Assignment 2

Steve Wilkins-Reeves

2018-02-28

Question 1

1. As in the lecture slides on profile likelihood, we require numerical methods to estimate σ_A^2, ϕ, Γ, & σ_U^2. Once these have been estimated, a closed form solution exists for the ML estimates of $\beta, \&\ \tau^2$.

2.

$$\text{since } Y_{ijk} = X_{ijk}\beta + U_i + V_{ij1} + V_{ij2}W_{ijk} + A_{ijk} + e_{ijk}$$

$$\text{Where } e_{ijk} \overset{IID}{\sim} N(0, \tau^2)$$

$$cov(Y_{ijk}, A_{ijn}) = cov(A_{ijk}, A_{ijn}) = \sigma_A^2 exp(-\frac{|t_{ijk} - t_{ijn}|}{\phi})$$

$$cov(Y_{ijk}, V_{ij1}) = cov(V_{ij1} + V_{ij2}W_{ijk}, V_{ij1}) = \Gamma_{12} + W_{ijk}\Gamma_{22}$$

$$\begin{aligned}
cov(Y_{ijk}, Y_{imn}) \ & (m,j) \neq (n,k),\ \text{If } m = j, n \neq k \\
&= cov(U_i, U_i) + cov(V_{ij1} + V_{ij2}W_{ijk}, V_{im1} + V_{im2}W_{imn}) + cov(A_{ijk}, A_{imn}) \\
&= \sigma_U^2 + \Gamma_{11} + (W_{ijk} + W_{imn})\Gamma_{12} + W_{ijk}W_{imn}\Gamma_{22} + \sigma_A^2 exp(-\frac{|t_{ijk} - t_{imn}|}{\phi})
\end{aligned}$$

If $m \neq j$

$$cov(Y_{ijk}, Y_{imn}) = \sigma_U^2$$

$$cov(Y_{ijk}, Y_{ijn}) = \sigma_U^2 + \Gamma_{11} + (W_{ijk} + W_{imn})\Gamma_{12} + W_{ijk}W_{imn}\Gamma_{22} + \sigma_A^2 exp(-\frac{|t_{ijk} - t_{imn}|}{\phi})$$

Which is a subset of the last question

$$cov(Y_{ijk}, Y_{ljk}) = 0$$

If $i \neq l$ due to the independence of individuals

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

$E(exp(Y_{ijk}))$ = This is a lognormal distribution with parameters $\mu_{ijk}$and τ^2

We must consider the additivity of the variances of the U_i, V_{ij1}, V_{ij2}, & A_{ijk}

$$= exp(X_{ijk}\beta + \frac{\tau^2 + \sigma_U^2 + \sigma_A^2 + \Gamma_{11} + 2W_{ijk}\Gamma_{12} + W_{ijk}^2\Gamma_{22}}{2})$$

3. If $W_{ijk} \sim 1000$ - 3000 then σ_U^2 and σ_A^2 become insignificant. A more suitable model is therefore:

$$
\begin{aligned}
Y_{ijk}|V &\sim N(\mu_{ijk}, \tau^2) \\
\mu_{ijk} &= X_{ijk}\beta + V_{ij1} + V_{ij2}W_{ijk} \\
\text{where } (V_{ij1}, V_{ij2})' &\sim MVN(0, \Gamma)
\end{aligned}
$$

Question 2 (Math)

The decision on whether or not to treat school as a random or fixed effect may depend on whether or not we wish to generalize our results to all schools, or only the schools in this data set. Using random effects allows for generalization to all schools.

In this example we are looking at the effect of school over many samples of a population (160 schools). This would not be an exhaustive list of all schools and we can assume this effect is not correlated with the Sex, Minority status, or SES of a student.

Table 1: Coefficients Modelling Math Scores With School As A Random Effect

	MLE	Std.Error	DF	t-value	p-value
(Intercept)	12.885	0.193	7022	66.794	0
MinorityYes	-2.962	0.206	7022	-14.400	0
SexMale	1.230	0.163	7022	7.562	0
SES	2.091	0.106	7022	19.782	0
σ	1.907	NA	NA	NA	NA
τ	5.991	NA	NA	NA	NA

Modelling school as a random effect we find that the minority status, sex, and SES were all significant predictors of the mathematics scores of the individuals. It was found that male students were predicted to have scored on average 1.230 points higher than female students, minority students were found to have scored on average 2.962 lower than non-minority students, and for every additional point on the SES students were found to have scored on average 2.091 points higher. The within school standard deviation τ was estimated to be 5.991, while the standard deviation between schools σ was estimated to be 1.907.

To test if the differences within schools are greater than can be explained by within school variation, we conduct a likelihood ratio test for the hypothesis of the variance for the random effects of school is 0. ($H_0 : \sigma_U^2 = 0$)

Table 2: Likelihood Ratio Test for Random Effects Inclusion

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
With Random Effects	6	-23193.42	NA	NA	NA
Without Random Effects	5	-23374.79	-1	362.742	7.12e-81

We see that even when accounting for the fact that we use this one-sided test (ie. multiplying the two-sided p-value by 2), the result is still significant. The adjusted p-value was found to be 1.42×10^{-80}. Thus the variance between schools cannot solely be accounted for by the variance between individuals.

Cystic Fibrosis Data

Cystic Fibrosis is an autosomal recessive genetic disease which affects 1 in 3600 children born in Canada. It is the most common fatal genetic disease affecting Canadian children. Though the disease is known to be caused by a single gene, there may be other genetic factors affecting its severity. This analysis seeks to discover if the F508 gene affects the lung function of individuals with Cystic Fibrosis. Lung function in this study will be measured as FEV1, the volume of air which an individual can exhale in 1 second. This analysis will account for age and gender effects, as those can affect the FEV1, as well as their interaction with the F508 gene. Conclusions will be drawn from looking at three different models: A random intercept model, a random slope and a serial correlation model. The first model treats the patient identifier as a random effect, meaning some individuals will naturally have a higher or lower FEV1 over all time, and these follow a normal distribution. The second model expands upon this by introducing a random slope, meaning some individuals may have a slower or fast lung decline, which is also related to the random intercept by a multivariate normal distribution, where the age related lung decline is constant at the individual level. The last model includes a random intercept but also includes a serial correlation between ages. The correlation between ages in lung function is more closely related to their FEV1 at nearby ages.

The models in question can be expressed as following:

$$Y_{ij} \sim N(\mu_{ij}, \tau^2)$$

Model 1 (Random Intercept):

$$\begin{aligned}\mu_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \beta_4 X_{ij1} X_{ij2} + \beta_5 X_{ij1} X_{ij3} +\\ \beta_6 X_{ij2} X_{ij3} + \beta_7 X_{ij1} X_{ij2} X_{ij3} + \beta_8 X_{ij4} + U_{i1}\end{aligned}$$

Model 2 (Random Slope):

$$\begin{aligned}\mu_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \beta_4 X_{ij1} X_{ij2} + \beta_5 X_{ij1} X_{ij3} +\\ \beta_6 X_{ij2} X_{ij3} + \beta_7 X_{ij1} X_{ij2} X_{ij3} + \beta_8 X_{ij4} + U_{i1} + U_{i2} X_{ij3}\end{aligned}$$

Model 3 (Serial Correlation):

$$\begin{aligned}\mu_{ij} = \beta_0 + \beta_1 X_{ij1} + \beta_2 X_{ij2} + \beta_3 X_{ij3} + \beta_4 X_{ij1} X_{ij2} + \beta_5 X_{ij1} X_{ij3} +\\ \beta_6 X_{ij2} X_{ij3} + \beta_7 X_{ij1} X_{ij2} X_{ij3} + \beta_8 X_{ij4} + U_{i1} + A_{jk}\end{aligned}$$

Where:

$$\begin{aligned} Y_{ij} &= \text{FEV1 Lung Function}\\ X_{ij1} &= \text{Gender}\\ X_{ij2} &= \text{F508 Genotype}\\ X_{ij3} &= \text{Age (From 18 years Old)}\\ X_{ij4} &= \text{Infection with Pseudomonas Aeruginosa}\\ (U_{i1}\ , U_{i2})' &\overset{IID}{\sim} MVN(0, \Gamma)\\ U_{i1} &\overset{IID}{\sim} N(0, \sigma_U^2)\\ \text{For the random intercept model } cov(A_{ij}, A_{nk}) &= \delta_{in}\sigma_A^2 exp(-\frac{|X_{ij3} - X_{nk3}|}{\phi})\\ \delta_{in} &\text{ is the Kroenecker delta function}\end{aligned}$$

The following output of parameter estimates for the second and third model were obtained.

Table 3: Random Slope Model

	Value	2.5%CI	97.5%CI	p.value
(Intercept)	68.563	60.706	76.421	0.000
GENDERfemale	-5.409	-15.946	5.127	0.316
F508heterozygous	3.651	-6.981	14.283	0.502
F508none	8.939	-6.585	24.463	0.260
ageC	-1.687	-2.419	-0.955	0.000
PSEUDOAyes	-2.841	-4.852	-0.830	0.006
GENDERfemale:F508heterozygous	-2.405	-17.375	12.565	0.753
GENDERfemale:F508none	2.838	-20.351	26.028	0.811
GENDERfemale:ageC	-0.188	-1.210	0.834	0.718
F508heterozygous:ageC	0.599	-0.415	1.612	0.247
F508none:ageC	1.349	-0.184	2.881	0.085
GENDERfemale:F508heterozygous:ageC	-0.618	-2.067	0.830	0.403
GENDERfemale:F508none:ageC	-1.001	-3.305	1.304	0.395

Table 4: Serial Correlation Model

	Value	2.5%CI	97.5%CI	p.value
(Intercept)	66.066	58.635	73.497	0.000
GENDERfemale	-0.381	-10.167	9.404	0.939
F508heterozygous	7.938	-1.804	17.679	0.112
F508none	8.883	-5.193	22.960	0.218
ageC	-2.049	-2.753	-1.345	0.000
PSEUDOAyes	-3.005	-4.965	-1.046	0.003
GENDERfemale:F508heterozygous	-8.321	-21.884	5.241	0.231
GENDERfemale:F508none	0.639	-20.709	21.988	0.953
GENDERfemale:ageC	0.347	-0.634	1.328	0.489
F508heterozygous:ageC	0.981	0.040	1.923	0.041
F508none:ageC	1.457	0.001	2.914	0.050
GENDERfemale:F508heterozygous:ageC	-1.046	-2.400	0.307	0.130
GENDERfemale:F508none:ageC	-1.430	-3.665	0.805	0.210

All three models have some degree of similarities. They all predict the FEV1 lung function of an individual, given the effects of their age, sex and F508 genotype, as well as interaction effects between the three. Additionally the presence of an infection with Pseudomonas Aeruginosa is a linear variable established as a confounder. All models also include a random effects term for each person. In other words, a person with high lung function would be expected to have high lung function at some later age, and there would be some variance among the general population. This ensures generalizability of the results to the population of all individuals with cystic fibrosis rather than only the individuals in the study. The first model assumes than all individuals will have the same response with the effect of age on lung function.

The second model expands upon this. This model includes a random slope in the age of the individual. Simply, the rate of decline of lung function over time varies between individuals in a normal distribution, but is constant within the individual across time provided all other factors such as Pseudomonas Aeruginosa infection remain constant. This slope random effect is also related to the random intercept through their covariance.

Lastly, the third model involves a serial correlation. This assumes that there is a correlation of the lung function at one time and another time which is exponentially decreasing within an individual. The effect of age within an individual is no longer required to be constant. Explicitly we may state:

$$cov(A_{ij}, A_{nk}) = \delta_{in}\sigma_A^2 exp(-\frac{|X_{ij3} - X_{nk3}|}{\phi})$$

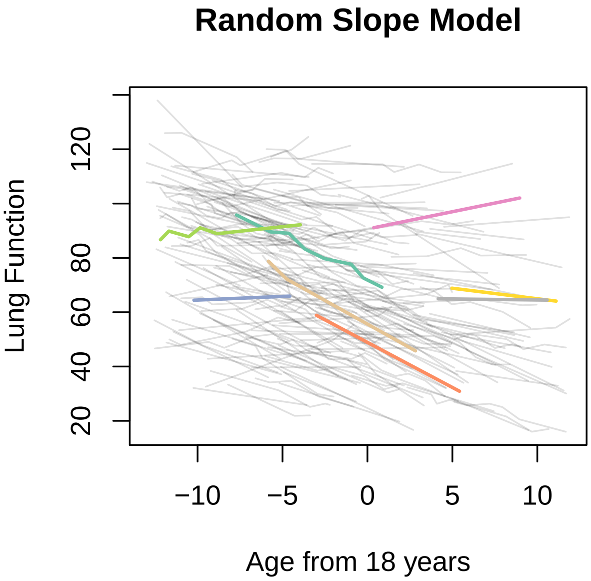
The serial correlation model is more general than the random slope model which is more general than the random intercept model. All models include the usual assumptions of a linear model: a linear relationship, multivariate normality, no or little multicollinearity, no auto-correlation and homoscedasticity.

The Akaike information criterion (AIC) is a measure of the quality of a statistical model for a given set. The model with the lowest AIC "minimizes" the information lost and is thus the most probable model.

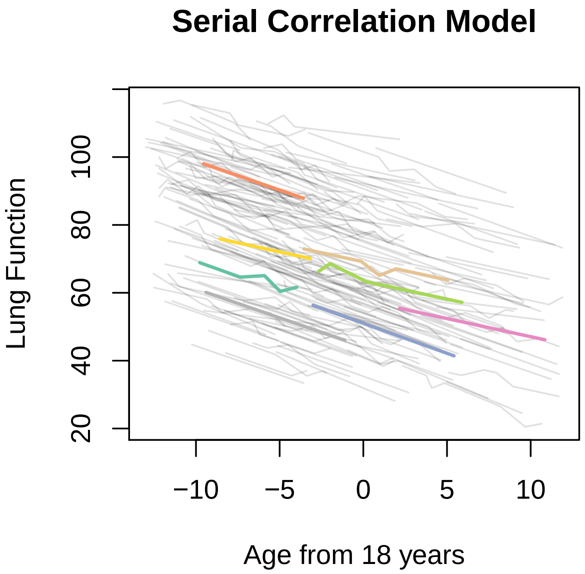
Table 5: Models and AIC Values

Model	AIC
Random Intercept	12531.65
Random Slope	12444.77
Serial Correlation	12382.48

Since the minimum AIC is achieved with the serial correlation model, it will be used for the analysis.



FEV1 Lung Function Prediction Over Age From Random Slope Model



FEV1 Lung Function Prediction Over Age From Serial Correlation Model

The research hypotheses were the following:
1. the rate at which lung function declines for CF patients depends on the F508 gene; and
2. the effect of the F508 gene on lung function decline differs for females and males.

A likelihood ratio test can be applied to test these hypotheses. The results are displayed below.

Table 6: Likelihood Ratio Test for F508 Gene Factor in Serial Correlation Model

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
Full Model	17	-6174.241	NA	NA	NA
No Combination Gender, Age, F508 Interaction	13	-6176.322	-4	4.162	0.385

	#Df	LogLik	Df	Chisq	Pr(>Chisq)
No Age, F508 Interaction	11	-6177.613	-2	2.583	0.275

As seen in **table 6.** using the serial correlation model, there is no significant effect of the F508 gene on the rate of age, nor was a difference found between genders FEV1 function over time. Thus we cannot confirm the F508 gene had a function on the rate of lung function decline. It is noteworthy however that the significance of the coefficients of the interaction between the F508 gene and age were significant at the $\alpha = 0.05$ level. When comparing to a person who is homozygous with the F508 gene, a heterozygous individual was found to have lung function decline of 0.981 FEV1 units per year (95% $CI = [0.040, 1.923]$) less than the homozygous positive case ($p = 0.041$). An individual which had no F508 gene was found to have lung function decline of 1.457 FEV1 units per year (95% $CI = [0.001, 2.914]$) less than the homozygous positive case ($p = 0.05$). Both of these intervals nearly overlapped 0, and thus these effects were not confirmed by a likelihood ratio test ($p = 0.385$, **table 6.**). There may be an effect that more copies of the F508 gene cause a faster lung function decline, however this may be worth further investigation.

Moss in Galicia

1. A two dimensional spatial model is used according to the following. Y_i is the box-cox transformation of the measurement of lead in the soil at a particular location based on the population of the area, rainfall per year, and predominant soil type.

$$\begin{aligned} Y_i &\sim N[\lambda(s_i), \tau^2] \\ \lambda(s_i) &= \mu + \beta_1 X_{log(\text{Pop})} + \beta_2 X_{\text{Rain}} + \beta_3 X_{\text{Soil}} + U(s_i) \\ \text{where } Cov(U(s_i), U(s_i + h)) &= \sigma^2 \rho(h/\phi, \nu) \end{aligned}$$

   where ρ is the Matérn family function with range parameter ϕ and shape parameter ν.

2. It does not seem plausible that rain or soil type influence the lead content in the moss in Galicia. This is due to the fact that the displayed 95% confidence intervals for these variables overlap with 0, indicating that there is a greater than 5% chance that the the estimates are due to chance assuming a normal distribution of the variable. This was true for the continuous variable of rain as well as each of the soil types. Additionally, using the method of the likelihood ratio test, the P value obtained for the nested model without these parameters was 0.0919 which is greater than the usual $\alpha = 0.05$ threshold.

3. Population was found to have a significant effect on the lead levels of the Moss in Galicia. Though the 95% confidence interval for the coefficient of the logarithm of the population overlapped 0. the likelihood ratio test however, when compared against the nested model without this parameter, gave a p-value of 0.00553. This suggests that the improvement of prediction from including the parameter is significant when compared to a model without this parameter.

4. The second order statistical properties of the lead in Galicia are the variances associated with the model. This includes the observation variance (τ^2) and residual spatial variation σ^2. In this case it was estimated that the observation variance was 0.04 with a 95% confidence interval of $[0.00, 65.22]$ and a residual spatial variation of 0.25 and a 95% confidence interval of $[0.18, 0.35]$. Additionally, in the model, it was set that the shape parameter of the correlation is 1, as well as an isotropic correlation. The observation variance is the error associated with the measurement of lead or very localized factors. It was found that there was a lower effect of the total variance of the data from the model, however there is a large variation in this estimate, when considering the 95% CI of this measurement there may in fact be a larger error in measurement or local factors. The range parameter was found to be 34.22 *km*. This indicates that the correlation in soil location was found to be effective on the order of 35 *km*.