PHP2516: Applied Longitudinal Data Analysis

Homework 2

Antonella Basso

Febuary 26, 2022

Part I

The 'calcium_all' data is a subset of the 'calcium' dataset that can be found in the 'lava' R package. This dataset contains information collected from an RCT aimed at comparing calcium vs placebo with respect to changes in bone mineral density measures (g/cm2) over time. For this purpose BMD measurements were taken on girls at approximately every 6th months in 3 years. The "calcium_all" dataset has information on the following variables:

- **bmd**: bone mineral density (BMD) in g/cm2
- **group**: treatment group ('C'=calcium, 'P'=placebo)
- **person**: person (girl) ID
- **visit**: visit number (time point)
- **age**: age at each visit in years

Question 1:

Using the 'calcium_allL' dataset please answer the following questions, assuming unstructured covariance pattern:

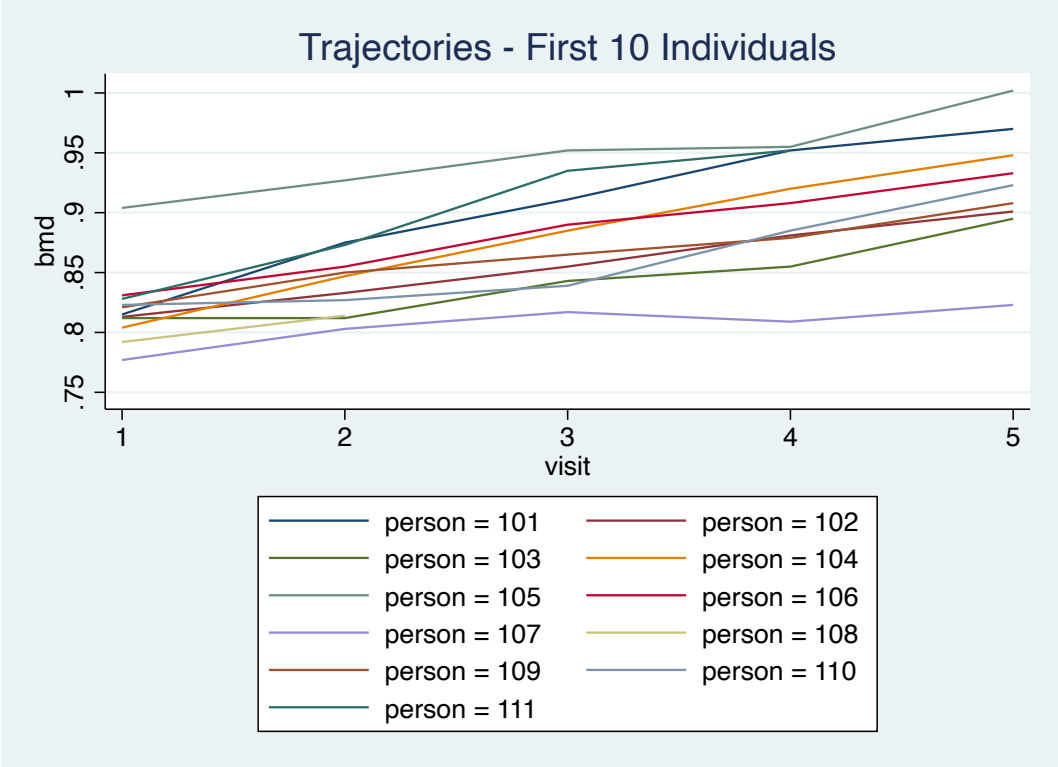
a) Describe the missing patterns you see in the data (if any).
b) Plot the observed data (means and spaghetti plots). What do you observe?
c) Conduct a Mean Response Profiles analysis (Model 1) with only time and treatment group. What is your overall conclusion about the changes in the mean response over time and the effect of the treatment group on those changes?
d) Suppose that you are mainly interested in describing the trends in the mean responses over time adjusting for age. Find a model for the mean that best fits your data. [HINT: Consider only up to two-way interactions of treatment with time and age. Start with the 'full' model and follow a backward elimination procedure based on the partial p-values of the regression coefficients.]

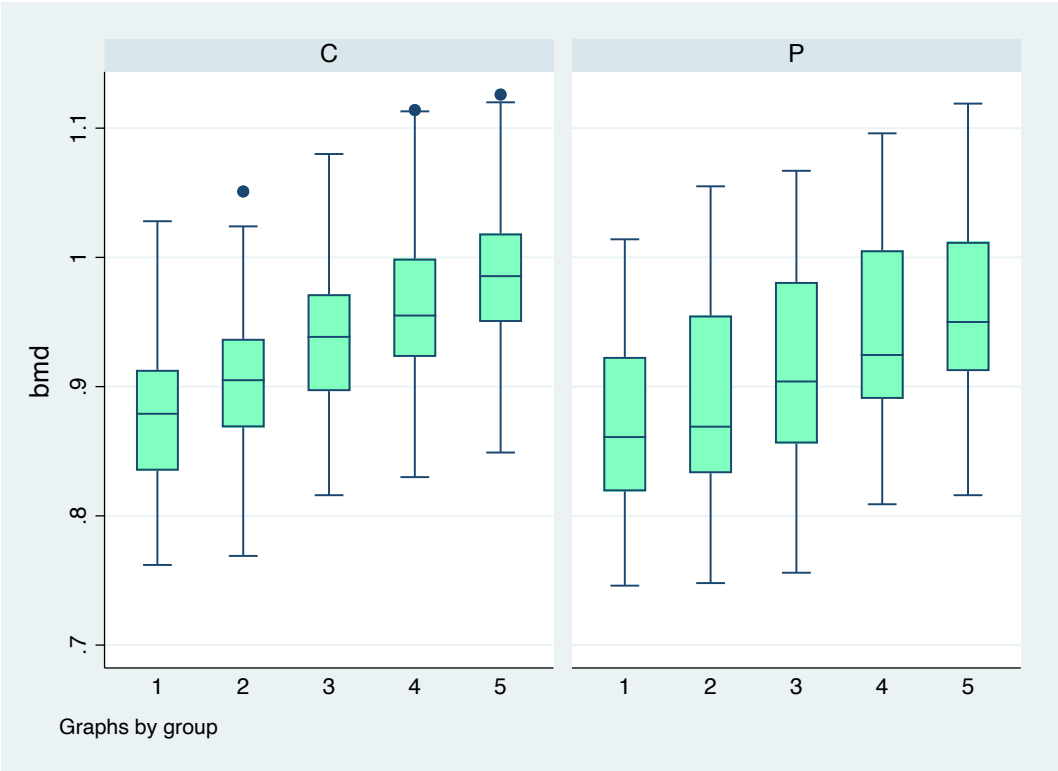
Solution

a) Looking at the data using the 'xtset' and 'xtdescribe' commands in STATA, we see that the data is unbalanced with respect to the number of bone mineral density (BMD) observations for each subject. That is, while some subjects had observations for all five visits, others did not. Specifically, out of 112 individuals there were, 7, 6, 5, and 3, who only had observations up to the first, second, third, and fouth visits, respectively.

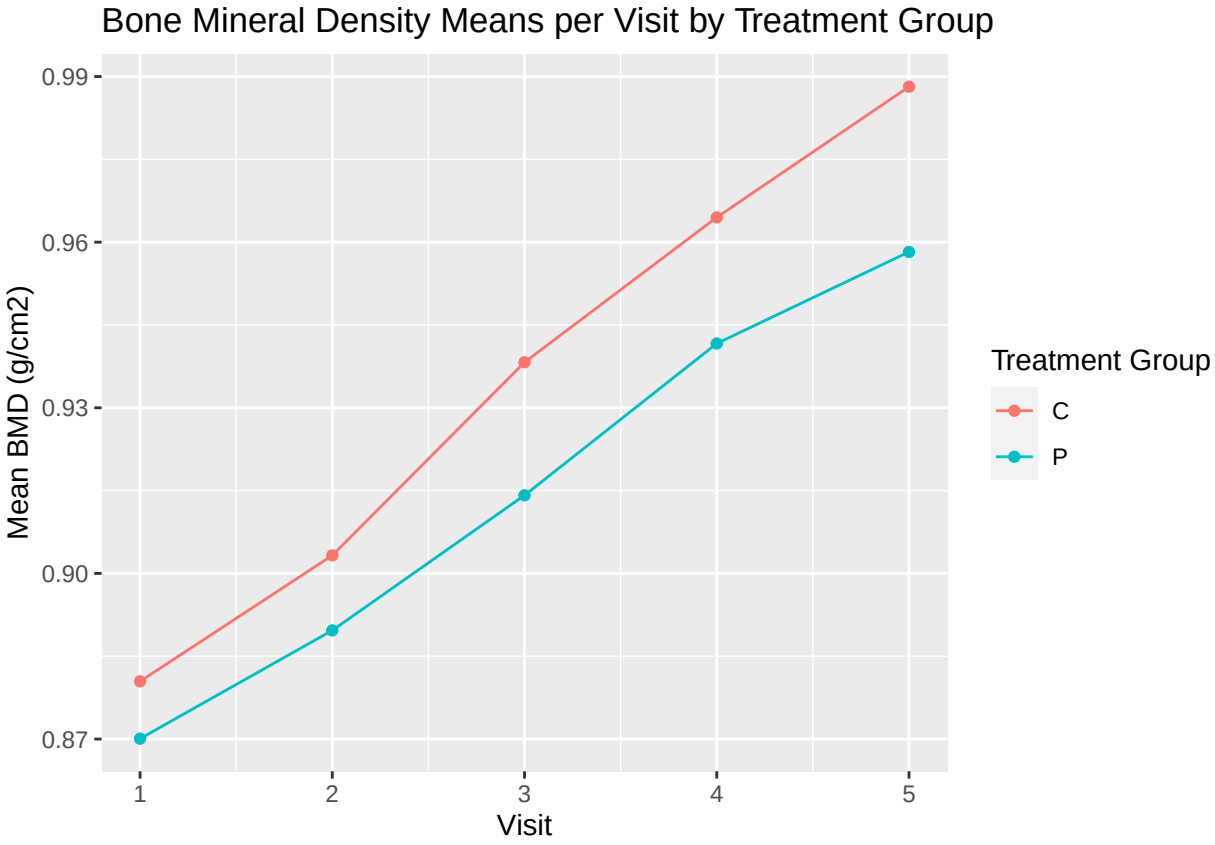
b) The first two plots below, graphed in STATA, display the trajectories (in BMD over time) for the first ten individuals and the spread of BMD at each visit for each treatment group (in the form of boxplots), respectively. The spaghetti plot shows us that despite variations in response becoming slightly greater with each visit, individuals display a noticeable range of BMD values at baseline, which follow very similar increasing trends over time. Similarly, the boxplots suggest that despite the placebo group

having overall smaller BMD's, the change in response over time is strikingly similar between both groups.





These observations are further confirmed by the graph below, which shows more explicitly the change in mean responses over time for each treatment group. Precisely, we notice that BMD means in both groups grow at very similar rates, although the treatment group shows a slightly more rapid increase, making the difference in means at the last visit (5) a bit larger than at baseline.



c) A Mean Response Profiles analysis with only time and treatment group was fitted using STATA, the code for which is given below. This marginal (population-average) model assumes no random effects; unstructured residuals; a discrete measure of time (visits); and uses restricted maximum likelihood estimation to derive coefficients for time, treatment, and their interaction.

Let:

- Y_{ij} be the outcome of interest (random variable), bone mass density (BMD), for the i^{th} individual in the study on the j^{th} visit, where $i = 1, 2, ..., 112$ and $j = 2, 3, ..., 5$
- V_{ij} be the discrete time covariate denoting the j^{th} visit in the study with V_{i1} as the baseline and reference group
  - *Note: each V_{ij} is binary and indicates the j^{th} visit for the i^{th} individual when $V_{ij} = 1$
- X_{i1} be the binary predictor for group, where $X_{i1} = 0$ indicates treatment (reference group), and $X_{i1} = 1$ indicates placebo

Model 1:

```
mixed bmd tx##visit || person: , noconst residuals(unstructured, t(visit)) reml
nolog
```

$$E[Y_{ij}] = \beta_0 + \beta_1 \cdot V_{i2} + \beta_2 \cdot V_{i3} + \beta_3 \cdot V_{i4} + \beta_4 \cdot V_{i5} - \beta_5 \cdot X_{i1} - \beta_6 \cdot (V_{i2} * X_{i1}) - \beta_7 \cdot (V_{i3} * X_{i1}) - \beta_8 \cdot (V_{i4} * X_{i1}) - \beta_9 \cdot (V_{i5} * X_{i1})$$

$$E[Y_{ij}] = 0.8804545 + 0.0270333 \cdot V_{i2} + 0.056471 \cdot V_{i3} + 0.0830392 \cdot V_{i4} + 0.1060277 \cdot V_{i5}$$
$$-0.0103844 \cdot X_{i1} - 0.0069691 \cdot (V_{i2} * X_{i1}) - 0.0123047 \cdot (V_{i3} * X_{i1})$$

$$-0.0124964 \cdot (V_{i4} * X_{i1}) - 0.0189692 \cdot (V_{i5} * X_{i1})$$

From these coeffients we see that with each visit, BMD increases by $\beta_k - \beta_{k-1}, k \neq 0$ for the treatment group. That is, BMD increases by:

- 0.0270333 from the first (baseline) to the second visit;
- $0.056471 - 0.0270333 = 0.0294377$ from the second to the third visit;
- $0.0830392 - 0.056471 = 0.0265682$ from the third to the fourth visit; and
- $0.1060277 - 0.0830392 = 0.0229885$ from the fourth to the fifth visit.

For the placebo group, we see similarly that BMD increases by:

- $0.0270333 - 0.0069691 = 0.0200642$ from the first (baseline) to the second visit;
- $(0.056471 - 0.0123047) - (0.0270333 - 0.0069691) = 0.0241021$ from the second to the third visit;
- $(0.0830392 - 0.0124964) - (0.056471 - 0.0123047) = 0.0263765$ from the third to the fourth visit; and
- $(0.1060277 - 0.018969) - (0.0830392 - 0.0124964) = 0.0165159$ from the fourth to the fifth visit.

Thus, based on this model, the rate of increase in mean response over time is almost constant for both groups.

d) To find a model for the mean response (BMD) over time (adjusting for age) that best fits the data, we use a backward elimination based on the partial p-values of the regression coefficients, starting with the largest ("full") model possible for our purposes (M1). Using STATA, we obtain the following parametric curve models (continuous time) sequentially, via this model selection procedure:

M1: (two-way interactions of treatment with time and age)

```
mixed bmd tx##c.visit tx##c.age || person:, noconst residuals(unstructured,
t(visit)) reml nolog
```

M2: (two-way interaction of treatment with time, age)

```
mixed bmd tx##c.visit age || person:, noconst residuals(unstructured,
t(visit)) reml nolog
```

Given the statistical significance of the treatment and time (visit) interaction and age coefficients, the best fitting model for the mean response (BMD) over time is M2.

Question 2:

Start with a simple model assuming only a linear trend over time, the main effect of treatment, no interaction between time and treatment, and no adjustment for other covariates.

a) Select the model that best fits the covariance structure of the data between unstructured and exchangeable.
b) Then fit the model that best describes the trends in the mean responses over time.
c) Is the final "best" model for your data the same with the one resulted from the process followed in Question 1?

NOTE: In this exercise note how the standard errors and the quantities tightly connected with them (p-values, and CIs) change with the choice of the covariance model. Always remember that the choice of the covariance model may affect the model selection procedure for the mean and vice versa.

Solution

a) Starting with a simple parametric curve model for the mean response, we fit unstructured and exchangeable covariance structures in STATA:

Unstructured Covariance:

```
mixed bmd tx visit || person: , covariance(unstr) reml nolog
```

Exchangeable Covariance:
`mixed bmd tx visit || person: visit, covariance(exch) reml nolog`

Given the reulting BIC values, we find that the covariance structure that best fits the model is that of unstructured covariance.

b) Holding this covariance pattern constant, we use backward elimination and BIC values to choose between the following models:

M1: `mixed bmd tx##c.visit tx##c.age || person: , covariance(unstr) reml nolog`

M2: `mixed bmd visit tx##c.age || person: , covariance(unstr) reml nolog`

M3: `mixed bmd tx##c.age || person: , covariance(unstr) reml nolog`

Based on the partial p-values of the regression coefficients, as well as BIC comparisons, we find that M3 is the best fitting model. That is, a model that uses restricted maximum likelihood estimation, takes age as a continuous measure of time (finding significance between the interaction of age and treatment), and assumes an unstructured covariance pattern (random intercept and constant slope).

c) This final "best" model for the data differs from the one resulted from the process followed in Question 1, primarily in that it accounts for between-subject variation. That is, this model is a mixed effects model with an unstructured covariance pattern, while the other is marginal (population-average) model with no covariance structure (random effects) and unstructured residuals. Moreover, the model from Question 1.d finds significance in using "visit" as a (continuous) measure of time and includes age, while the one fit here prioritizes "age" in the same way, excluding the "visit" variable altogether.

Question 3:

Interpret the regression coefficients from Model 1 and the best model(s) for the mean resulted from Question 1.d and Question 2. What is your final conclusion about the changes in the response over time between the two treatment groups?

Solution

Let:

- Y_{ij} be the outcome of interest (random variable), bone mass density (BMD), for the i^{th} individual in the study on the j^{th} visit, where $i = 1, 2, ..., 112$ and $j = 2, 3, ..., 5$
- V_{ij} be the discrete time/visit covariate (in the first model) denoting the j^{th} visit in the study with V_{i1} as the baseline and reference group
  - *Note: each V_{ij} is binary and indicates the j^{th} visit for the i^{th} individual when $V_{ij} = 1$
- V_i be the continuous time/visit covariate (in the second model) denoting the measurement time point (visit) for the i^{th} individual, with $V_i = 1$ as the baseline
- X_{i1} be the binary predictor for group, where $X_{i1} = 0$ indicates treatment (reference group), and $X_{i1} = 1$ indicates placebo
- X_{i2} be the continuous covariate for age
- $Z_i * b_i$ be the random effects part of the model, where Z_i (if present) are time-varying predictors (subset of X_i) for which the coefficients, b_i, give the i^{th} individual's deviation from the population mean response (varying slopes and/or intercepts)

Model 1:
`mixed bmd tx##visit || person: , noconst residuals(unstructured, t(visit)) reml nolog`

$$E[Y_{ij}] = \beta_0 + \beta_1 \cdot V_{i2} + \beta_2 \cdot V_{i3} + \beta_3 \cdot V_{i4} + \beta_4 \cdot V_{i5} - \beta_5 \cdot X_{i1} - \beta_6 \cdot (V_{i2} * X_{i1}) - \beta_7 \cdot (V_{i3} * X_{i1}) - \beta_8 \cdot (V_{i4} * X_{i1}) - \beta_9 \cdot (V_{i5} * X_{i1})$$

$$E[Y_{ij}] = 0.8804545 + 0.0270333 \cdot V_{i2} + 0.056471 \cdot V_{i3} + 0.0830392 \cdot V_{i4} + 0.1060277 \cdot V_{i5}$$

$$-0.0103844 \cdot X_{i1} - 0.0069691 \cdot (V_{i2} * X_{i1}) - 0.0123047 \cdot (V_{i3} * X_{i1})$$
$$-0.0124964 \cdot (V_{i4} * X_{i1}) - 0.0189692 \cdot (V_{i5} * X_{i1})$$

Model 2: (Question 1.d)

```
mixed bmd tx##c.visit age || person:, noconst residuals(unstructured, t(visit)) reml
nolog
```

$$E[Y_{ij}] = \beta_0 + \beta_1 \cdot V_i - \beta_2 \cdot X_{i1} + \beta_3 \cdot X_{i2} - \beta_4 \cdot (V_i * X_{i1})$$
$$E[Y_{ij}] = 0.4471778 + 0.0074499 \cdot V_i - 0.0062377 \cdot X_{i1} + 0.0383463 \cdot X_{i2} - 0.0042865 \cdot (V_i * X_{i1})$$

Model 3: (Question 2)

```
mixed bmd tx##c.age || person: , covariance(unstr) reml nolog
```

$$E[Y_{ij}|b_i] = \beta_0 + \beta_1 \cdot X_{i1} + \beta_2 \cdot X_{i2} - \beta_3 \cdot (X_{i1} * X_{i2}) + b_{i0}$$
$$E[Y_{ij}|b_i] = 0.274753 + 0.0869336 \cdot X_{i1} + 0.0547895 \cdot X_{i2} - 0.0089102 \cdot (X_{i1} * X_{i2}) + b_{i0}$$

Where:

- $Var(b_{i0}) = 0.0045157$
- Thus, a 95% CI for β_0 is given by $0.274753 \pm 1.96\sqrt{0.0045157} = 0.274753 \pm 0.1317075 \approx [0.143043, 0.406463]$

Coefficient Interpretation:

In the first model, β_0 represents the population mean response (bone mass density) at baseline (the first visit) for those in the treatment group. Since time is continuous and measured differently in the last two models however, with the baseline not being equal to 0 in either case (and age included, despite being constant among subjects), the interpretation of β_0 is not so straightforward. That is, the population mean response for the treatment group at baseline is given by $\beta_0 + \beta_1 + \beta_3 \cdot X_{i2}$, and $\beta_0 + \beta_2 \cdot X_{i2}$ in the second and third models, respectively. In Model 3 however, since it is not a marginal model, we see specifically that the i^{th} individual's BMD at baseline is given by $\beta_0 + \beta_2 \cdot X_{i2} + b_{0i}$, where b_{0i} is the individual's specific deviation from the population mean response at the starting age (for the treatment group). With $Var(b_{i0}) = 0.0045157$, it follows that the 95% CI for β_0, namely $[0.143043, 0.406463]$, gives the between-subject variation in the model's intercept. As this assumes an age of 0, it is more reasonable to understand a subject i's response at baseline as $0.0547895 \cdot X_{i2}$ plus some number between $[0.143043, 0.406463]$ 95% of the time for those in the treatment group. Looking at the expected mean responses at baseline (visit 1 or age ≈ 11) between all three models, we see that predictions are relatively consistent, with discrepancies likely resulting from a difference in measures of time, and the inclusions of age and a covariance structure (transition from marginal to mixed-effects models).

Model 1:

- treatment : 0.8804545
- placebo : $0.8804545 - 0.0103844 = 0.8700701$

Model 2: (Question 1.d)

- treatment : $0.4471778 + 0.0074499 + 0.0383463 * (11) = 0.876437$
- placebo : $0.4471778 + 0.0074499 - 0.0062377 + 0.0383463 * (11) - 0.0042865 = 0.8659128$

Model 3: (Question 2)

- treatment : $0.274753 + 0.0547895 * (11) + b_{0i} = 0.8774375 + b_{0i}$, for $b_{0i} \in [-0.1317075, 0.1317075]$
- placebo : $0.274753 + 0.0869336 + 0.0547895 * (11) - 0.0089102 * (11) + b_{0i} = 0.8663589 + b_{0i}$, for $b_{0i} \in [-0.1317075, 0.1317075]$

The remaining coefficients in the first model, as mentioned previously, show us that with each subsequent visit, BMD increases by $\beta_k - \beta_{k-1}, k \neq 0$ for the treatment group. That is, BMD increases by 0.0270333 from the first (baseline) to the second visit, by $0.056471 - 0.0270333 = 0.0294377$ from the second to the third visit, and so on. The same observation can be made with regards to the placebo group which display overall smaller outcomes, but changing in a very similar way over time (as shown in Question 1). The remaining coefficients in the second model show that BMD increases by roughly 0.0383463 g/cm^2 for every unit increase in age for both treatment groups, and by an additional 0.0074499 g/cm^2 and $0.0074499 - 0.0042865 = 0.0031634$ g/cm^2 for every unit increase in visit, for those in the treatment and placebo groups, respectively. The remaining coefficients in the third model on the other hand, show that BMD increases by roughly 0.0547895 g/cm^2 for every unit increase in age in the treatment group, and by approximately $0.0547895 - 0.0089102 = 0.0458793$ g/cm^2 for every unit increase in age in the placebo group. Thus, given all three models' coefficients, we see that changes in the response over time between the two treatment groups is very similar. Although, it may be safe to assume a slightly higher rate of increase for the treatment group, which also demonstrated slightly larger BMD measurements at baseline. The third and best fitting model outlines this observation more explicitly, which we notice in the plots shown above.

Part II

The 'cd4.dta' dataset includes data from a randomized, double-blind study of AIDS patients with advanced immune suppression (CD4 counts of ≤ 50 cells/mm^3) (Henry et al., 1998). Detailed description of this study can be found in your 'Applied Longitudinal Analysis' text book (p.228). For the purposes of the analyses we will collapse the treatment groups into two broader categories: a dual therapy (regimens 1, 2, and 3) versus a triple therapy (regimen 4).

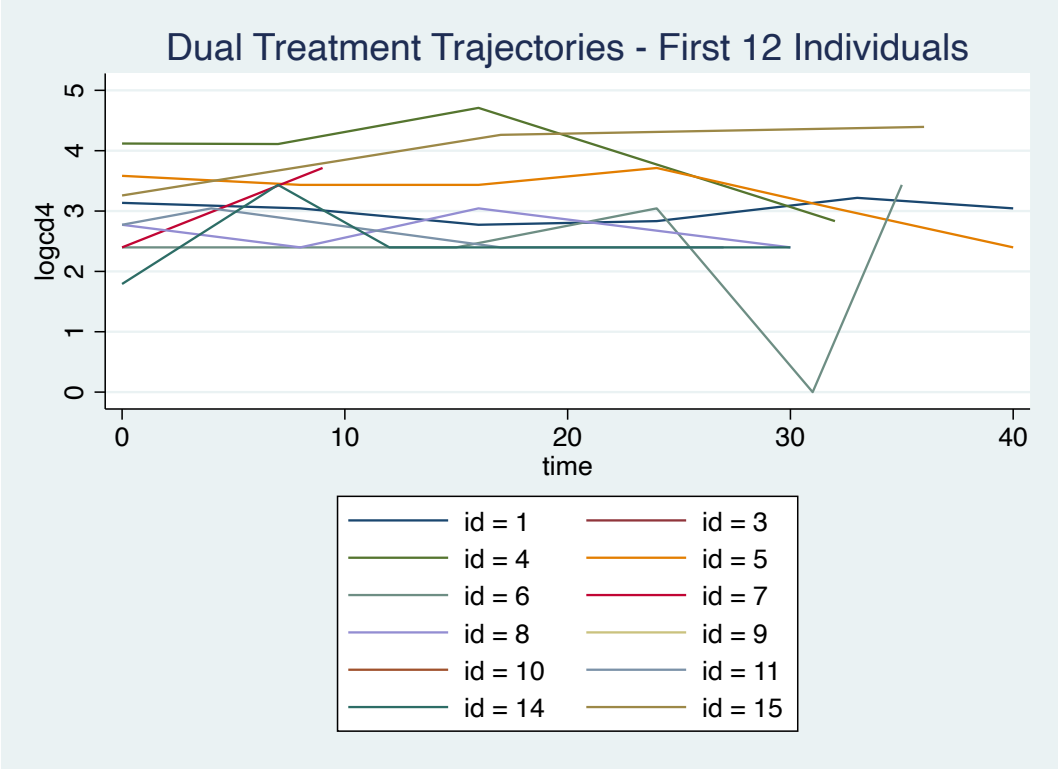
Question 4:

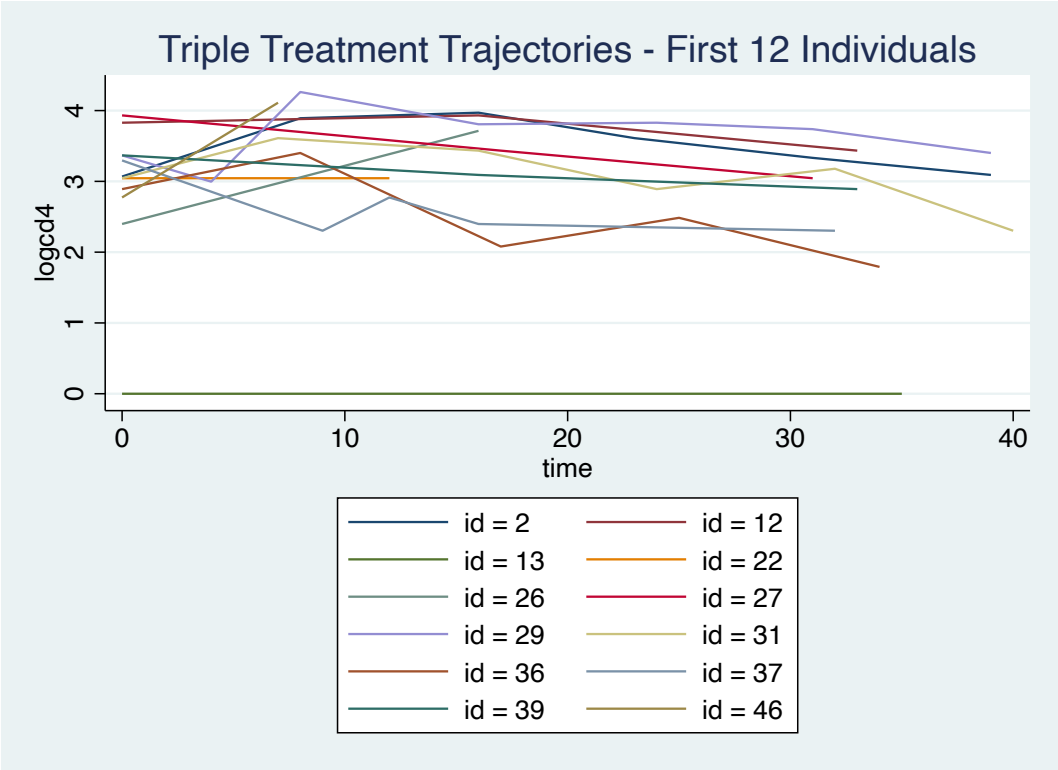
Using the 'cd4' dataset please answer the following questions:

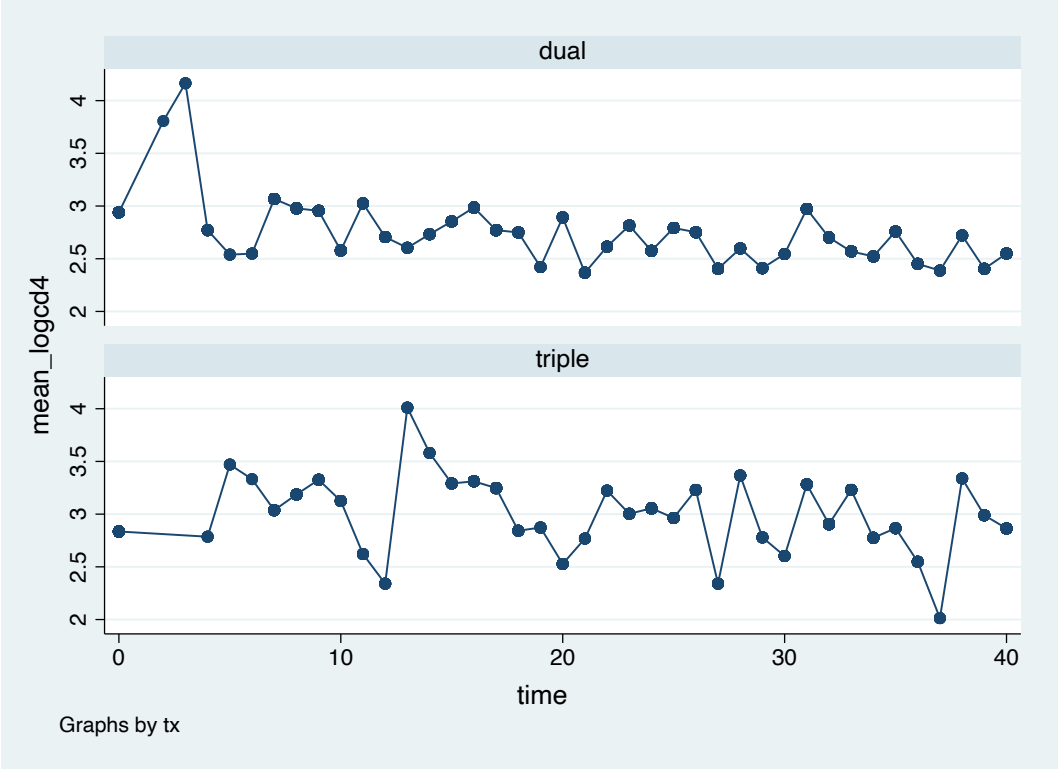
a) Create spaghetti plots to explore the patterns of the individual trajectories over time, by treatment group. What do you observe?
b) Based on the plots of the observed data what model would you prefer to fit to express the changes in the primary outcome (log of CD4 counts) over time and any differences between the two treatment groups? Explain.

Solution

a) The first two spaghetti plots below give us a glimpse into the considerable amount of data missing. It is evident that both treatment groups have several individuals lacking measurements in more than one visit, which makes this data significantly unbalanced. Moreover, these graphs highlight appreciable differences in subject observations at baseline as well as fluctuating responses over time, making it difficult to identify specific patterns between groups. The two-way scatter plot (third plot below), gives us a better overview of the general trends in the data based on treatment group. However, given the amount of missing data and between-subject variation, it is hard to tell whether these means could be biased or in fact representative of true population averages. Given all three plots, it seems that there could be overall higher log(cd4) levels with more sporadic behaviors/responses over time for subjects in the triple treatment group, but definitive claims regarding patterns in the outcome can't be made without a more extensive analysis of the data.


Dual Treatment Trajectories - First 12 Individuals
logcd4
0
1
2
3
4
5
0
10
20
30
40
time
id = 1
id = 3
id = 4
id = 5
id = 6
id = 7
id = 8
id = 9
id = 10
id = 11
id = 14
id = 15


Triple Treatment Trajectories - First 12 Individuals
logcd4
0
1
2
3
4
0
10
20
30
40
time
id = 2
id = 12
id = 13
id = 22
id = 26
id = 27
id = 29
id = 31
id = 36
id = 37
id = 39
id = 46



b) Not only would a marginal model be inappropriate and potentially misleading given the substantial amount of missing data observed, but fitting a population-average model would (as did the two-way scatter plot) fail to capture the notable between-subject differences we suspect are present and within-individual changes.Thus, based on what was gathered from the plots above, it seems that a mixed effects model with an unbalanced covariance structure would be more suitable for the data.

Question 5:

Try to fit a marginal model to describe the trends in the mean responses over time by treatment group. For the mean model consider only a linear trend, and test whether the changes over time are different among the treatment groups without adjusting for any other covariates in the model. For the covariance pattern assume unstructured and exchangeable. What do you observe? Do you think that a marginal model is appropriate for analyzing these data? Explain.

Solution

Marginal Model (Unstructured Covariance):

```
xi: mixed logcd4 i.tx*time || id: , noconst covariance(unstr) reml nolog
```

Marginal Model (Exchangeable Covariance):

```
xi: mixed logcd4 i.tx*time || id: , noconst covariance(exch) reml nolog
```

The marginal parametric curve models fit via restricted maximum likelihood estimation, assuming unstructured and exchangeable covariance patterns (in STATA), both yielded the same following output:

logcd4	Coefficient	Std. err.	z	P$>\|z\|$	[95% conf. interval]
_Itx_2	.0201841	.0547426	0.37	0.712	-.0871094 .1274776
time	-.0105755	.0013910	-7.60	0.000	-.0133019 -.0078491
_ItxXtime_2	.0115745	.0027262	4.25	0.000	.0062312 .0169178
_cons	2.983828	.0276616	107.87	0.000	2.929613 3.038044

As marginal models make no assumptions about covariance patterns or subject-specific outcomes, it is not surprising that both models would produce identical results. Specifically, the randomness found within the data, not accounted for by the fixed effects of these models is given by the residuals, which were estimated to have the following variance:

Random-effects parameters	Estimate	Std. err.	[95% conf. interval]
var(Residual)	1.142763	.0227734	1.098988 1.188281

Although p-values demonstrate statistically significant coefficients in this model, the high residual variance indicates that there is a substantial amount of information that is not being accounted for by a mere "population-average". This sheds light on the limitations of a marginal model for explaining this data and provides sufficient evidence in favor of a mixed-effects model.

Question 6:

Assume the following mixed-effects model to the data:

$$\begin{aligned} E[Y_{ij}|b_i] = \beta_0 + \beta_1 \cdot t_i + \beta_2 \cdot t_i^2 + \beta_3 \cdot \text{trt} + \beta_4 \cdot \text{age} + \beta_5 \cdot \text{gender} \\ +\beta_6 \cdot (t_i * \text{trt}) + \beta_7 \cdot (t_i^2 * \text{trt}) \\ +\beta_8 \cdot (\text{age} * \text{trt}) + \beta_9 \cdot (\text{gender} * \text{trt}) \\ +(Z_i * b_i) \end{aligned}$$

Where:

- t_{ij} is the time point at which the j^{th} measurement is taken for individual i (with t_{ij} at baseline)
- trt is the treatment group (1 =triple, 0 =dual therapy)
- Y_{ij} is the log(cd4) for individual i
- $Z_i * b_i$ is the random effects part of the model

Fit four different models for the covariance in the data, assuming the following random effects:

a) a random intercept
b) a random slope for time t
c) a random intercept and random slopes for time t, and t2
d) random intercept and random slopes for time t, t2, and age

Answer the following questions:

1) Can you assume a random slope for gender? Explain.
2) Plot the observed trajectories of four randomly selected individuals; two from the dual and two from the triple therapy. Compare these observed trajectories with the respective model predictions.
3) Which of the four models best fits the data? Explain.

Solution

The following models were fit in STATA:

Model (a):

```
xi: mixed logcd4 i.tx*time i.tx*time2 i.tx*age i.tx*i.gender || id: , covariance(unstr)
reml nolog
```

Model (b):

```
xi: mixed logcd4 i.tx*time i.tx*time2 i.tx*age i.tx*i.gender || id: time,
noconst covariance(unstr) reml nolog
```

Model (c):

```
xi: mixed logcd4 i.tx*time i.tx*time2 i.tx*age i.tx*i.gender || id: time time2,
covariance(unstr) reml nolog
```

Model (d):

```
xi: mixed logcd4 i.tx*time i.tx*time2 i.tx*age i.tx*i.gender || id: time time2
age, covariance(unstr) reml nolog
```

1) A random slope could not be assumed for gender as it is not a time-varying covariate. Under the assumption that all subjects preserve their gender throughout the course of the study, it would not make sense to look for variances or subject-specific deviations in the outcome for "non-varying" predictors. Since time-invariant covariates only pertain to between- and not within-subject differences, their inclusion in the the Z_i matrix of covariates would not impact a subject-specific rate of change in response over time (slope), but perhaps rather the intercept term.

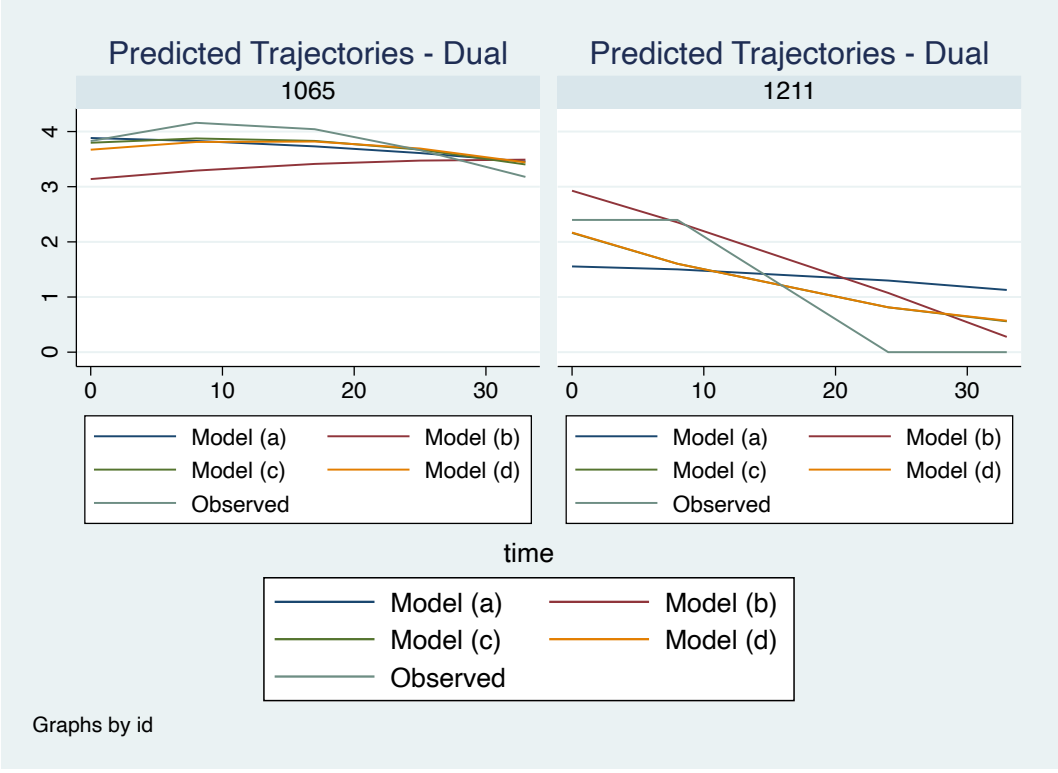
2) To select two random individuals from each group, the following lines were executed in STATA:

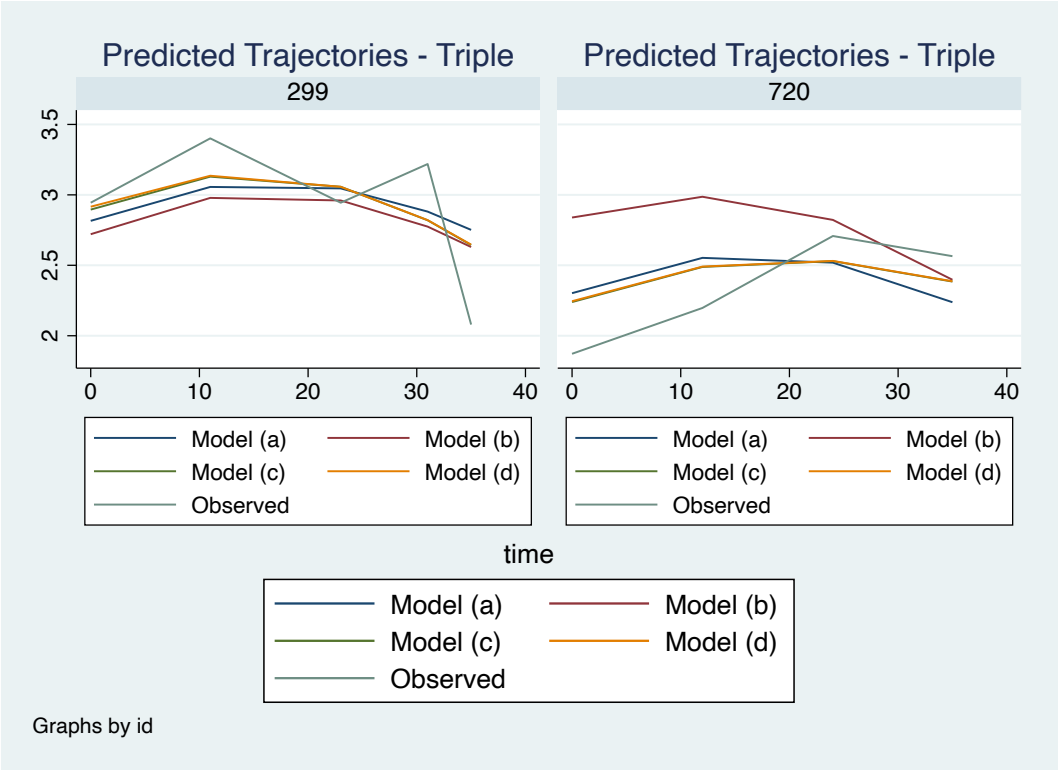
```
generate  dt = (tx == "dual") generate  tt = (tx == "triple")

generate random = runiform()

sort dt random generate sample_d = dt & (_N - _n) < 2

sort tt random generate sample_t = tt & (_N - _n) < 2

list if sample_d == 1 list if sample_t == 2
```

The individuals from the dual treatment group sampled in this way had ID's of 1065 and 1211, while those in the triple treatment group has ID's of 299 and 720. Moreover, to store all predicted values for which to subset (based on random individuals selected) and graph later on, the following commands were run after fitting each model:

```
predict pred_x, fitted
```

Using "`xtline`" to graph each "`pred_x`" value for each treatment group along with the observed ("true") values, we obtain the four graphs below. The individual predicted trajectories demonstrate very similar predictions, with the most extreme values being obtained from Model (b) (which may be due to the fact that it has a constant slope and places more weight on randomized slopes). Moreover, we see that the least extreme and more generalized model is Model (a). This makes sense as it is a marginal model that predicts population-based means of the primary outome over time. When compared to the observed trajectories, we notice that the closest in approximating the observed data is either Model (c) or Model (d), both of which produce nearly identical predictions.


Predicted Trajectories - Dual
1065
Predicted Trajectories - Dual
1211
0
1
2
3
4
0
10
20
30
Model (a)
Model (b)
Model (c)
Model (d)
Observed
time
Graphs by id


Predicted Trajectories - Triple
299
Predicted Trajectories - Triple
720
2
2.5
3
3.5
0
10
20
30
40
Model (a)
Model (b)
Model (c)
Model (d)
Observed
time
Graphs by id

3) Based the plots above, it is likely that the best fitting model is either Model (c) or Model (d). The fact that these models produce such similar results is most likely due to "age" having little (if any) influence on subject-specific variations over time. To decide between the two models, we store estimates using "`estimate store model_x`" after fitting each model and compare their BIC values produced after executing "`estimates stats model_a model_b model_c model_d`" in STATA:

Model	N	ll(null)	ll(model)	df	AIC	BIC
model_a	5,036	.	-6092.327	12	12208.65	12286.95
model_b	5,036	.	-6893.403	12	13810.81	13889.10
model_c	5,036	.	-5973.919	17	11981.84	12092.75
model_d	5,036	.	-5962.123	10	11944.25	12009.49

Despite Models (c) and (d) having very similar BIC values, the fact that Model (c)'s is lower indicates that it is the better fitting model. Specofically, Model (c) takes the following form:

Model (c):
`xi: mixed logcd4 i.tx*time i.tx*time2 i.tx*age i.tx*i.gender || id: time time2, covariance(unstr) reml nolog`

$$E[Y_{ij}|b_i] = \beta_0 + \beta_1 \cdot t_i + \beta_2 \cdot t_i^2 + \beta_3 \cdot \text{trt} + \beta_4 \cdot \text{age} + \beta_5 \cdot \text{gender}$$
$$+\beta_6 \cdot (t_i * \text{trt}) + \beta_7 \cdot (t_i^2 * \text{trt})$$
$$+\beta_8 \cdot (\text{age} * \text{trt}) + \beta_9 \cdot (\text{gender} * \text{trt})$$
$$+b_{i0} + b_{i1} \cdot t_i + b_{i2} \cdot t_i^2$$

$$E[Y_{ij}|b_i] = 2.55892 - 0.0042885 \cdot t_i - 0.0002801 \cdot t_i^2 + 0.3198502 \cdot \text{trt} + 0.0101967 \cdot \text{age} + 0.0182811 \cdot \text{gender}$$
$$+0.0372233 \cdot (t_i * \text{trt}) - 0.0007155 \cdot (t_i^2 * \text{trt})$$
$$-0.0009088 \cdot (\text{age} * \text{trt}) - 0.4143223 \cdot (\text{gender} * \text{trt})$$
$$+b_{i0} + b_{i1} \cdot t_i + b_{i2} \cdot t_i^2$$

Where:

- the reference group for treatment is "dual therapy" (trt = 0)
- the reference group for gender is 0
- $Var(b_{i0}) = 0.5755263$
- $Var(b_{i1}) = 0.0017772$
- $Var(b_{i2}) = 9.09e-07$

Question 7:

Based on the best model from the previous question what is your overall conclusion about the effect of the two treatments on the change of the log(CD4) over time?

Solution

The "best" fitting model identified in Question 7 not only indicates that the "age" covariate has little to no effect on subject-specific changes of log(CD4) over time, but that a mixed-effects model is far more representative of the patterns and overarching complexity present in the data than is a marginal model. Specifically, Model (c)'s coefficients show that adjusting for age and gender, the change in log(CD4) over time is negative for the dual therapy treatment group, while generally positive for the triple therapy treatment group. For this reason, it is likely that the triple treatment is the more effective of the two in increasing log(CD4).