Dominic Addulli Homework 6 BIOS 6643

Question 1

Below are two spaghetti plots of these individuals using two different x-axes. Assess the between and within subject variation in 3 sentences or less.

Left Plot

There is a noticeable amount of variation between subjects, with subjects covering a wide range at specific time points. There also seems to be a decent amount of variation within subjects. As time progresses subjects have a decrease along the y-axis.

Right Plot

There is a lot of variation between subjects, with each timepoint having a wide range between subjects. There is less variation within a subject, with not much change along the y-axis for a subject as time changes.

Dominic Abbucci Homework 6 BIOS 6643

Question 2

You decide to model these data using a linear mixed effects model with a linear trend for observation year and use a marginal compound symmetry model to incorporate correlation between the cognitive measures on the same person. Write out the individual (subject) level mixed effects model for this model. Define all matrix dimensions and indices. Include the distribution of model error and any random effects components including the structure of their variance terms.

$Y_i = \overbrace{X_i \beta}^{\text{fixed}} + \varepsilon_i$ Because we are using a compound symmetry model we are not including a random effect. This model will be equivalent to a LMM with a random intercept.

Age of 0 would be the reference.

$Y_i = (n_i \times 1)$; n_i is the number of observation years.

$X_i = (n_i \times p)$; where n_i is the number of observation years and $p = 2$ because of the fixed intercept and the effect of linear time. For subjects with less than 8, n_i will be < 8.

$\beta = (p \times 1)$; There is only 1 covariate, age, so $p = 2$.

$\varepsilon_i = (n_i \times 1)$; n_i is the number of observation years for a subject.

$i = 1, 2, \ldots, n$, where n is the number of subjects.

$\varepsilon_i \sim MVN(0, R_i)$,

where $R_i = \begin{bmatrix} \sigma^2 + \sigma_{12} & \cdots & \sigma_{12} \\ \sigma_{12} & \ddots & \sigma^2 + \sigma_{12} \end{bmatrix}$

Dominic Adducci Homework 6 BIOS 6643

Question 3

Show that the minimum of the GEE objective function $\sum_{i=1}^{N} [y_i - \mu_i(\beta)]^T V_i^{-1} [y_i - \mu_i(\beta)]$ is $\sum_{i=1}^{N} D_i^T V_i^{-1} [y_i - \mu_i(\beta)] = 0$

$$\sum_{i=1}^{N} (y_i - \mu_i(\beta))^T V_i^{-1} (y_i - \mu(\beta)) = \sum_{i=1}^{N} (y_i^T - \mu_i(\beta)^T) V_i^{-1} (y_i - \mu_i(\beta))$$

moving V_i^{-1} into first set of parenthesis

$$\sum_{i=1}^{N} (y_i^T V_i^{-1} - \mu_i^T(\beta) V_i^{-1})(y_i - \mu_i(\beta))$$

Then multiplying the two parenthesis together

$$\sum_{i=1}^{N} (y_i^T V_i^{-1} y_i - \mu_i^T(\beta) V_i^{-1} y_i - y_i^T V_i^{-1} \mu_i(\beta) + \mu_i^T(\beta) V_i^{-1} \mu_i(\beta))$$

Now evaluating all the terms:

$-y_i^T V_i^{-1} \mu_i(\beta) \rightarrow (1 \times N)(N \times N)(N \times 1) = -1$

$-\mu_i^T(\beta) V_i^{-1} y_i \rightarrow (1 \times N)(N \times N)(N \times 1) = -1$

$(\mu_i^T(\beta) V_i^{-1} y_i)^T = y_i^T (V_i^{-1})^T \mu_i(\beta) = y_i^T V_i^{-1} \mu_i(\beta)$

This works because V_i^{-1} is a symmetric matrix, meaning $(V_i^{-1})^T = V_i^{-1}$

This then evaluates to:

$$\sum_{i=1}^{N} (y_i^T V_i^{-1} y_i - 2\mu_i^T(\beta) V_i^{-1} y_i + \mu_i^T(\beta) V_i^{-1} \mu_i(\beta))$$

Taking the first derivative and setting to 0 will find the minimum.

$$\frac{\partial}{\partial \beta} \sum_{i=1}^{N} (y_i^T V_i^{-1} y_i - 2\mu_i^T(\beta) V_i^{-1} y_i + \mu_i^T(\beta) V_i^{-1} \mu_i(\beta)) = 0$$

$$\sum_{i=1}^{N} (0 - 2[\mu_i^T(\beta)'] V_i^{-1} y_i + 2[\mu_i(\beta)']^T V_i^{-1} \mu_i(\beta)) = 0$$

$$\sum_{i=1}^{N} ([\mu_i(\beta)']^T V_i^{-1} [y_i - \mu_i(\beta)]) = 0$$

$$\Rightarrow \sum_{i=1}^{N} D_i^T V_i^{-1} (y_i - \mu_i(\beta)) = 0$$

where $D_i = \mu_i(\beta)' = \frac{\partial}{\partial \beta} \mu_i(\beta)$

Dominic Abucci Homework 6 BIOS 6643

Question 4

Example: We received electronic health records data on 90,000 patients who had a positive test result or were administered monoclonal antibodies (mAbs) for COVID treatment in the CU Health system. All patients were eligible to receive mAbs treatment. The investigator was interested in whether mAbs treatment reduced 28-day hospitalization after adjusting for important precision and compounding variables of age, insurance status, and number of comorbid conditions.

A) Approach

I interpreted the investigators question as wanting to know if treatment was less than 28 days if mAbs treatment was used. The precision and confounding variables of age, insurance status, and number of comorbid conditions are not longitudinal. mAbs is also not longitudinal, as patients either received mAbs or they did not. A binomial generalized linear model will be used to determine if the probability of 28-day hospitalization is reduced with mAbs treatment. A logit link will be used and the previously noted covariates will be controlled for.

B) Outcome

The outcome for this is the odds ratio of 28 day hospitalization.

C) Distributional assumption for this model

The distributional assumption is a Bernoulli distribution for the outcome.

D) Write out the distribution of the outcome with correct subscripts and definitions of indices. Write out the expected value and variances for this distribution.

$Y_i \sim \text{Bernoulli}(\mu_i)$; i refers to the subject. μ_i is probability of < 28 day hospitalization.

$E[Y_i] = \mu_i$; $\text{Var}[Y_i] = \mu_i(1-\mu_i)$

E) What is the linear model for these data (i.e. the systematic component)? Be specific using the covariates you were given.

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \eta_i = \beta_0 + \beta_1 \text{mAbs}_i + \beta_2 \text{Age}_i + \beta_3 \text{InsuranceStatus}_i + \beta_4 \text{Number of Comorbid Cond}_i$$

(systematic component)

F) What is the link function?

$$\eta_i = \underbrace{\log\left(\frac{\mu_i}{1-\mu_i}\right)}_{\text{Link Function (logit)}}$$

G) How will you interpret the coefficient on treatment for your model? Write a formal statistical interpretation.

The mAbs coefficient is the log odds of less than 28 days of hospitalization if treatment is received. The exponentiated coefficient will give the odds of hospitalization within 28 days. 95% CI will give estimates of the bounds and a p-value will assess significance.

H) How will this model be estimated?

The coefficients will be estimated using maximum likelihood.

Dominic Addesi Homework 6 BIOS 6643

Question 4 [Continued]

I) Write pseudo R code for estimating your model.

```
Less_than_28 = glm( Days_28 ~ Most + Age + Insurance_Status + Number_of_Comorbs,
                    Family = binomial, data = Electronic_Health_Data)
```

Dominic Adducci Homework 6 BIOS 6643

Question 5

Example 2: We received electronic health records data on 90,000 patients who had a positive test result or were administered monoclonal antibodies (mAbs) for COVID treatment in the CU health system. All patients were eligible to receive mAbs treatment. The investigator was interested in whether mAbs treatment reduced the number of emergency department (ED) visits after adjusting for important precision and compounding variables of age, insurance status, and number of comorbid conditions.

A) Approach

In this analysis we want to determine if the number of ED visits (counts) were reduced if mAbs treatment was utilized. The precision and confounding variables of age, insurance status, and number of co morbid conditions are not longitudinal. mAbs is also not longitudinal, as patients either received mAbs or they did not. A Quasi-Poisson generalized linear model will be determined if ED visits will be reduced with mAbs treatment. The number of ED visits will be the outcome, and the previously mentioned covariates will be controlled for.

B) Outcome

The number of ED visits will be the outcome.

C) Distributional assumption for this model

The distributional assumption is Poisson. A Quasi-Poisson model will account for overdispersion.

E) Write out the distribution of the outcome with correct subscripts and definitions of indices. Write out the expected value and variance for this distribution

$Y_i \sim \text{Poisson}(\lambda_i)$; i refers to the subject, λ_i is the rate of hospitalization.

$E[Y_i] = \lambda_i$; $\text{var}[Y_i] = \phi \lambda_i$; ϕ accounts for overdispersion

G) What is the linear model for these data (i.e. the systematic component)?

$Y_i = \beta_0 + \beta_1 \text{mlbs}_i + \beta_2 \text{hx}_i + \beta_3 \text{Insurance Status}_i + \beta_4 \text{Number of comorbid cond.}_i$

A) What is the link function?

$\kappa_i = \log(\lambda_i)$

C) How will you interpret the coefficient on treatment for your model?

The mlbs coefficient is the log of rate for someone who received mlbs treatment versus someone who did not receive the treatment. Less than 1 will mean a lower number of ED visits. The exponentiated coefficient will show the difference rate of ED visits between the treatment and control group. 95% CI will give the bounds, and a p-value will assess significance.

H) How will this model be estimated?

The coefficients will be estimated using maximum likelihood.

Dominic Abucci Homework 6 BIOS 6643

Question 5 [continued]

I) Write pseudo R code for estimating your model.

```
ED_visits_treatment = glm(ED_visits ~ mAbs + Age + Insurance_Status + Number_of_Comorbs,
                          family = quasipoisson, data = Electronic_Health_Data)
```

Dominic Abucci Homework 6 BIOS 6643

Question 6

Example 3: We received electronic health records data on 90,000 patients who had a positive test result or were administered monoclonal antibodies (mAbs) for COVID treatment in the CU Health system who were followed every 3 months for a year (so baseline (first 28 days), 3 mo, 6 mo, 9 mo, and 12 mo). All patients were eligible to receive mAbs treatment. The investigator was interested in whether mAbs treatment modifies the number and pattern of post covid doctor visits over the year after COVID. We need to adjust for important precision and confounding variables of age, insurance status, and number of comorbid conditions at baseline.

A.) Approach

This analysis has a longitudinal component where subjects are assessed every 3 months. Each subject will have multiple observations, where the primary variable will be mAbs treatment and confounding variables will be age, insurance status, time, and number of comorbid conditions. A marginal GEE will be used to determine if the subjects who receive mAbs treatment have improvements in required doctors visits over time. The mAbs variable will assess the treatment effect, and the time variable will assess the patten of visits. The Gee will use a log link.

B.) Outcome

The outcome will be the number of doctors visits (count).

C.) Distributional assumption for this model

A GEE does not require a distributional assumption, but in this case a log link function will be used which will model the results as a Poisson.

D.) Write out the distribution of the outcome with correct subscripts and definitions of indices. Write out the expected value and variance for this distribution and any covariance structure.

$E[Y_{ij}] = \lambda_{ij}$; where λ_{ij} is the rate parameter, i is the subject, and j is the timepoint.
$Var[Y_{ij}] = \phi \lambda_{ij}$, ϕ accounts for over dispersion.
$Corr(Y_{ij}, Y_{ik}) = \alpha_{jk}$; Compound symmetry like relationship

E.) What is the linear model for these data (i.e. systematic component)?

$$\eta_{ij} = \beta_0 + \beta_1 mAbs_{ij} + \beta_2 Time_{ij} + \beta_3 Age_{ij} + \beta_4 InsuranceStatus_{ij} + \beta_5 NumberofComorbid_{ij}$$

F.) What is the link function?

$$\eta_{ij} = \log(\lambda_{ij})$$

G.) How will you interpret the coefficient of treatment on your model?

The coefficient of treatment will be the log rate rate of someone who received the mAbs treatment versus someone who did not receive the treatment. The rate in this case will be the rate of Doctors visits. A negative coefficient will mean mAbs treatment reduces Doctors visits.

H.) How will this model be estimated?

The exponentiated treatment coefficient will assess the effect of mAbs on the number of Doctors visits. A rate ratio < 1 will mean mAbs reduces the number of Doctors visits. A 95% CI will assess the range of the effect and a p-value will assess significance. The exponentiated time coefficient will assess the change in Doctors visits over time, where 95% CI will show the range of effect and the p-value will determine if the effect is significant.

Dominic Alluci Homework 6 BIOS 6643

Question 6 [Continued]

I) Write pseudo R code for estimating your model.

```
library(gee)
mtbs_treat = gee(Doctors_visits ~ mtbs + Time + Age + Insurance_Status +
                 Number_of_comorbids, Family = quasipoisson,
                 Corstr = unstructured, data = Electronic_Health_Data)
```

Dominic Alducci Homework 7 BIOS 6643

Question 7

Example 4: We received brain volume metrics and biomarker data from a neuropsychologist who is interested in how the biomarker influences the trajectory of brain volumes over 5 years in those with mild cognitive impairment. Patients received an MRI each year and different brain volumes were computed. For this project the investigator is interested in the hippocampus (one brain region implicated in Alzheimer's). The biomarker data was only measured at baseline. Hippocampal volume can be considered a continuous measure with an infinite number of values possible, meaning it isn't count or binary data. All analysis will be adjusted for baseline age, sex, and SES.

A) Pseudo code (Provided)

```
lmm.un.fit ← gls(hippvol ~ visit + visitsq + biomarker + visit * biomarker + visitsq * biomarker +
                 age + sex + SES, data = braindata, correlation = corSymm(form = 1|ID),
                 weights = varIdent(form = 1|visit))
lmm.un.summary ← summary(lmm.un.fit)
lmm.un.fit2 ← gls(hippvol ~ as.factor(visit) + biomarker + as.factor(visit) * biomarker + age +
                  sex + SES, data = braindata, correlation = corSymm(form = 1|ID),
                  weights = varIdent(form = 1|as.factor(visit))
```

B) Approach

Both models fit a linear regression using the generalized least squares method. For the first model visit is considered a continuous variable and there is a squared visit term to account for a non-linear relationship between visit and hippocampus volume. An interaction between the biomarker variable and the visit and visit squared term was included in the model. An unstructured correlation for each subject was assumed and different variance were assumed for different visits. The second model has many of the same attributes except that the visit variable is factored. There is still an interaction between factored visit and the biomarker and the weights parameter assigned different variances for different visits.

C1) What is the outcome for these analyses?

The outcome in both models is hippocampus volume.

D2) What is the distributional assumption you will make for this outcome?

The gls model makes no assumptions on the distribution of the outcome, but does assume the residuals are distributed normally with constant variance for each visit group.

G1) How is time modeled in each of these models? How did you determine this?

For the first model time is a continuous variable. I determined this by the fact that there is a squared time term. The model includes a visit term, a visit squared term, as well as an interaction between both of these terms and the biomarker variable. For the second model time is factored making it discrete. This was simple to determine, as a factored variable is inherently discrete. There is also an interaction in this model between factored time and the biomarker.

F3) Write out the distribution of the outcome with correct subscripts and definitions of indices. Write out the expected value and variance-covariance for this distribution for each of these models

Model 1 expected value:

$$E[Y_{ij}] = \beta_0 + \beta_1 \text{visit}_{ij} + \beta_2 \text{visitsq}_{ij} + \beta_3 \text{biomarker}_{ij} + \beta_4 \text{visit} * \text{biomarker}_{ij} + \beta_5 \text{visitsq} * \text{biomarker}_{ij} + \beta_6 \text{age}_{ij} + \beta_7 \text{sex}_{ij} + \beta_8 \text{SES}_{ij}$$

where i is the subject; $i = 1, \ldots, n$ n = number of subjects

where j is the observation; $j = 1, \ldots, n_i$ n_i = number of observations for a subject

Domini Adduci Homework 6 BIOS 6643

Question 7 [Continued]

F (Continued)

model 2 expected value (visit 1 reference)

$$E[Y_{ij}] = \beta_0 + \beta_1 Visit2_{ij} + \beta_2 Visit3_{ij} + \beta_3 Visit4_{ij} + \beta_4 Visit5_{ij} + \beta_5 Visit6_{ij} + \beta_6 biomarker_{ij} + \beta_7 Visit2 * biomarker_{ij} + \beta_8 Visit3 * biomarker_{ij} + \beta_9 Visit4 * biomarker_{ij} + \beta_{10} Visit5 * biomarker_{ij} + \beta_{11} age_{ij} + \beta_{12} Sex_{ij} + \beta_{13} SES_{ij}$$

Variance-Covariance $= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} & \sigma_{35} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 & \sigma_{45} \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_5^2 \end{bmatrix} = R_i$

where $\sigma_{12} = \sigma_{21}$; $\sigma_{13} = \sigma_{31}$, etc.

symmetric var-cov structure

R_i is the same for both models

c) what is the linear model for these data in observation level, subject level, and complete data notation for the first model.

Observation level:

$$Y_{ij} = \beta_0 + \beta_1 Visit_{ij} + \beta_2 Visitsq_{ij} + \beta_3 biomarker + \beta_4 Visit * biomarker_{ij} + \beta_5 Visitsq * biomarker_{ij} + \beta_6 age_{ij} + \beta_7 sex_{ij} + \beta_8 SES_{ij} + \epsilon_{ij}$$

where $\epsilon_{ij} \sim N(0, R_i)$; R_i defined above

Subject level:

$$\begin{bmatrix} y_{i1} \\ y_{i2} \\ y_{i3} \\ y_{i4} \\ y_{i5} \end{bmatrix} = \begin{bmatrix} 1 & visit_{i1} & visitsq_{i1} & biomarker_{i1} & visit*biomarker_{i1} & visitsq*biomarker_{i1} \\ 1 & visit_{i2} & visitsq_{i2} & biomarker_{i2} & visit*biomarker_{i2} & visitsq*biomarker_{i2} \\ 1 & visit_{i3} & visitsq_{i3} & biomarker_{i3} & visit*biomarker_{i3} & visitsq*biomarker_{i3} \\ 1 & visit_{i4} & visitsq_{i4} & biomarker_{i4} & visit*biomarker_{i4} & visitsq*biomarker_{i4} \\ 1 & visit_{i5} & visitsq_{i5} & biomarker_{i5} & visit*biomarker_{i5} & visitsq*biomarker_{i5} \end{bmatrix}$$

$$\begin{bmatrix} age_{i1} & sex_{i1} & SES_{i1} \\ age_{i2} & sex_{i2} & SES_{i2} \\ age_{i3} & sex_{i3} & SES_{i3} \\ age_{i4} & sex_{i4} & SES_{i4} \\ age_{i5} & sex_{i5} & SES_{i5} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \\ \beta_9 \end{bmatrix} + \begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \\ \epsilon_{i4} \\ \epsilon_{i5} \end{bmatrix} \qquad \epsilon_{ij} \sim N(0, R_i)$$

Complete data notation

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\epsilon_i \sim N\left(0, \begin{bmatrix} R_1 & & 0 \\ & \ddots & \\ 0 & & R_n \end{bmatrix}\right)$$

ϵ_i: $(n \times 5) \times (n \times 5)$

Dominic Adducci Homework 6 BIOS 6643

Question 7 [Continued]

H.) How will you interpret the coefficient on the interaction with visit in the first model?

If there is a positive interaction coefficient this means that as the visit number increases then the biomarker has an increase on Hippocampus volume volume. If there is a negative coefficient this means as visits increase in number the biomarker has a decrease on Hippocampus volume

I.) How would you interpret the coefficients on the interaction with visit in the second model?

	visit 2	visit 3	visit 4	visit 5
Biomarker	visit2 * biomarker	visit3 * Biomarker	visit4 * biomarker	visit5 * biomarker

Because visit is factored each visit will have its own interaction term with biomarker. For visit 2 * biomarker this coefficient would account for the interaction between visit 2 and the biomarker on the outcome of Hippocampus volume. If positive this would mean the biomarker would have had a positive (increased) effect on Hippocampus volume. If negative this would mean the biomarker would have decreased Hippocampus volume by visit 2.

J.) How will this model be estimated?

The generalized least squares method will be used to estimate the model

Dominic Addeo Homework 6 BIOS 6643

Question 8

Example 5: We received brain volume metrics and biomarker data from a neuropsychologist who is interested in how the biomarker influences the trajectory of brain volumes over 5 years in those with mild cognitive impairment. Patients received an MRI each year and different brain volumes were computed. For this project the investigator is interested in the hippocampus (one brain region implicated in Alzheimer's). The biomarker data was only measured at baseline. Hippocampal volume can be considered a continuous measure with an infinite number of values possible, meaning it isn't count or binary data.

A.) Pseudocode

```
lmm.AR.fit ← gls(hippvol ~ visit + biomarker + visit * biomarker + age + sex + SES, data = brainData,
                 correlation = corAR1(form = 1|ID), method = "ML")
lmm.AR.summary ← summary(lmm.AR.fit)
lmm.AR.fit2 ← gls(hippvol ~ visit + biomarker + visit * biomarker + age + sex + SES, data = brainData,
                  correlation = corAR1(form = 1|ID), method = "ML", weights = varIdent(form = 1|visit))
lmm.AR2.summary ← summary(lmm.AR.fit2)
```

B.) Approach

The first model uses the generalized least squares method to estimate the beta coefficients. The relationship between visit and the biomarker in Changing Hippocampal volume will be assessed using the visit * biomarker interaction term. The model will also control for age, sex, and SES. An autoregressive correlation structure and maximum likelihood will be used (as opposed to REML).

The second model uses the generalized least squares method to estimate the beta coefficients. The relationship between visit and the biomarker in changing Hippocampal volume will be assessed using the visit * biomarker interaction term. The model will also control for age, sex, and SES. An autoregressive correlation structure and maximum likelihood will be used. The weights of the variances will be adjusted between visits.

C.) Distributional Assumptions

The gls method assumes a multivariate normal distribution.

D.) Write out the distribution of the outcome with correct subscripts and definitions of indices. Write out the expected value and variance-covariance for this distribution for each model.

$E[Y_{ij}] = \beta_0 + \beta_1 visit_{ij} + \beta_2 biomarker_{ij} + \beta_3 visit * biomarker_{ij} + \beta_4 age_{ij} + \beta_5 sex_{ij} + \beta_6 SES_{ij}$

Expected value is the same for each model.

Where i is the subject; $i = 1, \ldots, n$ n = number of subjects

Where j is the observation; $j = 1, \ldots, n_i$ n_i = number of observations for a subject.

First model variance-covariance.

$$R_i = \begin{bmatrix} \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 & \sigma^2\rho^3 & \sigma^2\rho^4 \\ \sigma^2\rho & \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 & \sigma^2\rho^3 \\ \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 & \sigma^2\rho & \sigma^2\rho^2 \\ \sigma^2\rho^3 & \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 & \sigma^2\rho \\ \sigma^2\rho^4 & \sigma^2\rho^3 & \sigma^2\rho^2 & \sigma^2\rho & \sigma^2 \end{bmatrix}$$

ρ represents correlation.

Second model variance - covariance

$$R_i = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho^2\sigma_1\sigma_3 & \rho^3\sigma_1\sigma_4 & \rho^4\sigma_1\sigma_5 \\ \rho\sigma_2\sigma_1 & \sigma_2^2 & \rho\sigma_2\sigma_3 & \rho^2\sigma_2\sigma_4 & \rho^3\sigma_2\sigma_5 \\ \rho^2\sigma_3\sigma_1 & \rho\sigma_3\sigma_2 & \sigma_3^2 & \rho\sigma_3\sigma_4 & \rho^2\sigma_3\sigma_5 \\ \rho^3\sigma_4\sigma_1 & \rho^2\sigma_4\sigma_2 & \rho\sigma_4\sigma_3 & \sigma_4^2 & \rho\sigma_4\sigma_5 \\ \rho^4\sigma_5\sigma_1 & \rho^3\sigma_5\sigma_2 & \rho^2\sigma_5\sigma_3 & \rho\sigma_5\sigma_4 & \sigma_5^2 \end{bmatrix}$$

The subscripts on variances and standard deviations represents time points

E) What differed between these two model choices?

The variance-covariance matrix was different. For the second model each timepoint was allowed to have its own variance.

F) What is the linear model for these data in subject level and complete data notation for the first model?

Subject-level

$$\begin{bmatrix} Y_{i1} \\ Y_{i2} \\ \vdots \\ Y_{in_i} \end{bmatrix} = \begin{bmatrix} 1 & visit_{i1} & biomarker_{i1} & visit*biomarker_{i1} & age_{i1} \\ 1 & visit_{i2} & biomarker_{i2} & visit*biomarker_{i2} & age_{i2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & visit_{in_i} & biomarker_{in_i} & visit*biomarker_{in_i} & age_{in_i} \end{bmatrix}$$

$$\begin{bmatrix} age_{i1} & sex_{i1} & SES_{i1} \\ age_{i2} & sex_{i2} & SES_{i2} \\ \vdots & \vdots & \vdots \\ age_{in_i} & sex_{i3} & SES_{in_i} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix} + \underbrace{\begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{in_i} \end{bmatrix}}_{= \varepsilon_i}$$

Where i is the subject; $i = 1, \ldots, n$ n = number of subjects

where j is the observation; $j = 1, \ldots, n_i$ n_i is the number of observations for a subject.

$\varepsilon_i \sim N(0, R_i)$; R_i defined previously.

Complete Data

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

y_i defined in subject level model
X_i defined in subject level model

$$\varepsilon_i \sim N\left(0, \begin{bmatrix} R_1 & & 0 \\ & \ddots & \\ 0 & & R_{n_i} \end{bmatrix}\right)$$

(c.) How will this model be estimated?

This model will be estimated using maximum likelihood.