Chapter 4

Marginal Models for Discrete Data

4.1 Review of Generalized Linear Models

- So far we have concentrated on continuous/normal repeated measurements data
  - ▷ serum bilirubin and serum cholesterol in the PBC dataset
  - ▷ CD4 cell counts in the AIDS dataset
  - ▷ prothrombin time in the Prothro dataset

However often we may want to analyze other types of repeatedly measured outcomes that are not normally distributed

4.1 Review of Generalized Linear Models (cont'd)

- Examples:
  - ▷ in colon cancer studies the iFOBT test (presence or not of blood in stool) is used to monitor patients ⇒ *dichotomous data*
  - ▷ after a heart transplantation patients report their quality of life in frequent intervals, with the categories 'Poor', 'Moderate', 'Good' and 'Very Good' ⇒ *ordinal data*
  - ▷ in asthma studies often interest is in the number of asthma attacks patients have in a period of time ⇒ *Poisson data*
  - ▷ ...

4.1 Review of Generalized Linear Models (cont'd)

- Suppose we have a dichotomous outcome Y measured *cross-sectionally*
  - ▷ Example: The serum cholesterol levels from the PBC dataset at baseline (i.e., time $t = 0$) that are higher than 210 mg/dL

- We are interested in making statistical inferences for this outcome, e.g.,
  - ▷ is there any difference between placebo and D-penicillamine corrected for the age and sex of the patients?
  - ▷ which factors best predict serum cholesterol levels higher than 210 mg/dL?

$$\Downarrow$$

Generalized Linear Models

4.1 Review of Generalized Linear Models (cont'd)

- Reminder: The General Linear Model for continuous data

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

- There are two issues with applying this model for the dichotomous outcome $y_i = I(\texttt{serChol}_i > 210)$ ($I(\cdot)$ is the indicator function: $I(A) = 1$ if A is true & $I(A) = 0$ when A is false)

  1. y_i does not follow a normal distribution, it follows a *Bernoulli* distribution

  2. the mean of the Bernoulli distribution is π_i, *the probability*, of having serum cholesterol higher than the threshold – the mean in the linear regression model is $\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$

4.1 Review of Generalized Linear Models (cont'd)

- Say we use a naive strategy and consider a linear regression model for $y_i = I(\texttt{serChol}_i > 210)$ – the mean would be

$$\pi_i = E(y_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

- What is the problem?
  - ▷ π is a probability and is restricted to values between 0 and 1
  - ▷ $\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$ can take values below 0 and above 1, i.e., in $(-\infty, +\infty)$

- We have to bring π in the scale $(-\infty, +\infty)$ $\rightarrow$ a classic option is:

$$0 < \quad \pi \quad < 1$$

$$0 < \quad \frac{\pi}{1-\pi} \quad < +\infty$$

$$-\infty < \log \frac{\pi}{1-\pi} < +\infty$$

- This gives rise to the *logistic regression model*

$$\text{log odds of success} = \log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

$$\text{odds of success} = \frac{\pi_i}{1-\pi_i} = \exp\left(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}\right)$$

$$\text{probability of success} = \pi_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}$$

which respects the constraint that π takes values between 0 and 1

- A unit change in X_1 from x to $x+1$ (while all other covariates are held fixed) corresponds to

$$\log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x + \ldots + \beta_p x_{ip}$$

$$\log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 (x+1) + \ldots + \beta_p x_{ip}$$

- Thus,

$$\beta_1 = \log \frac{\pi_i}{1-\pi_i} - \log \frac{\pi_i}{1-\pi_i} = \log \left\{ \frac{\pi_i}{1-\pi_i} \Big/ \frac{\pi_i}{1-\pi_i} \right\}$$

$$\exp(\beta_1) = \frac{\pi_i}{1-\pi_i} \Big/ \frac{\pi_i}{1-\pi_i}$$

4.1 Review of Generalized Linear Models (cont'd)

- Notes:
  - ▷ The relationship between log odds of success and the covariates is linear
  - ▷ The relationship between π and the covariates is non-linear

    $\Rightarrow$ Interpretation of parameters is different than in linear regression models

4.1 Review of Generalized Linear Models (cont'd)

- For dichotomous X_1

  $\Rightarrow \beta_1$ is the log odds ratio of 'success' between the two levels of X_1 given that all other covariates remain constant

  $\Rightarrow \exp(\beta_1)$ is the odds ratio between the two levels of X_1 given that all other covariates remain constant

- For continuous X_1

  $\Rightarrow \beta_1$ is change in log odds of 'success' for a unit change in X_1 given that all other covariates remain constant

  $\Rightarrow \exp(\beta_1)$ is the odds ratio for a unit change in X_1 given that all other covariates remain constant

4.1 Review of Generalized Linear Models (cont'd)

- Relationships

$$\exp(\beta_1) = \begin{cases} = 1, & \text{the two odds are the same} \\ > 1, & \text{increased odds of success} \\ < 1, & \text{decreased odds of success} \end{cases}$$

- As π increases
  - ▷ odds of success increases
  - ▷ log odds of success increases

4.1 Review of Generalized Linear Models (cont'd)

- Example: In the PBC dataset we are interested in investigating how the probability of serum cholesterol higher than 210 mg/dL is associated with age, sex and the treatment the patients received – the model has the form

$$\log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 \texttt{Age}_i + \beta_2 \texttt{Female}_i + \beta_3 \texttt{D-penicil}_i$$

where $\pi_i = \Pr(\texttt{serChol}_i > 210)$

4.1 Review of Generalized Linear Models (cont'd)

	Value	Std.Err.	z-value	p-value
β_0	3.739	1.306	2.862	0.004
β_1	-0.038	0.020	-1.879	0.060
β_2	0.313	0.563	0.556	0.578
β_3	0.512	0.420	1.218	0.223

- $\beta_0 = 3.7$ is the log odds of excess levels of serum cholesterol for a male patient in the control group who is 0 years old
- $\beta_1 = -0.04$ is the log odds ratio for a unit increase in age for patients of the same sex who receive the same treatment
- $\beta_2 = 0.3$ is the log odds ratio of females versus males of the same age who receive the same treatment

4.2 Generalized Estimating Equations

- We return our focus on repeated measurements data, namely, repeated categorical data

  ▷ **we need to account for the correlations**

- Reminder: In the marginal models for continuous multivariate data (Chapter 2) we took account of the correlations by *incorporating a correlation matrix in the error terms*

- For categorical data it is not straightforward to do that because there are no clear multivariate analogues of the univariate distributions

  ▷ we will do something similar, **not** in the *error terms* but in the *score equations*

4.2 Generalized Estimating Equations (cont'd)

- Liang and Zeger (1986, Biometrika) made the following important contribution
  - ▷ The parameters of Generalized Linear models are estimated using the *maximum likelihood* approach
  - ▷ *Key idea:* Finding the top of the log-likelihood mountain is equivalent to finding the parameter values for which the slope of the mountain is flat (i.e., zero)
  - ▷ The slope of the log-likelihood mountain is given by the score vector

$$S(\beta) = \sum_i \frac{\partial \mu_i}{\partial \beta} V_i^{-1} (Y_i - \mu_i)$$

4.2 Generalized Estimating Equations (cont'd)

where in standard logistic regression

- ▷ μ_i the mean of Y_i, e.g., for dichotomous data $\mu_i = \pi_i$
- ▷ V_i is a diagonal matrix with the variance of Y_i, e.g., for dichotomous data $V_i = \text{diag}\{\pi_i(1-\pi_i)\}$

- The idea of Liang and Zeger was to replace the diagonal matrix V_i with a full covariance matrix

$$V_i = A_i^{1/2} R_i(\alpha) A_i^{1/2}$$

where

- ▷ $A_i = \text{diag}\{\text{var}(Y_i)^{1/2}\}$ a diagonal matrix with the standard deviations
- ▷ $R_i(\alpha)$ a 'working' assumption for the pairwise correlations

4.2 Generalized Estimating Equations (cont'd)

- If the assumed mean structure μ_i is correctly specified, then

$$\widehat{\beta} \sim \mathcal{N}\{\beta, \text{var}(\widehat{\beta})\}$$

where

$\text{var}(\widehat{\beta}) = V_0^{-1} V_1 V_0^{-1}$ is called the **Sandwich** or **Robust** estimator

with

$$\underbrace{V_0 = \sum_i \frac{\partial \mu_i}{\partial \beta^\top} V_i^{-1} \frac{\partial \mu_i}{\partial \beta}}_{\textbf{bread}} \quad \text{and} \quad \underbrace{V_1 = \sum_i \frac{\partial \mu_i}{\partial \beta^\top} V_i^{-1} \text{var}(Y_i) V_i^{-1} \frac{\partial \mu_i}{\partial \beta}}_{\textbf{meat}}$$

4.2 Generalized Estimating Equations (cont'd)

- **Sandwich/Robust** vs **Naive/Model Based** standard errors
  - ▷ software often also report the **Naive/Model Based** standard errors
  - ▷ these standard error assume that the working correlation matrix is correctly specified
  - ▷ the **Sandwich/Robust** corrects for a possible misspecification of the correlation structure
    * though at the expense of power

4.2 Generalized Estimating Equations (cont'd)

GEE is not a likelihood-based approach (i.e., a model)

$\Downarrow$

it is an estimation method

- No assumptions for the joint distribution of repeated measurements
  $\Rightarrow$ *Semi-parametric approach*

- Three components

  1. Model for mean response $E(Y_i) = \mu_i$, e.g., binary data $E(Y_i) = \pi_i$

$$\log \frac{\pi_i}{1-\pi_i} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$$

4.2 Generalized Estimating Equations (cont'd)

2. Variance of $Y_i \Rightarrow$ follows from GLM assumption for each measurement, e.g., binary data

$$\text{var}(Y_i) = \phi \pi_i (1 - \pi_i)$$

with ϕ a scale parameter that models over-dispersion

3. Pairwise correlations $\Rightarrow$ we make a "working" assumption that possibly depends on parameters to be estimated

The mean and the correlations are **separately** defined!
This is in contrast to the GLMMs we will see in the next chapter

4.2 Generalized Estimating Equations (cont'd)

Interest is primarily in the βs, the covariance structure is considered as "nuisance"

$\Downarrow$

Assumptions for the variance and correlation are <u>not</u> supposed to be correct

- This has implications for
  - ▷ Hypothesis testing
    - $\Rightarrow$ Likelihood ratio test or score test not applicable
    - $\Rightarrow$ The Wald test can be used
  - ▷ Performance when we have missing data

4.3 Interpretation

- Interpretation of β is the same as in classic GLMs
  - ▷ β_j denotes the change in the average y_i when x_j is increased by one unit and all other covariates are fixed

- Example: In the PBC dataset we are interested in the effect of treatment on the average longitudinal evolutions for the probability of serum cholesterol higher than 210 mg/dL – we fit the GEE model with
  - ▷ different average longitudinal evolutions per treatment group ($X\beta$ part)
  - ▷ exchangeable working correlation matrix

4.3 Interpretation (cont'd)

- The model has the form:

$$\log \frac{\pi_{ij}}{1-\pi_{ij}} = \beta_0 + \beta_1 \texttt{Time}_{ij} + \beta_2 \texttt{D-penicil}_i + \beta_3 \{\texttt{Time}_{ij} \times \texttt{D-penicil}_i\}$$

	Value	Std.Err.	z-value	p-value
β_0	2.058	0.250	67.658	< 0.001
β_1	-0.114	0.044	6.798	0.009
β_2	0.166	0.341	0.237	0.626
β_3	0.003	0.054	0.004	0.949

4.3 Interpretation (cont'd)

- We found no indication of a difference between the two treatments groups and the odds of having excess cholesterol levels during follow-up

- Interpretation of parameters (note that we have an interaction term)
  - ▷ $\exp(\beta_1) = 0.9$ is the odds ratio for a year increase for patients receiving placebo
  - ▷ $\exp(\beta_2) = 1.2$ is the odds ratio of D-penicil to placebo at baseline
  - ▷ $\exp(\beta_3) = 1.003$ is the relative difference between the odds ratios for a year increase in the two treatment groups
    - * the odds ratio for a year increase in patients receiving D-penicil is 0.3% higher than the corresponding odds ratio of the placebo patients

4.3 Interpretation (cont'd)

- As we have previously seen, to effectively communicate complex models we can use effect plots

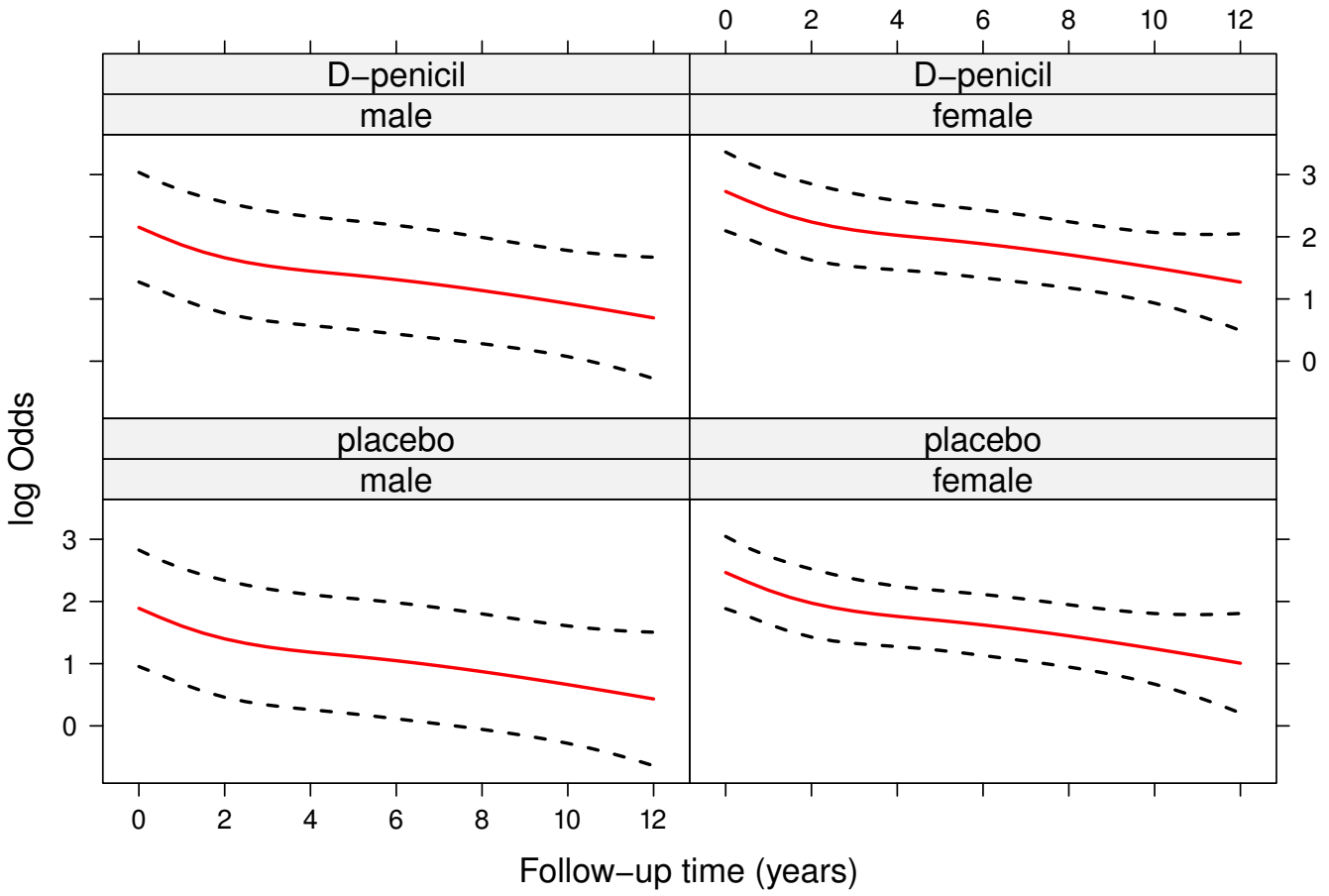
- Example: In the PBC dataset we are interested in the probability of having excess serum cholesterol levels
  - ▷ we allow for nonlinear time and baseline age effects using natural cubic splines with 3 degrees of freedom
  - ▷ we also correct for treatment and gender

4.3 Interpretation (cont'd)

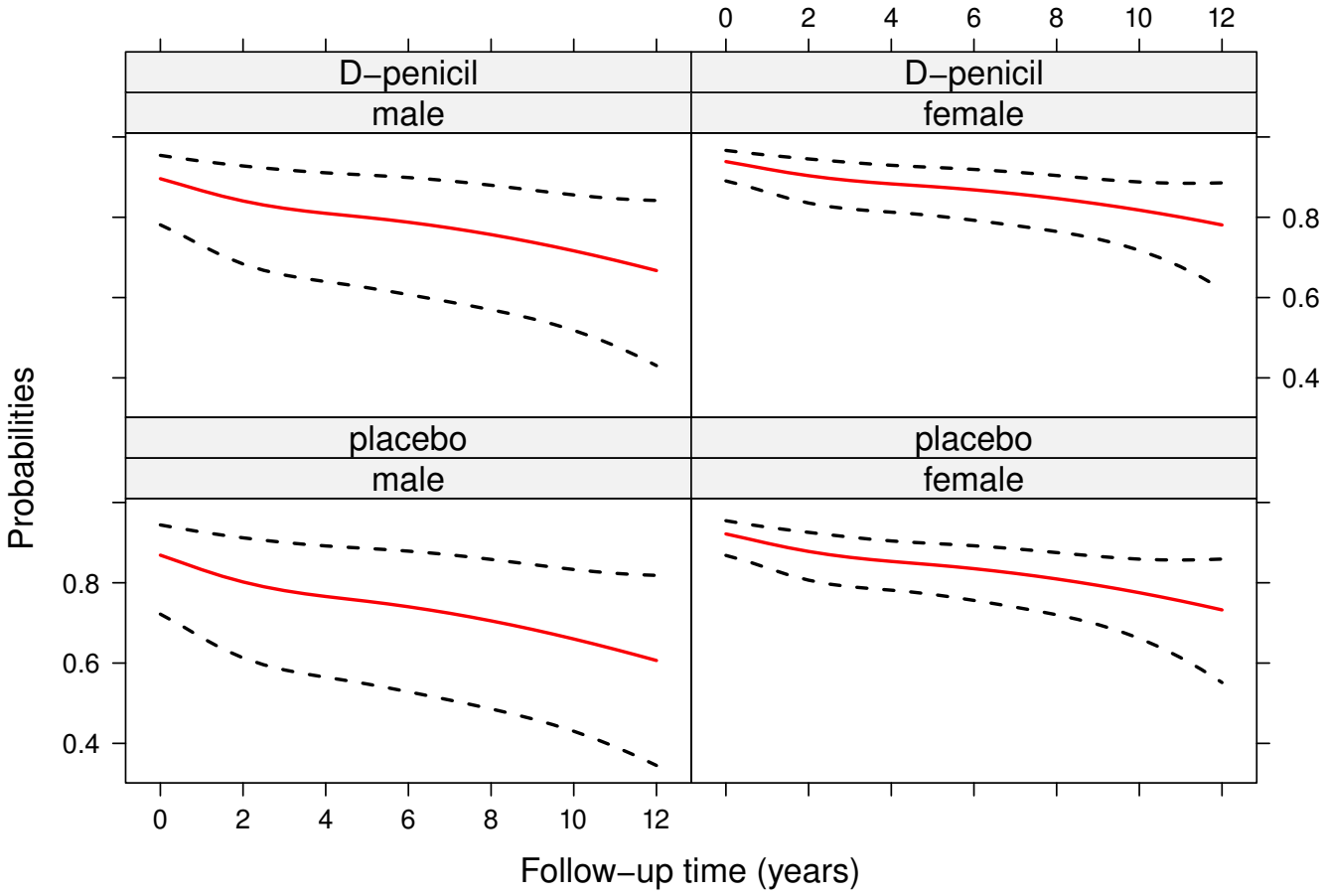
- The following two figures depict the relationship between
  - ▷ the log odds of excess serum cholesterol level
  - ▷ the probability of excess serum cholesterol level

  for male and female patients receiving either the active drug or placebo, who are 49 years old (in the app different ages can be selected)

4.3 Interpretation (cont'd)



4.3 Interpretation (cont'd)



4.4 Generalized Estimating Equations in R

R> In R there are two main packages for GEE analysis, namely **gee** and **geepack** – in this course we will only use **geepack**

- ▷ The main function to fit GEEs is `geeglm()` – this has similar syntax as the `glm()` function of base R that fits GLMs

R> Main arguments of `geeglm()`

- ▷ `formula`: An R formula specifying the response variable and the predictors
- ▷ `family`: a description of the error distribution and link function to be used in the model
- ▷ `id`: the variable denoting which measurements belong to the same group (e.g., to the same subject)

4.4 Generalized Estimating Equations in R (cont'd)

R> Main arguments of `geeglm()`

- `data`: a data frame that contains all these variables; **important**: the rows of this data frame must be ordered with respect to `id`, and the rows within the same `id` should be ordered with respect to time
- `corstr`: the assumed working correlation matrix; options are `"independence"`, `"exchangeable"`, `"ar1"`, `"unstructured"`, and `"userdefined"`

4.4 Generalized Estimating Equations in R (cont'd)

R> The following code fits a GEE model for serum cholesterol from the PBC dataset with an exchangeable working correlation matrix

```
geeFit <- geeglm(serCholD ~ year * drug, family = binomial(),
                 data = pbc2, id = id, corstr = "exchangeable")

summary(geeFit)
```

4.5 Working Correlation Matrix

- As we have seen, the GEEs are a *semiparametric* approach
  - ▷ using the *sandwich estimator* we obtain valid inferences *even if* the working correlation matrix is misspecified

- Hence, one could wonder: Why not always use the sandwich estimator in order not to have to care about the correlation structure?
  - ▷ in other words, why do we need this course?

The sandwich estimator does have important limitations!

4.5 Working Correlation Matrix (cont'd)

- About the sandwich estimator:
  - ▷ is based on asymptotic arguments ⇒ it only works for big samples
  - ▷ more specifically, it works better under the following settings
    - * when the number of subjects is considerably larger than the number or repeated measurements,
    - * balanced designs in which all subjects provide measurements at the same time points
    - * we do not have too many covariates (especially continuous)

- **Therefore**, the choice of the working correlation matrix is of particular importance
  - ▷ this is why all statistical software offer several alternative options for this matrix

4.5 Working Correlation Matrix (cont'd)

- Some standard options are
  - ▷ Independence

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

* repeated measurements are assumed uncorrelated
* even though an *unrealistic* assumption, in some circumstances it is an appropriate route to follow (see e.g., pp.209)

4.5 Working Correlation Matrix (cont'd)

▷ Exchangeable

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

* constant correlation over time
* more appropriate for short time intervals

4.5 Working Correlation Matrix (cont'd)

▷ First-order autoregressive

$$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

* correlations decrease over time
* more appropriate for longer time intervals

4.5 Working Correlation Matrix (cont'd)

▷ General/Unstructured

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \rho_3 \\ \rho_1 & 1 & \rho_4 & \rho_5 \\ \rho_2 & \rho_4 & 1 & \rho_6 \\ \rho_3 & \rho_5 & \rho_6 & 1 \end{bmatrix}$$

* the most flexible correlation structure
* it can only be fitted (i.e., without numerical problems) with balanced data and big sample sizes (rule of thumb: when $n \gg p(p-1)/2$, with n denoting the number of subjects and p the number of repeated measurements)

4.5 Working Correlation Matrix (cont'd)

- Example: A very low CD4 count (less than 150 cells/mm^3) is an indicator for opportunistic infections

  ▷ In the following analysis we dichotomize the CD4 cell counts from the AIDS dataset using this threshold

  ▷ We fit GEEs for this dichotomous response and only the categorical version of the time as covariate – for the working correlation matrix we assume "Independence", "Exchangeable", "AR1" and "Unstructured"

  ▷ We compare parameter estimates and standard errors (model-based and sandwich)

4.5 Working Correlation Matrix (cont'd)

- The model has the form:

$$\log \frac{\pi_{ij}}{1-\pi_{ij}} = \beta_0 + \beta_1\{\texttt{Time}_{ij} = 2\} + \beta_2\{\texttt{Time}_{ij} = 6\} + \beta_3\{\texttt{Time}_{ij} = 12\} + \beta_4\{\texttt{Time}_{ij} = 18\}$$

where

  - $\pi_{ij} = \Pr(\texttt{CD4}_{ij} < 150)$
  - $\{\texttt{Time}_{ij} = 2\}$ denotes the dummy variable for month 2, $\{\texttt{Time}_{ij} = 6\}$ the dummy variable for month 6, and so on

4.5 Working Correlation Matrix (cont'd)

- The estimated regression coefficients under the four GEEs are:

	Ind.	Exch.	AR1	Unstr.
β_0	1.474	1.474	1.477	1.533
β_1	−0.043	0.013	0.029	0.026
β_2	0.151	0.295	0.332	0.291
β_3	−0.110	0.465	0.415	0.660
β_4	0.541	1.173	0.652	−0.240

  - ▷ We observe considerable differences in the magnitude of some parameters

4.5 Working Correlation Matrix (cont'd)

- The estimated standard errors under the four GEEs are:

	Sandwich				Model-based			
	Ind.	Exch.	AR1	Unstr.	Ind.	Exch.	AR1	Unstr.
$s.e.(\beta_0)$	0.119	0.119	0.118	0.139	0.119	0.127	0.126	0.129
$s.e.(\beta_1)$	0.106	0.100	0.102	0.105	0.178	0.099	0.077	0.123
$s.e.(\beta_2)$	0.140	0.139	0.138	0.146	0.194	0.114	0.113	0.139
$s.e.(\beta_3)$	0.146	0.178	0.174	0.243	0.204	0.136	0.148	0.074
$s.e.(\beta_4)$	0.524	0.627	0.315	1.031	0.545	0.423	0.343	NaN

  - ▷ We also observe differences in the magnitudes of the estimated standard errors
  - ▷ These are smaller in the sandwich than in the model-based estimator

4.5 Working Correlation Matrix (cont'd)

- How to choose the most appropriate working correlation matrix?
  - ▷ unfortunately, there are no generally accepted formal tests for choosing the working correlation matrix
  - ▷ the choice **should not** be based on the grounds of statistical significance
  - ▷ consider appropriate choices based on the features of the data
    - * for balanced data with big sample size ⇒ Unstructured
    - * unbalanced data ⇒ Exchangeable or AR1
    - * when more than one options plausible ⇒ sensitivity analysis, *report all results*

Similarly to what we have seen in Chapter 2 and 3, **a prerequisite is that the mean structure is correctly specified**

4.6 Hypothesis Testing

- Having fitted a GEE model often scientific interest lies in testing specific hypotheses

- Due to the fact that GEE models are semiparametric models, we can **only** employ Wald tests to test the hypothesis of interest

  ▷ score and likelihood ratio tests are not available

- In addition, in standard GEEs, these tests are only available for the mean parameters (i.e., the regression coefficients) and not the parameters of the working correlation matrix

4.6 Hypothesis Testing (cont'd)

- For individual parameters the Wald test has the form:

$$\widehat{\beta} \Big/ s.e.(\widehat{\beta}) \sim \mathcal{N}(0,1)$$

  where

  ▷ $\widehat{\beta}$ denotes the estimated regression coefficient, and

  ▷ $s.e.(\widehat{\beta})$ the sandwich or model-based standard error of this regression coefficient

- We have seen an example of these tests in the analysis of excess cholesterol levels in the PBC dataset (see pp.276–277)

4.6 Hypothesis Testing (cont'd)

- When interest is more than one parameters, then we use the multivariate version of the Wald test, i.e.,

$$\begin{aligned} H_0: &\quad L\beta = 0 \\ H_a: &\quad L\beta \neq 0 \end{aligned}$$

  where L is the contrasts matrix of interest

- Example: We extend the GEE model for the AIDS dataset we have seen in pp.292
  - ▷ time is treated as categorical variable and we also take interactions with treatment
  - ▷ we are interested in the overall treatment effect
  - ▷ the working correlation matrix is assumed to have an AR1 structure

4.6 Hypothesis Testing (cont'd)

- The models under the null and alternative hypothesis have the form:

$$
\begin{cases}
H_0: \log\dfrac{\pi_{ij}}{1-\pi_{ij}} = \beta_0 + \beta_1\{\texttt{Time}_{ij}=2\} + \beta_2\{\texttt{Time}_{ij}=6\} + \\
\qquad\qquad \beta_3\{\texttt{Time}_{ij}=12\} + \beta_4\{\texttt{Time}_{ij}=18\} \\
\\
H_a: \log\dfrac{\pi_{ij}}{1-\pi_{ij}} = \beta_0 + \beta_1\{\texttt{Time}_{ij}=2\} + \beta_2\{\texttt{Time}_{ij}=6\} + \\
\qquad\qquad \beta_3\{\texttt{Time}_{ij}=12\} + \beta_4\{\texttt{Time}_{ij}=18\} + \beta_5\texttt{ddI}_i + \\
\qquad\qquad \beta_6\left[\texttt{ddI}_i \times \{\texttt{Time}_{ij}=2\}\right] + \beta_7\left[\texttt{ddI}_i \times \{\texttt{Time}_{ij}=6\}\right] + \\
\qquad\qquad \beta_8\left[\texttt{ddI}_i \times \{\texttt{Time}_{ij}=12\}\right] + \beta_9\left[\texttt{ddI}_i \times \{\texttt{Time}_{ij}=18\}\right]
\end{cases}
$$

4.6 Hypothesis Testing (cont'd)

- Hence, the parameters we wish to test are:

$$H_0: \quad \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = 0$$
$$H_a: \quad \text{at least one coefficient is different from zero}$$

- The Wald test gives:
  - ▷ W $= 8.5$
  - ▷ df $= 5$
  - ▷ p-value $= 0.131$

Hence, no evidence of a treatment effect

4.6 Hypothesis Testing (cont'd)

- Note that we could also apply a multivariate test not only with a contrast matrix L, but also with a design matrix of interest X

- Example: We fit a GEE model for excess serum cholesterol levels in which we include the following terms
  - ▷ nonlinear effect of time with natural cubic splines with 3 degrees of freedom
  - ▷ main effect of treatment
  - ▷ interaction of treatment with nonlinear effect of time
  - ▷ nonlinear effect of baseline age with natural cubic splines with 3 degrees of freedom

4.6 Hypothesis Testing (cont'd)

- As we have previously discussed, when we include nonlinear terms parameters do not have a straightforward interpretation

- But we could still test meaningful hypothesis
  - ▷ more specifically, we are interested in testing for a treatment effect at year 7 for patients who are 49 years old

- Let X_1 denote the design matrix of the placebo patient at year 7 and age 49, and X_2 the analogous design matrix for the patient who receives D-penicillamine

4.6 Hypothesis Testing (cont'd)

- Hence, we want to test

$$\begin{aligned} H_0: &\quad X_1\beta = X_2\beta \\ H_a: &\quad X_1\beta \neq X_2\beta \end{aligned}$$

which is equivalent to

$$\begin{aligned} H_0: &\quad (X_1 - X_2)\beta = 0 \\ H_a: &\quad (X_1 - X_2)\beta \neq 0 \end{aligned}$$

4.6 Hypothesis Testing (cont'd)

- The Wald test gives:
  - ▷ Estimated difference $= -0.2$ (s.e. $= 0.4$)
  - ▷ W $= 0.18$
  - ▷ df $= 1$
  - ▷ p-value $= 0.675$

4.7 Review of Key Points

- Generalized Estimating Equations: What they are
  - ▷ extension of the marginal models we have seen in Chapter 2, in the setting of categorical data
  - ▷ Three components: (i) a model for the mean, (ii) a model for the variance, & (iii) an assumption for the pairwise correlations
  - ▷ semiparametric approach $\Rightarrow$ no assumptions for the distribution of the data

- Generalized Estimating Equations: Features
  - ▷ important to correctly specify the mean
  - ▷ sandwich estimator protects against misspecification of the working correlation but works (satisfactorily) under specific settings

4.7 Review of Key Points (cont'd)

- Generalized Estimating Equations: Features
  - ▷ only Wald tests available
  - ▷ strict assumptions with respect to incomplete data