STAT 8320 Spring 2015 Assignment 4

March 18, 2015

▶ **1. Solution.** (a). Define

$$
\begin{aligned}
\boldsymbol{Y}_i &= (Y_i1, Y_i2)' \\
\boldsymbol{\beta} &= (\beta_0, \beta_0)' \\
\boldsymbol{b}_i &= (b_{0i}, b_{1i})' \\
\boldsymbol{W}_i &= \begin{pmatrix} 1 & W_{i1} \\ 1 & W_{i2} \end{pmatrix} \\
\boldsymbol{\epsilon}_i &= (\epsilon_{i1}, \epsilon_{i2})' \\
\boldsymbol{D} &= \begin{pmatrix} 4 & 1 \\ 1 & 2 \end{pmatrix} \\
\boldsymbol{\Sigma} &= \begin{pmatrix} \sigma^2 & \\ & \sigma^2 \end{pmatrix}
\end{aligned}
$$

Then the model can be written as

$$\boldsymbol{Y}_i = \boldsymbol{\beta} + \boldsymbol{W}_i \boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

The marginal variance/covariance matrix of $\boldsymbol{Y}_i$ is

$$
\begin{aligned}
Var(\boldsymbol{Y}_i) &= Var(\boldsymbol{\beta} + \boldsymbol{W}_i \boldsymbol{b}_i + \boldsymbol{\epsilon}_i) = Var(\boldsymbol{W}_i \boldsymbol{b}_i) + Var(\boldsymbol{\epsilon}_i) \\
&= \boldsymbol{W}_i Var(\boldsymbol{b}_i) \boldsymbol{W}_i' + \boldsymbol{\Sigma} \\
&= \boldsymbol{W}_i \boldsymbol{D} \boldsymbol{W}_i' + \boldsymbol{\Sigma} \\
&= \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 4 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix} + \begin{pmatrix} 2 & \\ & 2 \end{pmatrix} \\
&= \begin{pmatrix} 10 & 11 \\ 11 & 18 \end{pmatrix}
\end{aligned}
$$

(b) The conditional variane/covariance matrix of $\boldsymbol{Y}_i$ is

$$Var(\boldsymbol{Y}_i|\boldsymbol{b}_i) = Var(\boldsymbol{\epsilon}_i) = \boldsymbol{\Sigma} = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$$

(c). The hypotheses are

$$H_0 : cov(b_{0i}, b_{1i}) = 0 \quad \text{v.s.} \quad H_a : cov(b_{0i}, b_{1i}) \neq 0$$

Then statistic is

$$\Lambda = -2(\ell(Reduced\ Model) - \ell(Full\ Model)) = 426 - 420 = 6 \sim \chi^2(1)$$

Then the value of statistic is greater than $\chi^2_{0.95}(1) = 3.84$ with the P-value=0.014. Thus, we reject the null hypothesis, that is, we should favor the model for which the random effects parameters are dependent.

▶ **2. Solution.** (a). We have the form of model

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\epsilon}$$

where $\boldsymbol{Y} = (y_1, y_2)'$, $\boldsymbol{X} = (X_1, X_2)'$, $\boldsymbol{\beta} = \beta$, $\boldsymbol{b} = (b_1, b_2)'$, $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2)'$, and

$$\boldsymbol{Z} = \begin{pmatrix} 1 & \\ & 2 \end{pmatrix}$$

$$\boldsymbol{D} = \begin{pmatrix} \tau^2 & \frac{\phi\tau^2}{1+\phi^2} \\ \frac{\phi\tau^2}{1+\phi^2} & \tau^2 \end{pmatrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma^2 & \\ & \sigma^2 \end{pmatrix}$$

(b). The marginal variance/covariance matrix of $\boldsymbol{Y}$ is that

$$Var(\boldsymbol{Y}) = \boldsymbol{ZDZ}^T + \boldsymbol{\Sigma} = \begin{pmatrix} \tau^2 + \sigma^2 & \frac{2\phi\tau^2}{1+\phi^2} \\ \frac{2\phi\tau^2}{1+\phi^2} & 4\tau^2 + \sigma^2 \end{pmatrix}$$

Then the marginal variance of Y_2 is $4\tau^2 + \sigma^2$ and the marginal covariance between Y_1 and Y_2 is

$$cov(Y_1,\ Y_2) = \frac{2\phi\tau^2}{1+\phi^2}$$

(c). Nothing. Because in the restricted likelihood function there is no parameters other than those from variance and covariance matrix, we can only test the variances and covariances, but no the parameters of fixed and random effects based on REML.

▶ **3. Solution.** (a). The intercepts should be same among the different graphs, but the increments of different graphs should be different, and the time points of the maximum weights should be different.

(b). We have the form of model

$$\boldsymbol{Y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{Z}_i\boldsymbol{b}_i + \boldsymbol{e}_i$$

where

$$\begin{aligned}
\boldsymbol{Y}_i &= (y_{i1}, \dots, y_{in_i})', \\
\boldsymbol{X}_i &= \begin{pmatrix} 1 & t_{i1} & t_{i1}^2 \\ \vdots & \vdots & \vdots \\ 1 & t_{in_i} & t_{in_i}^2 \end{pmatrix}, \\
\boldsymbol{Z} &= (t_{i1}, \dots, t_{in_i})', \\
\boldsymbol{\beta} &= (\beta_0, \beta_1, \beta_2)', \\
\boldsymbol{b}_i &= b_{1i}, \\
\boldsymbol{e} &= (e_1, \dots, e_{in_i})', \\
var(\boldsymbol{e}_i) = \boldsymbol{\Sigma} &= \begin{pmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{pmatrix} = \sigma^2 \boldsymbol{I}_{n_i \times n_i}, \\
var(\boldsymbol{b}_i) = \boldsymbol{D} &= \begin{pmatrix} \sigma_b^2 & & \\ & \ddots & \\ & & \sigma_b^2 \end{pmatrix} = \sigma_b^2 \boldsymbol{I}_{n_i \times n_i},
\end{aligned}$$

(c). The marginal variance/covariance matrix of $\boldsymbol{Y}$ is that

$$\begin{aligned}
Var(\boldsymbol{Y}_i) &= \boldsymbol{Z}_i \boldsymbol{D} \boldsymbol{Z}_i^T + \boldsymbol{\Sigma} \\
&= \begin{pmatrix} t_{i1} \\ \vdots \\ t_{in_i} \end{pmatrix} \begin{pmatrix} \sigma_b^2 & & \\ & \ddots & \\ & & \sigma_b^2 \end{pmatrix} (t_{i1}, \ldots, t_{in_i}) + \begin{pmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{pmatrix} \\
&= \begin{pmatrix} t_{i1}^2 \sigma_b^2 + \sigma^2 & t_{i1} t_{i2} \sigma_b^2 & \cdots & t_{i1} t_{in_i} \sigma_b^2 \\ t_{i2} t_{i1} \sigma_b^2 & t_{i2}^2 \sigma_b^2 + \sigma^2 & \cdots & t_{i2} t_{in_i} \sigma_b^2 \\ \vdots & \vdots & \ddots & \vdots \\ t_{in_i} t_{i1} \sigma_b^2 & t_{in_i} t_{i2} \sigma_b^2 & \cdots & t_{in_i}^2 \sigma_b^2 + \sigma^2 \end{pmatrix}_{n_i \times n_i}
\end{aligned}$$

(d). Because the marginal covariance $\boldsymbol{Y}$ is

$$cov(Y_{ij}, Y_{ik}) = t_{ij} t_{ik} \sigma_b^2$$

then the correlation of Y_i is that

$$\begin{aligned}
corr(Y_{ij}, Y_{ik}) &= \frac{cov(Y_{ij}, Y_{ik})}{\sqrt{var(Y_{ij})}\sqrt{var(Y_{ik})}} \\
&= \frac{t_{ij} t_{ik} \sigma_b^2}{\sqrt{t_{ij}^2 \sigma_b^2 + \sigma^2}\sqrt{t_{ik}^2 \sigma_b^2 + \sigma^2}} \\
&= \frac{jk}{\sqrt{j^2+1}\sqrt{k^2+1}} \\
&= \frac{1}{\sqrt{1/j^2+1}\sqrt{1/k^2+1}}
\end{aligned}$$

The correlations will increase with the increase of time, j and k, but no trend just with temporal separation. This is not so realistic. In common sense, we usually may think that the correlations may be smaller with large temporal separation than the correlations with small small temporal separation, because status of one time point is more likely to affect or to be affected by the status of the near time point. The reason causing this unrealistic result may be we simply assume the conditional independence while the data may not have this property.

(e). There are two advantages of the marginal covariance derived hierarchically. First, compared to the unstructured covariance structure, the hierarchical marginal covariance have less unknown parameters to estimate, so it can reduce the computation, and avoid suffering overfitting problem. Secondly, it easily to understand and interpret the variance components, we can know that which parts of variation come from random effect and which parts come from the violation of conditional independence.

▶ **4. Solution.** (a) We have the split-plot design model

$$Y_{ijk} = \mu + \rho_i + \alpha_j + e_{ij} + \beta_k + (\alpha\beta)_{jk} + \epsilon_{ijk}$$

where ρ is plot effect, α is pasture effect and β is mineral effect. The random effects are ρ, e and ϵ. From the ANOVA table, we have that

$$\sigma^2_{\text{plot}} = 12.74, \quad \sigma^2_e = 1.05, \quad \sigma^2_\epsilon = 2.25$$

In addition, the test for the significance for interaction of the pasture and mineral effects yields a P-value of 0.4981, the Factor pasture effect yields a P-value of 0.0377, and the Factor mineral effect yields a P-value of 0.0932. So only pasture effect are significant at $\alpha = 0.05$.

(b). Because the data may come from different distribution, so the degree of freedom of the variance of random components may need some modification, like Satterthwaite method. The Kenwardroger method give us a more conservative distribution about t-test or F-test than Satterthwaite when sample size is not large enough, making the assumption seem more appropriate. For example, we assume that

(c). From the output "Differences of Least Squares Means", we can see that the only significance difference is the difference between pasture 1 and pasture 4 based on the Tukey-Kramer adjustment.

▶ **5. Solution.** (a). From the two graphs, we can see that there are a lot of distinct line, which indicate the difference of different subject, the random effects. The mean intercepts of the lines from different groups almost same, and they should be because the experimental subjects should be randomly assigned to one group and they all come from a same population. Furthermore, the trends of different groups are different show that there may be some interaction between group effect and visits. Finally, the score trends with respect to time seems to be linear, and two lack of fit tests show that it is reasonable assumption with P-values are 0.9127 and 0.9131.

To sum up, we should have 3 assumptions from the graphs: i) same intercept; 2) different slope; 3) linearity.

(b). Firstly, we will define some notations. We denote i as the index of group with $i = 0, 1$; j as the index of visit with $j = 1, 2, 3, 4, 5$ and k as the index of subject with $k = 1, 2, \cdots, 47$. $\boldsymbol{Y}_k = (y_{i1k}, y_{i2k}, y_{i3k}, y_{i4k}, y_{i5k})'$ denotes the repeated responses of a single subject. X_1 is a indictor variable of placebo group, where $X_1 = 1$ when the subject in placebo group. X_2 is a indictor variable of lecithin group, where $X_2 = 1$ when the subject in lecithin group. t is treated as a continuous variable of visit. $X_1 t$ and $X_2 t$ denote the interaction of group and visit, which will show the different slopes of different groups. β_0 is the identical intercept, β_1 and β_2 are the coefficients of $X_1 t$ and $X_2 t$. b_{0k} is the random component of the intercept, where $b_{0k} \sim \text{N}(0, \sigma_b^2)$. ϵ_{ijk} is error term, and $\epsilon_{ijk} \sim \text{N}(0, \sigma^2)$ Now we can construct the model based on the general structure,

$$\boldsymbol{Y}_k = \boldsymbol{X_k}\boldsymbol{\beta} + \boldsymbol{Z_k}\boldsymbol{b_k} + \boldsymbol{\epsilon_k}$$

where the $\boldsymbol{X\beta}$ are fixed effects, and $\boldsymbol{Zb}$ are random effects, and

- $\boldsymbol{Y}_k = (y_{i1k}, y_{i2k}, y_{i3k}, y_{i4k}, y_{i5k})'$;
- $\boldsymbol{X}_k = (\boldsymbol{1}, \boldsymbol{X_1 t}, \boldsymbol{X_2 t}) = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ X_1 & 2X_1 & 3X_1 & 4X_1 & 5X_1 \\ X_2 & 2X_2 & 3X_2 & 4X_2 & 5X_2 \end{pmatrix}'$;
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$;
- $\boldsymbol{Z}_k = \boldsymbol{1} = (1, 1, 1, 1, 1)'$;
- $\boldsymbol{b}_k = b_{0k}$.

We can also know that,

$$var(\boldsymbol{b}_k) = \boldsymbol{D} = \sigma_b^2, \qquad var(\boldsymbol{\epsilon_k}) = \boldsymbol{\Sigma} = \left\{_d \{_d \sigma^2\}_{j=1}^5 \right\}_{k=1}^{25}$$

where $\{_d \cdot\}$ denotes a diagonal matrix.

In this model, we assume that

- different groups have same intercept, that is, there exists significant;
- intercept has random component;
- $\boldsymbol{b}_k \sim \text{N}(\boldsymbol{0}, \boldsymbol{D})$, $\boldsymbol{\epsilon}_k \sim \text{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$, that is, variance-covariance structure is VC(variance component), and residual structure is normal and independent;
- the relationship between score and visit is linear.

(c). According to the model in part (2), we can get the marginal variance-covariance matrix of $\boldsymbol{Y}_k$ is

$$var(\boldsymbol{Y}_k) = \boldsymbol{ZDZ}' + \boldsymbol{\Sigma} = \sigma_b^2 \boldsymbol{ZZ}' + \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_b^2 + \sigma^2 & \sigma_b^2 & \cdots & \sigma_b^2 \\ \sigma_b^2 & \sigma_b^2 + \sigma^2 & \sigma_b^2 & \cdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_b^2 & \cdots & \cdots & \sigma_b^2 + \sigma^2 \end{pmatrix}_{5\times 5}$$

(d). To fit the model, the SAS code is as blow,

```
proc mixed data=alzheim method=ml noitprint;
class idno group;
model score=group*visit/ s outp=rdint;
random int/subject=idno;
run;
```

i. The estimates for variance components are

$$\sigma_b^2 = 20.0242, \quad \sigma^2 = 4.0691$$

ii. What we want to test is

$$H_0 : \beta_1 = \beta_2 \quad \text{v.s.} \quad H_a : \beta_1 \neq \beta_2$$

From the Type 3 test of fixed model, we have a significant P-value which is less than 0.0001. So we we reject null hypothesis, and get $\beta_1 \neq \beta_2$ which means that treatment effect is significant. And from the estimates, we have $\beta_1 = -0.6133 < 1.7514 = \beta_2$, which means the scores in the lecithin group are increasing over time, while scores in placebo are decreasing. thus, treatment helps.

iii. From the plot, firstly it shows the differences of slopes of groups, or in another word, the difference of treatment effects. Secondly, it shows that the status of patients are different at the initial of the experiment, which causes the different intercepts of the regression lines. However, from the graphs, the reactions of patients under the same treatment seem same, because different lines in same group have same slope. This is not so realistic, since we know that the difference of patient's physical or mental status may also cause the effect of treatment to be different. So the different lines should have different slopes.

(e). We can rewrite the model in part (b),

$$\boldsymbol{Y}_k = \boldsymbol{X_k \beta} + \boldsymbol{Z_k b_k} + \boldsymbol{\epsilon_k}$$

The only differences here from part (b) are

$$\boldsymbol{Z}_k = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix}'; \quad \boldsymbol{b}_k = (b_{0k}, b_{1k})'.$$

and

$$var(\boldsymbol{b}_k) = \boldsymbol{D} = \begin{pmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{pmatrix}$$

The assumptions of this model almost same, but we also need to assume that the coefficients also have random components.

(f). The SAS code is as blow,

```
proc mixed data=alzheim method=ml noitprint covtest;
class idno group;
model score=group*visit/ s outp=rdcoe;
random int visit/subject=idno type=un;
run;
```

i. The estimates for variance components are

$$var(\boldsymbol{b}) = \begin{pmatrix} 26.2524 & -1.8274 \\ -1.8274 & 0.4208 \end{pmatrix}$$
$$\sigma^2 = 3.1035$$

The P-value of d_{11} is less the 0.0001, the P-value of $d_{12} = d21 = 0.296$, the P-value of d_{22} is 0.0049 and the P-value of σ is less than 0.0001. Thus, all variance components are significant. The correlation between intercepts and slopes is

$$corr(\beta_{0k}, \beta_{1k}) = \frac{cov(\beta_{0k}, \beta_{1k})}{\sqrt{var(\beta_{0k})}\sqrt{var(\beta_{1k})}} = \frac{-1.8274}{\sqrt{26.2524}\sqrt{0.4208}} = -0.5498$$

ii. What we want to test is

$$H_0 : \beta_1 = \beta_2 \quad \text{v.s.} \quad H_a : \beta_1 \neq \beta_2$$

From the Type 3 test of fixed model, we still have a significant P-value which is less than 0.0001. So we we reject null hypothesis, and get $\beta_1 \neq \beta_2$ which means that treatment effect is significant. And from the estimates, we have $\beta_1 = -0.5173 < 1.6424 = \beta_2$, thus, the treatment helps. This is a same result from part (d), just different estimations of parameters when adjusted by randomness of subject.

iii. To compare the fitness of models, we can perform a likelihood ratio test,

$$H_0: \text{ Reduced Model in part (b)} \quad \text{v.s.} \quad H_a: \text{ Full Model in part (e)}$$

In the reduced model, we have $d_{12} = 0$ and $d_{22} = 0$. The statistic is

$$\Lambda = -2(\ell(Reduced\ Model) - \ell(Full\ Model)) = 1149.1 - 1134.8 = 14.3 \sim \chi^2(2)$$

The critical value of $\chi^2_{0.95}(2)$ is 5.99, which is less than 14.3. And the P-value is 0.0008. So we reject the null hypothesis. It is reasonable to believe that the model with random coefficient and unstructured variance-covariance structure is better fit than the random intercept model.

iv. From the plots, the different lines have different slopes compared with the parallel lines in the plots of random intercept model. It is more like the data plot, so we can think that it is a more realistic model.

(g). First, we should change the form of the model in part (e),

$$\boldsymbol{Y}_k = \boldsymbol{X_k\beta} + \boldsymbol{e_k}$$

Here, we put the random effect $\boldsymbol{Z_k b_k}$ and residual ϵ_k together to create the new residual side matrix. This time we want to control the residual matrix directly, other than the variance-covariance matrix of random effect. So here, we will define

$$\boldsymbol{e}_k \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_k^2), \quad \text{where the structure of } \boldsymbol{\Sigma}_k^2 \text{ is unknown}$$

To refit the model assuming the residual side matrix structure is the first order autoregression or unstructured covariance. we use the code below using SAS,

```
proc mixed data=alzheim noitprint contest;
class group visit;
model score=group|visit/ s outp=repmar;
repeated visit / subject=idno type=ar(1);
run;

proc mixed data=alzheim noitprint covtest;
```

```
class idno group visit;
model score=group|visit/ s outp=repun;
repeated visit / subject=idno type=un;
run;
```

For repeated measures model here, it is more like a nested experimental design, since it does not focus on the form of the relation between scores and visit, but just to compare the difference between scores under different combinations of group and visit. So SAS will give us 10 fixed parameter estimates rather than 3 (2 slopes and 1 intercept) as before. For the first order autoregression structure, we will get the estimate of correlation coefficient between two sequential visits

$$\rho = 0.8422, \quad \text{with the P-value} < 0.0001$$

The significant correlation coefficient implies that the score is related with the previous score, so the independent assumption may not be appropriate.

So the residual variance-covariance matrix is

$$\boldsymbol{\Sigma}_k = \begin{pmatrix} 20.8085 & 17.5250 & 14.7596 & 12.4306 & 10.4691 \\ 17.5250 & 20.9095 & 17.5250 & 14.7596 & 12.4306 \\ 14.7596 & 17.5250 & 20.9095 & 17.5250 & 14.7596 \\ 12.4306 & 14.7596 & 17.5250 & 20.9095 & 17.5250 \\ 10.4691 & 12.4306 & 14.7596 & 17.5250 & 20.9095 \end{pmatrix}$$

while the marginal variance-covariance matrix of iid error model is

$$\begin{aligned} \boldsymbol{V} &= \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}' + \boldsymbol{\Sigma} \\ &= \begin{pmatrix} 26.1219 & 21.6118 & 20.2052 & 18.7986 & 17.3920 \\ 21.6118 & 23.7295 & 19.6402 & 18.6544 & 17.6686 \\ 20.2052 & 19.6402 & 22.1787 & 18.5102 & 17.9452 \\ 8.7986 & 18.6544 & 18.5102 & 21.4695 & 18.2218 \\ 17.3920 & 17.6686 & 17.9452 & 18.2218 & 21.6019 \end{pmatrix} \end{aligned}$$

There is no surprising that they are different. As mentioned before, the repeated measure model is more like ANOVA, so here we should firstly test the interaction effect,

$$H_0: \text{ The interaction is insignificant} \quad \text{v.s.} \quad H_a: \text{ The interaction is significant}$$

then the interaction have s significant P-value less than 0.001. With significant interaction effect, we cannot test the main effect of group. But we can contrast them

$$H_0: \mu_{\text{group 1}} - \mu_{\text{group 2}} = 0 \quad \text{v.s.} \quad H_a: \mu_{\text{group 1}} - \mu_{\text{group 2}} \neq 0$$

The absolute difference is 3.0556 with P-value=0.0125, thus, we can still say that the treatment helps. Moreover the -2 times REML, AIC and BIC of AR(1) model is better than the model in part (e). Thus, we can believe that AR(1) model here is a better model than iid models. (Here we cannot perform a likelihood ratio test to support our conclusion, because these two models are not nested models)

Finally, we refit the repeated measure model by assuming unstructured covariance as the code above. We can see that the AIC of AR(1) model is 1148.8, while AIC of UN model is 1097.1. Thus we may choose unstructured model as a more appropriate. This also gives us a hint that the assumption of first order autoregression correlation structure may be not the real correlation of this longitude data, the correlation structure may be more complicated.