BIOS 6110
Applied Categorical Data Analysis

Instructor: Yuan Huang, Ph.D.

Department of Biostatistics

Fall 2017

Part IX

Correlated Data: GEE

Objective

- Correlated data, Longitudinal study (previously match pairs)
- Generalized Estimating Equation (GEE) approach for building marginal models
    - working correlation
- Use SAS/R to fit GEE models and understand outputs

Motivating examples:

- Longitudinal study on schizophrenia (a mental disorder), measuring illness (level 1-7).
  – introduce GEE
- Longitudinal study on depression, dichotomous assessment (0, 1)
  – R and SAS coding + outputs
- More examples on sas coding page.

Correlated or Clustered Data

Sharing location and resources often lead to clustered/correlated data.

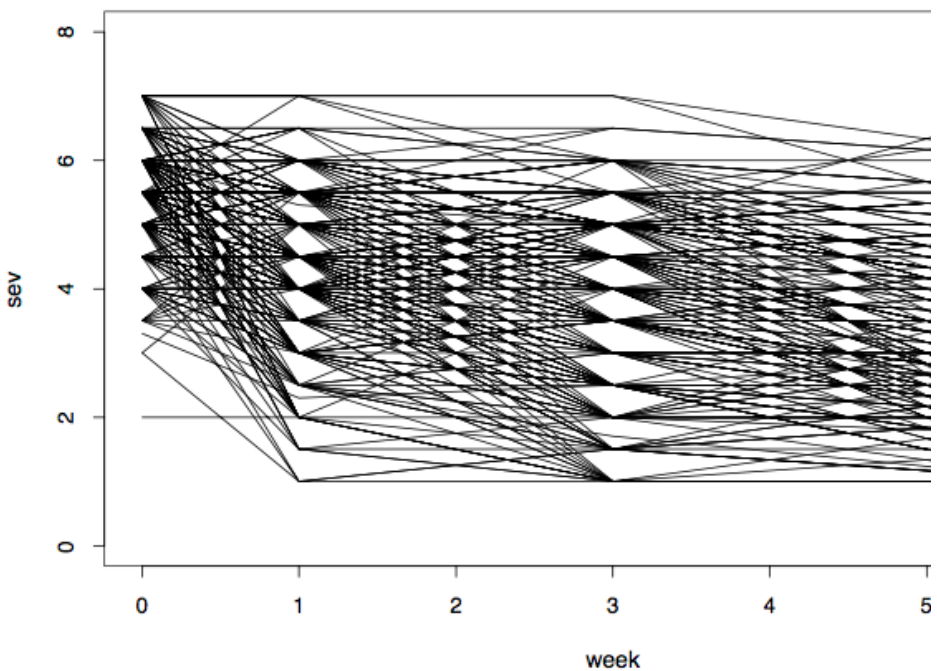
- A study of water-borne diseases in several African villages. We would expect a positive correlation among the disease statuses of subjects using the same well
- A study of high cholesterol in a community. We would expect correlation among the cholesterol levels of subjects from the same family
- A study of the flu in eighth grade classrooms across Iowa. We would expect correlation among the students from the same classroom
- Outcome variables measured on twins or husbands and wives are typically treated as correlated data. In general, studies involving matching give rise to correlated data

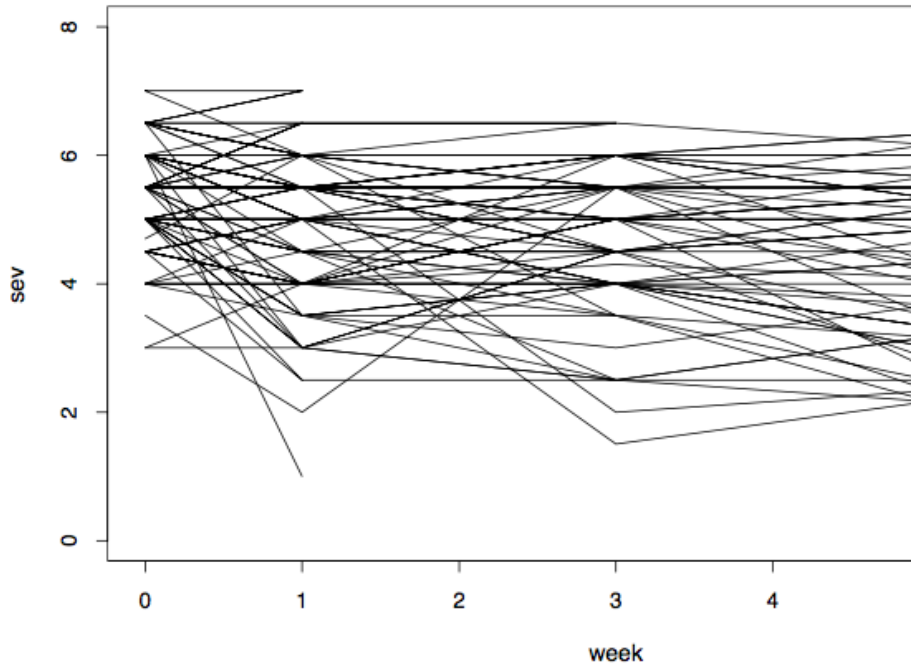
Longitudinal data is a common type of clustered data in which subjects are repeatedly measured at different points in time.

Longitudinal study on Schizophrenia

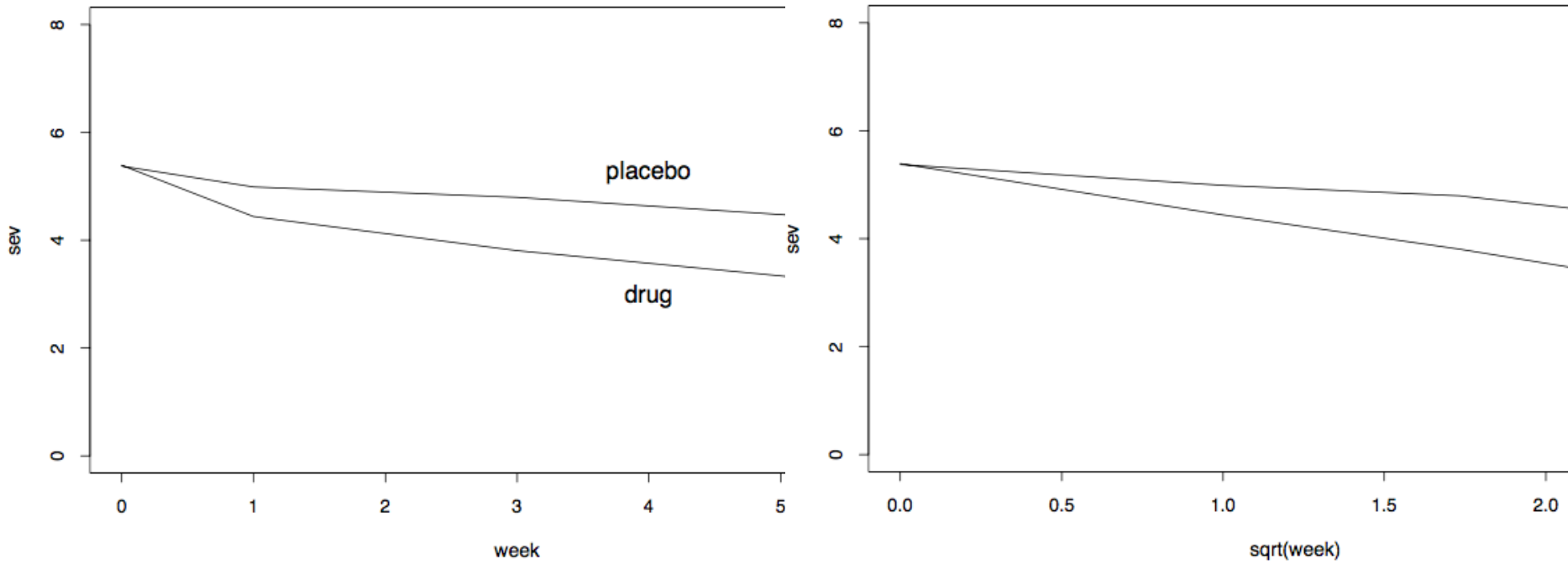
For example, Data analyzed by Hedeker and Gibbons (1997). A randomized trial for schizophrenia:

- 312 patients received drug therapy and 101 received placebo
- measurements at weeks 0, 1, 3, 6, but some subjects have missing data due to dropout
- outcome: severity of illness (1=normal, . . . , 7=extremely ill)





- Left: "Spaghetti plot" of response curves for drug patients
- Right: "Spaghetti plot" of response curves for placebo patients
- For these plots, every line depicts trajectory for one subject across different evaluation times.



- Left: average trajectories for the placebo and drug groups plotted against week
- Right: average trajectories for the placebo and drug groups plotted against square root of week
- After taking average, these plots do not contain subject-specific information anymore. It is marginal.
- At baseline (week 0), the two groups have very similar averages. This makes sense. In a randomized trial, the groups are initially just a random division of the subjects.

Based on the above observation, it makes sense to fit a model for mean response with

- an intercept
- a main effect for treatment group
- a main effect for $\sqrt{\text{week}}$
- an interaction term between treatment group and $\sqrt{\text{week}}$

This allows the two groups to have different intercepts and slopes.

Because the intercepts are defined as the average responses at week 0, we expect that the main effect for group (i.e. the difference in intercepts) will be small.

Question: How can we fit this model, taking into account the fact that the multiple observations for a subject are correlated?

Longitudinal Study of Treatments for Depression

- Subjects were classified into two groups based on their diagnosis severity
- In each group, subjects were randomly assigned to one of two drugs
- Dichotomous assessment of each subject's extent of suffering from mental depression was made at weeks 1, 2, and 4

Table 9.1. Cross-classification of Responses on Depression at Three Times (N = Normal, A = Abnormal) by Treatment and Diagnosis Severity

Diagnosis Severity	Treatment	Response at Three Times							
		NNN	NNA	NAN	NAA	ANN	ANA	AAN	AAA
Mild	Standard	16	13	9	3	14	4	15	6
Mild	New drug	31	0	6	0	22	2	9	0
Severe	Standard	2	2	8	9	9	15	27	28
Severe	New drug	7	2	5	2	31	5	32	6

Source: Reprinted with permission from the Biometric Society (G. G. Koch et al., *Biometrics*, **33**: 133–158, 1977).

There are 340 subjects and $3 \times 340 = 1020$ responses

We want to model Pr(response="normal") in terms of other covariates. So look at the sample proportions. Note that, the table and plots do not contain subject-specific information anymore. It is marginal.

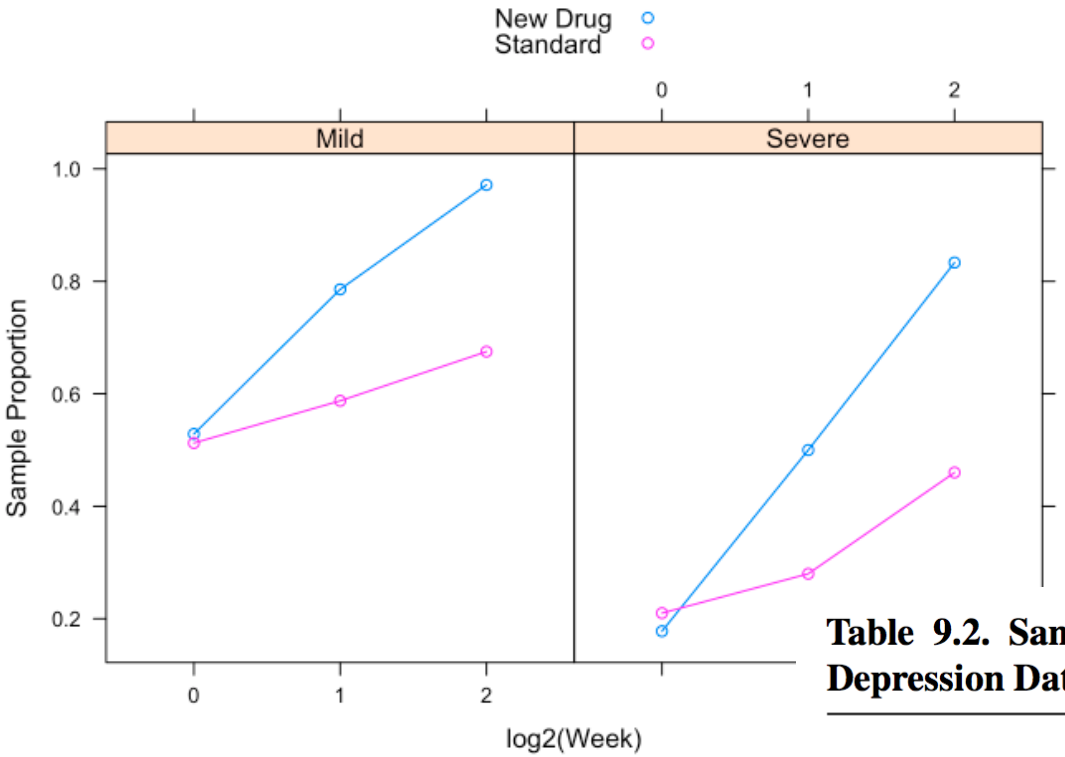


Table 9.2. Sample Marginal Proportions of Normal Response for Depression Data of Table 9.1

Diagnosis Severity	Treatment	Sample Proportion Week 1	Week 2	Week 4
Mild	Standard	0.51	0.59	0.68
	New drug	0.53	0.79	0.97
Severe	Standard	0.21	0.28	0.46
	New drug	0.18	0.50	0.83

Marginal Models and Conditional Models

Marginal Models

$$\text{logit}(\pi) = \alpha + \beta_1 \text{severity} + \beta_2 \text{drug} + \beta_3 \text{time} + \beta_4 (\text{drug} \times \text{time})$$

- This model doesn't make a distinction the observations are from the same subject or not, even though responses from the same subject tend to be similar.
- Effects are population-averaged

Conditional Models

$$\text{logit}(\pi) = \alpha_i + \beta_1 \text{severity} + \beta_2 \text{drug} + \beta_3 \text{time} + \beta_4 (\text{drug} \times \text{time})$$

- Each subject has his/her own α_i
- Effects are subject-specific

From α to α_i, a small step in notation, a big step in modeling

Marginal Modeling: The Generalized Estimating Equations (GEE)

- There is no likelihood function for GEE
- It does not assume a joint distribution on the responses from a cluster. It only assumes a particular distribution (e.g., binomial) for each responses.
- GEE is a quasi-likelihood method that assumes a relationship between $E(Y)$ and $Var(Y)$, where Y represents the response
- GEE estimates of model parameters are obtained by solving generalized estimating equations
- It requires an educated guess of the correlation structure among Ys, or the working correlation matrix
- GEE provides a robust (aka empirical) estimate of the correlation matrix
- Hence better guess is helpful but not essential, especially when sample size is large (Think about working correlation matrix as a prior and the robust estimate as the updated estimate given data)

* An outline of GEE

- ith cluster of responses: $y_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^t$
- Mean of y_{ij} is μ_{ij} which is assumed to satisfy the following model:

$$g(\mu_{ij}) = \beta_0 + x_{i1}\beta_1 + \ldots + x_{ip}\beta_p$$

  where g is the link function: identity link; logit link

- Variance of y_{ij} is

$$Var(y_{ij}) = v(\mu_{ij}) \cdot \phi$$

  - $v(\cdot)$ is a known function: $v(\mu_{ij}) = 1$ for linear regression; $= \mu_{ij}(1 - \mu_{ij})$ for binary response; $= \mu_{ij}$ for Poisson response
  - ϕ is possibly unknown. $\phi = 1$ for binary and Poisson response

* An outline of GEE (cont.)

- Let
$$\mathbf{A_i} = \text{Diag}(Var(y_{i1}), Var(y_{i2}), \ldots, Var(y_{in_i}))$$
and
$$\mathbf{V}_i = A_i^{1/2}\mathbf{R}_i A_i^{1/2}$$
where $\mathbf{R}_i$ is the correlation matrix within the cluster. The generalized estimating equation (GEE) is given by
$$\sum_{i=1}^{n} \frac{\partial \mu_{ij}}{\partial \beta_k} \mathbf{V_i}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = 0, \quad k = 0, 1, \ldots, p$$
βs are estimated by the solutions to these equations.
- Let $D = \frac{\partial \mu}{\partial \beta}$ and $\mathbf{V} = \text{Diag}(\mathbf{V}_1, \ldots, \mathbf{V}_n)$, the GEE can also be written as
$$U = D'\mathbf{V}^{-1}(y - \mu) = 0.$$
- Often times, we don't know the true correlation matrix and instead use a "working" correlation matrix $\tilde{\mathbf{R}}_i$ specified by users ($\tilde{\mathbf{R}}_i \leftrightarrow \tilde{\mathbf{V}}_i$).

Correlation Structures

The key to analyzing clustered data is to characterize the correlation structure in the measured response variable. We will assume the following:

- The data can be arranged into clusters such that there is correlation among the observed responses within clusters, but not between clusters

  In the Longitudinal Study of Treatment for Depression, the clusters are defined by the individual subjects. We assume that observations from a given subject are correlated over time, but that they are not correlated with the observations from other subjects

- The correlation structure is the same within each cluster

  The correlations between each of week 1 and 2, week 2 and 4, and weeks 2 and 4 are the same from subject-to-subject

In general, we summarize all the pairwise correlations within a cluster using a correlation matrix. The correlation matrix is **symmetric**.

For each subject in our example, we have a 3×3 correlation matrix:

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

The correlation terms correspond to the following:

Notation	Correlation between observations at
1	The same week
$\rho_{12} = \rho_{21}$	weeks 1 and 2
$\rho_{13} = \rho_{31}$	weeks 1 and 4
$\rho_{23} = \rho_{32}$	weeks 2 and 4

Currently, this correlation matrix is unstructured. Depending on the study design, we may decide to make various assumptions about the structure of the correlation matrix. There are many different types of correlation structures

Independence Correlation Structure

An independence correlation assumption implies that there is no correlation within clusters.

The corresponding correlation matrix is

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Exchangeable Correlation Structure

An exchangeable or compound symmetric structure implies a constant correlation within clusters. That is, any given pair of observations is no more or less correlated than any other pair

In terms of the example, this would imply that the correlations are equal between all time points.

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

This is a rather strong assumption for longitudinal data. It essentially implies that the correlation between observations taken at adjacent time points is the same as those taken 2, 3, or more time points apart

It is useful when

- There is no distinct ordering within clusters
- Observations can be considered a random sample within a cluster

Auto-Regressive Correlation Structure

The auto-regressive structure allows the correlation to vary as a function of the "distance" between the observations within a cluster.

This is attractive for longitudinal data since it allows for the correlation to decrease as observations are taken further apart in time.

In general, the correlation between the observation in the ith row and jth column is $\rho^{|i-j|}$. In our example, the corresponding correlation matrix is

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

It is useful when

- There is a natural ordering to the observations within clusters
- Assuming a constant correlation between adjacent observations
- The correlation strictly decreases as a function of the distance between observations

Unstructured Correlation Structure

An unstructured correlation assumption places no restrictions on the correlation matrix. In this case, correlation is allowed to vary between all observations in the cluster.

The correlation matrix has the form

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

It is useful when

- Do not want to assume a constant correlation between adjacent observations
- Do not want to specify a functional form that relates the correlation to the distance between observations

Properties of GEE

- The estimated $\hat{\beta}$ is consistent and asymptotically unbiased estimate of β, even if $\tilde{\mathbf{V}}_i \neq \mathbf{V}_i$. It is also asymptotically normal.
- If $\tilde{\mathbf{V}}_i \neq \mathbf{V}_i$, the $\hat{\beta}$ is not efficient. The asymptotic variance of $\hat{\beta}$ is the lowest when $\tilde{\mathbf{V}}_i = \mathbf{V}_i$.
- The $(D'\tilde{\mathbf{V}}^{-1}D)^{-1}$ is called the "model-based" or "naive" estimator of the $Var(\hat{\beta})$. When $\tilde{\mathbf{V}}_i = \mathbf{V}_i$, the naive estimator is not a consistent estimator for $Var(\hat{\beta})$.
- The "naive" estimator can be corrected by "robust" or "sandwich" estimator. The "robust" estimator is a consistent estimate of $Var(\hat{\beta})$ even if $\tilde{\mathbf{V}}_i = \mathbf{V}_i$.

$$(D'\tilde{\mathbf{V}}^{-1}D)^{-1}(D'\tilde{\mathbf{V}}^{-1}W\tilde{\mathbf{V}}^{-1}D)(D'\tilde{\mathbf{V}}^{-1}D)^{-1},$$

where $W = \text{Diag}((y_1 - \mu_1)^2, \ldots, (y_n - \mu_n))$.

Further Comments

GEE does not provide SE for the off-diagonal elements so a test can not be conducted to test the true correlation matrix.

In the presence of clustering, specification of the independence correlation structure seems like a poor choice. Indeed, it is the least desirable option for describing within-cluster correlation. However, when working with large or complex data sets, it is not always possible to obtain GEE estimates for all of the correlation structures. In practice, the independence structure may be the only structure for which GEE estimates can be obtained

SAS output for the Depression Study

Working Correlation Matrix

	Col1	Col2	Col3
Row1	1.0000	-0.0034	-0.0034
Row2	-0.0034	1.0000	-0.0034
Row3	-0.0034	-0.0034	1.0000

Exchangeable Working Correlation

Correlation -0.003432732

GEE Fit Criteria

QIC	1172.0189
QICu	1171.9405

Analysis Of GEE Parameter Estimates
Empirical Standard Error Estimates

Parameter		Estimate	Standard Error	95% Confidence Limits		Z	Pr > |Z|
Intercept		-0.0281	0.1742	-0.3695	0.3133	-0.16	0.8718
severity	Seve	-1.3139	0.1460	-1.6000	-1.0278	-9.00	<.0001
severity	Mild	0.0000	0.0000	0.0000	0.0000	.	.
drug	New Drug	-0.0593	0.2286	-0.5072	0.3887	-0.26	0.7954
drug	Standard	0.0000	0.0000	0.0000	0.0000	.	.
time		0.4825	0.1199	0.2474	0.7175	4.02	<.0001
time*drug	New Drug	1.0172	0.1877	0.6493	1.3851	5.42	<.0001
time*drug	Standard	0.0000	0.0000	0.0000	0.0000	.	.

Answer the following questions:

- What is the estimated time effect (slope) for standard and new drugs?
- Is there strong evidence of faster improvement for the new drug?
- How to interpret $\hat{\beta}_1 = -1.314$?
- How to interpret $\hat{\beta}_2 = -0.059$?
- How to interpret $\exp(-0.059 + 1.017t)$?

The GENMOD results labeled "Empirical Standard Error Estimates" are referred to as robust estimates. These standard errors are valid even if the specified correlation structure is not appropriate for the given data set

The estimated correlations between time points are given in the working correlation matrix.

- The estimated time effect is $\hat{\beta}_3 = 0.482$ for the standard drug and $\hat{\beta}_3 + \hat{\beta}_4 = 1.5$ for the new one
- The change of slope due to new drug is $\hat{\beta}_4 = 1.017$ (Robust SE $= 0.188$). The Wald test of no interaction, $H_0 : \beta_4 = 0$, tests a common time effect for each drug. Its z test statistic $= 1.017/0.188 = 5.4$ (p-value < 0.0001). There is strong evidence of faster improvement for the new drug
- The severity of depression estimate is $\hat{\beta}_1 = -1.314$ (Robust SE $= 0.146$). For each drug-time combination, the estimated odds of a normal response when the initial diagnosis was severe is $\exp(-1.314) = 0.27$ times the estimated odds when the initial diagnosis was mild

- The estimate $\hat{\beta}_2 = -0.059$ (Robust SE $= 0.228$) for the drug effect applies only when time $= 0$ (i.e., after one week, as we have taken log2 for week), for which the interaction term does not contribute to the drug effect. It indicates an insignificant difference between the drugs after 1 week.
- At time t, the estimated odds of normal response with the new drug are $\exp(-0.059 + 1.017t)$ times the estimated odds for the standard drug, for each initial diagnosis level. By the final week ($t = 2$), this estimated odds ratio has increased to 7.2
- In summary, severity, drug treatment, and time all have substantial effects on the probability of a normal response. The chance of a normal response is similar for the two drugs initially and increases with time, but it increases more quickly for those taking the new drug than the standard drug
- This conclusion is consistent with the graphical representation of the sample proportions

Using Independence working correlation matrix

```
## Coefficients:
##                     Estimate Naive S.E.    Naive z Robust S.E.   Robust z
## (Intercept)       -0.02798843  0.1627083 -0.1720160   0.1741865 -0.1606808
## severitySevere    -1.31391092  0.1453432 -9.0400569   0.1459845 -9.0003423
## drugNew Drug      -0.05960381  0.2205812 -0.2702126   0.2285385 -0.2608042
## time               0.48241209  0.1139224  4.2345663   0.1199350  4.0222784
## drugNew Drug:time  1.01744498  0.1874132  5.4288855   0.1876938  5.4207709
```

Using exchangeable working correlation matrix

```
## Coefficients:
##                     Estimate Naive S.E.    Naive z Robust S.E.   Robust z
## (Intercept)       -0.02809866  0.1625499 -0.1728617   0.1741791 -0.1613205
## severitySevere    -1.31391033  0.1448627 -9.0700418   0.1459630 -9.0016667
## drugNew Drug      -0.05926689  0.2205340 -0.2687427   0.2285569 -0.2593091
## time               0.48246420  0.1141154  4.2278625   0.1199383  4.0226037
## drugNew Drug:time  1.01719312  0.1877051  5.4191018   0.1877014  5.4192084
##
## Working Correlation
##              [,1]         [,2]         [,3]
## [1,]  1.000000000 -0.003432732 -0.003432732
## [2,] -0.003432732  1.000000000 -0.003432732
## [3,] -0.003432732 -0.003432732  1.000000000
```

The off-diagonal elements are pretty close to 0, indicating weak correlation among responses from the same subject

GEE summary

GEE provides marginal model and its parameter estimates are population-averaged rather than subject-specific

Advantages

- The algorithm is easily accessible in PROC GENMOD and may be used with any of the regression models available in the procedure (e.g. linear, logistic, and Poisson)
- Inferences are valid even if the wrong correlation structure is specified

Disadvantages

- Does not provide standard error estimates for the parameters in the correlation matrix
- The auto-regressive structure in GENMOD assumes that longitudinal observations are made at fixed, equally-spaced time points

The disadvantages of GEE could be overcome by using a mixed-effects model which provides conditional models. Mixed models, however, are sensitive to the chosen correlation structure.