BIST P8110: Applied Regression II

22. GEE

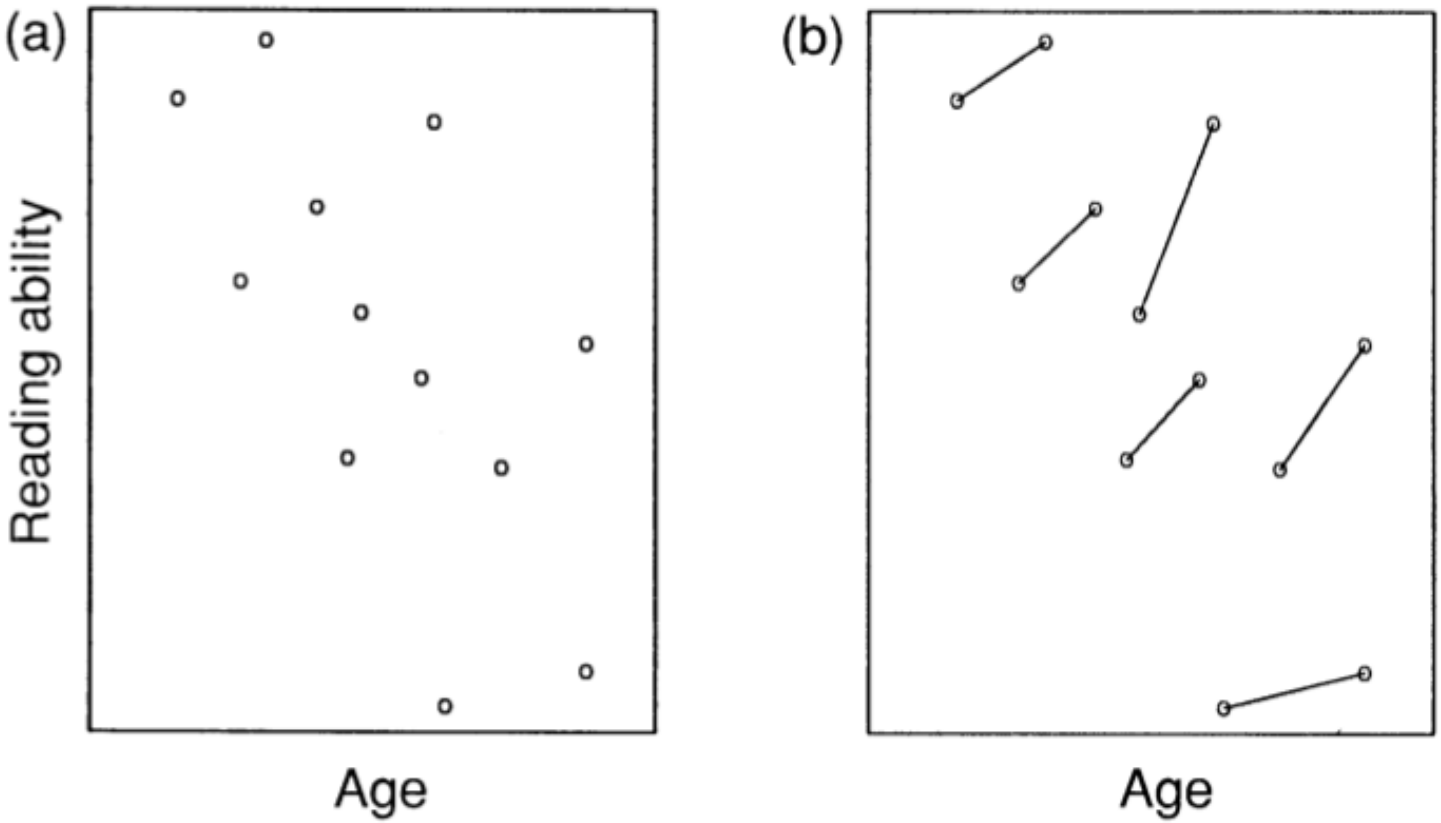
Qixuan Chen

Department of Biostatistics
Columbia University

Introduction to Repeated Measures

- Repeated measures analysis is concerned with study designs in which the same patient is observed repeatedly over time.
- Such data is also referred to as longitudinal data.
- In analyzing these data, we must take into consideration the fact that the multiple observations on the same patient are usually correlated.

Some hypothetical data



An Example of Repeated Measures

- An example of longitudinal data
  - The blood pressure of the same patient at different times in a day.
  - Measurements on the same patient at different times may be more alike than measurements on different patients.
  - To make valid statistical inference, models that take into account the correlation of multiple measurements on the same patient are needed.

Data Structure

- The data structure of the repeated measures data of i^{th} patient's blood pressure looks like

$$Y_i = \begin{bmatrix} BP_{i1} \\ BP_{i2} \\ BP_{i3} \end{bmatrix}$$

$$X_i = \begin{bmatrix} t_{i1} & age_i \\ t_{i2} & age_i \\ t_{i3} & age_i \end{bmatrix} = \begin{bmatrix} 1 & 40 \\ 2 & 40 \\ 3 & 40 \end{bmatrix}$$

where, age is the baseline age and it is time-invariant.

Statistical Models for Repeated Measures

- There are two approaches on modeling repeated measures data
  - The population-averaged approach: **estmating equations**
  - The subject-specific approach: **mixed models**

GEE

- Generalized Estimating Equation (GEE) extends the generalized linear models so that it can handle repeated measures data (Liang and Zeger, 1986).
- To do GEE analysis, we need to specify
  - Mean response of a data vector Y_i,
  - Variance of each element of Y_i,
  - Working correlation matrix among observations of Y_i on the same unit.

GEE: Mean Response

- The mean response of a data vector y_i is modeled as a function of time, other covariates, and parameters β by using a **generalized linear model**-type mean structure to represent the mean response of each element of Y_i
  - Continuous data Y_{ij}:

$$E(Y_{ij}) = \beta_0 + \beta_1 \times t_{ij} + \beta_2 \times \text{age}_i$$

  - Binary data Y_{ij}:

$$\text{logit}\{Pr(Y_{ij} = 1)\} = \beta_0 + \beta_1 \times t_{ij} + \beta_2 \times \text{age}_i$$

  - Count data Y_{ij}:

$$\log\{E(Y_{ij})\} = \log(n_{ij}) + \beta_0 + \beta_1 \times t_{ij} + \beta_2 \times \text{age}_i$$

GEE: Variance

- The variance of each element of Y_i is
  - Continuous data Y_{ij}:

$$Var(Y_{ij}) = \sigma^2$$

  - Binary data Y_{ij}:

$$Var(Y_{ij}) = E(Y_{ij})(1 - E(Y_{ij}))$$

  - Count data Y_{ij}:

$$Var(Y_{ij}) = E(Y_{ij})$$

- The variance are often modified to allow for greater variation by the addition of a dispersion parameter ϕ.

GEE: Working Correlation Matrix

- The correlation among pairs of observations on the same data vector needs to be specified.
- The often used correlation structures are
  - Independence
  - Unstructured correlation
  - Compound symmetry correlation
  - AR(1) correlation
- The chosen correlation structure is referred to as the "working correlation matrix", because there is no formal way to check whether the correlation structure is correct.

GEE: Unstructured Correlation

- Unstructured correlation places no restriction on the nature of association among elements of a data vector

$$\begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{21} & 1 & \rho_{23} \\ \rho_{31} & \rho_{32} & 1 \end{bmatrix}$$

  where, $\rho_{ij} = \rho_{ji}$.
- The unstructured working correlation assumption in this example depends on $k(k-1)/2 = 3 \times 2/2 = 3$ distinct correlation parameters.

GEE: Compound Symmetry Correlation

- Compound symmetry (exchangeable) correlation assumes that the correlation between distinct observations on the same unit is the same regardless of when in time the observations were taken.

$$\begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}$$

- Compound symmetry correlation assumption depends on only one correlation parameter.

GEE: AR(1) Correlation

- AR(1) correlation assumes that correlation among observations "tails off"

$$\begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix}$$

or, more generally

$$\begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{k-1} \\ \rho & 1 & \rho & \dots & \rho^{k-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{k-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \rho^{k-1} & \dots & \rho^2 & \rho & 1 \end{bmatrix}$$

GEE: Estimate β

- We won't talk about how, but GEE can be used to estimate the regression coefficients and variance parameters in GLMs for repeated measures.
- Two variance estimation methods
  - model-based variance estimate
  - robust ("sandwich" or "empirical") variance estimate

GEE: Wald Test for β

- We can use the Wald test to test the hypothesis
$$H_0 : \beta_1 = 0 \text{ vs. } H_\alpha : \beta_1 \neq 0$$
- The test statistic
$$\frac{(\hat{\beta}_1 - 0)^2}{V_\beta} \sim \chi^2_1$$

Case Study I: Repeated Measures for Normal Data

Blood Flow Data

- Example: effect of race and dose of isoproterenol on blood flow (Dupont 2009)
  - “Lang et al. (1995) studied the effect of isoproterenol on forearm blood flow in a group of 22 normotensive men. Nine of the study subjects were blacks and 13 were whites. Each subject’s blood flow was measured at baseline and then at escalating doses of isoproterenol."

    Lang, C.C., Stein, C.M., Brown, R.M., Deegan, R., Nelson, R., He, H.B., et al. Attenuation of isoproterenol-mediated vasodilatation in blacks. *N. Engl. J. Med.* 1995; **333:** 155-60

Data Structure

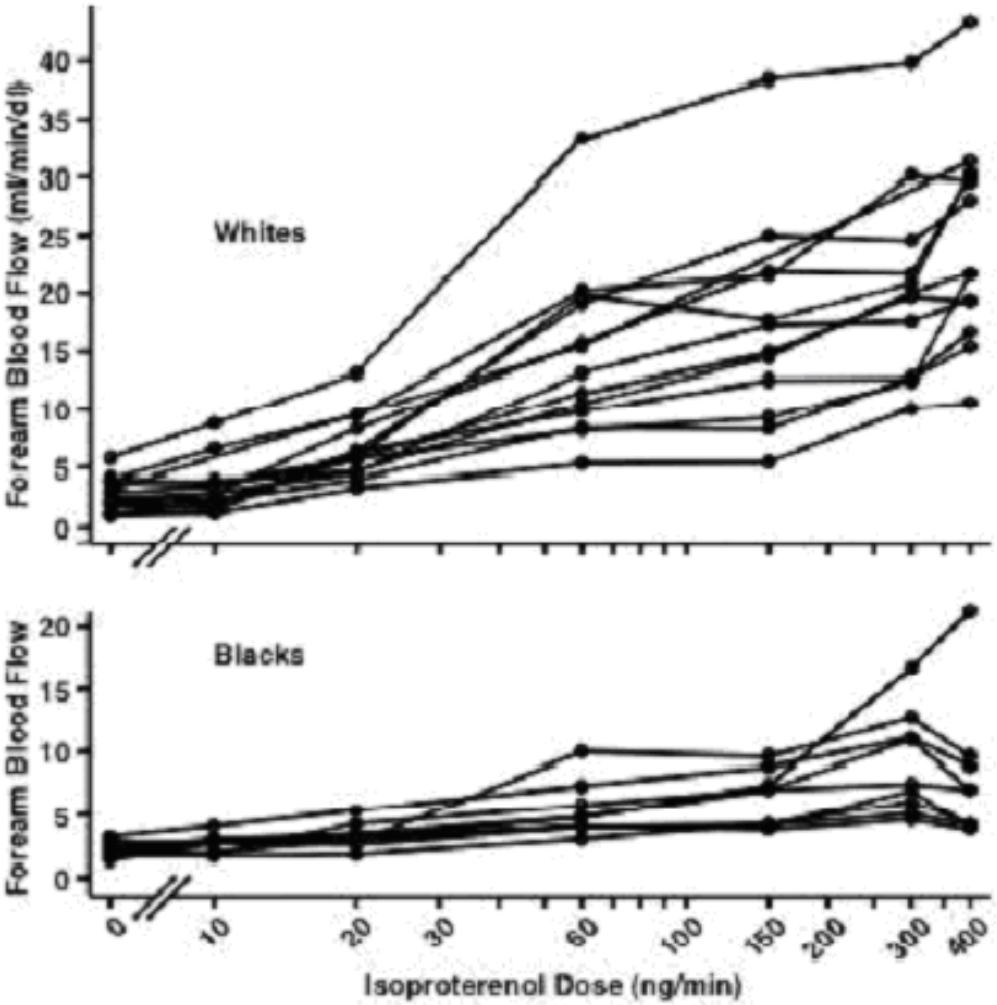
The repeated measures data for the first two subjects are displayed as follow:

id	dose	race	fbf
1	0	1	1.0000
1	10	1	1.4000
1	20	1	6.4000
1	60	1	19.1000
1	150	1	25.0000
1	300	1	24.6000
1	400	1	28.0000
2	0	1	2.1000
2	10	1	2.8000
2	20	1	8.3000
2	60	1	15.7000
2	150	1	21.9000
2	300	1	21.7000
2	400	1	30.1000

Research Questions

- The key research questions to address in this study are
  1. whether forearm blood flow increases as the dose of isoproterenol increases,
  2. and whether the effect of dose of isoproterenol on blood flow is different between normotensive black and white men.

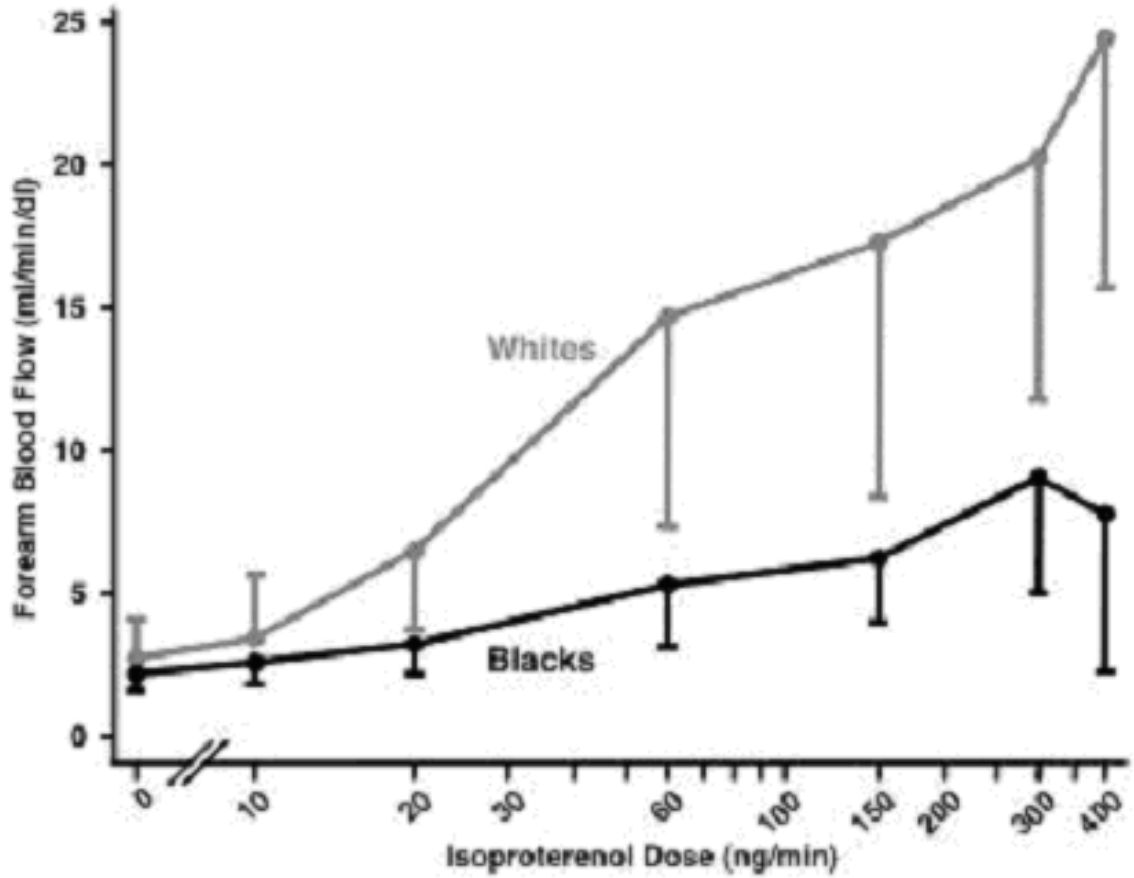
Descriptive Figures



- Figure 1: Plot of forearm blood flow against isoproterenol dose for white and black men. Straight lines connect observations from the same study subjects.
- This plot is useful when the number of subjects is small.
- This is called "spaghetti plot".

Descriptive Figures

- Figure 2: Mean rates of forearm blood flow in black and white men in response to different doses of isoproterenol. The vertical bars indicate the estimated standard deviation within each racial group at each dose.
- The information in this figure can also be displayed in a summary table.



Descriptive Figures

- Figures 1 and 2 show that the the rate of forearm blood flow increases with isoproterenol dose, and this increase is faster among white men than among black men.
- Next, we want to fit a statistical model to estimate the association between isoproterenol dose and forearm blood flow.

Statistical Model

- The mean function of the repeated measures model for the forearm blood flow is

  $$E(Y_{ij}) = \beta_0 + \beta_1 \times dose_{ij} + \beta_2 \times race_i + \beta_3 \times dose_{ij} \times race_i$$

  where,
  - Y_{ij} = forearm blood flow (ml/min/dl)
  - race = 1 if white, = 2 if black (0 if white, 1 if black, after using the CLASS statement on next page)
  - dose = 0, 10, 20, 60, 150, 300, and 400 (ng/min)

- Use GEE method to estimate β, by assuming some working correlation structure.

SAS Codes: Compound Symmetry

The following SAS statements invoke the GENMOD procedure to perform the analysis

```
proc genmod data=isoproterenol;
class race id /param=ref ref=first;
model fbf = race dose race*dose /dist=nor link=identity
                                 type3 wald;
repeated subject=id /type=cs modelse;
run;
```

NOTE: "type=cs" is to request compound symmetry correlation matrix and "modelse" is to request the output of model-based standard error estimates. The default is to only output the empirical (robust) standard error estimates. "subject=" is to specify the variable defining the subject effect. For example, it is "id" in this data set, which needs to be specified in the CLASS statement as well.

SAS Output: CS

Model Information

Data Set	WORK.ISOPROTERENOL
Distribution	Normal
Link Function	Identity
Dependent Variable	fbf
Number of Observations Read	154
Number of Observations Used	150
Missing Values	4

Algorithm converged.

GEE Model Information

Correlation Structure	Exchangeable
Subject Effect	id (22 levels)
Number of Clusters	22
Clusters With Missing Values	1
Correlation Matrix Dimension	7
Maximum Cluster Size	7
Minimum Cluster Size	3

SAS Output: CS

GEE Fit Criteria

QIC	162.1743
QICu	154.0000

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		6.1797	0.8432	4.5270	7.8324	7.33	<.0001
race	2	-3.0334	0.9042	-4.8055	-1.2612	-3.35	0.0008
dose		0.0495	0.0046	0.0405	0.0586	10.72	<.0001
dose*race	2	-0.0343	0.0063	-0.0467	-0.0220	-5.44	<.0001

QIC: Quasi-likelihood Information Criterion, a modification of the Akaike information criterion (AIC) to apply to models fit by GEEs. A model with smaller QIC is preferred. QIC can be used to select regression models and working correlations.

SAS Output: CS

Analysis Of GEE Parameter Estimates
Model-Based Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		6.1797	1.2344	3.7603	8.5991	5.01	<.0001
race	2	-3.0334	1.9204	-6.7974	0.7306	-1.58	0.1142
dose		0.0495	0.0031	0.0435	0.0556	16.06	<.0001
dose*race	2	-0.0343	0.0048	-0.0437	-0.0249	-7.18	<.0001
Scale		5.7260	.	.	.	.	.

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

Wald Statistics For Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
race	1	11.25	0.0008
dose	1	114.86	<.0001
dose*race	1	29.63	<.0001

SAS Codes: AR(1) Correlation Matrix

The same GEE model can be fitted using AR(1) working correlation matrix:

```
proc genmod data=isoproterenol;
class race id /param=ref ref=first;
model fbf = race dose race*dose /dist=nor link=identity
                                 type3 wald;
repeated subject=id /type=AR(1) modelse;
run;
```

NOTE: "type=AR(1)" is to request AR(1) correlation matrix.

Test for the interaction term

- Hypothesis: $H_0 : \beta_3 = 0$ vs $H_\alpha : \beta_3 \neq 0$
- Test statistic: $T_W = 29.63$ with 1 degree of freedom
- p-value: $Pr(\chi^2_1 \geq 29.63) < 0.0001$
- Conclusion: Reject H_0 at $\alpha = 0.05$. There is sufficient evidence to conclude that the effect of isoproterenol dose on blood flow is different between black men and white men.

Interpret β

- How to interpret β_1?
  - mean Y among whites with a dose of (a+1):
    $E(Y|race = 0, dose = (a+1)) =$
    $\beta_0 + \beta_1 \times (a+1) + \beta_2 \times 0 + \beta_3 \times (a+1) \times 0 = \beta_0 + \beta_1 \times (a+1)$
  - mean Y among whites with a dose of a: $E(Y|race = 0, dose = a) = \beta_0 + \beta_1 \times a + \beta_2 \times 0 + \beta_3 \times a \times 0 = \beta_0 + \beta_1 \times a$
  - mean increase in Y among whites for each unit increase in dose:
    $E(Y|race = 0, dose = (a+1)) - E(Y|race = 0, dose = a) = \beta_1$

Interpret β

- How to interpret $\beta_1 + \beta_3$?
  - mean Y among blacks with a dose of (a+1):
    $E(Y|race = 1, dose = (a+1)) =$
    $\beta_0+\beta_1\times(a+1)+\beta_2\times 1+\beta_3\times(a+1)\times 1 = \beta_0+\beta_2+(\beta_1+\beta_3)\times(a+1)$
  - mean Y among blacks with a dose of a:
    $E(Y|race = 1, dose = a) = \beta_0 + \beta_1 \times a + \beta_2 \times 1 + \beta_3 \times a \times 1 =$
    $\beta_0 + \beta_2 + (\beta_1 + \beta_3) \times a$
  - mean increase in Y among blacks for each unit increase in dose: $E(Y|race = 1, dose = (a+1)) - E(Y|race = 1, dose = a) = \beta_1 + \beta_3$

Interpret β

- Interpretation of β_3:
  - For each unit increase in isoproterenol dose, on average, the increase in forearm blood flow rate among whites is 0.0343 faster than that among blacks.

Case Study II: Repeated Measures of Binary Data

Repeated Measures of Binary Data

- Example: The "wheezing" data
  - We consider a data set on 32 children to study the association between maternal smoking and respiratory health of children.
  - Each child was examined once a year at a clinic visit (visits at ages 9, 10, 11, and 12) for evidence of "wheezing" (0=wheezing absent, 1=wheezing present). In addition, the mother's current smoking status was recorded (0=none, 1=moderate, 2=heavy).

Data Structure

The repeated measures data in an external file have the following wide format:

child	city	a1	s1	y1	a2	s2	y2	a3	s3	y3	a4	s4	y4
1	portage	9	0	1	10	0	1	11	0	1	12	0	0
2	kingston	9	1	1	10	2	1	11	2	0	12	2	0
3	kingston	9	0	1	10	0	0	11	1	0	12	1	0
4	portage	9	0	0	10	0	1	11	0	1	12	1	0
5	kingston	9	0	0	10	1	0	11	1	0	12	1	0
6	portage	9	0	0	10	1	0	11	1	0	12	1	0

NOTE: Each child has a single line of observation.

a1- age at first visit; s1- maternal smoking status at first visit; y1- wheezing status at first visit; a2, s2, and y2 - information at second visit; a3, s3, y3 - information at third visit; a4, s4, y4 - information at fourth visit.

Data Structure

The dataset needs to be transposed to have the following long format:

Obs	child	city	age	smoke	wheeze
1	1	portage	9	0	1
2	1	portage	10	0	1
3	1	portage	11	0	1
4	1	portage	12	0	0
5	2	kingston	9	1	1
6	2	kingston	10	2	1
7	2	kingston	11	2	0
8	2	kingston	12	2	0

NOTE: Each child has multiple lines of observations.

SAS Code to Read in Data

The following SAS codes can be used to read in the data and transpose the data into the long format:

```
data wheeze;
 infile "C:\wheeze.txt";
 input child city $ @@;
 do i=1 to 4;
  input age smoke wheeze @@;
 output;
 end;
run;
```

Research Questions

- The key research question to address in this study is whether the maternal smoking is associated with the respiratory health of children (measured as evidence of wheezing).

Statistical Model

- The mean function of the GEE model for the wheezing status is

$$\text{logit}\{E(Y_{ij}|X_{ij})\} = \beta_0+\beta_1 x_{1ij}+\beta_2 x_{2ij}+\beta_3 x_{3ij}+\beta_4 x_{4ij}+\beta_5 x_{5ij}+\beta_6 x_{6i}$$

or,

$$E(Y_{ij}|X_{ij}) = \frac{e^{\beta_0+\beta_1 x_{1ij}+\beta_2 x_{2ij}+\beta_3 x_{3ij}+\beta_4 x_{4ij}+\beta_5 x_{5ij}+\beta_6 x_{6i}}}{1+e^{\beta_0+\beta_1 x_{1ij}+\beta_2 x_{2ij}+\beta_3 x_{3ij}+\beta_4 x_{4ij}+\beta_5 x_{5ij}+\beta_6 x_{6i}}}$$

where,

  - Y = 0 if wheezing absent, = 1 if wheezing present
  - x_1 = 1 if moderate smoking, = 0 otherwise
  - x_2 = 1 if heavy smoking, = 0 otherwise
  - x_3 = 1 if age=10, = 0 otherwise
  - x_4 = 1 if age=11, = 0 otherwise
  - x_5 = 1 if age=12, = 0 otherwise
  - x_6 = 1 if Portage, = 0 if Kingston

Missing Data

- Missing data is an issue here: 3 missed 1 visit, 6 missed 2 visits, and 4 missed 3 visits among 32 children.
- The GEE analysis using completed cases assumes that the data are missing completely at random.

SAS Code

The following SAS statements invoke the GENMOD procedure to perform the analysis:

```
proc genmod data=wheeze descending;
class child city smoke age/param=ref ref=first;
model wheeze  = smoke age city /dist=bin link=logit
                                type3 wald;
repeated subject = child /ty=AR(1) within=age;
```

NOTE: “within=age" defines an effect specifying the order of measurements within subjects. If some measurements do not appear in the data for some subjects, this option properly orders the existing measurements and treats the omitted measurements as missing values. If the WITHIN= option is not used in this situation, measurements might be improperly ordered and missing values assumed for the last measurements in a cluster.

SAS Output

GEE Model Information

Correlation Structure	AR(1)
Within-Subject Effect	age (4 levels)
Subject Effect	child (32 levels)
Number of Clusters	32
Correlation Matrix Dimension	4
Maximum Cluster Size	4
Minimum Cluster Size	1

Algorithm converged.

GEE Fit Criteria

QIC	130.2296
QICu	130.4161

SAS Output

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-0.3219	0.5801	-1.4589	0.8150	-0.55	0.5789
smoke	1	-0.2614	0.4278	-1.0998	0.5770	-0.61	0.5412
smoke	2	0.4332	0.5609	-0.6661	1.5325	0.77	0.4399
age	10	-0.2853	0.5821	-1.4262	0.8556	-0.49	0.6240
age	11	-0.8289	0.6113	-2.0271	0.3692	-1.36	0.1751
age	12	-0.4038	0.6068	-1.5931	0.7856	-0.67	0.5058
city	portage	-0.3085	0.4726	-1.2347	0.6178	-0.65	0.5139

Wald Statistics For Type 3 GEE Analysis

Source	DF	Chi-Square	Pr > ChiSq
smoke	2	1.03	0.5962
age	3	2.03	0.5656
city	1	0.43	0.5139

Case Study III: Repeated Measures for Count Data

Repeated Measures for Count Data

- Example: The epileptic seizure data
  - Fifty-eight subjects suffering from epileptic seizures were assigned at random to receive either a placebo or the anti-seizure drug progabide in addition to a standard chemotherapy regimen all were taking.
  - On each subject, the investigators recorded the subject's age, a baseline number of seizures experienced by each subject over the 8-week period prior to the start of the study, and then the number of seizures over a 2 week period for four visits following initiation of assigned treatment.

Data Structure

The repeated measures data for two subjects are displayed as follow:

subject	seize	visit	trt	base	age
104	11	0	0	11	31
104	5	1	0	11	31
104	3	2	0	11	31
104	3	3	0	11	31
104	3	4	0	11	31
107	6	0	0	6	25
107	2	1	0	6	25
107	4	2	0	6	25
107	0	3	0	6	25
107	5	4	0	6	25

Research Questions

- The key research questions to addess in this study are
  1. whether the epileptic seizure count changes pre- and post-treatment
  2. whether this change is different between the placebo group and the progabide group

Descriptive Statistics

- Before we specify the model, we consider some summary statistics.
- This was a randomized study, so we would expect subjects in the two groups to be similar in their characteristics prior to administration of the treatment.

	Age (SD)	Baseline (SD)	number of subjects
Placebo	29.0 (6.0)	30.8 (26.1)	28
Progabide	27.9 (6.6)	27.6 (17.4)	30

NOTE: the baseline seizure counts are based on 8 weeks, and the counts at each visits 1-4 are based on 2 weeks.

Descriptive Statistics

- The sample mean seizure counts (SD) at baseline and each visit time are

Visit	Placebo	Progabide
0	7.7 (6.5)	6.9 (4.3)
1	9.4 (10.1)	5.5 (5.8)
2	8.3 (8.2)	6.5 (5.6)
3	8.8 (14.7)	6.0 (7.4)
4	8.0 (7.6)	4.8 (4.3)

NOTE: the baseline seizure counts are divided by 4, so that the number presented here are all based on 2 weeks

Statistical Model

- We adopt a model for mean seizure counts that allows the possibility of a different mean count at baseline and visits 1-4, where the mean at visits 1-4 is the same, and these might be different by group.
- The repeated measures model is

$$log\{E(Y_{ij})\} = log\{n_{ij}\} + \beta_0 + \beta_1 \times v_{ij} + \beta_2 \times trt_i + \beta_3 \times v_{ij} \times trt_i$$

where,

  - Y_{ij} =seizure count
  - v = 0 if baseline, = 1 after treatment
  - trt = 0 if placebo, = 1 if progabide

SAS Codes: Compound Symmetry

The following SAS statements invoke the GENMOD procedure to perform the analysis

```
data seizure; set seizure;
if visit = 0 then v = 0;
else v = 1;
if visit = 0 then log_n = log(8);
else log_n = log(2);
run;

proc genmod data=seizure;
 class subject;
 model seize = trt v trt*v /dist=poi link=log offset=log_n;
 repeated subject=subject /type=cs modelse;
 run;
```

SAS Output: CS

Model Information

Data Set	WORK.SEIZURE
Distribution	Poisson
Link Function	Log
Dependent Variable	seize
Offset Variable	log_n

Algorithm converged.

GEE Model Information

Correlation Structure	Exchangeable
Subject Effect	subject (58 levels)
Number of Clusters	58
Correlation Matrix Dimension	5
Maximum Cluster Size	5
Minimum Cluster Size	5

SAS Output: CS

Exchangeable Working Correlation

Correlation	0.5941485833

GEE Fit Criteria

QIC	-1052.5376
QICu	-1060.3906

SAS Output: CS

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	1.3476	0.1574	1.0392	1.6560	8.56	<.0001
trt	-0.1080	0.1937	-0.4876	0.2716	-0.56	0.5770
v	0.1108	0.1161	-0.1168	0.3383	0.95	0.3399
trt*v	-0.3016	0.1712	-0.6371	0.0339	-1.76	0.0781

Analysis Of GEE Parameter Estimates
Model-Based Standard Error Estimates

Parameter	Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept	1.3476	0.1106	1.1309	1.5644	12.19	<.0001
trt	-0.1080	0.1579	-0.4176	0.2015	-0.68	0.4940
v	0.1108	0.1233	-0.1308	0.3524	0.90	0.3687
trt*v	-0.3016	0.1936	-0.6811	0.0779	-1.56	0.1193
Scale	3.2469	.	.	.	.	.

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

Interpret β

- How to interpret β_1?
  - rate at baseline in the placebo group:
    $E(Y_{ij}|trt = 0, v = 0)/n_{ij} = e^{\beta_0+\beta_1\times 0+\beta_2\times 0+\beta_3\times 0\times 0} = e^{\beta_0}$
  - rate after treatment in the placebo group:
    $E(Y_{ij}|trt = 0, v = 1)/n_{ij} = e^{\beta_0+\beta_1\times 1+\beta_2\times 0+\beta_3\times 1\times 0} = e^{\beta_0+\beta_1}$
  - rate ratio between post and pre-treatment in the placebo group: $\frac{E(Y_{ij}|trt=0,v=0)/n_{ij}}{E(Y_{ij}|trt=\text{placebo},v=1)/n_{ij}} = e^{\beta_1}$

Interpret β

- How to interpret $\beta_1 + \beta_3$?
  - rate at baseline in the progabide group:
    $E(Y_{ij}|trt = 1, v = 0)/n_{ij} = e^{\beta_0+\beta_1\times 0+\beta_2\times 1+\beta_3\times 0\times 1} = e^{\beta_0+\beta_2}$
  - rate after treatment in the progabide group:
    $E(Y_{ij}|trt = 1, v = 1)/n_{ij} = e^{\beta_0+\beta_1\times 1+\beta_2\times 1+\beta_3\times 1\times 1} = e^{\beta_0+\beta_1+\beta_2+\beta_3}$
  - rate ratio between post and pre-treatment in the progabide group: $\frac{E(Y_{ij}|trt=1,v=1)/n_{ij}}{E(Y_{ij}|trt=1,v=0)/n_{ij}} = e^{\beta_1+\beta_3}$

Interpret β

- Interpretation of β_3:
  - Compared to the baseline, after treatment there is 26% ($(1 - e^{\beta_3})100\%$) more reduction in the incident rate of epileptic seizure among patients in the progabide group than those in the placebo group.

Conclusion: GEE

- GEE models make no distributional assumptions and only require three specifications: a mean function, a variance function, and a correlation structure.
- GEE models assume correlation within a subject but independence across subjects.
- If mean is correctly specified but either variance or correlation structure is misspecified, GEE models still provide consistent estimates of the parameters and thus the mean function.
- If all three components are correctly specified, the estimates of the parameters are more efficient.

Conclusion: GEE

- For a large sample size, the robust estimator gives consistent estimates of the standard errors even when the correlations are misspecified.
- When sample size is small, one can use the model-based standard errors, which assume that the specified correlations are closely approximate to the true underlying correlations.

Summary: Key Points

- What is GEE model?
- What are the three components in a GEE model?
- What is the working correlation matrix?
- Why is called “working correlation structure”?
- What is the number of parameters in each type of correlation structures?
- What kinds of descriptive statistics we can produce for repeated measures of data?
- How to write a GEE model?
- How to code a GEE model in SAS?

Summary: Key Points (Cont.)

- How to interpret intercept and slopes in GEE models, especially the coefficient for the interaction with time?
- How to use QIC to choose between working correlation structures and to conduct models comparison?
- How sensitive the estimates of regression coefficients subject to the misspecification in mean function, variance function, or correlation structure?
- What is the assumption on the correlations between measurements from different subjects and measurements from the same subject?

Suggested Readings

- Chapter 11 (Dupont)