BIO 226, Spring 2015: Lab 7

Generalized Estimating Equations

Yoonyoung Park

April 28, 2015

Introduction

We now extend Generalized Linear Models (GLMs) to Longitudinal Data. This allows us to model an outcome that is not distributed as a normal random variable.

There are three possible approaches:

- Marginal Models
- Mixed Effect Models
- Transitional Models

Today will use PROC GENMOD to fit and interpret Marginal Models.

Marginal Models

With Marginal Models:

- the mean response is modeled conditional on covariates only (not on random effects or other responses)
- the model for the mean and within-subject association are specified separately.

The First Dataset

We follow the Muscatine Coronary Risk Factor (MCRF) example described in Fitzmaurice, Laird, and Ware (FLW); pg. 364-374.

- Longitudinal survey of school-aged children to examine risk factors for coronary disease in children.
- Weight and height of five cohorts of children, 5-7, 7-9, 9-11, 11-13, and 13-15 years, were obtained in 1977, 1979, and 1981.
- 4856 boys and girls were collected.
- The data are incomplete for many children.

Analysis Questions

Analyze the binary obesity indicator for each child at each occasion.

The goal of the analyses is to determine whether the risk of obesity increases with age and whether patterns of change in obesity are the same for boys and girls.

Data Summarized

The percentage of children classified as obese are given in the table below. What patterns do you notice?

Gender	age	Percentage Obese 1977	1979	1981
Males	5–7	7.9	15.4	21.2
	7–9	18.8	20.5	23.7
	9–11	21.2	22.7	22.5
	11–13	24.3	21.8	19.4
	13–15	19.2	21.1	18.2
Females	5–7	14.0	17.2	25.1
	7–9	16.5	24.0	24.9
	9–11	25.4	26.2	22.2
	11–13	23.8	22.1	19.9
	13–15	22.9	25.8	20.9

Specifying the Marginal Model

A marginal model for longitudinal data has three-part specification:

1. Conditional mean of each response, $E(Y_{ij}|X_{ij}) = \mu_{ij}$ is assumed to depend on the covariates through a known link function g:

$$g(\mu_{ij}) = \beta_1 X_{1ij} + \beta_2 X_{2ij} + ... + \beta_p X_{pij}$$

2. Conditional variance of each Y_{ij}, given the covariates, is assumed to depend on the mean according to

$$\text{Var } (Y_{ij}|X_{ij}) = \phi v(\mu_{ij})$$

where $v(\mu_{ij})$ is a known 'variance function' and ϕ is a scale parameter that may be known or may need to be estimated depending upon our assumptions

Specifying the Marginal Model - cont.

3 Given covariates, conditional within-subject association among repeated responses is assumed to be a function of the means and an additional set of association parameters α

Model for the Data - Part 1

- Logistic model with linear and quadratic age, gender, and both gender-age interactions.
- Model age as the midpoint of the relevant age cohort (e.g., 6 years for those in the 5-7 years group).
- Let $Y_{ij} = 1$ if the i^{th} child is classified as obese at the j^{th} occasion.
- Marginal probability of obesity at each occasion then follows the logistic model.

Model for the data - Part 1 cont.

$$\log\left\{\frac{\Pr(Y_{ij}=1)}{\Pr(Y_{ij}=0)}\right\} = \beta_1 + \beta_2 * \text{gender}_i + \beta_3 * \text{age}_{ij} + \beta_4 * \text{age}_{ij}^2 + \beta_5 * \text{gender}_i * \text{age}_{ij} + \beta_6 * \text{gender}_i * \text{age}_{ij}^2$$

- age_{ij} = midpoint of i^{th} subject's age cohort at j^{th} occasion - 12 (centered)
- $gender_i = 1$ if the i^{th} child is female

Model for the Data - Part 2

Assume that:

1. $v(\mu_{ij}) = \mu_{ij}(1 - \mu_{ij})$
   - Direct result of the Bernoulli distribution
2. $\phi = 1$
   - Corresponds to our assumption that the responses do not have *overdispersion*, or variability that exceeds that predicted by the bernoulli distribution.
   - Note that when we have a binary outcome, overdispersion can't exist.

Model for the Data - Part 3

- Need to make assumptions about the pairwise within-subject associations among the binary responses.
- Correlation is not the best metric for association when the data are binary.
- Thus specify the association in terms of pairwise log odds ratios. (Within-subject association is assumed to have an unstructured pairwise log odds ratio pattern)

$$\log \text{OR}(Y_{ij}, Y_{ik}) = \log\left(\frac{\Pr(Y_{ij}=1, Y_{ik}=1)\Pr(Y_{ij}=0, Y_{ik}=0)}{\Pr(Y_{ij}=1, Y_{ik}=0)\Pr(Y_{ij}=0, Y_{ik}=1)}\right) = \alpha_{jk}$$

SAS Code

```
DATA obesity;
 INFILE 'muscatine.dat';
 INPUT id gender baseage age occasion y;
 cage=age - 12;
 cage2=cage*cage;
RUN;

PROC GENMOD DATA=obesity DESCENDING;
      CLASS id occasion;
      MODEL y=gender cage cage2 gender*cage gender*cage2  / DIST=bin LINK=logit;
      CONTRAST 'Age X Gender Interaction' gender*cage 1, gender*cage2 1 /WALD;
      REPEATED SUBJECT=id / WITHINSUBJECT=occasion LOGOR=FULLCLUST;
RUN;
```

SAS Code Described

- PROC GENMOD statement
  1. DESCENDING option reverses the levels of the *response* variable so that the smallest value is the baseline category
- MODEL statement
  1. Specifies entire form of the marginal model for the mean
  2. DIST option specifies appropriate distribution for the response
  3. LINK option specifies the link function [default: canonical link]
  4. WALD option gives Wald statistics rather than score statistics.

SAS Code Described cont.

- REPEATED statement
  1. Specify independent units of observation via SUBJECT variable; must be also listed in CLASS statement
  2. LOGOR option specifies assumed form of odds ratios
  3. WITHINSUBJECT option specifies order of measurement within subjects

     With missing data, this option properly orders the existing measurements and treats the omitted measurements as missing values. If the WITHINSUBJECT option is not used in this situation, measurements may be improperly ordered and missing values assumed for the last measurements in a cluster. Variable used must be listed in CLASS statement.
- CONTRAST statement
  1. Useful for tests involving multiple mean model parameters
  2. Uses same form as PROC GLM and PROC MIXED

Contrast Statement

We have also specified (via. CONTRAST statement) a test of the hypothesis that changes in the log odds of obesity are the same for boys and girls, $H_0 : \beta_5 = \beta_6 = 0$.

In contrast form, this can be written:

$$H_0 : \mathbf{L}\boldsymbol{\beta} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{pmatrix} = \begin{pmatrix} \beta_5 \\ \beta_6 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

Selected Output

Log Odds Ratio Parameter Information

Parameter	Group
Alpha1	(1, 2)
Alpha2	(1, 3)
Alpha3	(2, 3)

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	−1.2135	0.0506	−1.3126	−1.1144	−24.00	<.0001
gender	0.1159	0.0711	−0.0235	0.2553	1.63	0.1033
cage	0.0378	0.0133	0.0118	0.0638	2.85	0.0043
cage2	−0.0175	0.0034	−0.0241	−0.0109	−5.19	<.0001
gender∗cage	0.0075	0.0182	−0.0282	0.0433	0.41	0.6795
gender∗cage2	0.0039	0.0046	−0.0051	0.0130	0.85	0.3949
Alpha1	3.1528	0.1280	2.9019	3.4037	24.63	<.0001
Alpha2	2.5975	0.1353	2.3323	2.8627	19.20	<.0001
Alpha3	2.9868	0.1236	2.7446	3.2291	24.17	<.0001

Contrast Results for GEE Analysis

Contrast	DF	Chi-Square	Pr > ChiSq	Type
Age ∗ Gender Interaction	2	0.91	0.6356	Wald

Output Interpretation

What do we conclude about changes in the log odds of obesity over time for boys and girls?

Note the pairwise log odds ratios between adjacent occasions (the $\hat{\alpha}_{jk}$s) are similar and approximately 3. This indicates the odds ratio for within-subject association is approximately e^3 or 20. How do we interpret these parameters?

Model without Interactions

Finally consider a marginal logistic regression model without the gender $\times$ age interactions (since they are not significant in the previous model). We propose the following model for the log odds of obesity:

$$\log\left\{\frac{\Pr(Y_{ij}=1)}{\Pr(Y_{ij}=0)}\right\} = \beta_1 + \beta_2 * \text{gender}_i + \beta_3 * \text{age}_{ij} + \beta_4 * \text{age}_{ij}^2$$

Model without Interactions cont.

We fit this model with the following code:

```
PROC GENMOD DATA=obesity DESCENDING;
     CLASS id occasion;
     MODEL y=gender cage cage2 / DIST=bin LINK=logit;
     REPEATED SUBJECT=id / WITHINSUBJECT=occasion LOGOR=FULLCLUST;
RUN;
```

and obtain the following output:

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	−1.2283	0.0477	−1.3218	−1.1348	−25.75	<.0001
gender	0.1449	0.0627	0.0221	0.2678	2.31	0.0208
cage	0.0418	0.0091	0.0240	0.0596	4.60	<.0001
cage2	−0.0155	0.0023	−0.0200	−0.0110	−6.73	<.0001
Alpha1	3.1496	0.1280	2.8987	3.4004	24.61	<.0001
Alpha2	2.5931	0.1352	2.3281	2.8582	19.17	<.0001
Alpha3	2.9878	0.1236	2.7456	3.2300	24.18	<.0001

Interpreting Results

- The effect of gender is significant and is estimated to be 0.1449. This represents the change in the log odds of being obese for girls compared to boys holding age constant.
- Equivalently, the odds ratio of being obese for girls compared to boys is $e^{0.1449} = 1.16$.
- The estimated effect of age^2 is significant and these results provide evidence that the log odds of obesity increases from 6 to 12 years, levels off between age 12 and 14, and declines between 14 to 18 years.
- Although the rates of obesity are significantly higher for girls at all ages, the patterns of change in rates of obesity over time do not depend on gender.

Interpreting Results cont.

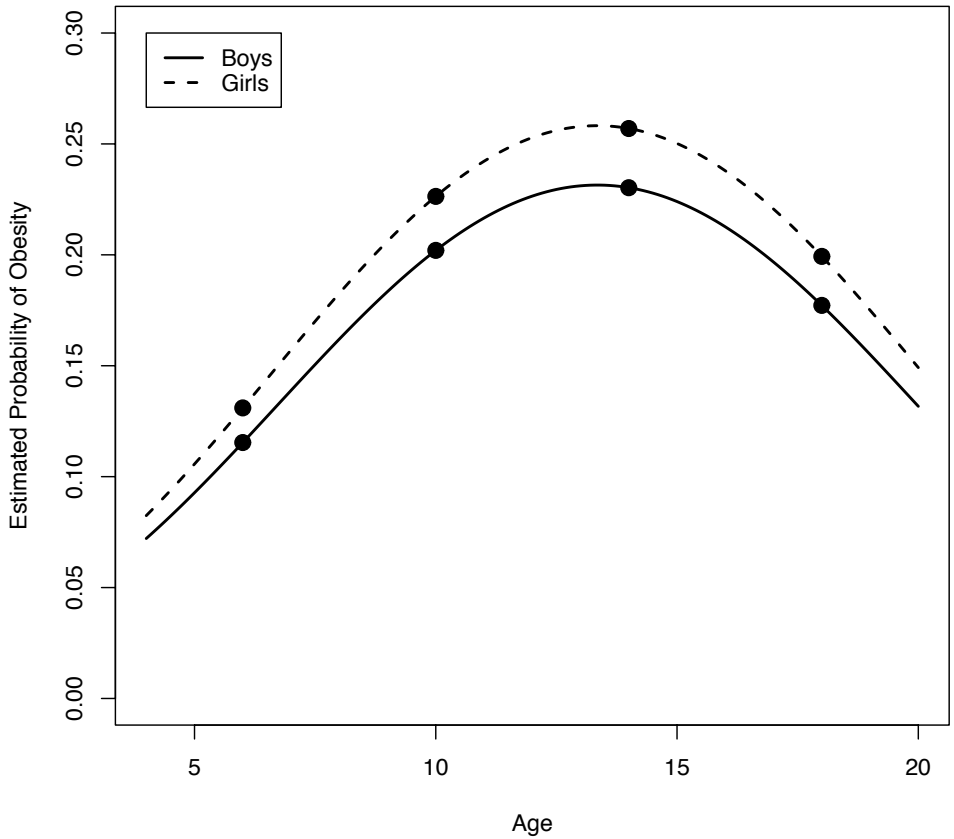
Can also estimate the probability of obesity at each age for boy and girls,

$$P(Y_{ij} = 1 | age_{ij}, gender_i) = \frac{e^{\widehat{\beta}_1 + \widehat{\beta}_2 gender_i + \widehat{\beta}_3 age_{ij} + \widehat{\beta}_4 age_{ij}^2}}{1 + e^{\widehat{\beta}_1 + \widehat{\beta}_2 gender_i + \widehat{\beta}_3 age_{ij} + \widehat{\beta}_4 age_{ij}^2}}$$

To have SAS automatically calculate these probabilities, add "output out=pprobs p=pred;" to your code.

e.g. Estimated probabilities of obesity at ages 6, 10, 14, and 18 are 0.12, 0.20, 0.23, and 0.18 respectively (boys); 0.13, 0.22, 0.26, and 0.20 respectively (girls).
Additive effect of gender (on log odds scale) does not translate into a constant difference over time in the probability of obesity (for logistic model).

Results, graphically



Including Cohort Effects

- Data are from five cohorts of children.
- Analysis to this point has assumed that there are no cohort effects, i.e. that the cross-sectional and longitudinal effects of aging are identical (see slides 22 - 30 in Lecture 13).
- We can formally test this assumption by including linear and quadratic effects of both baseline age and current age minus baseline age (and also their interactions with gender if we wanted).

$$\log\left\{\frac{\Pr(Y_{ij}=1)}{\Pr(Y_{ij}=0)}\right\} = \beta_1 + \beta_2 * \text{gender}_i + \beta_3 * \text{age}_{i1} + \beta_4 * \text{age}_{i1}^2 + \beta_5 * (\text{age}_{ij} - \text{age}_{i1}) + \beta_6 * (\text{age}_{ij}^2 - \text{age}_{i1}^2)$$

Testing Cohort Effects

A test of the equality of the cross-sectional (β_3, β_4) and longitudinal (β_5, β_6) effects of aging is

$$H_0 : (\beta_3 - \beta_5) = (\beta_4 - \beta_6) = 0$$

In contrast form, this can be written:

$$H_0 : \mathbf{L}\boldsymbol{\beta} = \begin{pmatrix} 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{pmatrix} = \begin{pmatrix} \beta_3 - \beta_5 \\ \beta_4 - \beta_6 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

We fit this model and test the above hypothesis with the following code:

```
DATA obesity;
 SET obesity;
 cbaseage=baseage-12;
 cbaseage2=cbaseage*cbaseage;
 agediff=cage-cbaseage;
 agediff2=cage2-cbaseage2;
RUN;

PROC GENMOD DATA=obesity DESCENDING;
     CLASS id occasion;
     MODEL y=gender cbaseage cbaseage2 agediff agediff2/ DIST=bin LINK=logit;
     CONTRAST 'cohort effects' cbaseage 1 agediff -1,
                               cbaseage2 1 agediff2 -1 /wald;
     REPEATED SUBJECT=id / WITHINSUBJECT=occasion LOGOR=FULLCLUST;
RUN;
```

Contrast Results for GEE Analysis

Contrast	DF	Chi-Square	Pr > ChiSq	Type
cohort effects	2	1.74	0.4187	Wald

What do we conclude?

Example 2: Count Data

We now change our focus to another dataset that contains **count data**, the seizure data. Recall this data,

- A randomized clinical trial of 58 epileptic patients
- Response is the number of epileptic seizures
- Measured over an 8-week baseline period (prior to any treatment)
- Then measured in each of four 2-week treatment periods
- Patients received either a placebo or the drug progabide in addition to other therapy.

Goal: is mean seizure rate different in the two groups over time?

Inputting the Data

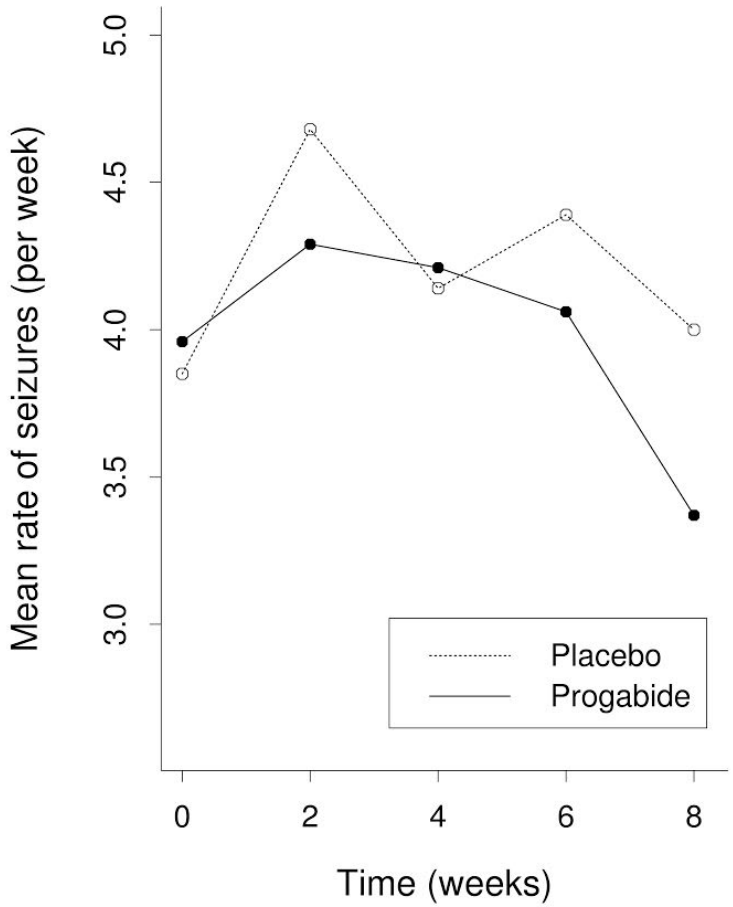
```
DATA seizure;
INFILE "seizure.txt";
INPUT id trt age y1 y2 y3 y4 y5;
y=y1; occasion=1; time=0; OUTPUT;
y=y2; occasion=2; time=1; OUTPUT;
y=y3; occasion=3; time=2; OUTPUT;
y=y4; occasion=4; time=3; OUTPUT;
y=y5; occasion=5; time=4; OUTPUT;
DROP y1-y5;
RUN;


PROC MEANS DATA=seizure maxdec=1 n mean var;
VAR y;
CLASS trt occasion;
RUN;
```

The Data, Summarized

trt	time	Obs	N	Mean	Variance
0	0	28	28	30.8	681.4
	1	28	28	9.4	102.8
	2	28	28	8.3	66.7
	3	28	28	8.8	215.3
	4	28	28	8.0	57.9
1	0	30	30	27.7	303.1
	1	30	30	5.5	33.2
	2	30	30	6.5	31.4
	3	30	30	6.0	54.3
	4	30	30	4.9	18.3

The Data, Graphically



Specifying the Marginal Model

Recall from lecture, a marginal model for longitudinal data has three-part specification:

1. Conditional mean of each response, $E(Y_{ij}|X_{ij}) = \mu_{ij}$ is assumed to depend on the covariates through a known link function g:

$$g(\mu_{ij}) = \beta_1 X_{1ij} + \beta_2 X_{2ij} + ... + \beta_p X_{pij}$$

2. Conditional variance of each Y_{ij}, given the covariates, is assumed to depend on the mean according to

$$\text{Var } (Y_{ij}|X_{ij}) = \phi v(\mu_{ij})$$

3. Given covariates, conditional within-subject association among repeated responses is assumed to be a function of the means and an additional set of association parameters α

Our Marginal Model

1. Conditional mean of each response,

$$\log \mathrm{E}(Y_{ij}) = \log(t_{ij}) + \beta_1 + \beta_2 * \mathrm{TIME}_{ij} + \beta_3 * \mathrm{TRT}_{ij} + \beta_4 * \mathrm{TRT}_{ij} * \mathrm{TIME}_{ij}$$

- We fit a linear trend in time
- TRT is an indicator variable for the progabide group.
- $\log(t_{ij})$ is an offset term (Lecture 15, slides 7-8)
  Our offset depends on j because the time period before baseline is different from the times between each of the repeated measurements

Our Marginal Model cont.

2. Conditional variance model, we assume $\text{Var}(Y_{ij}|X_{ij}) = \phi\mu_{ij}$
   - ϕ allows for overdispersion,

3. We assume the off-diagonal elements of the covariance matrix are all the same, so the within-subject association is accounted for by assuming a common correlation (i.e. 'exchangeable' in SAS):

$$\text{Corr}(Y_{ij}, Y_{ik}) = \alpha$$

SAS code

```
DATA seizure;
   SET seizure;
   IF occasion=1 THEN ltime=log(8);
   ELSE ltime=log(2);
RUN;


PROC GENMOD DATA=seizure;
    CLASS id occasion;
    MODEL y = time trt time*trt / DIST=POISSON LINK=LOG SCALE=PEARSON OFFSET=ltime;
    REPEATED SUBJECT=id / WITHINSUBJECT=occasion TYPE=exch CORRW MODELSE;
RUN;
```

SAS Code Described

- MODEL statement
  1. Specifies entire form of the marginal model for the mean
  2. DIST option specifies appropriate marginal variance for the response
  3. LINK option specifies the link function [default: canonical link]
  4. SCALE=PEARSON option allows for estimation of overdispersion parameter ϕ
  5. OFFSET option specifies variable for offset term in regression

SAS Code Described cont.

- REPEATED statement
  1. Specify independent units of observation via SUBJECT variable; must be also listed in CLASS statement
  2. TYPE option specifies assumed form of working correlation matrix
  3. CORRW option displays estimated working correlation matrix
  4. MODELSE option displays the model-based standard errors (in addition to default empirical standard errors)
  5. WITHINSUBJECT option specifies order of measurement within subjects, as described previously

Selected SAS output

Exchangeable Working Correlation
Correlation 0.5977939731

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		1.3341	0.1581	1.0243	1.6439	8.44	<.0001
time		0.0085	0.0430	−0.0758	0.0927	0.20	0.8442
trt	1	−0.0886	0.1950	−0.4708	0.2936	−0.45	0.6494
time*trt	1	−0.0706	0.0560	−0.1803	0.0391	−1.26	0.2072

Analysis Of GEE Parameter Estimates
Model−Based Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		1.3341	0.1110	1.1165	1.5517	12.01	<.0001
time		0.0085	0.0399	−0.0697	0.0866	0.21	0.8322
trt	1	−0.0886	0.1579	−0.3980	0.2208	−0.56	0.5744
time*trt	1	−0.0706	0.0615	−0.1911	0.0499	−1.15	0.2508
Scale		3.2565	.	.	.	.	.

NOTE: The scale parameter for GEE estimation was computed as the square root of the normalized Pearson's chi−square.

Output Interpretation

- From the estimated interaction term, we can see that there is no significant difference in the change in log mean seizure rates over time between the two treatment groups. This is true whether we use the empirical or model-based standard error estimates ($p_{emp} = 0.2508$ and $p_{mb} = 0.2072$).
- From the estimated scale parameter, it does appear that there is a fair bit of over-dispersion ($\hat{\phi} = 3.2565^2 = 10.60$). So, on average, variances at any timepoint are roughly 10 times the mean at that timepoint.
- Finally, the estimated correlation between any subject's pair of observations is $\hat{\alpha} = 0.5978$.

Fix the scale parameter

To illustrate how the results change if we fix the scale parameter, we will slightly adjust the previous analysis by fixing the scale parameter (ϕ) to be known and equal to 1.

- If $\phi = 1$, this implies that we do not have overdispersion relative to Poisson variability, **which is unlikely**
- In SAS we can fix the scale parameter using the SCALE option on the MODEL statement:

```
PROC GENMOD DATA=seizure;
    CLASS id occasion;
    MODEL y = time trt time*trt / DIST=POISSON LINK=LOG SCALE=1 OFFSET=ltime;
    REPEATED SUBJECT=id / WITHINSUBJECT=occasion TYPE=exch CORRW MODELSE;
RUN;
```

Results

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	1.3341	0.1581	1.0243	1.6439	8.44	<.0001
time	0.0085	0.0430	−0.0758	0.0927	0.20	0.8442
trt	−0.0886	0.1950	−0.4708	0.2936	−0.45	0.6494
time*trt	−0.0706	0.0560	−0.1803	0.0391	−1.26	0.2072

Analysis Of GEE Parameter Estimates
Model−Based Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	1.3341	0.0341	1.2673	1.4009	39.13	<.0001
time	0.0085	0.0123	−0.0156	0.0325	0.69	0.4902
trt	−0.0886	0.0485	−0.1836	0.0064	−1.83	0.0675
time*trt	−0.0706	0.0189	−0.1076	−0.0336	−3.74	0.0002
Scale	1.0000	.	.	.	.	.

NOTE: The scale parameter was held fixed.

Output Interpretation

- The mean model parameter estimates remain the same.
- However, the model-based standard errors are much smaller in the analysis that assumed a known $\phi = 1$. Why? Because the standard errors are multiplied by the estimated scale parameter.

Output Interpretation (cont.)

How does not allowing for overdispersion affect your inference?

- The standard errors have changed so much with scale set equal to one that the TIME*TRT interaction is significant in this analysis based on the model-based results. If we were to make inferences based on the corresponding model-based results, we would conclude that the change in log seizure rates over time between the two treatment groups is not the same.
- Since the sign associated with the interaction term is negative, we would conclude that change in the log rate of seizures over time for subjects in the progabide group is significantly lower than the change for subjects in the placebo group. Although all of the previous analyses suggest results that are in the same direction, none of those results were significant.

First Model without Interaction

Running our first model without the interaction term results in the following SAS output:

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	1.3180	0.1604	1.0037	1.6324	8.22	<.0001
time	−0.0232	0.0293	−0.0807	0.0342	−0.79	0.4282
trt	−0.0567	0.1985	−0.4457	0.3322	−0.29	0.7750

Parameter Interpretation

The expected seizure rate at baseline is $e^{1.3180} = 3.74$ in the placebo group and $e^{1.3180-0.0567} = 3.53$ in the Progabide group.

A two week (one unit) increase in time decreases the rate of seizure *multiplicatively* by a factor of $e^{-0.0232} = 0.977$.

Here, we see that the main effects are non-significant.