An Introduction to Generalized Estimating Equations

Cancer Prevention and Control Tutorial

16 October 2008

Repeated measures ANOVA limitations

- Unbalanced design (missing data) causes problems in estimation of expected mean squares $\Rightarrow$ F-tests
- Subjects with incomplete response profile deleted from analysis
- Constrained to continuous responses

Generalized linear model

- Unifies in a single method
  - Linear regression (continuous response)
  - Logistic regression (binary response)
  - Poisson regression (count response)
- Specify distribution of random component, Y
  - Linear regression $\Rightarrow Y \sim$ Normal
  - Logistic regression $\Rightarrow Y \sim$ Bernoulli
  - Poisson regression $\Rightarrow Y \sim$ Poisson
- Systematic component of model is linear combination of predictors - called *linear predictor*

$$\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

Generalized linear model (cont.)

- Relate mean of Y to linear predictor through *link function*
- $Y \sim$ Normal
  - Mean of Y is μ, the center of the distribution
  - Link function is *identity link* (i.e. no transformation of mean required)
  - $\mu = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$
- $Y \sim$ Bernoulli
  - Mean of Y is p, probability of success
  - Link function is usually *logit link*
  - $\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$

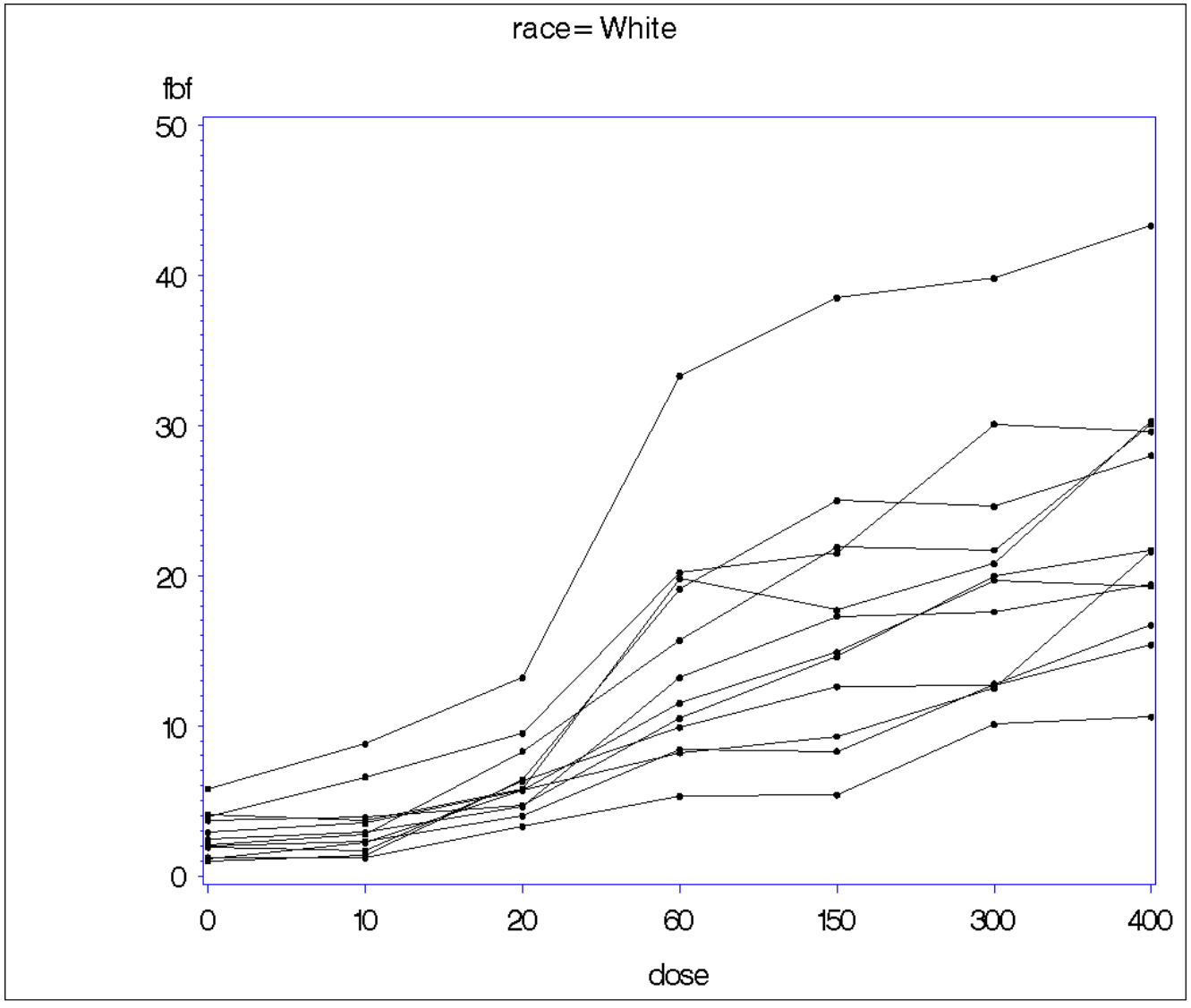
Generalized linear model (cont.)

- $Y \sim$ Poisson
  - Mean of Y is λ, rate per unit time of events
  - Link function is *log link*
  - $\log(\lambda) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$

Generalized Estimating Equations

- Extends generalized linear model to accommodate correlated Ys
  - Longitudinal (e.g. Number of cigarettes smoked per day measured at 1, 4, 8 and 16 weeks post intervention)
  - Repeated measures (e.g. Protein concentration sample from primary tumor and metastatic site)
- Need to specify distribution
- Link function
- Correlation structure

Visualizing correlation



Describing correlation mathematically

Exchangeable correlation: Responses within subjects are equally correlated

	cigs1	cigs2	cigs3	cigs4
cigs1	1	ρ	ρ	ρ
cigs2	ρ	1	ρ	ρ
cigs3	ρ	ρ	1	ρ
cigs4	ρ	ρ	ρ	1

Describing correlation mathematically (cont.)

First-order Auto Regressive (AR1): Correlation among responses within subjects decays exponentially

	cigs1	cigs2	cigs3	cigs4
cigs1	1	ρ	ρ^2	ρ^3
cigs2	ρ	1	ρ	ρ^2
cigs3	ρ^2	ρ	1	ρ
cigs4	ρ^3	ρ^2	ρ	1

Describing correlation mathematically (cont.)

Unstructured: Correlation among responses within subjects completely unspecified

	cigs1	cigs2	cigs3	cigs4
cigs1	1	$\rho_{1,2}$	$\rho_{1,3}$	$\rho_{1,4}$
cigs2	$\rho_{2,1}$	1	$\rho_{2,3}$	$\rho_{2,4}$
cigs3	$\rho_{3,1}$	$\rho_{3,2}$	1	$\rho_{3,4}$
cigs4	$\rho_{4,1}$	$\rho_{4,2}$	$\rho_{4,3}$	1

Describing correlation mathematically (cont.)

Independence: No correlation among responses within subjects

	cigs1	cigs2	cigs3	cigs4
cigs1	1	0	0	0
cigs2	0	1	0	0
cigs3	0	0	1	0
cigs4	0	0	0	1

GEE analysis

- Specify distribution
- Specify link function
- Specify correlation structure $\Rightarrow$ *working variance-covariance matrix*
- Estimate model parameters using *quasi-likelihood* $\Rightarrow \hat{\beta}$s
- Estimate variance-covariance matrix of model parameters using *sandwich estimator* $\Rightarrow$ confidence intervals, inference for the βs

The wonderful thing about GEEs... even if the working variance-covariance matrix is mis-specified, the sandwich estimator converges to the true variance-covariance matrix of the model parameters.

Caution!

- Convergence of sandwich estimator to true var-cov matrix requires
  - Diminishing fraction of missing data

    OR

  - Missing completely at random
- Asymptotics for inference about βs hold if
  - Number of subjects (n) is large

    AND

  - Cluster sizes (m) are small
- If n small relative to m, better to use generalized score tests as opposed to Wald tests for CIs and tests associated with βs

Data structure

Wide

ID	Cigs1	Cigs2	Cigs3	Cigs4	Cigs0	Trt	Sex
1	12	10	8	2	10	1	1
2	15	16	15	18	18	1	0

Long

ID	Cigs	Time	Cigs0	Trt	Sex
1	12	1	10	1	1
1	10	2	10	1	1
1	8	3	10	1	1
1	2	4	10	1	1
2	15	1	18	1	0
2	16	2	18	1	0
2	15	3	18	1	0
2	18	4	18	1	0