STATISTICS
TEXAS A&M UNIVERSITY

Handout 14

Statistical Analysis with the GLIMMIX Procedure

Applications Using the GLIMMIX Procedure:

Beta Regression

Repeated measures

Copyright (c) Derya Akleman 2017

An Example of Beta Regression

Objective

- Describe beta distributions.
- Fit a model for beta distribution using PROC GLIMMIX.

Copyright (c) Derya Akleman 2017

Beta Distribution

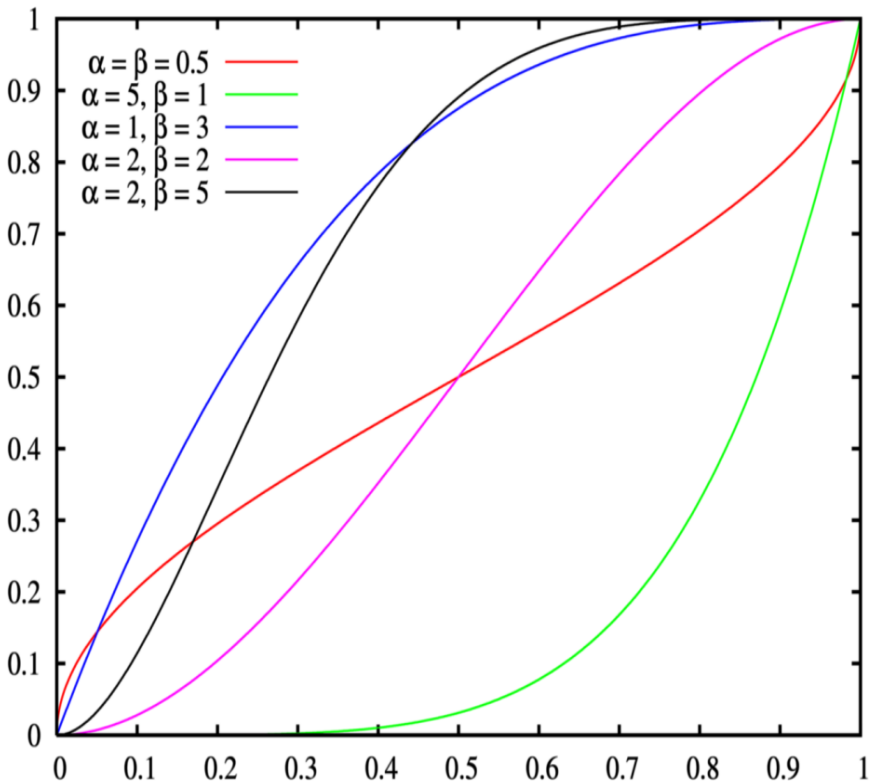
Beta distribution

- belongs to the exponential family of distributions
- is generally for a continuous variable between the values 0 and 1, such as an insurance claim rate
- is often used as a prior distribution for binomial proportions in Bayesian analysis
- is provided in the GLIMMIX procedure with the default link function of logit.
- Example of Variables:
  - Percent of recovery of insurance claims
  - Proportions of gas expenditure
  - Proportions of leaf area infected by a fungus

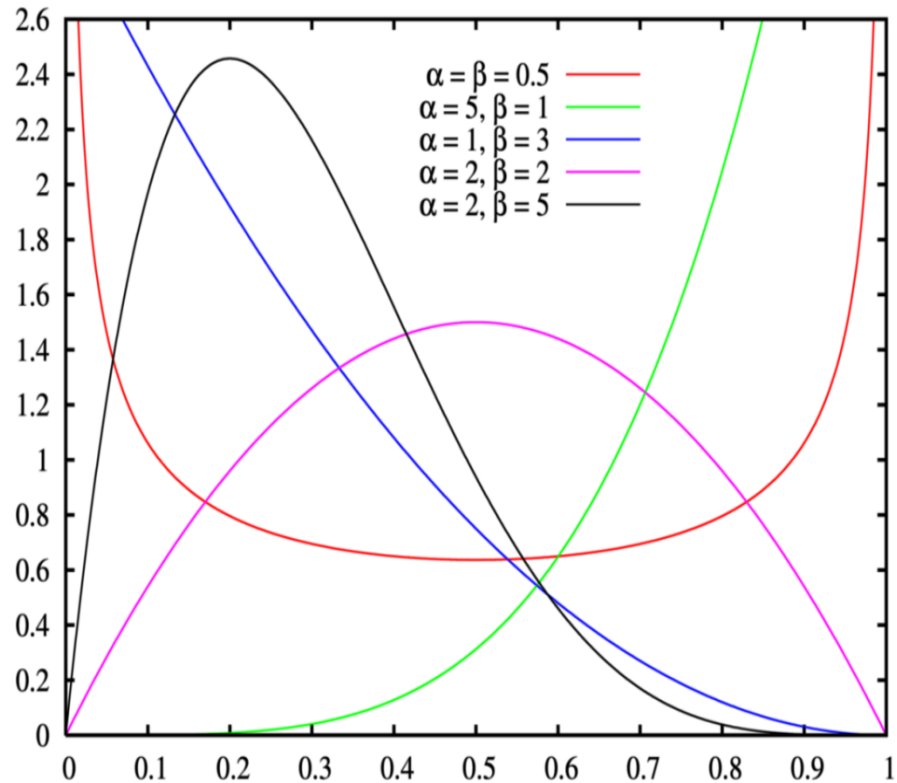
Copyright (c) Derya Akleman 2017

Beta Distribution

cdf function



pdf function



Beta distribution and binomial distribution are both used for proportions and can be used interchangeably.

- ❍ True
- ❍ False

Copyright (c) Derya Akleman 2017

Beta Distribution and Binomial Distribution

- Beta distribution can be used to model proportions as well as other positive values.
- Binomial distribution is used to model *n* independent Bernoulli trials with an event probability of *p*.
- They both use the logit as the canonical Link function.
- Beta distribution might be more general than the binomial distribution.

Copyright (c) Derya Akleman 2017

Beta Regression and Poisson Regression for Rates

For beta regression,

- the dependent variable follows a beta distribution
- typically only the proportion is known
- the proportion is not necessarily a result of count.

For Poisson regression for rate data,

- the dependent variable is a count variable and is assumed to follow a Poisson distribution
- the measure of exposure is known and is used as the offset variable.

Copyright (c) Derya Akleman 2017

Food Expenditure Example

Data was collected from randomly selected households in 20 randomly selected cities in USA. For each household, the proportion of household income spent on food is recorded, as well as an index measure of affluence of the household, including income, disposable income, value of family home, and so on.

As the index grows smaller, the family is less affluent. This data is stored in SAS dataset food.

City: the city identification code
Affluence: the affluence index of the household
Foodexp: proportion of household income spent on food

Food Expenditure Example

Affluence index

Proportions of food expenditure

City

The Data and The Model

Obs	city	affluence	foodexp
1	1	84.9278	0.09480
2	2	56.6979	0.48784
3	3	81.8640	0.04742
4	3	81.6557	0.04946
5	3	82.9491	0.05669
6	3	56.4912	0.14023
7	3	43.9401	0.06682
8	4	51.1889	0.46421
9	4	48.9796	0.40510
10	4	45.2729	0.24639
11	4	85.1834	0.21552
12	4	51.9527	0.45444
...			

$$\eta_{ij} \mid u_i = \log\left(\frac{p_{ij} \mid u_i}{1 - p_{ij} \mid u_i}\right) = \beta_0 + \beta_1 x_{ij} + u_i$$

Copyright (c) Derya Akleman 2017

The SGSCATTER Procedure

General form of the SCSCATTER procedure:

```
PROC SGSCATTER options;
      COMPARE X= variable Y= variable / options;
      MATRIX variable-1  ... variable-n / options;
      PLOT plot-request(s) / options;
RUN;
```

Copyright (c) Derya Akleman 2017

Fitting a Model for the Beta Distribution

This demonstration illustrates concepts discussed previously.

foodexample.sas

Copyright (c) Derya Akleman 2017

Question

Which of the following is **false**?

a. Uniform distribution is a special case of beta distribution.
b. Beta regression is not the same as Poisson regression for rate because the distributional assumption of the response variable is different.
c. Beta regression and logistic regression are the same. They use the logit link function and the dependent variables are both proportions.

Copyright (c) Derya Akleman 2017

Repeated Measures Data with Discrete Response

Objective

- List the issues involved with analyzing repeated measures data.
- Model the G-side and the R-side random effects.
- Use the GLIMMIX procedure to analyze repeated measures data with discrete responses.

Copyright (c) Derya Akleman 2017

Issues with Repeated Measures Data

- Repeated measures data refers to multiple measures on the same experimental unit (or subject).

- The basic issue in repeated measures data is the violation of the independence assumption.
  - Measurements taken on the same subjects tend to be more similar than measurements taken on different subjects.
  - Measurements taken close in time on the same subject tend to be more similar than measurements taken far apart in time.

Copyright (c) Derya Akleman 2017

What Happens If the Correlation Is Ignored?

- The inferences on the treatment effects (or the predictor variables) might be biased.

- The predictions might not be as accurate.

Copyright (c) Derya Akleman 2017

The Analysis Strategy

- Produce profile plots for continuous responses.
- Write a complex mean model for the treatment effects.
- Model the correlations among the repeated measurements.
- Make statistical inferences based on the model.

Copyright (c) Derya Akleman 2017

The Modeling Tool for Continuous Responses

- Use the RANDOM statement in PROC MIXED to model the variances between subjects (the **G** matrix).
- Use the REPEATED statement in PROC MIXED to model the correlations within a subject (the **R** matrix).
- Depending on your data and the model, you might need only one or both statements.

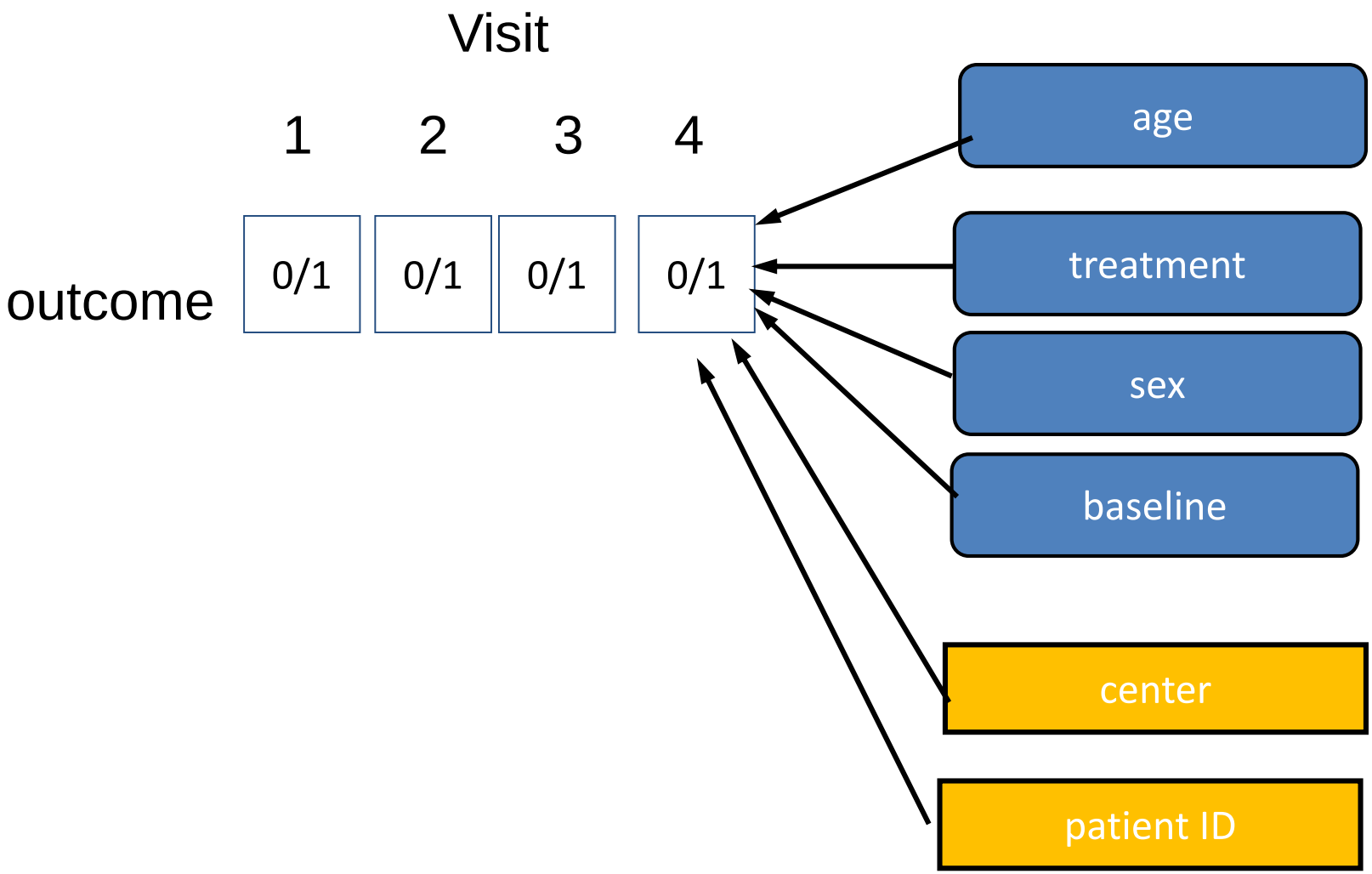
Copyright (c) Derya Akleman 2017

Repeated Measures with Discrete Response (resp1example.sas)

A clinical trial was conducted to compare two treatments for a respiratory illness in 4 randomly selected centers.
Eligible patients were randomly assigned to one of the two treatments: active treatment (A) or placebo (P).
During the treatment, respiratory status of good outcome (1) or poor outcome (0) was determined after 4 visits for each patient.
The variables are:
Outcome: respiratory status: 0-poor, 1-good
Visit: visits to the center where the measurements were taken
Age: patient's age at the beginning of the study
Treatment: A=active treatment, P=placebo
Sex: patient's gender
Baseline: baseline measurement of the respiratory status: 0=poor, 1=good
Center: clinic where the trial was conducted

Copyright (c) Derya Akleman 2017

Repeated Measures with Discrete Response



Copyright (c) Derya Akleman 2017

Exploratory Data Analysis Using Logit Plots

This demonstration illustrates the concepts discussed previously.

resp1example.sas

Copyright (c) Derya Akleman 2017

Question

While Exploratory Data Analysis provides a nice presentation of the data, it does not tell you the statistical significance of an effect.

- ❍ True
- ❍ False

Copyright (c) Derya Akleman 2017

GLMM Formulation and PROC GLIMMIX

$$g(\mu \mid \gamma) = \mathbf{X}\beta + \mathbf{Z}\gamma$$

LINK= option

MODEL statement

RANDOM statement

$Y|\gamma$ ~ exponential family

DIST= option

$\text{var}(\gamma) = G$

Options in the RANDOM statement

$\text{Var}(y|\gamma) = A_\mu R A_\mu$

RANDOM _RESIDUAL_ statement

Copyright (c) Derya Akleman 2017

G-Side and R-Side Random Effects

- G-side random effects are
  - contained in the **G** matrix for the covariance
  - modeled by the RANDOM statement in PROC GLIMMIX.

- R-side random effects are
  - contained in the **R** matrix for the covariance
  - modeled by the RANDOM statement with the _RESIDUAL_ keyword or the RESIDUAL option in PROC GLIMMIX for non-default structures
  - equivalent to a REPEATED effect in PROC MIXED.

Copyright (c) Derya Akleman 2017

G-side and R-side Random Effect Models for Repeated Measures Data

- G-side random effects accomplish the following:
  - model the random effects within the link function
  - provide subject-specific interpretations of the model
  - indirectly model the correlations among the repeated measurements
- R-side random effects accomplish the following:
  - model the random effects outside the link function
  - provide population-average interpretations of the model if no G-side random effects are present
  - directly model the correlations among the repeated measurements

Copyright (c) Derya Akleman 2017

G-side and R-side Random Effect Models for Repeated Measures Data

- For linear mixed models, you can obtain equivalent marginal models for some specifications of G-side and R-side random effects.

- For generalized linear mixed models, equivalent marginal models cannot be obtained between the G-side random effect models and the R-side random effect models.

Copyright (c) Derya Akleman 2017

One Fact about PROC MIXED

For linear mixed models, the following statements yield the same marginal model, and the results are identical. Is that true?

```
proc mixed;
    class A;
    model y=A;
    random int / subject=id;
run;
```

```
proc mixed;
    class A;
    model y=A;
    repeated / type=cs subject=id;
run;
```

Copyright (c) Derya Akleman 2017

One Fact about PROC GLIMMIX for Nonnormal Data

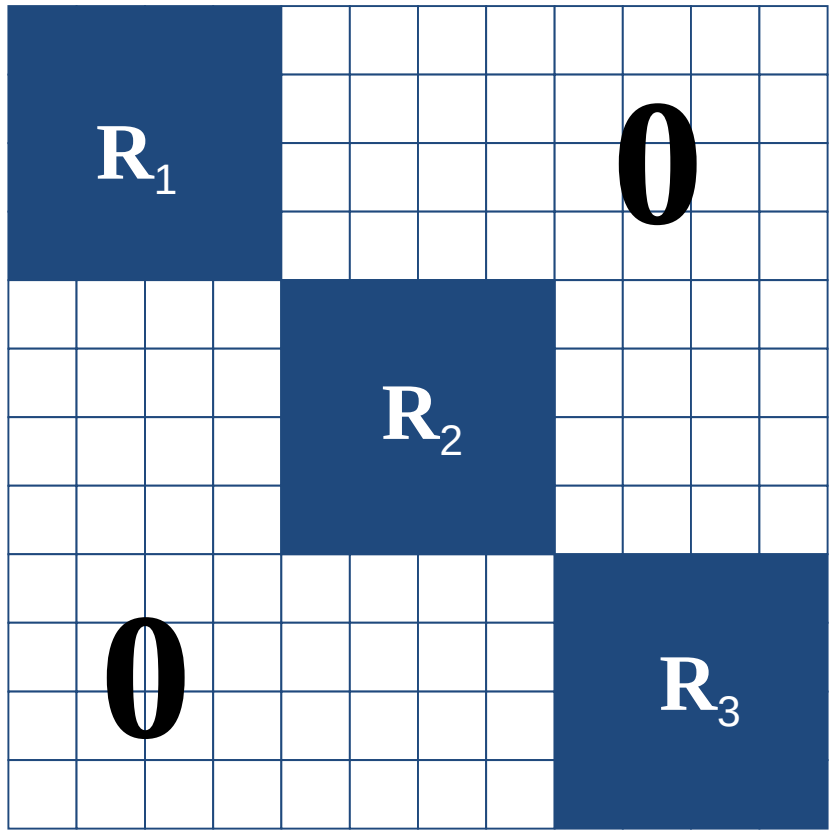
For generalized linear mixed models, the following statements do **not** yield the same marginal model, and the results are **not** identical. Is that true?

```
proc glimmix;
   class A;
   model y=A / dist=bin link=logit;
   random int / subject=id;
run;
```

```
proc glimmix;
   class A;
   model y=A / dist=bin link=logit;
   random _residual_ / type=cs subject=id;
run;
```

Copyright (c) Derya Akleman 2017

Block Diagonal Covariance Matrix




```
random _residual_ / subject= type= ;
```

or

```
random visit / subject= type= residual;
```

Copyright (c) Derya Akleman 2017

Covariance Matrix in Each Block:

Variance Component (VC)

σ_1^2			
	σ_1^2		
		σ_1^2	
			σ_1^2

0 0

Unstructured Covariance(UN)

σ_1^2	σ_{12}	σ_{13}	σ_{14}
	σ_2^2	σ_{23}	σ_{24}
		σ_3^2	σ_{34}
			σ_4^2

Unstructured Parametrized (UNR)

σ_1^2	ρ_{12}	ρ_{13}	ρ_{14}
	σ_2^2	ρ_{23}	ρ_{24}
		σ_3^2	ρ_{34}
			σ_4^2

Copyright (c) Derya Akleman 2017

Covariance Matrix in Each Block: Unstructured through its Cholesky Root (CHOL)

The Cholesky root of a positive definite matrix **A** is an upper triangular matrix **T**, such that **T'T=A**.

$$\mathbf{T}' = \begin{bmatrix} t_1 & 0 \\ t_{12} & t_2 \end{bmatrix}$$

$$\mathbf{T}'\mathbf{T} = \begin{bmatrix} t_1 & 0 \\ t_{12} & t_2 \end{bmatrix}\begin{bmatrix} t_1 & t_{12} \\ 0 & t_2 \end{bmatrix} = \begin{bmatrix} t_1^2 & t_1 t_{12} \\ t_1 t_{12} & t_{12}^2 + t_2^2 \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

Copyright (c) Derya Akleman 2017

Covariance Matrix in Each Block: Unstructured through Its Cholesky Root

- The Cholesky root (TYPE=CHOL) and the unstructured (TYPE=UN) are generally interchangeable.
- The covariance parameter estimates are different because they use different parameterizations.
- CHOL constrains the covariance matrix to be nonnegative definite, whereas TYPE=UN can occasionally produce an indefinite estimate.
- Using TYPE=CHOL can improve the convergence and stability in the model fitting process.

Copyright (c) Derya Akleman 2017

Covariance Matrix in Each Block:

Compound Symmetry (CS)

$\sigma_b^2 + \sigma_e^2$	σ_b^2	σ_b^2	σ_b^2
	$\sigma_b^2 + \sigma_e^2$	σ_b^2	σ_b^2
		$\sigma_b^2 + \sigma_e^2$	σ_b^2
			$\sigma_b^2 + \sigma_e^2$

First Order Autoregressive (AR(1))

σ^2

1.0	ρ	ρ^2	ρ^3
	1.0	ρ	ρ^2
		1.0	ρ
			1.0

Toeplitz (TOEP)

σ^2

1.0	ρ_1	ρ_2	ρ_3
	1.0	ρ_1	ρ_2
		1.0	ρ_1
			1.0

Spatial Power (SP(POW)(*time*))

σ^2

1.0	$\rho^{\lvert t_1 - t_2 \rvert}$	$\rho^{\lvert t_1 - t_3 \rvert}$	$\rho^{\lvert t_1 - t_4 \rvert}$
	1.0	$\rho^{\lvert t_2 - t_3 \rvert}$	$\rho^{\lvert t_2 - t_4 \rvert}$
		1.0	$\rho^{\lvert t_3 - t_4 \rvert}$
			1.0

Copyright (c) Derya Akleman 2017

Covariance Matrix in Each Block:

Spatial Exponential (SP(EXP)(*time*))

$$\sigma^2 \begin{bmatrix} 1.0 & e^{-\frac{d_{12}}{\rho}} & e^{-\frac{d_{13}}{\rho}} & e^{-\frac{d_{14}}{\rho}} \\ & 1.0 & e^{-\frac{d_{23}}{\rho}} & e^{-\frac{d_{24}}{\rho}} \\ & & 1.0 & e^{-\frac{d_{34}}{\rho}} \\ & & & 1.0 \end{bmatrix}$$

Question

Which of the following is **false**?

a. Compound symmetry is useful when the data is equally correlated.
b. Unstructured and Cholesky root structures are unrelated and are totally different.
c. Spatial covariance structures take into account the time values for which the repeated measures were taken.
d. Spatial covariance structures are particularly useful for repeated measures that have irregular time points.
e. AR(1) structure can be considered a special case of the Toeplitz structure.

Copyright (c) Derya Akleman 2017

Fitting a Model for Repeated Measures Data

This demonstration illustrates concepts discussed previously.

resp1.sas

Copyright (c) Derya Akleman 2017

Question

Which of the following is true?

a. The RANDOM statement can be used to model the G-side and/or the R-side covariance in PROC GLIMMIX.
b. You can use the fit statistics in PROC GLIMMIX to compare models for GLMMs.
c. You cannot model the correlations among the repeated measures using PROC GLIMMIX because there is no REPEATED statement.

Copyright (c) Derya Akleman 2017

The COVTEST Statement

The COVTEST statement enables you to obtain statistical inferences for the covariance parameters.

- Fit the model using PROC GLIMMIX.
- Specify hypotheses about the covariance parameters in the COVTEST statement.
- The procedure will
  - refit the model under the restriction on the covariance parameters
  - compare -2(restricted) log likelihoods
  - make *p*-value adjustments for testing on the boundary, if possible and necessary.

Copyright (c) Derya Akleman 2017

The COVTEST Statement

COVTEST <'*label*'> <*test-specification*> </ *options*>**;**

```
covtest 'Ho: common variance' homogeneity;
covtest 'Ho: no random effects' GLM;
covtest 'Ho: independent random effects' diagG;
covtest 'Ho: no slope variance ' . . 0;
```

Copyright (c) Derya Akleman 2017

A Note on the COVTEST Statement

- When the model is estimated by ML or REML, the likelihood from the reduced model by the COVTEST statement is the same as the likelihood from fitting the reduced model directly.

- When the model is estimated by pseudo-likelihood, the pseudo-likelihood from the reduced model by the COVTEST statement is based on the pseudo-data for the full model, so the pseudo-likelihood is not the same as the pseudo-likelihood from fitting the reduced model directly.

resp1.sas

Copyright (c) Derya Akleman 2017

Answer the following Questions

- Is it possible to write a COVTEST statement to test whether the AR(1) correlation structure is sufficient for your data compared with the unstructured covariance model?
  - ❍ Yes
  - ❍ No

- The ODDSRATIO option can be specified in both the MODEL statement and the LSMEANS statement with the DIFF option.
  - ❍ True
  - ❍ False

Copyright (c) Derya Akleman 2017