





# Chapter 15 Generalized Estimating Equations

## 15.1 Introduction

- WLS methods can only accommodate categorical explanatory variables, can't easily handle missing values, and have restrictive sample size requirements.
- GEE approach is an extension of generalized linear models that provides a semi-parametric approach to longitudinal data analysis with univariate outcomes for which the quasi-likelihood formulation is sensible.


- 
- Scope for GEE strategy useful in many situations:
    1. a two-period crossover study in which researchers study the effects of two treatments and a placebo
    2. a longitudinal study on the efficacy of a new drug designed to prevent fractures in the elderly. The outcome of interest is the number of fractures that occur.
    3. a large study on the effects of air pollution on children in which measurements on respiratory symptoms are taken every year for three years. Many children have one or two measurements missing.

- 
- In this chapter, GEE for analysis of repeated data is discussed, with examples. GEE methods for analysis of some univariate response outcomes is also included.

## **15.2 Methodology**

### **15.2.1 Motivation**

Within-subject factors (visit, time) are likely to have correlated measurements, while between-subject factors (age, gender) are likely to have independent measurements. If correlation is not taken into account, it will lead to incorrect standard errors.



GEE can handle continuous explanatory variables, large number of categorical variables, missing response values, and/or time-dependent covariates, which the WLS approach cannot.

GEE introduced as a way of handling correlated data that, except for the correlation among responses, can be modeled with a GLM. Ideal for discrete response data such as binary outcomes and Poisson counts. They work for longitudinal studies and cluster sampling data.

Difference between GLM and GEE is that, with GEE, you account for the structure of the covariances of the responses through its specification in the estimating process.



### 15.2.2 Generalized Linear Models (GLM)

GLM relates a mean response to a vector of explanatory variables through a link function:

$$g(E(y_i)) = g(\mu_i) = \mathbf{x}_i' \boldsymbol{\beta}$$

where  $y_i$  is a response variable ( $i = 1, \dots, n$ ),  $\mu_i = E(y_i)$ ,  $g$  is a link function,  $\mathbf{x}_i$  is a vector of independent variables, and  $\boldsymbol{\beta}$  is a vector of regression parameters to be estimated.

- The variance of  $y_i$  is  $v_i = v_i(\mu_i)$  and is a specified function of its mean  $\mu_i$ .
- The  $y_i$  are from the exponential family (includes binomial, Poisson, normal, gamma, and inverse normal distributions). When normal distribution is assumed, the identity link function,  $g(\mu_i) = \mu_i$ , produces same model as GLM.





For logistic regression, the link and variance functions are:

$$g(\mu) = \log \left\{ \frac{\mu}{1-\mu} \right\} \text{ and } v(\mu) = \mu(1-\mu)$$

For Poisson regression, the link and variance functions are:

$$g(\mu) = \log(\mu) \text{ and } v(\mu) = \mu$$

Maximum likelihood estimator  $\hat{\beta}$  is obtained by solving estimating equations (usually by iteration), which are score equations shown below. These estimators also maximize the log likelihood.

$$\sum_{i=1}^n \frac{\partial \mu_i}{\partial \beta} v_i^{-1} (y_i - \mu_i(\beta)) = \mathbf{0}$$




### 15.2.3 GEE Methodology

Suppose:

- data obtained from each of  $n$  subjects at  $t_i$  timepoints
- $y_{ij}$  denotes the response from subject  $i$  at time  $j$ ,  
for  $i = 1, \dots, n$  and  $j = 1, \dots, t_i$ .
- $\mathbf{x}_{ij} = (x_{ij1}, \dots, x_{ijp})'$  denotes a  $p \times 1$  vector of explanatory variables.

Assume that you have chosen a model that relates a marginal mean to the linear predictor  $\mathbf{x}_{ij}'\boldsymbol{\beta}$  through a link function. The GEE for estimating  $\boldsymbol{\beta}$ :

$$\sum_{i=1}^n \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1} (\mathbf{Y}_i - \boldsymbol{\mu}_i(\boldsymbol{\beta})) = \mathbf{0}$$



Where  $\boldsymbol{\mu}_i$  is the corresponding vector of means  $\boldsymbol{\mu}_i = (\mu_{i1}, \dots, \mu_{it_i})'$ ,  $\mathbf{Y}_i = (y_{i1}, \dots, y_{it_i})$ , and  $\mathbf{V}_i$  is an estimator of the covariance of  $\mathbf{Y}_i$ .

$$\mathbf{V}_i = \phi \mathbf{A}_i^{1/2} \mathbf{R}_i(\boldsymbol{\alpha}) \mathbf{A}_i^{1/2}$$

Where  $\mathbf{A}_i$  is a  $t_i \times t_i$  diagonal matrix with  $v(\mu_{ij})$  as the  $j$ th diagonal element.

$\mathbf{R}_i(\boldsymbol{\alpha})$  is the working correlation matrix. The  $(j, j')$  element of  $\mathbf{R}_i(\boldsymbol{\alpha})$  is the known, hypothesized, or estimated correlation between  $y_{ij}$  and  $y_{ij'}$ . This working correlation matrix may depend on a vector of unknown parameters  $\boldsymbol{\alpha}$ , which is the same for all subjects. Assume  $\mathbf{R}_i(\boldsymbol{\alpha})$  is known except for a fixed number of parameters  $\boldsymbol{\alpha}$  that must be estimated from the data.






## Choosing the Working Correlation Matrix:

Fixed:  $\mathbf{R}_i(\alpha) = \mathbf{R}_0$

If one uses  $\mathbf{R}_0 = \mathbf{I}$ , one gets working independence.

$$\text{Exchangeable: } \text{Corr}(Y_{ij}, Y_{ij'}) = \begin{cases} 1 & j = j' \\ \alpha & j \neq j' \end{cases}$$



Autoregressive AR(1):  $Corr(Y_{ij}, Y_{i,j+s}) = \alpha^s$   
for  $s = 0, 1, \dots, t_i - j$

$$m\text{-dependent: } Corr(Y_{ij}, Y_{i,j+s}) = \begin{cases} 1 & s = 0 \\ \alpha_s & s = 1, 2, \dots, m \\ 0 & s > m \end{cases}$$

$$\text{Unstructured: } Corr(Y_{ij}, Y_{ij'}) = \begin{cases} 1 & j = j' \\ \alpha_{jj'} & j \neq j' \end{cases}$$



## **GEE Fitting algorithm:**

1. Compute an initial estimate of  $\beta$ , for example from a GLM model (or quasi-likelihood)
2. Compute the standardized Pearson residuals

$$r_{ij} = \frac{y_{ij} - \hat{\mu}_{ij}}{\sqrt{v(\hat{\mu}_{ij})}}$$

and get estimates for the nuisance parameters  $\phi$  and  $\alpha$  using moment estimation.



3. Compute  $\hat{V}_i = V_i(\hat{\boldsymbol{\beta}}, \hat{\phi}, \hat{\alpha})$

4. Update the estimate for  $\boldsymbol{\beta}$  with:

$$\hat{\boldsymbol{\beta}} - \left( \sum_{i=1}^n \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \boldsymbol{\beta}} \hat{V}_i^{-1} \frac{\partial \hat{\boldsymbol{\mu}}_i}{\partial \boldsymbol{\beta}} \right)^{-1} \left( \sum_{i=1}^n \frac{\partial \hat{\boldsymbol{\mu}}_i'}{\partial \boldsymbol{\beta}} \hat{V}_i^{-1} (\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i) \right),$$

a modification of the Fisher's scoring method

5. Iterate until convergence



The model-based (naïve) estimator for the covariance of  $\hat{\beta}$  :

$$\text{Cov}_M(\hat{\beta}) = \hat{I}_0^{-1}$$

$$\text{where } I_0 = \sum_{i=1}^n \frac{\partial \mu_i'}{\partial \beta} V_i^{-1} \frac{\partial \mu_i}{\partial \beta}$$

This is a consistent estimator only if the model is correctly specified (mean function and marginal variances) and the working correlation structure is correct.





The empirical (robust) estimator for the covariance of  $\hat{\beta}$ :

$$\text{Cov}_R(\hat{\beta}) = \hat{I}_0^{-1} \hat{I}_1 \hat{I}_0^{-1}$$

$$\text{where } \mathbf{I}_1 = \sum_{i=1}^n \frac{\partial \boldsymbol{\mu}_i'}{\partial \boldsymbol{\beta}} \mathbf{V}_i^{-1} \text{Cov}(\mathbf{Y}_i) \mathbf{V}_i^{-1} \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$$

and  $\text{Cov}(\mathbf{Y}_i)$  is estimated by  $(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{Y}_i - \hat{\boldsymbol{\mu}}_i)'$

This estimator is consistent when the mean function is correctly specified even if the working correlation model and/or the model for the marginal variances is incorrectly specified.

Of course, such misspecification causes a loss of efficiency of  $\hat{\beta}$ , but the asymptotic efficiency is expected to be high if the working correlation structure is approximately correct.



## Multiple Linear Regression


$$\mu_{ij} = \mathbf{x}'_{ij}\boldsymbol{\beta} \quad \text{where } \mu_{ij} = E(Y_{ij})$$

$$\text{Var}(Y_{ij}) = \sigma^2 \quad \text{Not a function of } \mu_{ij}$$

$$\mathbf{R}_i(\alpha) = \mathbf{I}_{t_i} \quad \text{"working independence"}$$

$$\mathbf{V}_i = \sigma^2 \mathbf{I}_{t_i}$$

$$\sum_{i=1}^n \mathbf{X}'_i (\sigma^2 \mathbf{I}_{t_i})^{-1} (\mathbf{Y}_i - \mathbf{X}_i \boldsymbol{\beta}) = 0$$


$$\sum_{i=1}^n X_i'(Y_i - X_i\beta) = 0$$

$$X'(Y - X\beta) = 0$$

$$X'Y = X'\hat{Y}$$

$$\text{Cov}_M(\hat{\beta}) = \hat{\sigma}^2 (X'X)^{-1}$$

$$\text{Cov}_R(\hat{\beta}) = (X'X)^{-1} \left( \sum_{i=1}^n X_i'(Y_i - \hat{Y}_i)(Y_i - \hat{Y}_i)' X_i \right) (X'X)^{-1}$$

## Logistic Regression

$$\text{logit}(\pi_{ij}) = \mathbf{x}_{ij}'\boldsymbol{\beta} \quad \text{where } \pi_{ij} = E(Y_{ij})$$

$$\text{Var}(Y_{ij}) = \pi_{ij}(1 - \pi_{ij}) \quad \phi = 1$$

$$\mathbf{R}_i(\alpha) = \mathbf{I}_{t_i} \quad \text{"working independence"}$$

$$\mathbf{V}_i = \mathbf{A}_i = \text{Diag}[\pi_{ij}(1 - \pi_{ij})]$$

$$\sum_{i=1}^n \mathbf{X}_i'(\mathbf{Y}_i - \boldsymbol{\pi}_i) = 0$$

$$X'(Y - \pi) = 0$$

$$X'Y = X'\hat{\pi}$$

$$Cov_M(\hat{\beta}) = (X'Diag[\hat{\pi}_{ij}(1 - \hat{\pi}_{ij})]X)^{-1}$$

$$Cov_R(\hat{\beta}) = (X'Diag[\hat{\pi}_{ij}(1 - \hat{\pi}_{ij})]X)^{-1} \times$$

$$\left( \sum_{i=1}^n X'_i(Y_i - \hat{\pi}_i)(Y_i - \hat{\pi}_i)'X_i \right) \times$$

$$(X'Diag[\hat{\pi}_{ij}(1 - \hat{\pi}_{ij})]X)^{-1}$$






## **15.4 Passive Smoking Example**

Hypothetical study of the effect of air pollution on children. Wheezing symptoms of 25 children recorded at ages 8, 9, 10, 11. Response is 1 for symptoms, 0 for no symptoms. Explanatory variables include age, city, and a passive smoking index with values 0, 1, 2 reflecting degree of smoking in the home.

See page 497-499 for dataset, SAS data manipulations.

```
proc genmod data=children descending;  
    class id city;  
    model symptom = city age smoke / link=logit  
        dist=bin type3;  
    repeated subject=id / type=exch covb corrw;  
run;
```



GEE analysis is requested with the REPEATED statement. SUBJECT=ID identifies the clustering variable. Note that ID must be listed in the CLASS statement, and needs to have a unique value for each cluster (subject). TYPE=EXCH specifies the exchangeable working correlation matrix, COVB requests the parameter estimate covariance matrix, and CORRW requests the final working correlation matrix.

**Output 15.1** Basic Model Information

Model Information	
Data Set	WORK.CHILDREN
Distribution	Binomial
Link Function	Logit
Dependent Variable	symptom
Observations Used	100
Probability Modeled	Pr( symptom = 1 )

## **Output 15.2** Class Levels and Response Profiles

### Class Level Information

Class	Levels	Values
id	25	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
city	2	greenhil steelcit

### Response Profile

Ordered Level	Ordered Value	Count
1	1	42
2	0	58

### **Output 15.3** Information about Parameters

#### Parameter Information

Parameter	Effect	city
Prm1	Intercept	
Prm2	city	greenhil
Prm3	city	steelcit
Prm4	age	
Prm5	smoke	

## **Output 15.4 Initial Parameter Estimates**

### Analysis of Initial Parameter Estimates

Parameter		DF	Estimate	Standard Error	Wald	95% Confidence Limits	Chi-Square
Intercept		1	2.4161	1.8673	-1.2438	6.0760	1.67
city	greenhil	1	0.0017	0.4350	-0.8508	0.8543	0.00
city	steelcit	0	0.0000	0.0000	0.0000	0.0000	.
age		1	-0.3283	0.1914	-0.7035	0.0468	2.94
smoke		1	0.5598	0.2952	-0.0188	1.1385	3.60
Scale		0	1.0000	0.0000	1.0000	1.0000	

### Analysis of Initial Parameter Estimates

Parameter	Pr > ChiSq
Intercept	0.1957
city greenhil	0.9968
city steelcit	.
age	0.0863
smoke	0.0579
Scale	

NOTE: The scale parameter was held fixed.





### **Output 15.5**    General GEE Model Information

#### GEE Model Information

Correlation Structure	Exchangeable
Subject Effect	id (25 levels)
Number of Clusters	25
Correlation Matrix Dimension	4
Minimum Cluster Size	4
Maximum Cluster Size	4


## **Output 15.6 and 15.7 GEE Parameter Estimates, Type 3 Analysis**

Analysis of GEE Parameter Estimates Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr >  z
Intercept		2.2615	2.0243	-1.7060	6.2290	1.12	0.2639
city	greenhil	0.0418	0.5435	-1.0234	1.1070	0.08	0.9387
city	steelcit	0.0000	0.0000	0.0000	0.0000	.	.
age		-0.3201	0.1884	-0.6894	0.0492	1.70	0.0893
smoke		0.6506	0.2821	0.0978	1.2035	2.31	0.0211

Score Statistics for Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
city	1	0.01	0.9388
age	1	2.74	0.0981
smoke	1	3.59	0.0583



The above results indicate that city is not a factor in wheezing status, but smoking exposure has a borderline significant association ( $p = 0.058$ ). Age is marginally influential ( $p = 0.098$ ).

Note that the  $p$ -value for the  $Z$  for smoking is 0.0211, compared to the 0.058 reported with the score statistic in the Type 3 analysis. In a strict testing situation, we used the more accurate score statistic. The  $Z$  and Wald statistics generally produce more inaccurate  $p$ -values.

Odds of symptoms for those with on higher category of smoking exposure are  $\exp(0.6506) = 1.9$  times the odds of symptoms for those children with the lower exposure.




## **15.5 Using a Modified Wald Statistic to Assess Model Effects**

Shah, Holt & Folsom (1977) describe a modification of the Wald Statistic based on a Hotelling  $T^2$  type of transformation of  $Q_C$ :

$$\frac{(n-c)Q_C}{c(n-1)} \text{ is distributed as } F_{c, n-c}$$

The quantity  $n$  is equal to the number of clusters, and  $c$  is equal to the number of rows of the contrast. Thus, for tests concerning effects of explanatory factors, it is equal to the corresponding number of df. This test is more conservative than the Wald test.



See page 555 (Chapter 15, Appendix B) for macro (GEEF) that produces the  $F$ -transform statistics and appends them to a table at the end of PROC GENMOD output.

```
%include 'macros.sas';  
proc genmod descending data=children;  
    class id city;  
    model symptom = city age smoke / link=logit dist=bin  
        type3 wald;  
    repeated subject=id / type=exch covb corrw;  
ods output GEEModInfo=clustout Type3=scoreout;  
run;  
%geef;
```




## **Output 15.8**    Type 3 Analysis

### Wald Statistics for Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
city	1	0.01	0.9387
age	1	2.89	0.0893
smoke	1	5.32	0.0211

### F-Statistics for Type 3 GEE Analysis

Source	DF	F Value	Pr > F
city	1	0.01	0.9393
age	1	2.89	0.1023
smoke	1	5.32	0.0300



Note that  $F$  statistics are same as Wald statistics for single df tests. However, all  $p$ -values are more conservative, which might be preferable with small numbers of clusters, especially if you have marginal significance.

## 15.6 Crossover Example

Subjects serve as their own controls and receive two or more treatments in two or more consecutive periods. Following data is from a two-period cross-over study investigating three treatments (analyzed with conditional logistic regression in Chapter 10):

Table 15.1 Crossover Design Data

Age	Sequence	Response Profiles				Total
		FF	FU	UF	UU	
older	A:B	12	12	6	20	50
older	B:P	8	5	6	31	50
older	P:A	5	3	22	20	50
younger	B:A	19	3	25	3	50
younger	A:P	25	6	6	13	50
younger	P:B	13	5	21	11	50



See pages 507-508 for data entry and manipulation.

```
proc genmod data=cross2;  
    class subject age drug carry;  
    model response = period age drug period*age  
        carry drug*age / link=logit dist=bin type3;  
    repeated subject=subject / type=unstr;  
run;
```

## Output 15.9 Class Level Information

Class	Levels	Values
Subject	300	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 ...
age	2	older younger
drug	3	A B P
carry	3	A B N

### Response Profile

Ordered Level	Ordered Value	Count
1	F	284
2	U	316

## Output 15.10    Parameter Information

Parameter Information			
Parameter	Effect	age	drug      carry
Prm1	Intercept		
Prm2	period		
Prm3	age	older	
Prm4	age	younger	
Prm5	drug		A
Prm6	drug		B
Prm7	drug		P
Prm8	period*age	older	
Prm9	period*age	younger	
Prm10	carry		A
Prm11	carry		B
Prm12	carry		N
Prm13	age*drug	older	A
Prm14	age*drug	older	B
Prm15	age*drug	older	P
Prm16	age*drug	younger	A
Prm17	age*drug	younger	B
Prm18	age*drug	younger	P



### **Output 15.11** GEE Model Information


#### GEE Model Information

Correlation Structure	Unstructured
Subject Effect	subject (300 levels)
Number of Clusters	300
Correlation Matrix Dimension	2
Minimum Cluster Size	2
Maximum Cluster Size	2

### **Output 15.12** Type 3 Analysis

#### Score Statistics for Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
period	1	4.61	0.0318
age	1	36.03	< .0001
drug	2	27.66	< .0001
period*age	1	4.69	0.0303
carry	2	1.15	0.5626
age*drug	2	0.72	0.6981



From Output 15.12, we see that we have no carry-over effect ( $p = 0.56$ ). In addition, the interaction of age  $\times$  drug appears to be unimportant ( $p = 0.70$ ).

We can also use a CONTRAST statement to get joint test for CARRY and AGE\*DRUG:

```
ods select contrasts;
proc genmod data=cross2;
    class subject age drug carry;
    model response = period age drug period*age carry drug*age
        / link=logit dist=bin type3;
    repeated subject=subject / type=unstr;
    contrast 'joint' carry 1 0 -1,
        carry 0 1 -1,
        age*drug 1 0 -1 -1 0 1,
        age*drug 0 1 -1 0 -1 1;
run;
```

### **Output 15.13**    Type 3 Analysis

#### Contrast Results for GEE Analysis

Contrast	DF	Chi-Square	Pr > ChiSq	Type
joint	4	1.31	0.8595	Score

### Reduced Model:

```
proc genmod data=cross2;  
    class subject age drug;  
    model response = period age drug period*age  
        / dist=bin type3;  
    repeated subject=subject / type=unstr corrw;  
run;
```

### **Output 15.14**    Type 3 Table

#### Score Statistics for Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
Period	1	24.98	< .0001
Age	1	35.53	< .0001
Drug	2	39.31	< .0001
Period*age	1	5.10	0.0240

### **Output 15.15**    Parameter Estimates

#### Analysis of GEE Parameter Estimates Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr >  Z
Intercept		0.5127	0.2063	0.1084	0.9170	2.49	0.0129
period		-1.1553	0.2304	-1.6069	-0.7037	-5.01	< .0001
age	older	-1.4994	0.2583	-2.0056	-0.9931	-5.80	< .0001
age	younger	0.0000	0.0000	0.0000	0.0000	.	.
drug	A	1.2542	0.2010	0.8602	1.6483	6.24	< .0001
drug	B	0.3404	0.2016	-0.0546	0.7355	1.69	0.0912
drug	P	0.0000	0.0000	0.0000	0.0000	.	.
period*age	older	0.7088	0.3131	0.0951	1.3224	2.26	0.0236
period*age	younger	0.0000	0.0000	0.0000	0.0000	.	.

### **Output 15.16** Working Correlation Matrix

Working Correlation Matrix		
	Col1	Col2
Row1	1.0000	0.2274
Row2	0.2274	1.0000

To compare two drugs (A vs. B):

```
ods select Contrasts;  
proc genmod data=cross2;  
    class subject age drug;  
    model response = period age drug period*age / dist=bin type3;  
    repeated subject=subject / type=unstr;  
    contrast 'A versus B' drug 1 -1 0;  
run;
```

### **Output 15.17** Contrast Results

Contrast Results for GEE Analysis				
Contrast	DF	Chi-Square	Pr > ChiSq	Type
A versus B	1	19.15	< .0001	Score



## **15.7 Respiratory Data**

Response is whether the outcome was good or excellent vs. all other responses; modeled as logit. Explanatory variables: treatment, center, visit. Also age, sex, baseline status couldn't be handled in the WLS repeated measurements setting (sample size inadequate). GEE enables us to take these variables into account.

First, use same model that resulted from WLS.

```
proc genmod descending;  
    class id center treatment visit;  
    model dichot = center treatment visit  
        / link=logit dist=bin type3;  
    repeated subject=id*center / type=unstr;  
  
run;
```



The cross-product of ID and CENTER is used in REPEATED statement because unique identification of clusters requires both.

**Output 15.18** Model Information

Model Information	
Data Set	WORK.RESP2
Distribution	Binomial
Link Function	Logit
Dependent Variable	dichot
Observations Used	444
Probability Modeled	Pr (dichot = 1)

**Output 15.19** Response Profiles

Response Profile		
Ordered Level	Ordered Value	Count
1	1	248
2	0	196

## **Output 15.20**   GEE Model Information

### GEE Model Information

Correlation Structure	Unstructured
Subject Effect	id*center (111 levels)
Number of Clusters	111
Correlation Matrix Dimension	4
Minimum Cluster Size	4
Maximum Cluster Size	4

## **Output 15.21**   Type 3 Analysis

### Score Statistics for Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
center	1	8.01	0.0047
treatment	1	9.89	0.0017
visit	3	3.47	0.3251

## **Output 15.22 Parameter Estimates**

Analysis of GEE Parameter Estimates  
Empirical Standard Error Estimates


Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr >  Z
Intercept		0.0732	0.2946	-0.5042	0.6506	0.25	0.8039
center	1	-0.9168	0.3157	-1.5355	-0.29982	-2.90	0.0037
center	2	0.0000	0.0000	0.0000	0.0000	.	.
treatment	A	1.0145	0.3165	0.3941	1.6349	3.21	0.0013
treatment	P	0.0000	0.0000	0.0000	0.0000	.	.
visit	1	0.2835	0.2094	-0.1269	0.6939	1.35	0.1757
visit	2	0.0804	0.2053	-0.3220	0.4829	0.39	0.6953
visit	3	0.2840	0.1932	-0.0946	0.6627	1.47	0.1415
visit	4	0.0000	0.0000	0.0000	0.0000	.	.

**Table 15.3 Comparison of WLS and GEE Estimates**

	WLS		GEE	
Parameter	Estimate	S.E.	Estimate	S.E.
Intercept	0.0168	0.2901	0.0732	0.2946
Treatment A	1.0434	0.3203	1.0145	0.3165
Visit 1	0.2216	0.1883	0.2835	0.2094
Visit 2	0.0201	0.1896	0.0804	0.2053
Visit 3	0.1811	0.1681	0.2840	0.1932
Center	-0.8803	0.3188	-0.9168	0.3157

Estimates and standard errors for WLS and GEE analyses with cluster sizes of at least 400 are very similar.

GEE analysis can handle additional variables reasonably well with 111 clusters.



Include age and sex as main effects, and visit\*treatment and treatment\*center interactions:

```
proc genmod descending;  
    class id center sex treatment visit;  
    model dichot = treatment sex age center di_base visit  
        visit*treatment treatment*center / link=logit dist=bin  
        type3;  
    repeated subject=id*center / type=exch;  
run;
```

### **Output 15.23** Model Information

Model Information	
Data Set	WORK.RESP2
Distribution	Binomial
Link Function	Logit
Dependent Variable	dichot
Observations Used	444
Probability Modeled	Pr (dichot = 1)

### **Output 15.24** GEE Model Information

#### GEE Model Information

Correlation Structure	Exchangeable
Subject Effect	id*center (111 levels)
Number of Clusters	111
Correlation Matrix Dimension	4
Minimum Cluster Size	4
Maximum Cluster Size	4

### **Output 15.25** Type 3 Tests for Model with Interactions

#### Score Statistics for Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
treatment	1	12.85	0.0003
sex	1	0.24	0.6247
age	1	2.23	0.1351
center	1	3.32	0.0683
di_base	1	23.06	< 0.0001
visit	3	3.33	0.3429
treatment*visit	3	3.10	0.3760
center*treatment	1	2.46	0.1169



Since both interactions not significant, drop them from the model:

```
proc genmod descending;  
    class id center sex treatment visit;  
    model dichot = center sex treatment age di_base visit  
        / link=logit dist=bin type3;  
    repeated subject=id*center / type=exch;  
run;
```

### **Output 15.26** Type 3 Tests for Reduced Model

#### Score Statistics for Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
center	1	3.24	0.0720
sex	1	0.10	0.7565
treatment	1	12.11	0.0005
age	1	2.24	0.1345
di_base	1	22.53	< 0.0001
visit	3	3.47	0.3251

Can further reduce model by removing VISIT:

```
proc genmod descending;  
    class id center sex treatment;  
    model dichot = center sex treatment age di_base  
        / link=logit dist=bin type3;  
    repeated subject=id*center / type=exch corrw;  
run;
```

**Output 15.27** Type 3 Tests for Final Model

Score Statistics for Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
center	1	3.11	0.0780
sex	1	0.10	0.7562
treatment	1	12.52	0.0004
age	1	2.28	0.1312
di_base	1	22.97	< 0.0001

Note the very significant treatment effect; as seen in Output 15.28, active treatment increases the odds of a good or excellent response. Baseline is also very influential. Sex and age remain non-significant, and center is marginally influential.

### **Output 15.28**    Parameter Estimates

Analysis of GEE Parameter Estimates Empirical Standard Error Estimates							
Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr >  Z
Intercept		-0.2066	0.5776	-1.3388	0.9255	-0.36	0.7206
center	1	-0.6495	0.3532	-1.3418	0.0428	-1.84	0.0660
center	2	0.0000	0.0000	0.0000	0.0000	.	.
sex	F	0.1368	0.4402	-0.7261	0.9996	0.31	0.7560
sex	M	0.0000	0.0000	0.0000	0.0000	.	.
treatment	A	1.2654	0.3467	0.5859	1.9448	3.65	0.0003
treatment	P	0.0000	0.0000	0.0000	0.0000	.	.
age		-0.0188	0.0130	-0.0442	0.0067	-1.45	0.1480
di_base		1.8457	0.3460	1.1676	2.5238	5.33	< .0001

### **Output 15.29** Working Correlation Matrix

Working Correlation Matrix				
	Col1	Col2	Col3	Col4
Row1	1.0000	0.3270	0.3270	0.3270
Row2	0.3270	1.0000	0.3270	0.3270
Row3	0.3270	0.3270	1.0000	0.3270
Row4	0.3270	0.3270	0.3270	1.0000

Try unstructured working correlation matrix, since there are 4 visits per subject. Note that this requires responses to be in a consistent order. That is, the first observation in a cluster contains the first response, the following the second response, etc.

```

proc genmod descending;
    class id center sex treatment visit;
    model dichot = center sex treatment age di_base
        / link=logit dist=bin type3;
    repeated subject=id*center / type=unstr corrw;
run;

```

### **Output 15.30**    Unstructured Working Correlation Matrix

Working Correlation Matrix				
	Col1	Col2	Col3	Col4
Row1	1.0000	0.3351	0.2140	0.2953
Row2	0.3351	1.0000	0.4429	0.3581
Row3	0.2140	0.4429	1.0000	0.3964
Row4	0.2953	0.3581	0.3964	1.0000



### **Output 15.31** Parameter Estimates for Unstructured Correlation Matrix

Analysis Of GEE Parameter Estimates  
Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr >  Z
Intercept		-0.2324	0.5763	-1.3620	0.8972	-0.40	0.6868
center	1	-0.6558	0.3512	-1.3442	0.0326	-1.87	0.0619
center	2	0.0000	0.0000	0.0000	0.0000	.	.
sex	F	0.1128	0.4408	-0.7512	0.9768	0.26	0.7981
sex	M	0.0000	0.0000	0.0000	0.0000	.	.
treatment	A	1.2442	0.3455	0.5669	1.9214	3.60	0.0003
treatment	P	0.0000	0.0000	0.0000	0.0000	.	.
age		-0.0175	0.0129	-0.0427	0.0077	-1.36	0.1728
di_base		1.8981	0.3441	1.2237	2.5725	5.42	< .0001

Note that estimates are very similar to those in Output 15.28, and standard errors are a little smaller. This gives very little gain in efficiency. The choice of working correlation structure depends on what you believe is most realistic for you data.





## **15.9 Using GEE for Count Data**

Examples:

- 1) number of acute pain episodes in a time interval
- 2) number of insurance claims during specific year
- 3) number of unscheduled medical visits during study

Can fit Poisson regression using GEE methods.

Example:

In a double-blind study, women past menopause were randomized to a new drug to treat osteoporosis or placebo. Both groups were given calcium supplements, nutritional counseling, and encouraged to exercise. The study ran for 3 years, and the number of fractures in each of those years was recorded. See pages 528-529 for data.

```

proc genmod;
    class id treatment center year;
    model fractures = center treatment age year treatment*center
        treatment*year / dist=poisson type3 offset=1months;
    repeated subject=id*center / type=exch corrw;
run;

```

### **Output 15.32** Model Information

Model Information	
Data Set	WORK.FRACTURE2
Distribution	Poisson
Link Function	Log
Dependent Variable	fractures
Offset Variable	1months
Observations Used	642

### **Output 15.33** GEE Model Information

#### GEE Model Information

Correlation Structure	Exchangeable
Subject Effect	id*center (214 levels)
Number of Clusters	214
Correlation Matrix Dimension	3
Minimum Cluster Size	3
Maximum Cluster Size	3

### **Output 15.34** Type 3 Tests for Model with Interactions

#### Score Statistics for Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
center	1	0.02	0.8750
treatment	1	4.69	0.0303
age	1	2.44	0.1180
year	2	7.64	0.0220
treatment*center	1	0.04	0.8364
treatment*year	2	3.15	0.2074

Same analysis repeated with main effects only:


**Output 15.35** Type 3 Tests for Reduced Model

Score Statistics for Type 3 GEE Analysis			
Source	DF	Chi-Square	Pr > ChiSq
center	1	0.02	0.8930
treatment	1	3.41	0.0647
age	1	2.22	0.1359
year	2	4.71	0.0948

## **Output 15.36 Parameter Estimates**

Analysis of GEE Parameter Estimates  
Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr >  Z
Intercept		-6.6379	1.1201	-8.8333	-4.4424	-5.93	< .0001
center	A	0.0400	0.2968	-0.5416	0.6216	0.13	0.8928
center	B	0.0000	0.0000	0.0000	0.0000	.	.
treatment	p	0.5715	0.3042	-0.0248	1.1678	1.88	0.0603
treatment	t	0.0000	0.0000	0.0000	0.0000	.	.
age		0.0223	0.0147	-0.0065	0.0512	1.52	0.1294
year	1	0.2763	0.2940	-0.2999	0.8524	0.94	0.3473
year	2	-0.3830	0.3747	-1.1173	0.3513	-1.02	0.3067
year	3	0.0000	0.0000	0.0000	0.0000	.	.



Placebo increases the log fracture rate by 0.5715; the test treatment lowers the log fracture rate by -0.5715.

In other words, placebo has a fracture rate that is  $\exp(0.5715)=1.77$  times that of test treatment, after controlling for center, age, and year. Alternatively, one could say the test treatment has a fracture rate that is  $\exp(-0.5715) = 0.56$  times the fracture rate for placebo.

The estimate of the exchangeable correlation is 0.1049, indicating a small (but non-ignorable) correlation among the respective years.






## **15.10 Fitting the Proportional Odds Model**

Recall the respiratory data from section 15.7. It contained an ordinal response that ranged from 0 (poor) to 4 (excellent). The proportional odds model takes into account the ordinality of the data (See Chapter 9 for univariate case).

Following requests a proportional odds model to be fit via GEE methods:

```
proc genmod data=resp descending;  
    class id center sex treatment visit;  
    model outcome = treatment sex center age baseline visit  
        visit*treatment / link=clogit dist=mult type3;  
    repeated subject=id*center / type=ind;  
run;
```



Since interest lies in assessing how much better the subjects receiving active treatment were, we form the cumulative logits that focus on the comparison of better to poorer outcomes, hence the DESCENDING option.

LINK=CLOGIT requests the cumulative logit link, and DIST=MULT requests the multinomial distribution. Together, these options specify the proportional odds model. TYPE=IND indicates the independence working correlation matrix (currently the only correlation structure available with the ordinal response model).

## **Output 15.37** Class Level and Response Information

### Class Level Information

Class	Levels	Values
id	56	1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56
center	2	1 2
sex	2	F M
treatment	2	A P
visit	4	1 2 3 4

### Response Profiles

Ordered Level	Ordered Value	Count
1	4	152
2	3	96
3	2	116
4	1	40
5	0	40

### **Output 15.38** GEE Model Information

#### GEE Model Information

Correlation Structure	Independent
Subject Effect	id*center (111 levels)
Number of Clusters	111
Correlation Matrix Dimension	4
Minimum Cluster Size	4
Maximum Cluster Size	4

### **Output 15.39** Type 3 Test Results

#### Score Statistics for Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
treatment	1	15.33	< .0001
sex	1	0.53	0.4664
center	1	1.33	0.2481
age	1	2.68	0.1016
baseline	1	21.60	< .0001
visit	3	0.66	0.8837
treatment*visit	3	10.47	0.0150

Reduce model by excluding AGE and SEX. Since CENTER was part of the study design, we will leave it in the model:

```
proc genmod data=resp2 descending;  
    class id center treatment visit;  
    model outcome = treatment center baseline visit  
                visit*treatment / link=clogit dist=mult type3;  
    repeated subject=id*center / type=ind;  
run;
```

**Output 15.40** Type 3 Test Results for Reduced Model

Score Statistics for Type 3 GEE Analysis			
Source	DF	Chi-Square	Pr > ChiSq
treatment	1	16.40	< .0001
center	1	1.25	0.2636
baseline	1	21.27	< .0001
visit	3	0.54	0.9106
treatment*visit	3	10.50	0.0148



## Output 15.41 Parameter Estimates

Analysis of GEE Parameter Estimates  
Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr >  Z
Intercept1		-3.3645	0.5766	-4.4945	-2.2345	-5.84	< .0001
Intercept2		-2.2049	0.5412	-3.2657	-1.1441	-4.07	< .0001
Intercept3		-0.6060	0.5193	-1.6329	0.4119	-1.17	0.2433
Intercept4		0.2929	0.5643	-0.8131	1.3988	0.52	0.6037
treatment	A	0.9995	0.3625	0.2891	1.7100	2.76	0.0058
treatment	P	0.0000	0.0000	0.0000	0.0000	.	.
center	1	-0.3491	0.3023	-0.9415	0.2434	-1.15	0.2482
center	2	0.0000	0.0000	0.0000	0.0000	.	.
baseline		0.8993	0.1670	0.5719	1.2266	5.38	< .0001
visit	1	0.2581	0.2501	-0.2321	0.7484	1.03	0.3021
visit	2	-0.2505	0.2303	-0.7019	0.2010	-1.09	0.2768
visit	3	-0.0360	0.1615	-0.3525	0.2806	-0.22	0.8238
visit	4	0.0000	0.0000	0.0000	0.0000	.	.
treatment*visit	A 1	-0.3049	0.3927	-1.0746	0.4648	-0.78	0.4375
treatment*visit	A 2	0.7247	0.3547	0.0296	1.4198	2.04	0.0410
treatment*visit	A 3	0.2990	0.3321	-0.3519	0.9500	0.90	0.3679
treatment*visit	A 4	0.0000	0.0000	0.0000	0.0000	.	.
treatment*visit	P 1	0.0000	0.0000	0.0000	0.0000	.	.
treatment*visit	P 2	0.0000	0.0000	0.0000	0.0000	.	.
treatment*visit	P 3	0.0000	0.0000	0.0000	0.0000	.	.
treatment*visit	P 4	0.0000	0.0000	0.0000	0.0000	.	.






## **15.11 GEE Analysis for Data with Missing Values**

GEE can handle data with missing values (i.e. varying number of responses per cluster, a common problem with loss to follow-up). GEE assumes that missing values are missing completely at random (MCAR).

### **15.11.1 Crossover Study with Missing Data**

Two-period crossover study for treatments for a skin disorder where patients (stratified by gender) given sequences of standard drug A, a new drug B, and a placebo. Investigators introduced a skin irritant, then applied the topical treatments. 300 patients participated in the first session, but 50 patients failed to attend the second session (1 week later) due to usual attrition reasons.



Analyze similar to crossover study in Section 15.6, except that in this case, 50 patients have a cluster size of 1. See page 538-539 for dataset. The 3-level variable CARRY is assigned a value of N for period 1. For period 2, CARRY=N if treatment=P in period 1, CARRY=A if treatment=A in period 1, and CARRY=B if treatment=B in period 1.

```
proc genmod data=skincross2 descending;  
    class subject treatment period gender carry;  
    model response = treatment period carry gender  
                gender*period / type3 dist=bin link=logit;  
    repeated subject=subject / type=exch;  
run;
```

### **Output 15.42 GEE Model Information**


#### GEE Model Information

Correlation Structure	Exchangeable
Subject Effect	subject (300 levels)
Number of Clusters	300
Clusters With Missing Values	50
Correlation Matrix Dimension	2
Maximum Cluster Size	2
Minimum Cluster Size	1

### **Output 15.43 Type 3 Analysis**

#### Score Statistics for Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
treatment	2	29.38	< .0001
period	1	7.11	0.0077
gender	1	29.94	< .0001
carry	2	1.08	0.5841
period*gender	1	4.21	0.0401



The above suggests that the carryover effects are not influential. We therefore delete them from the model. The ESTIMATE statements specify that odds ratio estimates be computed to compare the effects of the 3 treatments.

```
proc genmod data=skincross2 descending;  
    class subject treatment period gender;  
    model response = treatment period gender  
                    gender*period / type3  
                    dist=bin link=logit;  
    repeated subject=subject / type=exch;  
    estimate 'OR:A-B' treatment 1 -1 0 / exp;  
    estimate 'OR:A-P' treatment 1 0 -1 / exp;  
    estimate 'OR:B-P' treatment 0 1 -1 / exp;  
    lsmeans treatment / pdiff exp cl;  
  
run;
```

### **Output 15.44** Type 3 Analysis

#### Score Statistics for Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
treatment	2	40.02	< .0001
period	1	20.97	< .0001
gender	1	28.89	< .0001
period*gender	1	3.93	0.0474



## **Output 15.45   Parameter Estimates**

Analysis Of GEE Parameter Estimates  
Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr >  Z
Intercept		-0.9287	0.2249	-1.3696	-0.4879	-4.13	< .0001
treatment	A	1.2622	0.2079	0.8548	1.6696	6.07	< .0001
treatment	B	0.1722	0.2141	-0.2473	0.5918	0.80	0.4210
treatment	P	0.0000	0.0000	0.0000	0.0000	.	.
period	1	-0.4520	0.2257	-0.8944	-0.0095	-2.00	0.0453
period	2	0.0000	0.0000	0.0000	0.0000	.	.
gender	f	1.4443	0.2816	0.8925	1.9961	5.13	< .0001
gender	m	0.0000	0.0000	0.0000	0.0000	.	.
period*gender	1 f	-0.6505	0.3289	-1.2951	-0.0059	-1.98	0.0480
period*gender	1 m	0.0000	0.0000	0.0000	0.0000	.	.
period*gender	2 f	0.0000	0.0000	0.0000	0.0000	.	.
period*gender	2 m	0.0000	0.0000	0.0000	0.0000	.	.



### Output 15.46 Odds Ratio Estimates

#### Contrast Estimate Results

Label	L'Beta Estimate	Standard Error	Alpha	L'Beta Confidence Limits		Chi-Square	Pr > ChiSq
OR:A-B	1.0899	0.2193	0.05	0.6601	1.5198	24.70	< .0001
Exp (OR:A-B)	2.9741	0.6523	0.05	1.9350	4.5713		
OR:A-P	1.2622	0.2079	0.05	0.8548	1.6696	36.87	< .0001
Exp (OR:A-P)	3.5331	0.7344	0.05	2.3508	5.3099		
OR:B-P	0.1722	0.2141	0.05	-0.2473	0.5918	0.65	0.4210
Exp (OR:B-P)	1.1880	0.2543	0.05	0.7809	1.8072		

Subjects on drug A had odds of improvement that were 3.5 times the odds of improvement for those on placebo, and almost 3 times the odds of improvements vs. those on drug B. Subjects on new drug B did not do significantly better than those on placebo.

treatment Least Squares Means									
treatment	Estimate	Standard Error	z Value	Pr> z	Lower	Upper	Exp	Exp Lower	Exp Upper
A	0.6670	0.1616	4.13	<.0001	0.3503	0.9836	1.9483	1.4195	2.6741
B	-0.4230	0.1650	-2.56	0.0104	-0.7463	-0.09961	0.6551	0.4741	0.9052
P	-0.5952	0.1574	-3.78	0.0002	-0.9037	-0.2867	0.5515	0.4050	0.7508

Differences of treatment Least Squares Means										
trt	_trt	Estimate	Standard Error	z Value	Pr> z	Lower	Upper	Exp	Exp Lower	Exp Upper
A	B	1.0899	0.2193	4.97	<.0001	0.6601	1.5198	2.9741	1.9350	4.5713
A	P	1.2622	0.2079	6.07	<.0001	0.8548	1.6696	3.5331	2.3508	5.3099
B	P	0.1722	0.2141	0.80	0.4210	-0.2473	0.5918	1.1880	0.7809	1.8072




## 15.12 Alternating Logistic Regression

The modeling approach in previous sections focuses on modeling correlation between repeated measures. However, the data can influence the range of the correlation since the estimates of correlation  $r_{jk}$  are constrained by the means  $\mu_{ij}$ :

$$\text{Corr}(Y_{ij}, Y_{ik}) = r_{jk} = \frac{\Pr(Y_{ij} = 1, Y_{ik} = 1) - \mu_{ij}\mu_{ik}}{\sqrt{\mu_{ij}(1 - \mu_{ij})\mu_{ik}(1 - \mu_{ik})}}$$

However, odds ratios are a more natural choice for modeling the association since they are not constrained by the means:

$$\text{OR}(Y_{ij}, Y_{ik}) = \frac{\Pr(Y_{ij} = 1, Y_{ik} = 1) \Pr(Y_{ij} = 0, Y_{ik} = 0)}{\Pr(Y_{ij} = 1, Y_{ik} = 0) \Pr(Y_{ij} = 0, Y_{ik} = 1)}$$

- 
- In GEE, the correlations are treated as nuisance parameters, so the modeling of correlations vs. odds ratio usually has little influence on inference for  $\beta$ .
  - The ALR algorithm models the log of the odds ratio as  $\psi_{ijk} = \mathbf{z}'\alpha = \log(\text{OR}(Y_{ij}, Y_{ik}))$ . The method alternates between first-order GEE estimation of  $\beta$  and a modified (with offset) logistic regression estimate of the  $\alpha$  until convergence.
  - You can specify whether the log odds ratio should be
    - A constant across clusters (option `logor=exch`)
    - A constant within levels of a blocking factor (option `logor=logorvar(variable name)` )
    - Fully parameterized within cluster, as one parameter per pair (option `logor=fullclust`)
  - Respiratory Data example provided on Pages 544-549






## **15.13 Using GEE to Account for Overdispersion: Univariate Outcome**

Overdispersion occurs when the observed variance is greater than the nominal variance for a particular distribution, and must be taken into account since it can have a major impact on inference.

Can manage overdispersion by assuming a more flexible distribution, or adjust the covariance matrix with a scale factor. Another option is to use GEE methods. Robust covariance matrix estimated by GEE is robust to the misspecification of the covariance structure (which occurs in case of overdispersion). With GEE estimation, we are using a subject-to-subject measure for variance estimation instead of a model-based one.



To study incidence of lower respiratory illness (LRI), repeated measurements (every 2 weeks) on 284 infants were taken over one year. Explanatory variables are passive smoking, socioeconomic status, and crowding, and the outcome is the total number of times of LRI recorded for the year. Can model with Poisson regression, but since it's reasonable to expect some overdispersion, we will consider GEE methods. See page 550-551 for data.

```
proc genmod data=lri;  
  class ses id race agegroup;  
  model count = passive crowding ses race agegroup/  
              dist=poisson link=log offset=logrisk type3;  
run;
```



### **Output 15.47** Model Information


#### Model Information

Data Set	WORK.LRI
Distribution	Poisson
Link Function	Log
Dependent Variable	count
Offset Variable	logrisk
Observations Used	284

### **Output 15.48** Goodness-of-Fit-Statistics

#### Criteria for Assessing Goodness of Fit

Criterion	DF	Value	Value/DF
Deviance	276	408.1549	1.4788
Scaled Deviance	276	408.1549	1.4788
Pearson Chi-Square	276	495.4493	1.7951
Scaled Pearson X2	276	495.4493	1.7951
Log Likelihood		-260.4117	



With Values of 1.4788 for the Deviance/df and 1.7951 for Pearson/df, there is evidence of overdispersion. (Ideally, these would be  $\approx 1$ ). Model-based estimates of standard errors may not be appropriate.

```
proc genmod data=lri;  
  class ses id race agegroup;  
  model count = passive crowding ses race agegroup /  
    dist=poisson link=log offset=logrisk type3;  
  repeated subject=id / type=ind;  
run;
```



### **Output 15.49**   GEE Model Information

#### GEE Model Information

Correlation Structure	Independent
Subject Effect	id (284 levels)
Number of Clusters	284
Correlation Matrix Dimension	1
Maximum Cluster Size	1
Minimum Cluster Size	1

## **Output 15.50   GEE Parameter Estimates**

Analysis of GEE Parameter Estimates  
Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr >  Z
Intercept		0.6047	0.5564	-0.4858	1.6952	1.09	0.2771
passive		0.4310	0.2105	0.0184	0.8436	2.05	0.0406
crowding		0.5199	0.2367	0.0559	0.9839	2.20	0.0281
ses	0	-0.3970	0.2977	-0.9805	0.1865	-1.33	0.1824
ses	1	-0.0681	0.2520	-0.5619	0.4258	-0.27	0.7871
ses	2	0.0000	0.0000	0.0000	0.0000	.	.
race	0	0.1402	0.2211	-0.2931	0.5736	0.63	0.5259
race	1	0.0000	0.0000	0.0000	0.0000	.	.
agegroup	1	-0.4792	0.6033	-1.6617	0.7033	-0.79	0.4270
agegroup	2	-0.9919	0.4675	-1.9082	-0.0756	-2.12	0.0339
agegroup	3	0.0000	0.0000	0.0000	0.0000	.	.

### **Output 15.51** Type 3 Analysis

#### Score Statistics for Type 3 GEE Analysis


Source	DF	Chi-Square	Pr > ChiSq
passive	1	3.90	0.0484
crowding	1	4.72	0.0298
ses	2	2.11	0.3478
race	1	0.42	0.5176
agegroup	2	2.79	0.2484




## **Lower respiratory illness during first year of life of children (LaVange et al [1994, Statistics in Medicine])**

1. Assessment of 294 children for one or more episodes of lower respiratory illness during consecutive two week intervals of risk for the first year of life; there were 6115 two week intervals.
2. Passive smoke exposure, socio-economic status, and crowding in home are explanatory variables for the child.
3. Age and season are time dependent covariables.




- 
4. SUDAAN Logistic Regression with child as primary sampling unit and equal weight for all two week intervals provides the following results:

Parameter	Estimate	Std. err.	<i>p</i> -value
Intercept	-4.85	0.24	
SES	0.33	0.26	0.21
Crowding	0.48	0.24	0.04
Age 4-6 months	0.79	0.22	< 0.001
Age > 6 months	0.62	0.20	0.002
Season	0.44	0.14	0.002
Passive smoking	0.45	0.22	0.04

- 
5. Predicted incidence per person year at risk was 0.92 with (0.70, 1.21) as 0.95 confidence interval for children exposed to passive smoking and 0.59 with (0.45, 0.78) confidence interval for those not exposed, where uniform distributions apply to other explanatory variables. (1.03, 2.40) is 0.95 confidence interval for corresponding odds ratio.

---

Tilley, B.B. et al [1996, Stroke, 27, 2136-42, “Use of a global test for multiple outcomes in stroke trials with application to the National Institute of Neurological Disorders and Stroke t-PA Stroke Trial (New England Journal of Medicine, 333 (24), December 14, 1995, 1581-87)]

- 
1. dichotomous outcomes from the Barthel Index, Modified Rankin Scale, Glasgow Outcome Scale and National Institutes of Health Stroke Scale (NIHSS) were integrated as composite endpoint through generalized estimating equations (GEE) for logistic regression
  2. the overall result had  $p = 0.008$  for 1.7 as the estimated odds ratio and (1.2, 2.6) as its 0.95 confidence interval
  3. the respective components had  $0.019 < p < 0.033$ , although a method to address their multiplicity (e.g., Lehmacher et al [1991]) was not specified
  4. global test was considered helpful for interpretation in a setting where no single measure is accepted and where evidence of efficacy should be a “consistent and persuasive” difference between treatments
  5. tests of homogeneity among components are possible with GEE through treatment  $\times$  component interaction