INTRODUCTION TO MODELING CORRELATED DATA

What are correlated data?

- We will be looking at "clustered" data – Observations within each "cluster" are *correlated* with each other.
  - Positive correlation → large measurements tend to cluster with large measurements.
  - Negative correlation → large measurements tend to cluster with small measurements.
- Examples of "clusters"?
  - Longitudinal data: Repeated measurements taken on the same individual over time.
  - Measurements taken on both a mother and daughter.
  - Measurements taken on all individuals in a household.

Cross-Sectional vs. Longitudinal

- In a **cross-sectional** study, measurements are obtained at only a single point in time.
  - It is not possible to assess individual changes across time.
- In a **longitudinal** study, participants are measured repeatedly throughout the duration of the study.
  - Permits direct assessment of changes in the response variable over time.

Terminology in Longitudinal Data

- Participants or units being studied = *individuals* or *subjects*.
- Individuals are measured repeatedly at different *times* or *occasions*.
  - Times need not be equally spaced.

Terminology in Longitudinal Data

- If all individuals have the same number of repeated measurements obtained at a common set of occasions, we say the study is ***"balanced" over time***.
- If repeated measurements are not obtained at a common set of occasions (or individuals have differing numbers of measurements), the study is ***"unbalanced" over time***.
  - Common when study is retrospective (e.g., data obtained from medical databases) or when times defined relative to some individual benchmark event, e.g., menarche study.
- If there are missing data (an intended measurement could not be obtained), the data set is called ***"incomplete."***

Goal of Longitudinal Studies

- There are two goals in longitudinal data analysis:
  - Assess *within-individual* (intra-individual) changes in the response variable.
    - How do we characterize the change in the response variable over time?
  - Assess *between-individual* (inter-individual) changes in the response variable.
    - Are the "response trajectories" of individuals related to certain covariates?
- Cross-sectional studies are only able to assess between-individual variation.

Linear Models for Longitudinal Data

Independent data:

- Assume Y has a normal distribution with mean $E(Y \mid X)$ and variance σ^2
- Model the mean of the response variable Y as some linear function (in the parameters) of covariates $X_1, X_2, \ldots, X_k$, e.g.

$$E(Y \mid X) = \beta_0 + \beta_1 X$$

$$E(Y \mid \mathbf{X}) = \beta_0 + \beta_1 \log(X_1) + \beta_2 X_2 + \beta_3 X_2^2$$

Linear Models for Longitudinal Data

Two primary extensions of the linear model:

- Non-normally distributed response variable
  - Generalized linear models (STAT 439)
- Dependent/correlated (not independent) observations
  - Generalized least squares and Linear mixed effects models (STAT 412, STAT 448)
- Both non-normal and correlated data
  - Marginal models (generalized estimating equations) and Generalized linear mixed effects models (Now!)

→ Need to spend some time thinking about modeling covariance structures.

Correlation in Longitudinal Data

Nature of correlations among repeated measures taken on one individual:

1. positive
2. decrease with increasing time separation
3. rarely approach zero for pairs of measurements taken far apart in time
4. rarely approach one for pairs of measurements taken very closely together in time

Variation in Longitudinal Data

- Between-subject heterogeneity in *mean response*
  - Some individuals consistently respond higher than average, and others lower.
    e.g., annual income, daily caloric intake, systolic blood pressure
  - Induces a positive correlation between repeated measurements
- Between-subject heterogeneity in *response trajectory*
  - Some individuals "improve" more quickly than others, and some may worsen.
    e.g., CD4 lymphocyte counts after antiviral treatment in AIDS patients, or rate of increase in annual income
  - Often induces decreasing correlations with increasing time separation, e.g., scores at times 1 and 4 often less correlated than scores at times 1 and 2.

Variation in Longitudinal Data

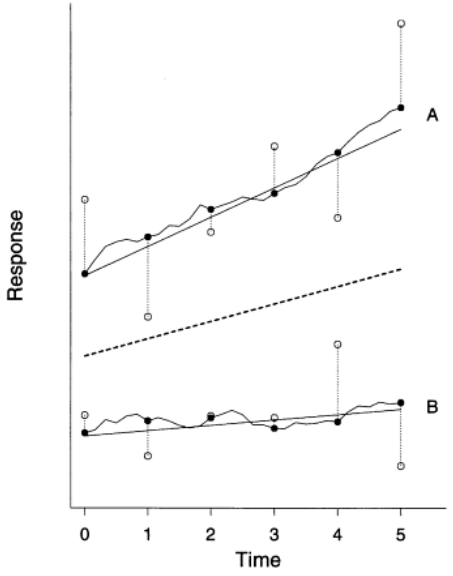
- Within-subject biological variation
  - Repeated measures are realizations of some biological process operating within the individual.
    e.g., weight, systolic blood pressure, serum cholesterol
  - Serial correlation: a stronger correlation for measurements that are closer together in time.
- Measurement error
  - Not to be confused with within-subject biological variation.
  - May *shrink* the correlation among repeated measures closer to zero.

Sources of Variation in Longitudinal Data

1. Between-individual heterogeneity
2. Within-individual biological variation
3. Measurement error

- Solid dot indicates true measure (free of measurement error); open dot denotes actual measurement with measurement error.
- Solid line represents true individual response trajectory (free of biological variation); jagged curve is within-individual biological variation from solid line.
- Dotted line is average true response trajectory between the two respondents.

Fitzmaurice, Laird and Ware (2011)



GENERALIZED LINEAR MIXED MODELS (GLMMS) AND MARGINAL MODELS

Modified from original slides by Scott Bartell, PhD

Marginal Models vs. GLMMs

- Both marginal models and GLMMs are models that handle non-normal response variables (e.g., binary, counts) with correlated data.
- GLMMs model *subject-specific* mean response: $E(Y_{ij} \mid b_i)$
- Marginal models are *population-averaged* models, and model the *marginal* mean response: $E(Y_{ij})$
  - There are no random effects in marginal models.
  - The mean structure and covariance structure are modeled separately.
  - Analogous to generalized least squares.
- The choice between GLMM and marginal model relies on the subject-matter and the scientific question of interest.

Marginal Models vs. GLMMs: Example

Hypothetical Data on Probability of a Disease

Individual	Baseline Risk	Post-Baseline Risk	Risk Difference (Post – Pre)	log(OR)
A (High)	0.80	0.67	–0.13	–0.678
B (Med)	0.50	0.33	–0.17	–0.708
C (Low)	0.20	0.11	–0.09	–0.704
Population Average	0.50	0.37	–0.13	?

Marginal Models vs. GLMMs: Example

How to estimate the effectiveness of the treatment?

Option 1: Average the **subject-specific** effects

$$\frac{-0.678 + (-0.708) + (-0.704)}{3} = -0.697 \quad \Rightarrow \quad e^{-0.697} = 0.498$$

Option 2: Calculate the **population-averaged** log odds ratio

$$\log\left(\frac{0.37/(1-0.37)}{0.50/(1-0.50)}\right) = -0.532 \quad \Rightarrow \quad e^{-0.532} = 0.587$$

Which one is better? Depends on the question!

Marginal Models vs. GLMMs: Example

Option 1 (subject-specific): There is an estimated 50.2% reduction in the odds of disease for any individual treated with the drug.

→ Of most interest to an individual and his/her physician in the physician-patient context.

Option 2 (population-averaged): There is an estimated 41.3% reduction in the odds of disease in the population if everyone were to be treated with the drug.

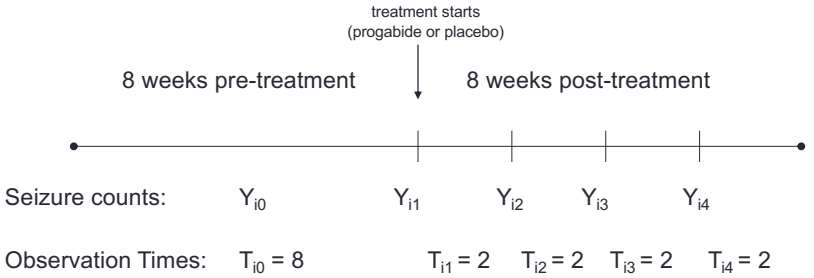
→ Of most interest to public health researchers interested in the potential benefits of the drug on the prevalence of disease in the population as a whole.

INTERPRETING GLMMS

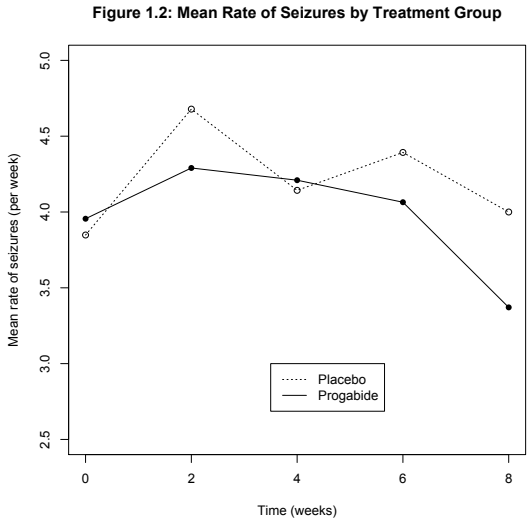
Example: Clinical Trial of an Anti-Epileptic Drug

- 59 epileptic patients randomized to progabide or placebo (Leppik et al., 1987)
  - Fitzmaurice et al., 2011, pp. 421-427
- Number of seizures in prior 8 weeks recorded (baseline), then treatment starts
- Number of seizures during the following four 2-week intervals (total of 8 weeks) after treatment
- Research question: Does treatment with progabide reduce the rate of epileptic seizures?

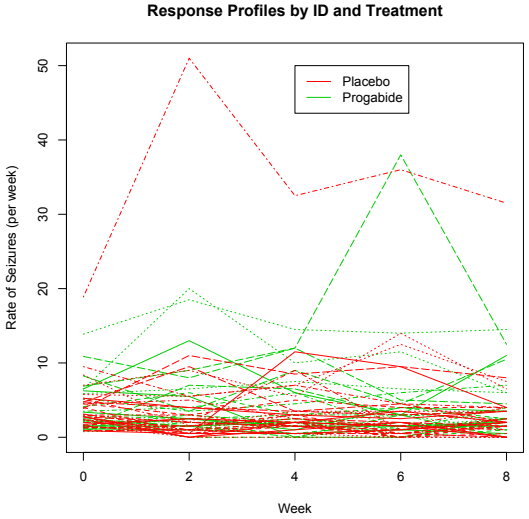
Anti-Epileptic Drug Trial Timeline



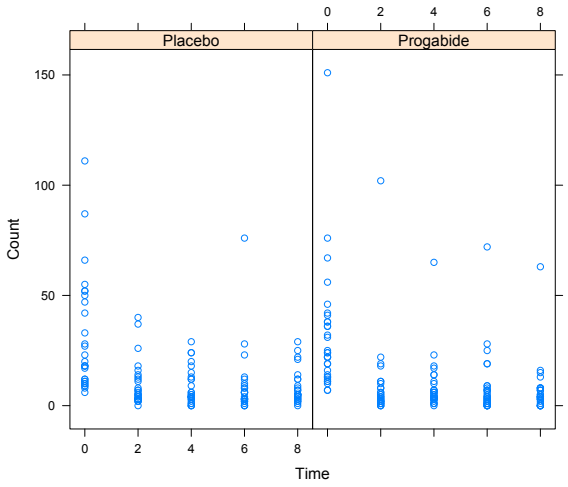
Exploratory Data Analysis



Exploratory Data Analysis



Exploratory Data Analysis



Does it seem reasonable to assume that the ***number*** of seizures has a normal distribution? what about the ***rate*** (no. of seizures per week)?

Generalized Linear Mixed Effects Models

$$g(E(\underline{Y}_i|\underline{b}_i)) = \mathbf{X}_i\underline{\beta} + \mathbf{Z}_i\underline{b}_i$$

where g(.) is a known "link function",

- $Y_{ij}|\underline{b}_i$ has an "exponential family" distribution (e.g., normal, Poisson, binomial, etc.),
- $Y_{ij}|\underline{b}_i$ and $Y_{ik}|\underline{b}_i$ are independent for all $j \neq k$ (conditional independence),
- $\mathbf{X}_i$ is the matrix of fixed effect covariates for subject i,
- $\underline{\beta}$ is the column vector of fixed effects,
- $\mathbf{Z}_i$ is the matrix of random effect covariates for subject i,
- $\underline{b}_i$ is the column vector of **multivariate normal** random effects.

Exponential family →

Var($Y_{ij}|b_i$) is decomposed into a product of the "variance function" $v(\mu_{ij})$ involving any terms dependent on $\mu_{ij} := E(Y_{ij}|b_i)$, and a constant "scale parameter" ϕ, either known or estimated.

Example: Poisson GLMM → Log Link

$\log(E(Y_{ij}|\underline{b}_i)) = (\beta_0 + b_{0i}) + \log(T_{ij}) + \beta_1 X_{1ij} + (\beta_2 + b_{2i})*X_{2ij} + \beta_3 X_{1ij} X_{2ij}$

where

- Y_{ij} is the seizure count for subject i in period j, with a Poisson distribution conditional on the subject's covariates and random effects
- $\underline{b}_i$ is the multivariate normal random effects vector $(b_{0i}, b_{2i})^T$,
- T_{ij} is the length of the observation period for subject i and period j=0,1,2,...,4 (T_{ij} is either 8 weeks or 2 weeks),
- X_{1ij} is the progabide indicator variable (1 if progabide, 0 placebo),
- X_{2ij} is the post-baseline indicator variable (1 if j>0, 0 else)

Example: Poisson GLMM → Log Link

$\log(E(Y_{ij}|\underline{b}_i)) = (\beta_0 + b_{0i}) + \log(T_{ij}) + \beta_1 X_{1ij} + (\beta_2 + b_{2i})*X_{2ij} + \beta_3 X_{1ij} X_{2ij}$

- The term $\log(T_{ij})$ is called the "offset" – the model can also be written as modelling the mean rate:

  $\log(E(Y_{ij}|\underline{b}_i)/T_{ij}) = (\beta_0 + b_{0i}) + \beta_1 X_{1ij} + (\beta_2 + b_{2i})*X_{2ij} + \beta_3 X_{1ij} X_{2ij}$

- This is a random intercept/random "slope" GLMM with a log link, variance function $v(\mu) = \mu$, and scale parameter $\phi = 1$.

- The model can also be written in hierarchical form:

  $Y_{ij}|\mu_{ij} \sim \text{Poisson}(\mu_{ij})$

  $\log(\mu_{ij}) \sim \text{Normal}(\beta_0 + \log(T_{ij}) + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{1ij} X_{2ij}, \text{var}(b_{0i} + b_{2i} X_{2ij}))$

R Functions to Fit GLMMs

- `glmer` (in lme4 library) – *We'll use this one*.
  - adaptive Gauss-Hermite quadrature approximation
  - `lmer` function in this library fits LMEs
- `nlme` (in nlme library)
- `glmmML` (in glmmML library)
  - Only allows for random intercept
- `glmPQL` (in mass library)
  - penalized quasi-likelihood
- `MCMCglmm` (in MCMCglmm library)
  - Bayesian estimation

Fitting the GLMM in R

```
> library(lme4)
> mod1 <- glmer(Count ~ trt*PostBase + (PostBase | ID), offset=log(Weeks),
                family=poisson, data=epi_long)
> summary(mod1)
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: poisson  ( log )
Formula: Count ~ trt * PostBase + (PostBase | ID)
   Data: epi_long
 Offset: log(Weeks)

Random effects:
 Groups Name        Variance Std.Dev. Corr
 ID     (Intercept) 0.500    0.707
        PostBase    0.232    0.482    0.16
Number of obs: 295, groups:  ID, 59

Fixed effects:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)            1.0708     0.1403    7.63  2.3e-14 ***
trtProgabide           0.0512     0.1927    0.27    0.790
PostBase              -0.0005     0.1091    0.00    0.996
trtProgabide:PostBase -0.3062     0.1504   -2.04    0.042 *
```

Interpreting Est. Coefs (Ignoring p-values…)

- exp{$\hat{\beta}_1$} = exp{0.051} = 1.05 is…
  - the estimated rate ratio of baseline seizure rates (per week) for a "typical" subject ($\mathbf{b}_i = \mathbf{0}$) taking progabide versus a "typical" subject taking the placebo
  - or, 5% higher estimated baseline seizure rate for a subject on progabide compared to a subject on placebo with *the same values of random effects*.
- exp{$\hat{\beta}_2$} = exp{-.0005} ≈ 1 is…
  - the estimated rate ratio of post-baseline to baseline seizure rates for an "typical" subject taking the placebo
  - i.e., a "typical" subject on placebo has approximately the same estimated rate of seizures before and after taking the placebo. (need "typical" due to b2i)

Interpreting Est. Coefs (Ignoring p-values…)

- exp{$\hat{\beta}_2 + \hat{\beta}_3$} = exp{-.0005 – .3062} = 0.74 is…
  - the estimated rate ratio of post-baseline to baseline seizure rates for a "typical" subject taking progabide
  - i.e., a "typical" subject has an estimated 26% fewer seizures while taking progabide than at baseline
- exp{$\hat{\beta}_3$} = exp{-0.3062} = 0.74 is…
  - the estimated ratio of post-baseline to baseline seizure rate ratios for a subject taking progabide versus a subject taking placebo *with the same values of random effects*
  - i.e., the estimated effect of progabide for a "typical" subject is a 26% reduction in the pre-/post-treatment seizure rate ratio compared to a "typical" subject taking placebo

Interpreting Random Effect Est. Variances

Estimated variance of random intercepts = 0.50 →

- Represents variability in baseline log rate of seizures.
- For example, approximately 95% of patients assigned to placebo have an estimated baseline seizure rate that varies from

$$\exp(1.071 \pm 2\sqrt{.500}) = (0.709, 12.003)$$

i.e. between 0.7 to 12.0 seizures per week.

Interpreting Random Effect Est. Variances

Estimated variance of random "slope" = 0.23:

- Variability in patient-to-patient *changes* in the log seizure rates from pre- to post-treatment.
- Approximately 95% of patients *treated with progabide* have estimated changes in the rates of seizures that vary from

$$\exp(-0.3066 \pm 2\sqrt{.23}) = (0.282, 1.920)$$

i.e., decrease of about 72% to an increase of about 92% after treatment.

Interpreting Covariate Effects for GLMMs: Marginal versus Conditional Means

In **GLMMs**, covariate effects often differ at the group and individual levels!

- Models are parameterized in terms of covariate effects on conditional means → within-subject or subject-specific changes in covariates
- But marginal means are not so easy to calculate!

  $E(\underline{Y}_i) = E(g^{-1}(\mathbf{X}_i\underline{\beta} + \mathbf{Z}_i\underline{b}_i)) = ?$
- Thus $\underline{\beta}$ is modeled as the effect of the covariates on $\mathbf{X}_i\underline{\beta} + \mathbf{Z}_i\underline{b}_i$ for any individual, but the effect on $g^{-1}(\mathbf{X}_i\underline{\beta} + \mathbf{Z}_i\underline{b}_i)$ generally differs from person to person depending on individual random effects ($\underline{b}_i$)

WARNING: *Interpret covariate effects (β's) from GLMMs cautiously, as conditional (subject-specific) effects only!*

- *Interpret as effects on "typical" subjects ($\boldsymbol{b}_i = \boldsymbol{0}$),*
- *or as effect sizes that only apply to subjects with the same $\boldsymbol{b}_i$ values*

Interpreting Covariate Effects for LMEs: Marginal versus Conditional Means

In **linear mixed effects models**, covariate effects (β's) are equivalent for groups and individuals

- recall that all methods studied to date describe subject-specific (conditional) expectations
- marginal (group) means are easy for LMEs because:

  $E(\underline{Y}_i) = E(\mathbf{X}_i\underline{\beta} + \mathbf{Z}_i\underline{b}_i + \underline{\varepsilon}_i) = \mathbf{X}_i\underline{\beta} + \mathbf{Z}_iE(\underline{b}_i) + E(\underline{\varepsilon}_i) = \mathbf{X}_i\underline{\beta}$
- so $\underline{\beta}$ is easily interpreted as *both* the effect of the covariates on an individual, and the same effect on any group of individuals with the same covariates

CAUTION

GLMM estimation algorithms are relatively new

- Rely on likelihood *approximations*
- Rapidly changing methods *even in the same software package*
- Parameter estimates may differ substantially
- Especially sensitive: low Poisson counts or rare dichotomous outcomes!
  - most epidemiologic analyses

INTERPRETING MARGINAL MODELS

Marginal Models

- Marginal models are also called "population-averaged models".
- They handle correlated data by specifying a separate covariance model (separate from the mean model).
- Only specify first two moments of the distribution (mean and variance/covariance) – do not assume an entire probability distribution.

Marginal Models

Same notation as before: For individual i ($i = 1, \ldots, N$):

- $n_i \times 1$ vector of responses (which need not be from a normal distribution – could be 0/1 or counts or whatever):

$$Y_i = \begin{pmatrix} Y_{i1} & Y_{i2} & \cdots & Y_{in_i} \end{pmatrix}'$$

- $n_i \times p$ matrix of covariates:

$$X_i = \begin{pmatrix} X_{i1}' \\ X_{i2}' \\ \vdots \\ X_{in_i}' \end{pmatrix} = \begin{pmatrix} X_{i11} & X_{i12} & \cdots & X_{i1p} \\ X_{i21} & X_{i22} & \cdots & X_{i2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{in_i1} & X_{in_i2} & \cdots & X_{in_ip} \end{pmatrix}$$

Marginal Models

Three-part specification:

1. The mean response (conditional on covariates) is assumed to depend on the covariates through a known **link function**:

"linear predictor"

$$g(E(Y_{ij})) = g(\mu_{ij}) = \eta_{ij} = X_{ij}'\beta$$

2. The variance (conditional on covariates) is assumed to depend on the mean according to

$$Var(Y_{ij}) = \phi \; v(\mu_{ij})$$

"Scale parameter" (either known or estimated)

Known "variance function"

Marginal Models

3. The within-subject association (conditional on covariates) among the vector of repeated responses is assumed to be a function of an additional set of "association parameters", α (and also depends on the means).
   - For continuous (e.g., normal) response, "association" can be specified in terms of correlations.
   - Correlations do not make much sense for binary data, and often associations are specified in terms of log odds ratios among repeated responses.

Generalized Estimating Equations

- Since there are no distributional assumptions in marginal models, there are no maximum likelihood estimates.
- Instead, marginal models use the method of estimation called **generalized estimating equations (GEE)**.
  - Generalized least squares is a special case of the GEE approach.
- Alternate between estimating mean and estimating variance/covariance parameters.

Revisiting the Anti-Epileptic Drug Example Using GEEs

$$\log(E(Y_{ij})) = \beta_0 + \log(T_{ij}) + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{1ij} X_{2ij}$$

where

- Y_{ij} is the seizure count for subject i in period j with a Poisson *variance function*, and Corr($\mathbf{Y}_i$) is compound symmetric,
- T_{ij} is the length of the observation period for subject i and period j=0,1,2,...,4 (T_{ij} is either 8 weeks or 2 weeks),
- X_{1ij} is the progabide indicator variable (1 if progabide, 0 placebo),
- X_{2ij} is the post-baseline indicator variable (1 if j>0, 0 else)

Fitting the Marginal Model in R

```
> library(gee)
> mod.gee <- gee(Count ~ trt*PostBase + offset(log(Weeks)),
                 id = ID, family = poisson(link = "log"),
                 corstr = "exchangeable", data = epi_long)
> summary(mod.gee)

Coefficients:
                     Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)           1.34761     0.1511  8.9188      0.1574   8.5640
trtProgabide          0.02753     0.2071  0.1330      0.2218   0.1241
PostBase              0.11184     0.1545  0.7238      0.1159   0.9647
trtProgabide:PostBase -0.10473     0.2197 -0.4767      0.2134  -0.4906
```


Interpreting GEE Results

The marginal model results have the interpretation we usually want—group comparisons.

- $\exp\{\hat{\beta}_1\} = \exp\{0.027\} = 1.03$ is…
  - the estimated rate ratio of baseline seizure rates (per week) for those taking progabide versus those taking the placebo
  - i.e., 3% higher estimated baseline seizure rate for those taking progabide

Interpreting GEE Results

- exp{$\hat{\beta}_2$} = exp{0.112} = 1.12 is…
  - the estimated rate ratio of post-baseline to baseline seizure rates for those taking the placebo
  - i.e., we estimate that subjects taking the placebo had 12% more seizures during the study than they did in the 8 weeks before it started
  - Note that this is much higher than our GLMM estimate…

Interpreting GEE Results

- exp{$\hat{\beta}_2 + \hat{\beta}_3$} = exp{0.112 – 0.105} = 1.01 is…
  - the estimated rate ratio of post-baseline to baseline seizure rates for those taking progabide
  - i.e., we estimate subjects taking progabide had 1% more seizures during the study than at baseline
- exp{$\hat{\beta}_3$} = exp{-0.105} = 0.90 is…
  - the esimated *ratio* of post-baseline to baseline seizure rate *ratios* for those taking progabide versus those taking the placebo
  - i.e., the estimated effect of progabide is a 10% reduction in the post-/pre-treatment seizure rate ratio compared to the placebo group

General Advice on GEEs/GLMMs

- Results of GEE often differ from GLMM when random effects variances are large.
  - If you want marginal interpretation, use GEEs.
  - If you want subject-specific interpretation, use GLMMs.
- Covariance/correlation matrices for GEE are more difficult to conceptualize than hierarchical random effects.
  - GEE parameter estimates are only moderately influenced by choice of model for the correlation structure.
  - Can determine or approximate the correlation structures implied by specific random effects models, with some serious mathematical effort. Such models are called “subject-specific GEEs”.

General Advice on GEEs/GLMMs

- GEE algorithms are fairly uniform across statistical packages, and more stable than GLMMs.
  - Ordinary (population-averaged) GEEs are common; subject-specific GEEs much less so.
- GLMMs delineate variance sources in the model, and may be more appropriate when variance components are of primary interest.
  - The same hierarchical models described by GLMMs can also be estimated using Bayesian methods (MCMC).

Summary of Types of Predictions

Model	Marginal: $E(Y_{ij})$	Subject-specific: $E(Y_{ij} \mid b_i)$
LME	$X_i\hat{\beta}$ `R: predict(…, level=0)`	$X_i\hat{\beta} + Z_i\hat{b}_i$ `R: predict(…, level=1)`
GLMM	None!	$g^{-1}\left(X_i\hat{\beta} + Z_i\hat{b}_i\right)$ or for "typical subject" ($b_i = 0$): $g^{-1}\left(X_i\hat{\beta}\right)$
GEE	$g^{-1}\left(X_i\hat{\beta}\right)$ `R: predict(…, type="response)`	None! (No random effects)