + Z'_{ij}b_i + \epsilon_{ij},$$

there is an important distinction between the conditional mean,

$$E(Y_{ij}|X_{ij}, b_i) = X'_{ij}\beta + Z'_{ij}b_i,$$

and the marginal mean,

$$E(Y_{ij}|X_{ij}) = E\left[E(Y_{ij}|X_{ij}, b_i)\right] = X'_{ij}\beta.$$

The former describes the mean response for an individual, the latter describes the mean response averaged over individuals.

In GLMMs there is an implied model for the marginal means.

Like the LMEM, this can be obtained by averaging over distribution of the random effects.

For concreteness, consider a logistic regression model:

$$\begin{aligned}
\mu_{ij} &= E(Y_{ij}|X_{ij}) \\
&= E\left[E(Y_{ij}|X_{ij}, b_i)\right] \\
&= E\left[\frac{\exp\left(X'_{ij}\beta^* + Z'_{ij}b_i\right)}{1+\exp\left(X'_{ij}\beta^* + Z'_{ij}b_i\right)}\right] \\
&= \int_{-\infty}^{\infty} \frac{\exp\left(X'_{ij}\beta^* + Z'_{ij}b_i\right)}{1+\exp\left(X'_{ij}\beta^* + Z'_{ij}b_i\right)} f(b_i)db_i.
\end{aligned}$$

This expression for $E(Y_{ij}|X_{ij})$ does not, in general, have a closed-form expression and, moreover,

$$E(Y_{ij}|X_{ij}) \neq \frac{\exp\left(X'_{ij}\beta\right)}{1+\exp\left(X'_{ij}\beta\right)}$$

for any β, e.g., logistic mixed effects model $\neq$ marginal logistic model.

That is, the marginalized model implied by the GLMM doesn't satisfy generalized linear model.

More generally in GLMMs there is an implied model for the marginal means.

This can be obtained by averaging over distribution of the random effects,

$$\begin{aligned}\mu_{ij} &= E(Y_{ij}|X_{ij}) \\ &= E[E(Y_{ij}|X_{ij}, b_i)] \\ &= E[h(X'_{ij}\beta^* + Z'_{ij}b_i)] \\ &= \int_{-\infty}^{\infty} h(X'_{ij}\beta^* + Z'_{ij}b_i) f(b_i) db_i.\end{aligned}$$

However, this expression for $E(Y_{ij}|X_{ij})$ does not, in general, have a closed-form expression and, moreover,

$$E(Y_{ij}|X_{ij}) \neq h(X'_{ij}\beta)$$

for any β, e.g., logistic mixed effects model $\neq$ marginal logistic model.

That is, marginalized model doesn't satisfy generalized linear model.

Simple Numerical Illustration

Consider hypothetical data on true propensity for disease, $\Pr(Y_{ij} = 1|b_i)$, for three individuals measured at baseline (pre) and following treatment with a new drug intended to reduce the risk of disease (post).

The three individuals are discernibly different in terms of their underlying propensity for disease at baseline.

This heterogeneity can be expressed in terms of random effects, b_i.

Individuals A, B, and C have "high", "medium" and "low" underlying risk for disease.

Assume target population is comprised of an equal number of individuals that fall into these three distinct risk groups.

Hypothetical data on the true propensity for disease, at baseline and post-baseline, for three individuals with heterogeneous propensities for disease.

Individual	Pre	Post	Difference	Log(OR)
A	0.80	0.67	-0.13	-0.68
B	0.50	0.33	-0.17	-0.71
C	0.20	0.11	-0.09	-0.70
Pop. Average	0.50	0.37	-0.13	

Final row of table contains the population averages (obtained as equally-weighted means).

For a linear function of the propensity for disease (i.e., the difference), the "difference of the averages" is equal to the "average of the differences".

Taking the average of the subject-specific effects (as a single number summary of the subject-specific effects),

$$\frac{-0.13 - 0.17 - 0.09}{3} = -0.13.$$

Alternatively, can compare the average propensity for disease at baseline (0.5) and post-baseline (0.37).

$$(0.37 - 0.50) = -0.13.$$

The latter can be thought of as a contrast of population averages.

With a non-linear function of the propensity for disease, a "non-linear contrast of the averages" is not equal to the "average of the non-linear contrasts".

Consider the log odds ratios:

Taking the average of the subject-specific effects (as a single number summary of the subject-specific effects),

$$\frac{-0.68 - 0.71 - 0.70}{3} = -0.697.$$

Alternatively, compare the log odds of disease in the population at baseline, $\log(0.5/0.5) = 0$ and post-baseline, $\log(0.37/0.63) = -0.532$.

This comparison yields a measure of effect, -0.532, which is approximately 25% smaller than the summary of the subject-specific effect, -0.697.

Individual	Pre	Post	Difference	Log(OR)
A	0.80	0.67	-0.13	-0.68
B	0.50	0.33	-0.17	-0.71
C	0.20	0.11	-0.09	-0.70
Pop. Average	0.50	0.37	-0.13	

In marginal models, the regression parameters describe the margins of the table.

In GLMMs, the fixed effects describe the interior of the table.

Next we consider a graphical illustration that highlights the differences between these two approaches.

Graphical Illustration

Suppose Y_i is a vector of binary responses and it is of interest to describe changes in the log odds of success over time.

A logistic regression model, with randomly varying intercepts, is given by

$$\text{logit}[E(Y_{ij}|b_i)] = \beta_1^* + \beta_2^* t_{ij} + b_i$$

where $t_{ij} = 0$ at baseline and $t_{ij} = 1$ post-baseline.

The b_i are assumed to have a normal distribution with zero mean and variance $\sigma_b^2 = \text{Var}(b_i)$.

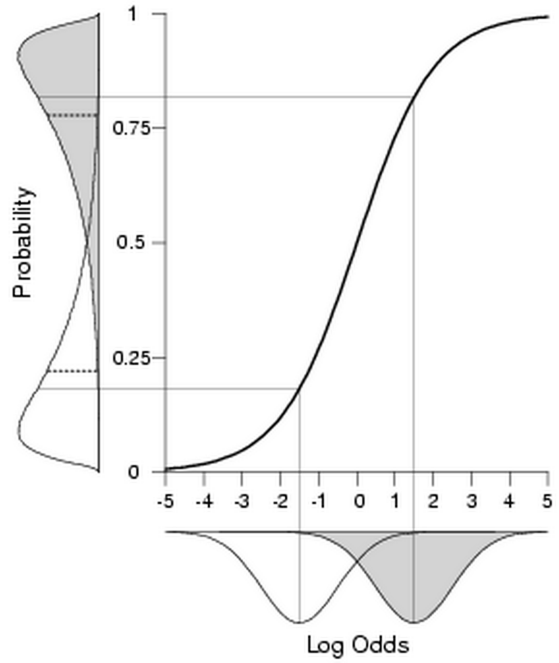
Let $\beta_1^* = 1.5$, $\beta_2^* = -3.0$, and $\text{Var}(b_i) = 1.0$.

At baseline, log odds has a normal distribution with mean = median = 1.5 (see shaded densities).

Note, however, that subject-specific probabilities of disease have a negatively skewed distribution with median, but not mean, of 0.82.

The mean of the subject-specific probabilities is 0.78.

Thus, probability of disease for a "typical" individual from the population (0.82) is not the same as the prevalence of disease in the same population (0.78).

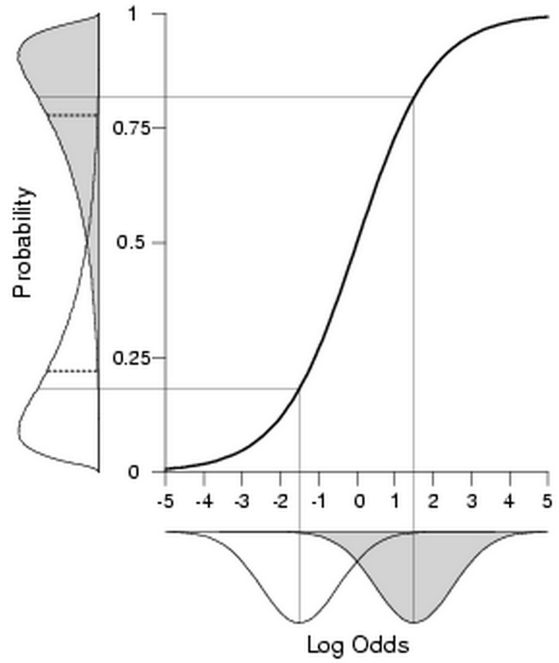


Similarly, the log odds of disease post-baseline has a normal distribution with mean = median = -1.5 (see unshaded densities).

However, subject-specific post-baseline probabilities of disease have a positively skewed distribution with median, but not mean, of 0.18.

The mean of the subject-specific probabilities is 0.22.

Thus, probability of disease post-baseline for a "typical" individual from the population (0.18) is not the same as the prevalence of disease in the same population (0.22).



The effect of treatment on the log odds of disease for a typical individual from the population, $\beta_2^* = -3.0$, is not the same as the contrast of population log odds.

The latter is what is estimated in a marginal model, say

$$\text{logit}\{E(Y_{ij})\} = \beta_1 + \beta_2 t_{ij},$$

and can be obtained by comparing the log odds of disease in the population at baseline, $\log(0.78/0.22) = 1.255$, with the log odds of disease in the population post-baseline, $\log(0.22/0.78) = -1.255$.

This yields a population-averaged measure of effect, $\beta_2 = -2.51$, which is approximately 15% smaller than β_2^*, the subject-specific effect of treatment.

Case Study

Cross-Over Trial of Cerebrovascular Deficiency

- Two-period cross-over trial comparing effects of active drug to placebo on cerebrovascular deficiency
- 67 patients randomly allocated to two treatment sequences
- 34 patients receiving Placebo → Active
- 33 patients receiving Active → Placebo
- Each patient has a bivariate binary response vector, $Y_i = (Y_{i1}, Y_{i2})$ denoting whether an electrocardiogram was normal (0) or abnormal (1).

Data from a two-period cross-over trial comparing the effects of active drug to placebo on cerebrovascular deficiency. The response indicates whether an electrocardiogram was normal (0) or abnormal (1).

	Response (Period 1, Period 2)			
Sequence	(1,1)	(1,0)	(0,1)	(0,0)
Sequence 1 (P $\rightarrow$ A)	6	0	6	22
Sequence 2 (A $\rightarrow$ P)	9	4	2	18

P: Placebo; A: Active drug.

First, consider marginal logistic model

$$\text{logit}(\mu_{ij}) = \text{logit}[\Pr(Y_{ij} = 1)] = \beta_1 + \beta_2 \text{Treatment} + \beta_3 \text{Period}$$

where Treatment (0 = Placebo, 1 = Active drug) and Period (0 = Period 1, 1 = Period 2).

The within subject association between the two responses was modelled in terms of a common log odds ratio, α,

$$\log \frac{\Pr(Y_{i1} = 1, Y_{i2} = 1) \Pr(Y_{i1} = 0, Y_{i2} = 0)}{\Pr(Y_{i1} = 1, Y_{i2} = 0) \Pr(Y_{i1} = 0, Y_{i2} = 1)} = \alpha.$$

Parameter estimates and standard errors from marginal logistic regression model for the cerebrovascular deficiency data.

Parameter	Estimate	SE	Z
Intercept	-1.2433	0.2999	-4.15
Treatment	0.5689	0.2335	2.44
Period	0.2951	0.2319	1.27
log OR (α)	3.5617	0.8148	4.37

The results indicate that treatment with the active drug is harmful, increasing the rates of abnormal electrocardiograms.

The odds of an abnormal electrocardiogram is 1.77 (or $e^{0.57}$) times higher when treated with active drug versus placebo.

The estimate of the within-subject association is $\widehat{\alpha} = 3.56$, indicating that there is very strong positive association (OR= 35.2).

Next, consider logistic regression model with a random patient effect,

$$\text{logit}[E(Y_{ij}|b_i)] = \beta_1^* + \beta_2^*\text{Treatment} + \beta_3^*\text{Period} + b_i$$

where the random effect b_i is assumed to have a normal distribution with zero mean and variance, $\sigma_b^2 = \text{Var}(b_i)$.

Parameter estimates and standard errors from mixed effects logistic regression model for the cerebrovascular deficiency data.

Parameter	Estimate	SE	Z
Intercept	-4.0817	1.6711	-2.44
Treatment	1.8631	0.9269	2.01
Period	1.0376	0.8189	1.27
$\sigma_b^2 = \text{Var}(b_i)$	24.4365	18.8500	1.30

ML based on 100-point adaptive Gaussian quadrature.

The results also indicate that treatment with the active drug is harmful, increasing the patient-specific risk of an abnormal electrocardiogram.

In particular, a patient's odds of an abnormal electrocardiogram is 6.4 (or $e^{1.86}$) times higher when treated with active drug than when treated with the placebo.

The estimate of the variance of b_i, $\widehat{\sigma}_b^2 = 24.4$, indicates that there is very substantial between-patient variability in their propensity for abnormal electrocardiograms.

Comparison of the two estimated effects of treatment, $e^{\widehat{\beta}_2} = 1.8$ and $e^{\widehat{\beta}_2^*} = 6.4$, from the marginal and mixed effects logistic regression models highlights the distinction between these two analytic approaches.

$\widehat{\beta}_2$ from marginal model describes how the average rates (expressed in terms of odds) of abnormal ECGs could be increased in the study population if patients are treated with the active drug.

$\widehat{\beta}_2^*$ from the mixed effects model describes how the odds of an abnormal ECG increases for any patient treated with the active drug.

Thus, a population-level analysis understates the individual risk, and vice versa.

In summary, the answer to the question "what are the side effects of the active drug" will depend on whether scientific interest is in its impact on the study population or on an individual drawn at random from that population.

With marginal models the main focus is on inferences about the study population.

With generalized linear mixed models the main focus is on inferences about individuals.

Aside

Does the very large estimate of variance, $\widehat{\sigma}_b^2 = 24.4$, accurately reflect between-patient variability in the risk of abnormal electrocardiogram?

In this example, a large proportion of subjects (82%) had same response, (0,0) or (1,1), at both occasions.

This feature can only be captured by a normal distribution for the log odds with large variance.

When number of repeated binary responses is small, and there is a large proportion of subjects with positive (negative) responses at all occasions, the normal assumption for b_i is questionable.

Concluding Remarks

Unlike linear models, where the concepts of regression analysis can be applied quite robustly, longitudinal analysis of categorical data raises many subtle issues.

Different models for categorical outcomes can give discernibly different results.

The choice and meaning of longitudinal models for categorical outcomes require somewhat greater care.

With different targets of inference, different models for categorical outcomes address subtly different questions regarding longitudinal change.

Choice among models?

- should be guided by specific scientific question of interest
- answers to different questions will usually demand that different models have to be applied
- different questions will often produce different, albeit compatible, answers
- "one size does not fit all"