Marginal Models and GEE: Examples

Biostatistics 653

Applied Statistics III: Longitudinal Data Analysis

Example: Blood Lead Study

- Exposure to lead can cause cognitive impairment. Airborne lead levels have been dramatically reduced by the discontinuation of leaded gasoline; however, a small fraction of children are exposed to high levels of lead through deteriorating lead-based paint (present in many homes built before 1978). Lead poisoning in children can be treated by helping children excrete the ingested lead. A new chelating agent, succimer, enhances urinary excretion of lead and may be given orally (as opposed to older, injection-only treatments).
- A randomized clinical trial was conducted in children with confirmed high blood levels, who were randomized to receive either succimer or placebo and were followed longitudinally (1, 4 and 6 weeks).

Example: Blood Lead Study

- Fitzmaurice et al. consider fitting a logistic model to the probability of a blood lead level below 20 ug per dL at the three post-randomization follow-up periods. The percentages of children with blood lead levels below the cutoff at the three post-treatment occasions are below.

	Succimer	Placebo
Week		
1	78	16
4	76	26
6	54	26

Marginal Model

- We will consider the mean model

$$logit(\mu_{ij}) = \beta_0 + \beta_1 time_{ij} + \beta_2 trt_i + \beta_3 time_{ij} trt_i$$

treating time as a continuous variable.

- Now, we must also make some assumptions about the variances and correlations. One natural choice for binary data is to assume that

$$V(Y_{ij}) = \mu_{ij}(1 - \mu_{ij})$$

- In addition, we might assume an exchangeable correlation, or that $Corr(Y_{ij}, Y_{ik}) = \alpha$.

SAS Code

- We can carry out the analysis in SAS PROC GENMOD as follows.

```
proc genmod data=lead2 descending;
class id;
/* d=bin tells SAS our data are binary */
model lowlead=succimer time
succimer*time/d=bin;
/* corrw requests printing of working
correlation matrix */
repeated subject=id/type=exch
corrw;
run;
```

SAS Output

The GENMOD Procedure

Model Information

Data Set	WORK.LEAD2
Distribution	Binomial
Link Function	Logit
Dependent Variable	lowlead

Number of Observations Read	300
Number of Observations Used	300
Number of Events	138
Number of Trials	300

Class Level Information

Class	Levels	Values
id	100	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 ...

Response Profile

Ordered Value	lowlead	Total Frequency
1	1	138
2	0	162

PROC GENMOD is modeling the probability that lowlead='1'.

Parameter Information

Parameter	Effect
Prm1	Intercept
Prm2	succimer
Prm3	time
Prm4	succimer*time

The GENMOD Procedure

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	296	337.7276	1.1410
Scaled Deviance	296	337.7276	1.1410
Pearson Chi-Square	296	300.8592	1.0164
Scaled Pearson X2	296	300.8592	1.0164
Log Likelihood		-168.8638	

Algorithm converged.

Analysis Of Initial Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square
Intercept	1	-1.6972	0.4342	-2.5481	-0.8463	15.28
succimer	1	3.3652	0.5995	2.1902	4.5402	31.51
time	1	0.1234	0.0980	-0.0688	0.3156	1.58
succimer*time	1	-0.3447	0.1345	-0.6083	-0.0811	6.57
Scale	0	1.0000	0.0000	1.0000	1.0000	

Analysis Of Initial Parameter Estimates

Parameter	Pr > ChiSq
Intercept	<.0001
succimer	<.0001
time	0.2082
succimer*time	0.0104
Scale	

NOTE: The scale parameter was held fixed.

GEE Model Information

Correlation Structure	Exchangeable
Subject Effect	id (100 levels)
Number of Clusters	100
Correlation Matrix Dimension	3
Maximum Cluster Size	3
Minimum Cluster Size	3

The GENMOD Procedure

Algorithm converged.

SAS Output

Working Correlation Matrix

	Col1	Col2	Col3
Row1	1.0000	0.4784	0.4784
Row2	0.4784	1.0000	0.4784
Row3	0.4784	0.4784	1.0000

Exchangeable Working Correlation

Correlation 0.4783943345

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	-1.6952	0.3935	-2.4665	-0.9239	-4.31	<.0001
succimer	3.3776	0.5711	2.2583	4.4970	5.91	<.0001
time	0.1233	0.0770	-0.0276	0.2742	1.60	0.1091
succimer*time	-0.3452	0.1045	-0.5500	-0.1404	-3.30	0.0010

SAS Output

- When looking at the output, note a couple of important points. The first set of parameter estimates, *Analysis of Initial Parameter Estimates*, does not account for the correlation in the data. These estimates should be ignored. The second set of estimates, *Analysis of GEE Parameter Estimates*, are the ones we want.

SAS Output

- We can see that both treatment group and time affect the probability that the blood lead levels fall below the cutoff. In particular, for subjects on placebo, we have

$$logit(\hat{\mu}_{ij}) = -1.6952 + 0.1233 time_{ij}$$

and for succimer subjects we have

$$logit(\hat{\mu}_{ij}) = -1.6952 + 3.3776 + (0.1233 - 0.3452) time_{ij}$$

- The estimated OR for blood lead levels below the cutoff comparing succimer to placebo at week 1 is

$$\frac{\exp(1.6824 - 0.2219)}{\exp(-1.6952 + 0.1233)} = 20.75$$

and we see that this estimated OR decreases over time to 7.37 by week 4 and 3.69 by week 6.

Modeling Association using Odds Ratios

- One drawback of using correlations to model association among binary variables was that the range of observed correlations is restricted by the means of the binary variables.
- For example, consider the blood lead data. If we assume the true proportion of samples that are below the cutoff at week 1 is 0.78 in the succimer group and 0.76 in the succimer group at week 4, then the correlation across these two times cannot be greater than 0.95 and cannot be less than -0.30. Similarly, if we assume the true proportion of samples below the cutoff at week 6 is 0.54 in the succimer group, then the correlation between Y_{i1} and Y_{i6} is constrained to the interval (-0.49, 0.58).

Modeling Association using Odds Ratios

- We can use a modified GEE with the association modeled in terms of odds ratios rather than correlations. Recall that the OR for any pair of binary responses, Y_j and Y_k, is given by

$$OR(Y_j, Y_k) = \frac{P(Y_j = 1, Y_k = 1)P(Y_j = 0, Y_k = 0)}{P(Y_j = 1, Y_k = 0)P(Y_j = 0, Y_k = 1)}$$

SAS Code

We can do this in PROC GENMOD as follows for the lead data

```
proc genmod data=lead2 descending;
class id;
/* d=bin tells SAS our data are binary */
model lowlead=succimer time succimer*time/d=bin;
/* covb requests printing of covariance
matrix of the betahats */
/* logor command asks for OR parameterization
of associations */
repeated subject=id/logor=fullclust covb;
run;
```

SAS Output

The GENMOD Procedure

Model Information

Data Set	WORK.LEAD2
Distribution	Binomial
Link Function	Logit
Dependent Variable	lowlead

Number of Observations Read	300
Number of Observations Used	300
Number of Events	138
Number of Trials	300

Class Level Information

Class	Levels	Values
id	100	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 ...

Response Profile

Ordered Value	lowlead	Total Frequency
1	1	138
2	0	162

PROC GENMOD is modeling the probability that lowlead='1'.

Parameter Information

Parameter	Effect
Prm1	Intercept
Prm2	succimer
Prm3	time
Prm4	succimer*time

The GENMOD Procedure

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	296	337.7276	1.1410
Scaled Deviance	296	337.7276	1.1410
Pearson Chi-Square	296	300.8592	1.0164
Scaled Pearson X2	296	300.8592	1.0164
Log Likelihood		-168.8638	

Algorithm converged.

SAS Output

Analysis Of Initial Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square
Intercept	1	-1.6972	0.4342	-2.5481	-0.8463	15.28
succimer	1	3.3652	0.5995	2.1902	4.5402	31.51
time	1	0.1234	0.0980	-0.0688	0.3156	1.58
succimer*time	1	-0.3447	0.1345	-0.6083	-0.0811	6.57
Scale	0	1.0000	0.0000	1.0000	1.0000	

Analysis Of Initial Parameter Estimates

Parameter	Pr > ChiSq
Intercept	<.0001
succimer	<.0001
time	0.2082
succimer*time	0.0104
Scale	

NOTE: The scale parameter was held fixed.

SAS Output

GEE Model Information

Log Odds Ratio Structure	Fully Parameterized Clusters
Subject Effect	id (100 levels)
Number of Clusters	100
Correlation Matrix Dimension	3
Maximum Cluster Size	3
Minimum Cluster Size	3

The GENMOD Procedure

Log Odds Ratio Parameter Information

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

Covariance Matrix (Model-Based)

	Prm1	Prm2	Prm3	Prm4
Prm1	0.1524936	-0.152494	-0.019544	0.0195437
Prm2	-0.152494	0.3078342	0.0195437	-0.041166
Prm3	-0.019544	0.0195437	0.0050138	-0.005014
Prm4	0.0195437	-0.041166	-0.005014	0.0101041

Covariance Matrix (Empirical)

	Prm1	Prm2	Prm3	Prm4
Prm1	0.1415312	-0.141531	-0.017937	0.017937
Prm2	-0.141531	0.3267618	0.017937	-0.04436
Prm3	-0.017937	0.017937	0.0049939	-0.004994
Prm4	0.017937	-0.04436	-0.004994	0.0106233
Alpha1	-0.003316	-0.012195	0.0009537	0.00538
Alpha2	-0.007901	-0.001199	0.0008857	0.0013964
Alpha3	0.0267384	-0.003371	-0.001118	-0.001082

Covariance Matrix (Empirical)

	Alpha1	Alpha2	Alpha3
Prm1	-0.003316	-0.007901	0.0267384
Prm2	-0.012195	-0.001199	-0.003371
Prm3	0.0009537	0.0008857	-0.001118
Prm4	0.00538	0.0013964	-0.001082

SAS Output

Alpha1	0.3491303	0.1428527	0.0737928
Alpha2	0.1428527	0.3895171	0.1085227
Alpha3	0.0737928	0.1085227	0.3824711

Algorithm converged.

The GENMOD Procedure

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	-1.6388	0.3762	-2.3761	-0.9014	-4.36	<.0001
succimer	3.4335	0.5716	2.3131	4.5538	6.01	<.0001
time	0.1072	0.0707	-0.0313	0.2457	1.52	0.1294
succimer*time	-0.3633	0.1031	-0.5653	-0.1613	-3.52	0.0004
Alpha1	2.0552	0.5909	0.8971	3.2132	3.48	0.0005
Alpha2	2.3938	0.6241	1.1706	3.6171	3.84	0.0001
Alpha3	3.2503	0.6184	2.0382	4.4624	5.26	<.0001

Modeling Association using Odds Ratios

- While the point estimates are not exactly the same, they are similar to those obtained before. The parameters α describe the association between the repeated measures.
- For example, the OR for lead below the cutoff at week 4 as a function of low lead at week 1 is exp(2.0552) = 7.8, the corresponding OR for week 6 based on low lead at week 1 is exp(2.3938) = 11.0, and the corresponding OR for week 6 based on low lead at week 4 is exp(3.2503) = 25.8.

Missing Data with GEE

- PROC GENMOD allows missing data that are MCAR (missing completely at random; which we will cover later in class). It may be necessary to define an effect in the model specifying the order of the measurements within individuals in this case:

```
data lead3; set lead2;
cattime=time;
run;

proc genmod data=lead3 descending;
class id cattime;
model lowlead=succimer time
succimer*time/d=bin;
repeated subject=id/within=cattime type=exch
corrw;
run;
```

Example: Skin Cancer Prevention Study

- We consider data from the Skin Cancer Prevention Study, a randomized, double-blind, placebo-controlled clinical trial of beta-carotene to prevent non-melanoma skin cancer in high risk subjects. A total of 1805 subjects were randomized to either placebo or 50mg of beta-carotene per day for 5 years.
- Subjects were examined once a year and biopsied if a cancer was suspected to determine the number of new skin cancers occurring since the last exam. The outcome variable Y_{ij} is a binary variable that takes value 1 if new skin cancers were detected at time j and 0 otherwise.

Example: Skin Cancer Prevention Study

- The categorical variable treatment is coded 1=beta-carotene, 0=placebo. The variable year denotes the year of follow-up. The categorical variable gender is coded 1=male, 0=female. The categorical variable skin denotes skin type and is coded 1=burns, 0=otherwise. The variable age is the age (in years) of each subject at randomization. In addition, a variable exposure contains the number of previously-diagnosed skin cancers. Complete data are available on 1683 subjects comprising a total of 7081 measurements.

Example: Skin Cancer Prevention Study

- Investigators in this randomized clinical trial are interested in whether betacarotene helps to prevent skin cancer. However, because doses in the same range had been anecdotally linked to development of other cancers, the investigators do not wish to conduct one-sided hypothesis tests. In addition, the investigators wish to know whether covariates gender, skin type, and age are also related to cancer risk.

Hypotheses and Model

- We will address the following questions of the investigator:

  -- What conclusions can we draw about the effect of beta carotene on the occurrence of skin cancers?

  -- What is the association between age, skin type, gender, and skin cancer occurrence?

  -- What is the association between number of previously diagnosed skin cancers and subsequent cancer occurrence?

Hypotheses and Model

- We fit the logistic mean model

$$logit\left(P\left(Y_{ij}=1\right)\right) = \beta_0 + \beta_1 trt_i + \beta_2 year_{ij} + \beta_3 (trt_i)\left(year_{ij}\right) + \beta_4 age_i + \beta_5 gender_i + \beta_6 skin_i + \beta_7 exposure_i + \beta_8 exposure_i^2$$

where $year_{ij}$ is the measurement occasion for subject i, $j = 1, \cdots, 5$

- Next, we specify the variance function as

$$Var\left(Y_{ij}\right) = \mu_{ij}(1-\mu_{ij})$$

- We will assume that the within-subject association among the five repeated binary responses has an unstructured pairwise log odds ratio pattern, where

$$logOR\left(Y_{ij}, Y_{ik}\right) = \alpha_{jk}$$

SAS Code

```
proc genmod data=new descending;
class id year;
model ybin=treatment yearcont treatment*yearcont agecent skin
          gender exposure exposure*exposure/d=bin;
repeated subject=id/withinsubject=year logor=fullclust covb;
/* gives generalized score statistic for testing treatment effect */
contrast 'treatment effect' treatment 1 -1, treatment*yearcont 1 -1/e;
contrast 'exposure' exposure 1 -1, exposure*exposure 1 -1/e;
run;
```

Note that the contrast statement has both 1 and -1 inside. This is because the treatment is a factor that contains two levels (control: β_0; treatment: $\beta_0 + \beta_1$), and in order to test $\beta_1 = 0$ we need to contrast $1 * (\beta_1 + \beta_0) - 1 * (\beta_0)$. The same applies to the interaction term where we need to test $\beta_3 = 0$.

SAS Output

Model Information

Data Set	WORK.NEW
Distribution	Binomial
Link Function	Logit
Dependent Variable	ybin

Number of Observations Read	7081
Number of Observations Used	7081
Number of Events	1165
Number of Trials	7081

Class Level Information

Class	Levels	Values
ID	1683	30 100012 100023 100034 100045 100056 100067 100078 100089 100102 100113 100124 100146 100157 100168 100179 100190 100203 100214 100236 100247 100258 100269 100280 100291 100304 100326 100348 100359 100370 100381 100392 100405 100416 100427 100460 ...
Year	5	1 2 3 4 5

Response Profile

Ordered Value	ybin	Total Frequency
1	1	1165
2	0	5916

PROC GENMOD is modeling the probability that ybin='1'.

Parameter Information

Parameter	Effect
Prm1	Intercept
Prm2	Treatment
Prm3	yearcont
Prm4	Treatment*yearcont
Prm5	agecent
Prm6	Skin
Prm7	Gender
Prm8	Exposure

The GENMOD Procedure

Parameter Information

Parameter	Effect
Prm9	Exposure*Exposure

SAS Output

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	7072	5625.9476	0.7955
Scaled Deviance	7072	5625.9476	0.7955
Pearson Chi-Square	7072	6885.7026	0.9737
Scaled Pearson X2	7072	6885.7026	0.9737
Log Likelihood		-2812.9738	

Algorithm converged.

Analysis Of Initial Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square
Intercept	1	-3.2124	0.1433	-3.4933	-2.9315	502.53
Treatment	1	-0.0371	0.1555	-0.3419	0.2677	0.06
yearcont	1	-0.0209	0.0371	-0.0937	0.0519	0.32
Treatment*yearcont	1	0.0401	0.0516	-0.0610	0.1413	0.61
agecent	1	0.0193	0.0038	0.0120	0.0267	26.49
Skin	1	0.2063	0.0690	0.0711	0.3415	8.95
Gender	1	0.6503	0.0818	0.4900	0.8106	63.23
Exposure	1	0.3944	0.0236	0.3480	0.4407	278.09

SAS Output

Analysis Of Initial Parameter Estimates

Parameter	Pr > ChiSq
Intercept	<.0001
Treatment	0.8113
yearcont	0.5740
Treatment*yearcont	0.4365
agecent	<.0001
Skin	0.0028
Gender	<.0001
Exposure	<.0001

SAS Output

The GENMOD Procedure

Analysis Of Initial Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square
Exposure*Exposure	1	-0.0133	0.0013	-0.0158	-0.0108	112.45
Scale	0	1.0000	0.0000	1.0000	1.0000	

Analysis Of Initial Parameter Estimates

Parameter	Pr > ChiSq
Exposure*Exposure	<.0001
Scale	

NOTE: The scale parameter was held fixed.

SAS Output

GEE Model Information

Log Odds Ratio Structure	Fully Parameterized Clusters
Within-Subject Effect	Year (5 levels)
Subject Effect	ID (1683 levels)
Number of Clusters	1683
Correlation Matrix Dimension	5
Maximum Cluster Size	5
Minimum Cluster Size	1

Log Odds Ratio
Parameter Information

Parameter	Group
Alpha1	(1, 2)
Alpha2	(1, 3)
Alpha3	(1, 4)
Alpha4	(1, 5)
Alpha5	(2, 3)
Alpha6	(2, 4)
Alpha7	(2, 5)
Alpha8	(3, 4)
Alpha9	(3, 5)
Alpha10	(4, 5)

SAS Output

The GENMOD Procedure

Covariance Matrix (Model-Based)

	Prm1	Prm2	Prm3	Prm4	Prm5
Prm1	0.022907	-0.011883	-0.003326	0.00325	-0.000014
Prm2	-0.011883	0.0234399	0.0032615	-0.006291	1.4598E-6
Prm3	-0.003326	0.0032615	0.0012259	-0.001225	5.7078E-7
Prm4	0.00325	-0.006291	-0.001225	0.0023658	-2.208E-7
Prm5	-0.000014	1.4598E-6	5.7078E-7	-2.208E-7	0.0000195
Prm6	-0.002844	-0.000098	0.0000224	0.0000121	-7.757E-7
Prm7	-0.007002	-0.000042	0.0000399	-0.000011	8.5775E-6
Prm8	-0.001694	-0.000076	8.7122E-6	4.6964E-6	-7.118E-6
Prm9	0.0000804	3.0438E-6	-3.991E-7	-7.08E-8	3.0756E-7

SAS Output

Covariance Matrix (Model-Based)

	Prm6	Prm7	Prm8	Prm9
Prm1	-0.002844	-0.007002	-0.001694	0.0000804
Prm2	-0.000098	-0.000042	-0.000076	3.0438E-6
Prm3	0.0000224	0.0000399	8.7122E-6	-3.991E-7
Prm4	0.0000121	-0.000011	4.6964E-6	-7.08E-8
Prm5	-7.757E-7	8.5775E-6	-7.118E-6	3.0756E-7
Prm6	0.0068486	0.0001858	-0.000286	0.0000105
Prm7	0.0001858	0.009227	-0.000046	-4.739E-7
Prm8	-0.000286	-0.000046	0.0008475	-0.000043
Prm9	0.0000105	-4.739E-7	-0.000043	2.4457E-6

SAS Output

Covariance Matrix (Empirical)

	Prm1	Prm2	Prm3	Prm4	Prm5
Prm1	0.0239305	-0.012103	-0.003374	0.0032828	-0.000019
Prm2	-0.012103	0.024205	0.003202	-0.006528	-0.00002
Prm3	-0.003374	0.003202	0.0011945	-0.001194	-3.326E-6
Prm4	0.0032828	-0.006528	-0.001194	0.0024672	2.488E-6
Prm5	-0.000019	-0.00002	-3.326E-6	2.488E-6	0.0000218
Prm6	-0.003009	-0.000214	-6.962E-6	0.000029	8.3034E-6
Prm7	-0.007871	-0.000139	0.0001365	-0.000033	0.0000252
Prm8	-0.001646	0.0000619	0.0000338	-0.00004	-7.309E-6
Prm9	0.0000783	-6.41E-6	-1.857E-6	2.6521E-6	2.4704E-7
Alpha1	0.0029538	-0.000365	-0.000396	0.0000953	0.0000202
Alpha2	0.0012302	0.0003723	0.0001521	-0.000153	-0.000052
Alpha3	-0.000935	-0.000234	0.0001314	-0.000033	0.0000296
Alpha4	0.000899	-0.00064	-0.00001	0.0001999	-7.453E-6
Alpha5	-0.000125	0.0004834	-0.000044	-0.000113	-3.947E-6
Alpha6	0.000859	-0.000152	0.0002236	0.0002418	-0.000038
Alpha7	0.0005622	-0.000655	0.0001152	0.0003173	0.0000689
Alpha8	0.0003835	-0.000693	0.0000695	-0.000222	0.0000399

The GENMOD Procedure

Covariance Matrix (Empirical)

	Prm1	Prm2	Prm3	Prm4	Prm5
Alpha9	0.00179	0.0005368	0.000413	-0.000667	-0.000036
Alpha10	-0.000826	0.0000219	-0.000479	0.0002945	1.3655E-6

Covariance Matrix (Empirical)

	Prm6	Prm7	Prm8	Prm9	Alpha1
Prm1	-0.003009	-0.007871	-0.001646	0.0000783	0.0029538
Prm2	-0.000214	-0.000139	0.0000619	-6.41E-6	-0.000365
Prm3	-6.962E-6	0.0001365	0.0000338	-1.857E-6	-0.000396
Prm4	0.000029	-0.000033	-0.00004	2.6521E-6	0.0000953
Prm5	8.3034E-6	0.0000252	-7.309E-6	2.4704E-7	0.0000202
Prm6	0.0070746	0.0002786	-0.000312	0.0000117	-0.000765
Prm7	0.0002786	0.0097256	0.0000236	-3.457E-6	-0.000381
Prm8	-0.000312	0.0000236	0.0007897	-0.000039	-0.000505
Prm9	0.0000117	-3.457E-6	-0.000039	2.2591E-6	0.0000234
Alpha1	-0.000765	-0.000381	-0.000505	0.0000234	0.0304239
Alpha2	-0.00021	-0.000766	-0.000321	0.0000146	0.0038682
Alpha3	-0.000182	0.0004031	0.000112	-9.343E-6	0.0036589
Alpha4	-0.001229	0.000037	-0.000137	0.0000107	0.0011796
Alpha5	-0.000158	0.0008918	-0.000143	5.8037E-6	0.0048728
Alpha6	-0.0002	-0.001428	-0.000082	9.8815E-7	0.003442
Alpha7	-0.001688	0.0007652	-0.000393	0.0000215	0.0015927
Alpha8	-0.000384	0.0005678	-0.00012	2.3524E-6	0.0022204
Alpha9	-0.001149	-0.00145	-0.000162	0.0000114	0.0025858
Alpha10	0.0004074	0.0011842	0.000142	-5.945E-6	-0.000994

SAS Output

Covariance Matrix (Empirical)

	Alpha2	Alpha3	Alpha4	Alpha5	Alpha6
Prm1	0.0012302	-0.000935	0.000899	-0.000125	0.000859
Prm2	0.0003723	-0.000234	-0.00064	0.0004834	-0.000152
Prm3	0.0001521	0.0001314	-0.00001	-0.000044	0.0002236
Prm4	-0.000153	-0.000033	0.0001999	-0.000113	0.0002418
Prm5	-0.000052	0.0000296	-7.453E-6	-3.947E-6	-0.000038
Prm6	-0.00021	-0.000182	-0.001229	-0.000158	-0.0002
Prm7	-0.000766	0.0004031	0.000037	0.0008918	-0.001428
Prm8	-0.000321	0.000112	-0.000137	-0.000143	-0.000082
Prm9	0.0000146	-9.343E-6	0.0000107	5.8037E-6	9.8815E-7
Alpha1	0.0038682	0.0036589	0.0011796	0.0048728	0.003442
Alpha2	0.0305868	0.0025942	0.0064352	0.0026454	0.0018494
Alpha3	0.0025942	0.0322865	0.0036476	0.0018684	0.0019924
Alpha4	0.0064352	0.0036476	0.0579072	0.0025618	-0.001075
Alpha5	0.0026454	0.0018684	0.0025618	0.0307501	0.001877
Alpha6	0.0018494	0.0019924	-0.001075	0.001877	0.0357694

The GENMOD Procedure

Covariance Matrix (Empirical)

	Alpha2	Alpha3	Alpha4	Alpha5	Alpha6
Alpha7	0.0027207	-0.000814	0.0017768	0.0024109	0.0055124
Alpha8	0.0038956	0.004934	0.0034022	0.0033056	0.0038331
Alpha9	0.0046123	0.003392	0.0113053	0.0005801	-0.00025
Alpha10	0.0025092	0.0023572	0.0016114	-0.000334	0.0035443

Covariance Matrix (Empirical)

	Alpha7	Alpha8	Alpha9	Alpha10
Prm1	0.0005622	0.0003835	0.00179	-0.000826
Prm2	-0.000655	-0.000693	0.0005368	0.0000219
Prm3	0.0001152	0.0000695	0.000413	-0.000479
Prm4	0.0003173	-0.000222	-0.000667	0.0002945
Prm5	0.0000689	0.0000399	-0.000036	1.3655E-6
Prm6	-0.001688	-0.000384	-0.001149	0.0004074
Prm7	0.0007652	0.0005678	-0.00145	0.0011842
Prm8	-0.000393	-0.00012	-0.000162	0.000142
Prm9	0.0000215	2.3524E-6	0.0000114	-5.945E-6
Alpha1	0.0015927	0.0022204	0.0025858	-0.000994
Alpha2	0.0027207	0.0038956	0.0046123	0.0025092
Alpha3	-0.000814	0.004934	0.003392	0.0023572
Alpha4	0.0017768	0.0034022	0.0113053	0.0016114
Alpha5	0.0024109	0.0033056	0.0005801	-0.000334
Alpha6	0.0055124	0.0038331	-0.00025	0.0035443
Alpha7	0.0610023	-0.000166	0.0031813	0.0089339
Alpha8	-0.000166	0.034758	0.0039237	0.0041518
Alpha9	0.0031813	0.0039237	0.0606733	0.0031197
Alpha10	0.0089339	0.0041518	0.0031197	0.0530195

Algorithm converged.

SAS Output

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	-3.2772	0.1547	-3.5804	-2.9740	-21.18	<.0001
Treatment	-0.0235	0.1556	-0.3285	0.2814	-0.15	0.8798
yearcont	-0.0131	0.0346	-0.0808	0.0547	-0.38	0.7051
Treatment*yearcont	0.0330	0.0497	-0.0644	0.1303	0.66	0.5068
agecent	0.0185	0.0047	0.0093	0.0276	3.96	<.0001
Skin	0.2247	0.0841	0.0599	0.3896	2.67	0.0075
Gender	0.6558	0.0986	0.4625	0.8491	6.65	<.0001

SAS Output

The GENMOD Procedure

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Exposure	0.4079	0.0281	0.3528	0.4630	14.52	<.0001
Exposure*Exposure	-0.0138	0.0015	-0.0168	-0.0109	-9.20	<.0001
Alpha1	0.6492	0.1744	0.3074	0.9911	3.72	0.0002
Alpha2	0.9520	0.1749	0.6092	1.2947	5.44	<.0001
Alpha3	0.7238	0.1797	0.3717	1.0760	4.03	<.0001
Alpha4	0.8082	0.2406	0.3365	1.2798	3.36	0.0008
Alpha5	0.7776	0.1754	0.4339	1.1213	4.43	<.0001
Alpha6	0.9193	0.1891	0.5486	1.2899	4.86	<.0001
Alpha7	0.7890	0.2470	0.3049	1.2731	3.19	0.0014
Alpha8	0.6189	0.1864	0.2535	0.9843	3.32	0.0009
Alpha9	1.1505	0.2463	0.6677	1.6333	4.67	<.0001
Alpha10	0.8703	0.2303	0.4190	1.3216	3.78	0.0002

SAS Output

Coefficients for Contrast treatment effect

Label	Row	Prm1	Prm2	Prm3	Prm4	Prm5	Prm6
		Prm7	Prm8	Prm9			
treatment effect	1	0	1	0	0	0	0
		0	0	0			
treatment effect	2	0	0	0	1	0	0
		0	0	0			

Coefficients for Contrast exposure

Label	Row	Prm1	Prm2	Prm3	Prm4	Prm5	Prm6
		Prm7	Prm8	Prm9			
exposure	1	0	0	0	0	0	0
		0	1	0			
exposure	2	0	0	0	0	0	0
		0	0	1			

Contrast Results for GEE Analysis

Contrast	DF	Chi-Square	Pr > ChiSq	Type
treatment effect	2	1.03	0.5985	Score
exposure	2	198.12	<.0001	Score

Hypothesis Testing

What conclusions can we draw about the effect of beta carotene on the occurrence of skin cancers?

- When we conduct a score test of $H_0: \beta_1 = \beta_3 = 0$, we obtain the test statistic 1.03, which has p-value 0.60 when compared to a χ^2_2 distribution. Thus we conclude that beta-carotene is not associated with skin cancer occurrence in these data.

Hypothesis Testing

What is the association between age, skin type, and gender and skin cancer occurrence?

- Age is significantly associated with skin cancer risk; each 10-year increase in age is associated with an odds ratio of 1.02 (1.01, 1.03) for developing a new skin cancer.
- Having skin that burns easily is also associated with skin cancer risk; those who burn easily have 1.25 times the odds (95% CI (1.06, 1.48)) of developing a new cancer as those who do not have skin that burns easily.
- Men are at increased risk, with an odds ratio of 1.93 (1.59, 2.34) compared to women.

Hypothesis Testing

What is the association between number of previously diagnosed skin cancers and subsequent cancer occurrence?

- Conducting a score test of $H_0: \beta_7 = \beta_8 = 0$, we obtain the test statistic 198.02, which has p-value < 0.0001 when compared to a χ^2_2 distribution. Thus we conclude that the number of prior cancers (which ranges from 1 to 21 in these data) is strongly associated with future cancer risk.

Modeling Counts instead of Binary Outcome

- In addition to modeling the binary occurrence of cancer over time, we may also be interested in directly modeling the tumor occurrence rates directly, using the count of new cancers at each follow-up time as the outcome variable. We will use the same form of the linear predictor but fit the model

$$log\left(E\left(Y_{ij}\right)\right) = \beta_0 + \beta_1 trt_i + \beta_2 year_{ij} + \beta_3 (trt_i)\left(year_{ij}\right) + \beta_4 age_i + \beta_5 gender_i + \beta_6 skin_i + \beta_7 exposure_i + \beta_8 exposure_i^2$$

where $year_{ij}$ is the measurement occasion for subject i, $j = 1, \cdots, 5$

Modeling Counts

- Next, we specify the variance function as

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

Recalling $\mu_{ij} = \exp(X_{ij}\beta)$ as we have used the log link function, and noting that we will estimate ϕ in case we have extra-Poisson variability. We do this because of the observed means and variances by time and treatment, presented in the table.

Table 1: Means (Variances) over Time of Cancer Counts

	Year 1	Year 2	Year 3	Year 4	Year 5
Beta-Carotene	0.3 (0.6)	0.3 (0.5)	0.3 (1.1)	0.3 (1.3)	0.3 (0.8)
Placebo	0.3 (0.8)	0.2 (0.5)	0.2 (0.6)	0.2 (0.6)	0.3 (0.7)

Modeling Counts

- While there appears to be some evidence of overdispersion, we must bear in mind that we have not yet adjusted for other explanatory variables including age, skin type, prior skin cancers, and gender.
- We will assume that the within-subject association among the five repeated count responses has an unstructured pattern.

SAS Code

```
proc genmod data=new descending;
class id year;
model y=treatment yearcont treatment*yearcont agecent skin
       gender exposure
        exposure*exposure/d=poisson link=log scale=deviance;
repeated subject=id/withinsubject=year type=un covb;
contrast 'treatment effect' treatment 1 -1, treatment*yearcont 1 -1/e;
contrast 'exposure' exposure 1 -1, exposure*exposure 1 -1/e;
run;
```

SAS Output

The GENMOD Procedure

Model Information

Data Set	WORK.NEW
Distribution	Poisson
Link Function	Log
Dependent Variable	Y

Number of Observations Read	7081
Number of Observations Used	7081

Class Level Information

Class	Levels	Values
ID	1683	30 100012 100023 100034 100045 100056 100067 100078 100089 100102 100113 100124 100146 100157 100168 100179 100190 100203 100214 100236 100247 100258 100269 100280 100291 100304 100326 100348 100359 100370 100381 100392 100405 100416 100427 100460 ...
Year	5	1 2 3 4 5

SAS Output

Parameter	Information Effect
Prm1	Intercept
Prm2	Treatment
Prm3	yearcont
Prm4	Treatment*yearcont
Prm5	agecent
Prm6	Skin
Prm7	Gender
Prm8	Exposure
Prm9	Exposure*Exposure

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	7072	6158.4434	0.8708
Scaled Deviance	7072	7072.0000	1.0000
Pearson Chi-Square	7072	11333.1967	1.6025
Scaled Pearson X2	7072	13014.3873	1.8403
Log Likelihood		-4132.5091	

Algorithm converged.

Analysis Of Initial Parameter Estimates

Parameter	DF	Estimate	Standard Error	Wald 95% Confidence Limits		Chi-Square
Intercept	1	-2.8285	0.0932	-3.0112	-2.6457	920.23
Treatment	1	0.0024	0.0976	-0.1889	0.1936	0.00
yearcont	1	-0.0044	0.0236	-0.0506	0.0418	0.03
Treatment*yearcont	1	0.0365	0.0322	-0.0266	0.0996	1.29
agecent	1	0.0146	0.0025	0.0098	0.0194	35.42
Skin	1	0.0570	0.0436	-0.0284	0.1425	1.71
Gender	1	0.5630	0.0564	0.4525	0.6734	99.81
Exposure	1	0.3156	0.0129	0.2903	0.3408	600.44
Exposure*Exposure	1	-0.0087	0.0006	-0.0099	-0.0075	202.96
Scale	0	0.9332	0.0000	0.9332	0.9332	

Analysis Of Initial Parameter Estimates

Parameter	Pr > ChiSq
Intercept	<.0001
Treatment	0.9805
yearcont	0.8523
Treatment*yearcont	0.2569
agecent	<.0001
Skin	0.1910
Gender	<.0001
Exposure	<.0001
Exposure*Exposure	<.0001
Scale	

NOTE: The scale parameter was estimated by the square root of DEVIANCE/DOF.

SAS Output

GEE Model Information

Correlation Structure	Unstructured
Within-Subject Effect	Year (5 levels)
Subject Effect	ID (1683 levels)
Number of Clusters	1683
Correlation Matrix Dimension	5
Maximum Cluster Size	5
Minimum Cluster Size	1

The GENMOD Procedure

Covariance Matrix (Model-Based)

	Prm1	Prm2	Prm3	Prm4	Prm5
Prm1	0.01070	-0.004650	-0.001241	0.001219	-0.000011
Prm2	-0.004650	0.008894	0.001222	-0.002301	-5.348E-7
Prm3	-0.001241	0.001222	0.0004781	-0.000478	1.307E-7
Prm4	0.001219	-0.002301	-0.000478	0.0008960	5.1492E-7
Prm5	-0.000011	-5.348E-7	1.307E-7	5.1492E-7	9.8043E-6
Prm6	-0.001262	0.0000332	5.9054E-6	9.3436E-6	1.821E-6
Prm7	-0.003982	-0.000040	0.0000110	-6.119E-6	4.2971E-7
Prm8	-0.000627	-0.000018	2.7489E-6	-9.63E-7	-2.365E-6
Prm9	0.0000272	5.9484E-7	-1.427E-7	9.6277E-8	8.1963E-8

SAS Output

Covariance Matrix (Model-Based)

	Prm6	Prm7	Prm8	Prm9
Prm1	-0.001262	-0.003982	-0.000627	0.0000272
Prm2	0.0000332	-0.000040	-0.000018	5.9484E-7
Prm3	5.9054E-6	0.0000110	2.7489E-6	-1.427E-7
Prm4	9.3436E-6	-6.119E-6	-9.63E-7	9.6277E-8
Prm5	1.821E-6	4.2971E-7	-2.365E-6	8.1963E-8
Prm6	0.003120	0.0000698	-0.000153	5.6391E-6
Prm7	0.0000698	0.005241	-0.000082	1.9628E-6
Prm8	-0.000153	-0.000082	0.0002695	-0.000012
Prm9	5.6391E-6	1.9628E-6	-0.000012	6.111E-7

Covariance Matrix (Empirical)

	Prm1	Prm2	Prm3	Prm4	Prm5
Prm1	0.02354	-0.01405	-0.002707	0.003263	-0.000025
Prm2	-0.01405	0.02214	0.002520	-0.005555	0.0000577
Prm3	-0.002707	0.002520	0.0009673	-0.000975	-7.522E-6
Prm4	0.003263	-0.005555	-0.000975	0.002355	-0.000036
Prm5	-0.000025	0.0000577	-7.522E-6	-0.000036	0.0000267
Prm6	-0.007516	0.004275	0.0000960	-0.001054	0.0000719
Prm7	-0.004969	0.0001739	0.0001883	-0.000400	-0.000012
Prm8	-0.002177	0.0007348	0.0000532	-0.000178	8.235E-6
Prm9	0.0001445	-0.000070	-4.287E-6	0.0000176	-1.337E-6

SAS Output

Covariance Matrix (Empirical)

	Prm6	Prm7	Prm8	Prm9
Prm1	-0.007516	-0.004969	-0.002177	0.0001445
Prm2	0.004275	0.0001739	0.0007348	-0.000070
Prm3	0.0000960	0.0001883	0.0000532	-4.287E-6

Covariance Matrix (Empirical)

	Prm6	Prm7	Prm8	Prm9
Prm4	-0.001054	-0.000400	-0.000178	0.0000176
Prm5	0.0000719	-0.000012	8.235E-6	-1.337E-6
Prm6	0.01083	-0.000618	0.001044	-0.000097
Prm7	-0.000618	0.008571	-0.000503	0.0000294
Prm8	0.001044	-0.000503	0.0007970	-0.000048
Prm9	-0.000097	0.0000294	-0.000048	3.2475E-6

Algorithm converged.

SAS Output

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	-2.8865	0.1534	-3.1872	-2.5857	-18.81	<.0001
Treatment	0.0152	0.1488	-0.2765	0.3068	0.10	0.9188
yearcont	0.0039	0.0311	-0.0571	0.0648	0.12	0.9007
Treatment*yearcont	0.0304	0.0485	-0.0647	0.1255	0.63	0.5311
agecent	0.0145	0.0052	0.0044	0.0247	2.81	0.0049
Skin	0.1046	0.1041	-0.0994	0.3085	1.00	0.3149
Gender	0.5756	0.0926	0.3941	0.7571	6.22	<.0001
Exposure	0.3213	0.0282	0.2659	0.3766	11.38	<.0001
Exposure*Exposure	-0.0091	0.0018	-0.0126	-0.0056	-5.05	<.0001

SAS Output

Coefficients for Contrast treatment effect

Label	Row	Prm1	Prm2	Prm3	Prm4	Prm5	Prm6
		Prm7	Prm8	Prm9			
treatment effect	1	0	1	0	0	0	0
		0	0	0			
treatment effect	2	0	0	0	1	0	0
		0	0	0			

The GENMOD Procedure

Coefficients for Contrast exposure

Label	Row	Prm1	Prm2	Prm3	Prm4	Prm5	Prm6
		Prm7	Prm8	Prm9			
exposure	1	0	0	0	0	0	0
		0	1	0			
exposure	2	0	0	0	0	0	0
		0	0	1			

Contrast Results for GEE Analysis

Contrast	DF	Chi-Square	Pr > ChiSq	Type
treatment effect	2	1.21	0.5463	Score
exposure	2	144.26	<.0001	Score

Hypothesis Testing

What conclusions can we draw about the effect of beta carotene on the occurrence of skin cancers?

- When we conduct a test of $H_0: \beta_1 = \beta_3 = 0$, we obtain the test statistic 1.21, which has p-value 0.55 when compared to a χ^2_2 distribution. Thus we conclude that beta-carotene is not associated with skin cancer occurrence in these data.

Hypothesis Testing

What is the association between age, skin type, and gender and skin cancer occurrence?

- Age is significantly associated with skin cancer risk; each 10-year increase in age is associated with an rate ratio of 1.01 (1.01, 1.03) for developing a new skin cancer.
- Having skin that burns easily is not associated with rate of skin cancer occurrence.
- Men are at increased risk, with an rate ratio of 1.78 (1.48, 2.13) compared to women.

Hypothesis Testing

What is the association between number of previously diagnosed skin cancers and subsequent cancer occurrence?

- Conducting a test of $H_0: \beta_7 = \beta_8 = 0$, we obtain the test statistic 144.26, which has p-value < 0.0001 when compared to a χ^2_2 distribution. Thus we conclude that the number of prior cancers (which ranges from 1 to 21 in these data) is strongly associated with future cancer risk.