

1. **Objective:** analyzing CD4 counts data using GEE marginal model and random effects model. Demonstrate the analysis using SAS and STATA.

2. **Scientific question:** estimate the average time course of CD4 cell depletion.

3. **Dataset:** CD4 data set (cd4.data)

To simplify the demonstration, we are only interested in how the CD4 counts change with time. Thus we only consider one covariate — time.

Outcome: CD4 counts, is a continuous variables and ranges from 10 to 3183 with mean 765.

Covariate: Time, is defined as the year after seroconversion. It ranges from -3 to 5.5.

4. **EDA:** we already did in previous lab.

The smooth plots indicate that the average CD4 cell is nearly constant until the time of seroconversion and then decrease very rapidly.

5. **Model:**

- **GEE marginal model**

Mean model is:

$$E[CD4] = b_0 + b_1 time + b_2 (time - 0)_+ \quad (1)$$

For the correlation structure we will consider exchangeable (uniform) correlation and autoregressive(1) correlation.

In above model 1, CD4 is the cell counts for each subject. $(time-0)_+$ is a spline term for the time and is equal to time if time >0 and equal to zero if time <0.

b_0 is the average CD4 cell counts at seroconversion, b_1 is the time slope until seroconversion, and $b_1 + b_2$ is the time slope after seroconversion.

- **Random effects model**

$$E[CD4_{ij}] = b_0 + a_i + b_1 time + b_2 (time - 0)_+ + e_{ij} \quad (2)$$

The interpretation of b_0 , b_1 , $b_1 + b_2$ are the same as model 1, α_i is the random intercept for the i th subject, and e_{ij} is assumed to follow $N(0, s^2)$.

6. **Model fitting using SAS**

- **SAS Macro**

```
*****
* Demonstration of SAS analysis using GEE *
* marginal model and Random effects model *
* on CD4 count data *
* Date: Feb, 6, 2002 *
***** ;

OPTIONS LS=80 FORMDLIM='*' NODATE;

**SPECIFY A LIBRARY PATH**;
```

```
LIBNAME lda 'c:/data';
```

```

DATA cd4data;
SET lda.cd4;
sptime= time;
IF sptime LE 0 THEN sptime=0;
RUN;

***SUMMARY OF THE DATASET***;
PROC CONTENTS DATA=cd4data;
TITLE 'SUMMARY OF THE CD4 DATA';
RUN;

*****
**GEE MARGINAL MODEL OF CD4 COUNT ON TIME**
*****
**EXCHANGEABLE CORRELATION STRUCTURE*****;
PROC GENMOD DATA=cd4data;
CLASS id;
MODEL cd4 = sptime time /DIST=NORMAL;
REPEATED SUBJECT=id/TYPE=CS;
ESTIMATE 'time slope' intercept 0 sptime 1 time 1;
TITLE 'GEE MARGINAL MODEL WITH EXCHANGEABLE CORRELATION';
RUN;

**AUTOREGRESSIVE(1) CORRELATION STRUCTURE*****;
PROC GENMOD DATA=cd4data;
CLASS id;
MODEL cd4 = sptime time /DIST=NORMAL;
REPEATED SUBJECT=id/TYPE=AR;
ESTIMATE 'time slope' intercept 0 sptime 1 time 1;
TITLE 'GEE MARGINAL MODEL WITH AUTOREGRESSIVE(1) CORRELATION';
RUN;

*****;
**RANDOM EFFECTS MODEL OF CD4 COUNT ON TIME*;
*****;
PROC MIXED DATA=cd4data METHOD=ML COVTEST NOCLPRINT;
CLASS id;
MODEL cd4 = sptime time /solution;
RANDOM INTERCEPT /SUBJECT=id g;
ESTIMATE 'time slope' intercept 0 sptime 1 time 1;
TITLE ' RANDOM EFFECTS MODEL WITH ONLY INTERCETP';
RUN;

```

▪ SAS Output

```

*****
SUMMARY OF THE CD4 DATA
45

The CONTENTS Procedure

Data Set Name: WORK.CD4DATA      Observations:      2376
Member Type:   DATA             Variables:         9
Engine:        V8                Indexes:          0
Created:       16:54 Monday, February 4, 2002  Observation Length: 56
Last Modified: 16:54 Monday, February 4, 2002  Deleted Observations: 0

```

Protection:	Compressed:	NO
Data Set Type:	Sorted:	NO
Label:		

-----Engine/Host Dependent Information-----

Data Set Page Size: 8192
 Number of Data Set Pages: 17
 First Data Page: 1
 Max Obs per Page: 145
 Obs in First Data Page: 112
 Number of Data Set Repairs: 0
 File Name: C:\Temp\SAS Temporary Files_TD1792\cd4data.sas7bdat
 Release Created: 8.0202M0
 Host Created: WIN_PRO

-----Alphabetic List of Variables and Attributes-----

#	Variable	Type	Len	Pos
3	Age	Num	8	8
2	CD4	Num	4	24
7	Cesd	Num	4	40
5	Drugs	Num	4	32
8	ID	Num	6	44
4	Packs	Num	4	28
6	Sex	Num	4	36
1	Time	Num	8	0
9	sptime	Num	8	16

GEE MARGINAL MODEL WITH EXCHANGEABLE CORRELATION

46

The GENMOD Procedure

Model Information

Data Set	WORK.CD4DATA
Distribution	Normal
Link Function	Identity
Dependent Variable	CD4
Observations Used	2376

Parameter Information

Parameter	Effect
Prm1	Intercept
Prm2	sptime
Prm3	Time

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
-----------	----	-------	----------

Deviance	2373	312511831.60	131694.8300
Scaled Deviance	2373	2376.0000	1.0013
Pearson Chi-Square	2373	312511831.60	131694.8300
Scaled Pearson X2	2373	2376.0000	1.0013
Log Likelihood		-17374.3293	

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > Z
Intercept	848.3204	16.9625	815.0746	881.5662	50.01	<.0001
sptime	-18.4783	15.9074	-49.6562	12.6997	-1.16	0.2454
Time	-86.2720	13.1366	-112.019	-60.5247	-6.57	<.0001

Contrast Estimate Results

Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
time slope	-104.750	6.3099	0.05	-117.118	-92.3831	275.59	<.0001

GEE MARGINAL MODEL WITH AUTOREGRESSIVE(1) CORRELATION 48

The GENMOD Procedure

Model Information

Data Set	WORK.CD4DATA
Distribution	Normal
Link Function	Identity
Dependent Variable	CD4
Observations Used	2376

Parameter Information

Parameter	Effect
Prm1	Intercept
Prm2	sptime
Prm3	Time

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	2373	312511831.60	131694.8300
Scaled Deviance	2373	2376.0000	1.0013
Pearson Chi-Square	2373	312511831.60	131694.8300
Scaled Pearson X2	2373	2376.0000	1.0013
Log Likelihood		-17374.3293	

Analysis Of GEE Parameter Estimates

Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > Z
Intercept	850.1748	17.8268	815.2349	885.1148	47.69	<.0001
sptime	-26.2818	17.7116	-60.9959	8.4324	-1.48	0.1378
Time	-73.2670	14.6756	-102.031	-44.5033	-4.99	<.0001

Contrast Estimate Results

Label	Estimate	Standard Error	Alpha	Confidence Limits		Chi-Square	Pr > ChiSq
time slope	-99.5488	6.9226	0.05	-113.117	-85.9806	206.79	<.0001

RANDOM EFFECTS MODEL WITH ONLY INTERCETP

50

The Mixed Procedure

Model Information

Data Set	WORK.CD4DATA
Dependent Variable	CD4
Covariance Structure	Variance Components
Subject Effect	ID
Estimation Method	ML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Containment

Estimated G Matrix

Row	Effect	ID	Col1
1	Intercept	10002	64717

Covariance Parameter Estimates

Cov Parm	Subject	Estimate	Standard Error	Z Value	Pr Z
Intercept	ID	64717	5662.44	11.43	<.0001
Residual		66394	2093.41	31.72	<.0001

RANDOM EFFECTS MODEL WITH ONLY INTERCETP

51

The Mixed Procedure

Fit Statistics

-2 Log Likelihood	33828.8
AIC (smaller is better)	33838.8
AICC (smaller is better)	33838.8
BIC (smaller is better)	33858.3

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > t
Intercept	848.32	16.3983	368	51.73	<.0001
sptime	-18.4092	11.9824	2005	-1.54	0.1246
Time	-86.2538	9.3837	2005	-9.19	<.0001

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
sptime	1	2005	2.36	0.1246
Time	1	2005	84.49	<.0001

Estimates

Label	Estimate	Standard Error	DF	t Value	Pr > t
time slope	-104.66	4.7389	2005	-22.09	<.0001

7. Model fitting using STATA

```
. sum cd4
```

Variable	Obs	Mean	Std. Dev.	Min	Max
cd4	2376	765.1313	399.3716	10	3184

```
. sum time
```

Variable	Obs	Mean	Std. Dev.	Min	Max
time	2376	.8284246	1.878228	-2.989733	5.459274

```
. gen sptime=time
```

```
. replace sptime=0 if sptime < 0
```

```
(863 real changes made)
```

```
. xtgee cd4 sptime time, i(id) f(gaussian) corr(exc)
```

```
Iteration 1: tolerance = 5.5773989
```

```
Iteration 2: tolerance = .00309178
```

```
Iteration 3: tolerance = .00001831
```

```
Iteration 4: tolerance = 1.094e-07
```

GEE population-averaged model		Number of obs	=	2376
Group variable:	id	Number of groups	=	369
Link:	identity	Obs per group: min	=	1
Family:	Gaussian	avg	=	6.4
Correlation:	exchangeable	max	=	12

Scale parameter: 132098.4 Wald chi2(2) = 851.16
 Prob > chi2 = 0.0000

cd4	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
sptime	-18.48101	11.89907	-1.55	0.120	-41.80275	4.840726
time	-86.27276	9.321519	-9.26	0.000	-104.5426	-68.00291
_cons	848.3205	16.5231	51.34	0.000	815.9358	880.7052

. lincom sptime+time
 (1) sptime + time = 0.0

cd4	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
(1)	-104.7538	4.706057	-22.26	0.000	-113.9775	-95.53007

. xtgee cd4 sptime time, i(id) f(gaussian) corr(ar1) t(time)
 note: observations not equally spaced
 modal spacing is delta time = .49828899
 364 groups omitted from estimation
 note: some groups have fewer than 2 observations
 not possible to estimate correlations for those groups
 5 groups omitted from estimation
 insufficient observations
 r(2000);

. xtreg cd4 sptime time, i(id) re
 Random-effects GLS regression Number of obs = 2376
 Group variable (i) : id Number of groups = 369

 R-sq: within = 0.2869 Obs per group: min = 1
 between = 0.0880 avg = 6.4
 overall = 0.1747 max = 12

 Random effects u_i ~ Gaussian Wald chi2(2) = 836.57
 corr(u_i, X) = 0 (assumed) Prob > chi2 = 0.0000

cd4	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
sptime	-18.37797	12.00121	-1.53	0.126	-41.8999	5.143964
time	-86.24575	9.397068	-9.18	0.000	-104.6637	-67.82784
_cons	848.3168	16.32439	51.97	0.000	816.3216	880.312
sigma_u	252.11599					
sigma_e	257.8637					
rho	.48873095	(fraction of variance due to u_i)				

. lincom sptime+time
 (1) sptime + time = 0.0

cd4	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
(1)	-104.6237	4.746328	-22.04	0.000	-113.9264	-95.32109

8. Model interpretation

From above analysis, we found the estimate for the regression coefficient b_1 is about -86 with p-value less than 0.001, and the estimate for the regression coefficient b_1+b_2 is about -105 with p-value less than 0.001. Therefore the regression results provide more informative estimation of the average time course than EDA: before seroconversion the average CD4 counts decrease 86/per year and is significant and after seroconversion the average CD4 counts decrease 105/per year and is statistical significant.