STA2201 Assignment 2

Matthew Scicluna

2016-02-23

Linear Mixed Model

We are given the following linear mixed model

$$Y_{ijk} \mid U, V, A \sim N(\mu_{ijk}, \tau^2), \quad where \quad \mu_{ijk} = X_{ijk}\beta + U_i + V_{ij1} + V_{ij2}W_{ijk} + A_{ijk}$$

$$cov(A_{ijk}, A_{lmn}) = \begin{cases} \sigma_A^2 e^{\frac{|t_{ijk} - t_{lmn}|}{\phi}} & i = l, \quad and \quad j = m, \\ 0 & i \neq l, \quad or \quad j \neq m, \end{cases}$$

$$\begin{pmatrix} V_{ij1} \\ V_{ij2} \end{pmatrix} \sim N(0, \Gamma), \quad U_i \sim N(0, \sigma_U^2)$$

Which Parameters Have Closed Form Solutions?

We wish to determine which parameters would have closed form Maximum Likelihood Estimates, and which would need to be estimated using a numerical optimizer. The parameters in the model are Γ, σ_U^2, σ_A^2, τ, ϕ, and β

We see that $Y_{ijk} \mid A, U, V$ is distributed normally, as is A, U and V. This means that Y_{ijk} is a normally distributed random variable also. This means that we can get maximum likelihood estimates for μ_{ijk} and τ. We notice that μ_{ijk} contains all paramters in the model other than τ. This term is dependant on the nonlinear term $cov(A_{ijk}, A_{lmn})$. This means that all parameters in the model except for τ are dependant on a nonlinear term and therefore must be evaluated using numeric optimization.

Some derivations

1

We are given $n \neq k$. For simplicity let $Y_{ijk} = \mu_{ijk} + \epsilon_{ijk}$ where $\epsilon_{ijk} \sim N(0, \tau^2)$ and are independant of eachother and everything else in the model. Let $\mu_{ijk} = X_{ijk}\beta + U_i + V_{ij1} + V_{ij2}W_{ijk} + A_{ijk}$ as before. Then:

$$\begin{aligned} cov(Y_{ijk}, A_{ijn}) &= cov(X_{ijk}\beta + U_i + V_{ij1} + V_{ij2}W_{ijk} + A_{ijk} + \epsilon_{ijk}, A_{ijn}) \\ &= cov(X_{ijk}\beta, A_{ijn}) + cov(U_i, A_{ijn}) + cov(V_{ij1}, A_{ijn}) + cov(V_{ij2}W_{ijk}, A_{ijn}) + cov(A_{ijk}, A_{ijn}) + cov(\epsilon_{ijk}, A_{ijn}) \\ &= cov(A_{ijk}, A_{ijn}) \\ &= \sigma_A^2 e^{\frac{|t_{ijk} - t_{lmn}|}{\phi}} \end{aligned}$$

Note that we used the fact that A_{ijn} is a constant with respect to the other random and fixed effects in our model.

2

Note that

$$\begin{pmatrix} V_{ij1} \\ V_{ij2} \end{pmatrix} \sim N_2(0, \Gamma) \quad \Rightarrow V_{ij1} \sim N(0, \gamma_{11}) \, and \, V_{ij2} \sim N(0, \gamma_{22})$$

where

$$\Gamma = \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix}$$

Since the marginal distribution of a Multivariate Normal density is Univatiate Normal.

$$\begin{aligned} cov(Y_{ijk}, V_{ij1}) &= cov(X_{ijk}\beta + U_i + V_{ij1} + V_{ij2}W_{ijk} + A_{ijk} + \epsilon_{ijk}, V_{ij1}) \\ &= cov(X_{ijk}\beta, V_{ij1}) + cov(U_i, V_{ij1}) + cov(V_{ij1}, V_{ij1}) + cov(V_{ij2}W_{ijk}, V_{ij1}) + cov(A_{ijk}, V_{ij1}) + cov(\epsilon_{ijk}, V_{ij1}) \\ &= cov(V_{ij1}, V_{ij1}) + cov(V_{ij2}W_{ijk}, V_{ij1}) \\ &= \gamma_{11} + W_{ijk}\gamma_{12} \end{aligned}$$

3

let $(j,k) \neq (m,n)$

$$\begin{aligned} cov(Y_{ijk}, Y_{imn}) &= cov(\mu_{ijk} + \epsilon_{ijk}, \mu_{imn} + \epsilon_{imn}) \\ &= cov(\mu_{ijk}, \mu_{imn}) \\ &= cov(X_{ijk}\beta + U_i + V_{ij1} + V_{ij2}W_{ijk} + A_{ijk}, X_{imn}\beta + U_i + V_{im1} + V_{im2}W_{imn} + A_{imn}) \\ &= cov(U_i + V_{ij1} + V_{ij2}W_{ijk} + A_{ijk}, U_i + V_{im1} + V_{im2}W_{imn} + A_{imn}) \\ &= cov(U_i, U_i) + cov(V_{ij1}, V_{im2}) + cov(V_{ij1}, V_{im2}W_{imn}) + cov(V_{ij2}W_{ijk}, V_{im1}) \\ &+ cov(V_{ij2}W_{ijk}, V_{im2}W_{imn}) + cov(A_{ijk}, A_{imn}) \\ &= \sigma_U^2 \end{aligned}$$

In Which the 4th line utilizes the bilinearity of the covariance operator.

4

given $n \neq k$

$$\begin{aligned} cov(Y_{ijk}, Y_{ijn}) &= cov(\mu_{ijk} + \epsilon_{ijk}, \mu_{ijn} + \epsilon_{ijn}) \\ &= cov(\mu_{ijk}, \mu_{ijn}) \\ &= cov(X_{ijk}\beta + U_i + V_{ij1} + V_{ij2}W_{ijk} + A_{ijk}, X_{ijn}\beta + U_i + V_{ij1} + V_{ij2}W_{ijn} + A_{ijn}) \\ &= cov(U_i + V_{ij1} + V_{ij2}W_{ijk} + A_{ijk}, U_i + V_{ij1} + V_{ij2}W_{ijn} + A_{ijn}) \\ &= cov(U_i, U_i) + cov(V_{ij1}, V_{im2}) + cov(V_{ij1}, V_{im2}W_{ijn}) + cov(V_{ij2}W_{ijk}, V_{ij1}) \\ &+ cov(V_{ij2}W_{ijk}, V_{im2}W_{ijn}) + cov(A_{ijk}, A_{ijn}) \\ &= \sigma_U^2 + \gamma_{11} + \gamma_{12}W_{ijn} + \gamma_{21}W_{ijk} + \gamma_{22}W_{ijn}W_{ijk} + \sigma_A^2 e^{\frac{|t_{ijk} - t_{lmn}|}{\phi}} \end{aligned}$$

5

$cov(Y_{ijk}, Y_{ljk}) = 0$ when $i \neq l$ trivially, due to random sampling assumption.

6

$$\begin{aligned} Var(Y_{ijk} \mid A, V) &= var(X_{ijk}\beta + U_i + V_{ij1} + V_{ij2}W_{ijk} + A_{ijk} + \epsilon_{ijk} \mid A, V) \\ &= var(X_{ijk}\beta \mid A, V) + Var(U_i \mid A, V) + var(V_{ij1} \mid A, V) + var(V_{ij2}W_{ijk} \mid A, V) \\ &+ var(A \mid A, V) + var(\epsilon_{ijk} \mid A, V) \\ &= Var(U_i \mid A, V) + var(\epsilon_{ijk} \mid A, V) \\ &= Var(U_i) + var(\epsilon_{ijk}) \\ &= \sigma_U^2 + \tau^2 \end{aligned}$$

7

$E(e^{Y_{ijk}})$ We note that

$$\begin{aligned} Y_{ijk} &= X_{ijk}\beta + U_i + V_{ij1} + V_{ij2}W_{ijk} + A_{ijk} + \epsilon_{ijk} \\ &= X_{ijk}\beta + N(0, \sigma_U^2) + V_{ij1} + V_{ij2}W_{ijk} + N(0, \sigma_A^2) + N(0, \tau^2) \end{aligned}$$

where

$$V_{ij1} + V_{ij2}W_{ijk} = \begin{pmatrix} 1 \\ w_{ijk} \end{pmatrix} N_2(0, \Gamma) = N(0, \gamma_{11} + 2W_{ijk}\gamma_{21} + W_{ijk}^2\gamma_{22})$$

So Y_{ijk} is a linear combination of Normal random variables with constant term $X_{ijk}\beta \Rightarrow Y_{ijk}$ has a univariate normal distribution with mean $X_{ijk}\beta$ and variance $\sigma_U^2 + \gamma_{11} + 2W_{ijk}\gamma_{21} + W_{ijk}^2\gamma_{22} + \sigma_A^2 + \tau^2$

Notice that since Y_{ijk} Normally distributed, $e^{Y_{ijk}}$ is lognormally distributed and has expected value $e^{X_{ijk}\beta + \frac{\sigma_U^2 + \gamma_{11} + 2W_{ijk}\gamma_{21} + W_{ijk}^2\gamma_{22} + \sigma_A^2 + \tau^2}{2}}$

A New Model

We are given $\widehat{\tau^2} = 2$, $\widehat{\sigma_U^2} = 0.001$, $\widehat{\theta}$=4, $\widehat{\sigma_A^2} = 0.00002$

$$\widehat{\Gamma} = \begin{pmatrix} 1 & 0.05 \\ 0.05 & 0.002 \end{pmatrix}$$

We can see that the W_{ijk}, U_i, and A_{ijk} terms do not contribute much since their means are 0 and variances are very small. Thus can be safely removed. We keep the Γ terms since they contribute about $\gamma_{11} + 2W_{ijk}\gamma_{21} + W_{ijk}^2\gamma_{22} = 1 + 0.1W_{ijk} + 0.002W_{ijk}^2$. This amount is substantial – it is a 8200 unit increase in the observed variable for an animal that weighs 2000 ounces!

The resultant reduced model is as follows:

$$Y_{ijk} \mid U, V \sim N(\mu_{ijk}, \tau^2), \quad where \quad \mu_{ijk} = X_{ijk}\beta + V_{ij1} + V_{ij2}W_{ijk}$$

where V_{ij1}, V_{ij2} are defined as before.

The Math Dataset

We analyzed the `MathAchieve` dataset from the MEMSS package. The decision to treat school as a fixed effect would depend on the aim of the study. If we were specifically interested in analyzing the differences in mathematics achievement scores between specific schools, then treating school as a random effect would make sense. Otherwise, if we were interested in just controlling for it, then treating it as a fixed effect would work and would give the weights more leverage in the overall design since random effects suffer from parameter shrinkage.

Table 1: Estimation of fixed effects in linear mixed model of math achievement dataset

	Value	Std.Error	t-value	p-value
(Intercept)	12.88	0.19	66.59	0
MinorityYes	-2.96	0.21	-14.39	0
SexMale	1.23	0.16	7.56	0
SES	2.09	0.11	19.77	0

We carried out an analysis treating school as a random effect. The results of the fixed effects are summarized in table 1. To check if differences between schools are greater then within-school, we check the random effects. We found that the variance between schools is smaller then the variance within. This is since the intercept of the random effect was about $\frac{1}{3}$ of the slope. Therefore, the differences between schools are not greater then what can be explained by within-school variation.

The Cystic Fibrosis Data

Introduction

We were tasked with helping a medical scientist interpret their findings regarding the effect of the F508 gene on the lungs of individuals with cystic fibrosis. The researchers were primarily interested in analyzing whether the rate at which lung function declines for these patients depends on the F508 gene and whether this effect differs between males and females.

Model Selection

We consider the following 3 models for this analysis:

Original Model

$$Y_{ij} \mid U \sim N(\mu_{ij}, \tau^2), \quad where \quad \mu_{ij} = X_{ij}\beta + U_i$$
$$U_i \sim N(0, \sigma^2)$$

where:

- Y_{ij} is the lung capacity of individual i at measurement j
- $X_{ij}\beta$ contains the intercept and whether the individual has Pseudo Aeruginosa. It also contains the individuals gender, gene type and age as well as all interactions between them.

- U_i is each persons lung capacity at the beginning of the study.

This model makes a very strong assumption, which is that each persons lung capacity changes in the same way over time, although it makes no assumptions in the structure of the change over time.

Random Slope Model

$$Y_{ij} \mid U \sim N(\mu_{ij}, \tau^2), \quad where \quad \mu_{ij} = X_{ij}\beta + U_{i1} + U_{i2}A_{ij}$$
$$\begin{pmatrix} U_{i1} \\ U_{i2} \end{pmatrix} \sim N(0, \Gamma)$$

where:

- Y_{ij} is the lung capacity of individual i at measurement j
- $X_{ij}\beta$ contains the intercept and whether the individual has Pseudo Aeruginosa. It also contains the individuals gender, gene type and the interaction between them.
- U_{i1} is each persons lung capacity at the beginning of the study.
- U_{i2} is the change in each persons lung capacity as they age.
- A_{ij} is the age of individual i at measurement j

This model is more general then the previous one since it allows for different individuals to have different lung capacity changes over time. It does however assumes that the change in lung trajectory over time follows a linear trend. Note that this means that observations of the same individual are correlated, but this correlation is the same regardless of the amount of time that passed.

Serial Correlation Model

$$Y_{ij} \mid U \sim N(\mu_{ij}, \tau^2), \quad where \quad \mu_{ij} = X_{ij}\beta + U_i + A_{ij}$$
$$U_i \sim N(0, \sigma_U^2)$$
$$A_{ij} \sim N(0, \Sigma)$$

where Σ is defined as follows:

$$cov(A_{ij}, A_{lm}) = \begin{cases} \sigma_A^2 e^{\frac{|t_{ij} - t_{lm}|}{\phi}} & i = l, \\ 0 & i \neq l \end{cases}$$

where:

- Y_{ij} is the lung capacity of individual i at measurement j
- $X_{ij}\beta$ contains the intercept and whether the individual has Pseudo Aeruginosa. It also contains the individuals gender, gene type and age as well as all interactions between them.
- U_i is each persons lung capacity at the beginning of the study.
- A_{ij} are random variables which model the correlations between ages and assumes that this correlation has an exponential correlation function structure.

This model makes the same assumptions as the previous model, but assumes that the change in lung trajectory over time is governed by an exponential correlation function rather than a linear one. This means that as measurements become exponentially less correlated the further they are taken across time. This assumption less strong and thus more reasonable than the random slopes model because of this.

Results

Original Model

Table 2: Likelihood Ratio Test of Significance of Gene-Age Interaction

#Df	LogLik	Df	Chisq	Pr(>Chisq)
15	-6237.03	NA	NA	NA
11	-6242.93	-4	11.8	0.02

To test the reseachers first hypothesis we used a likelihood ratio test with a model not including any gene-age interactions. We found that the F508 gene does affect the rate of lung decay ($\chi^2_4 = 11.8$, p = 0.02). Unfortunately, the model gives limited insight into how it does, since it only considers the marginal effects of the degradation over time over all subjects in the study. Also note that according to the model, gender significantly altered the effect of heterozygosity on lung decay (p = 0.04). It also effected the effect of not having the gene on lung decay (p = 0.02). This means that, under this model, there was evidence for the researchers first and second hypothesis.

Random Slope Model

Table 3: Output from random slope model

	MLE	Std.Error	t-value	p-value
(Intercept)	75.78	3.77	20.13	0.00
GENDERfemale	-5.64	5.11	-1.10	0.27
F508heterozygous	0.21	5.14	0.04	0.97
F508none	-0.69	7.68	-0.09	0.93
PSEUDOAyes	-4.02	1.11	-3.63	0.00
GENDERfemale:F508heterozygous	-2.25	7.29	-0.31	0.76
GENDERfemale:F508none	11.53	11.42	1.01	0.31
ageC	11.49	NA	NA	NA
ID	11.49	NA	NA	NA
τ	5.80	NA	NA	NA

Table 4: Likelihood Ratio Test of Significance of Gene-Age Interaction for random slopes model

#Df	LogLik	Df	Chisq	Pr(>Chisq)
10	-6321.24	NA	NA	NA
6	-6333.40	-4	24.32	0

The variance of the random intercepts were twice as large as the residuals, (11.49 vs 5.8), indicating that the variation in lung capacity over time within individuals was significantly different than between individuals. This directly contradicts the assumptions of our first model and so we do not use the first model. We see via a likelihood ratio test that the F508 gene is an important predictor and cannot be removed (p < 0.001). Notice that under this model the interaction between gender and gene type's effect on lung decay is insignificant. For heterozygotes we have (p = 0.76) and for total absence of the gene we have (p = 0.31). This means that

both genders have lung decay affected by gene type in the same way, and so there is insufficient evidence of the researchers' second hypothesis.

Serial Correlation Model

Table 5: Fixed effects from serial correlation model

	Value	Std.Error	t-value	p-value
(Intercept)	66.07	3.84	17.20	0.00
GENDERfemale	-0.39	5.06	-0.08	0.94
F508heterozygous	7.93	5.04	1.57	0.12
F508none	8.87	7.29	1.22	0.23
ageC	-2.04	0.36	-5.64	0.00
PSEUDOAyes	-2.99	1.00	-3.00	0.00
GENDERfemale:F508heterozygous	-8.33	7.02	-1.19	0.24
GENDERfemale:F508none	0.69	11.05	0.06	0.95
GENDERfemale:ageC	0.34	0.51	0.67	0.50
F508heterozygous:ageC	0.98	0.49	2.01	0.04
F508none:ageC	1.44	0.75	1.92	0.05
GENDERfemale:F508heterozygous:ageC	-1.04	0.70	-1.49	0.14
GENDERfemale:F508none:ageC	-1.41	1.15	-1.22	0.22

Table 6: Likelihood Ratio Test of Significance of Gene-Age Interaction for serial correlation model

#Df	LogLik	Df	Chisq	Pr(>Chisq)
17	-6158.60	NA	NA	NA
13	-6163.18	-4	9.15	0.06

This model assumes a structure to the temporal correlation of the data. Upon doing a likelihood ratio test removing the gene-age interaction, we see that gene type does not affect lung capacity either ($\chi^2_4 = 9.15$, p = 0.06). This means that we cannot confirm the researchers first hypothesis with this model. Note that under this model, individuals who are heterozygous or do not have the F508 gene do not have significantly different gender - lung decay interactions (p = 0.14 and p = 0.22 respectively). So we see that this model confirms neither of the researchers hypotheses.

Conclusion

The serial correlation model has the most reasonable assumptions, since it is the only model which captures that observations become less correlated the more time has passed between them. While neither of the new models have evidence that gender effects the gene effect on lung capacity over time, the first two models have evidence that the gene type has an effect. The serial correlation model has insufficient evidence to conclude either of the research hypotheses. Since it is most appropriate model, we must conclude that there is not sufficient evidence to conclude the reseach hypotheses, and this contradicts the the medical scientist's initial conclusions.

Moss in Galicia

The full model is as follows:

$$Y_i \sim N(\lambda(s_i), \tau^2)$$

$$\lambda(s) = \mu + X(s)\beta + U(s)$$

Where:

- Y_i is the lead levels taken from moss growing in or near the province of Galicia.
- $X(s)\beta$ is the logarithm of the population density, dominant soil types and average annual rainfall.
- $U(s)$ is the spacial random effect.

The reduced model takes the same form as the full model, except that the $X(s)\beta$ term only has the logarithm of the population density. It does not contain the enviornmental variables.

The third model has the same form as the other models but does not have the $X(s)\beta$ term.

The environmental variables seem do not significantly influence the lead content of moss in Galicia? We can see this from the likelihood ratio test between the first two models ($\chi^2_6 = 7.99$), p=0.092, which means that the reduced model (without the enviornmental covariates) is not significantly different from the full model.

It seems that humans do influence the lead content in the moss. This is since the likelihood ratio test between the reduced model with the logarithm of the population is significantly different than the model without any covariates in it ($\chi^2_5 = 7.7$, p = 0.005).

For the remainder of the analysis, we use the reduced model with the logarithm of the population density but no other covariates. This is because it contains all and only all the significant covariates. We find that for every unit increase of logarithmic population density, the lead content in moss increases by 0.06 units. Areas 31 units away are uncorrelated, according to the model (since $\frac{range}{1000} = 30.91$).