P8157_midterm_rw2844

Ryan Wei

2023-10-20

Question 1

(a)

Figure 1 shows the relationship between the outcome $Y_{ki} = \frac{\text{FEV}_{ki}}{\text{height}_{ki}^2}$ and age, with four random selected individuals' trajectories.

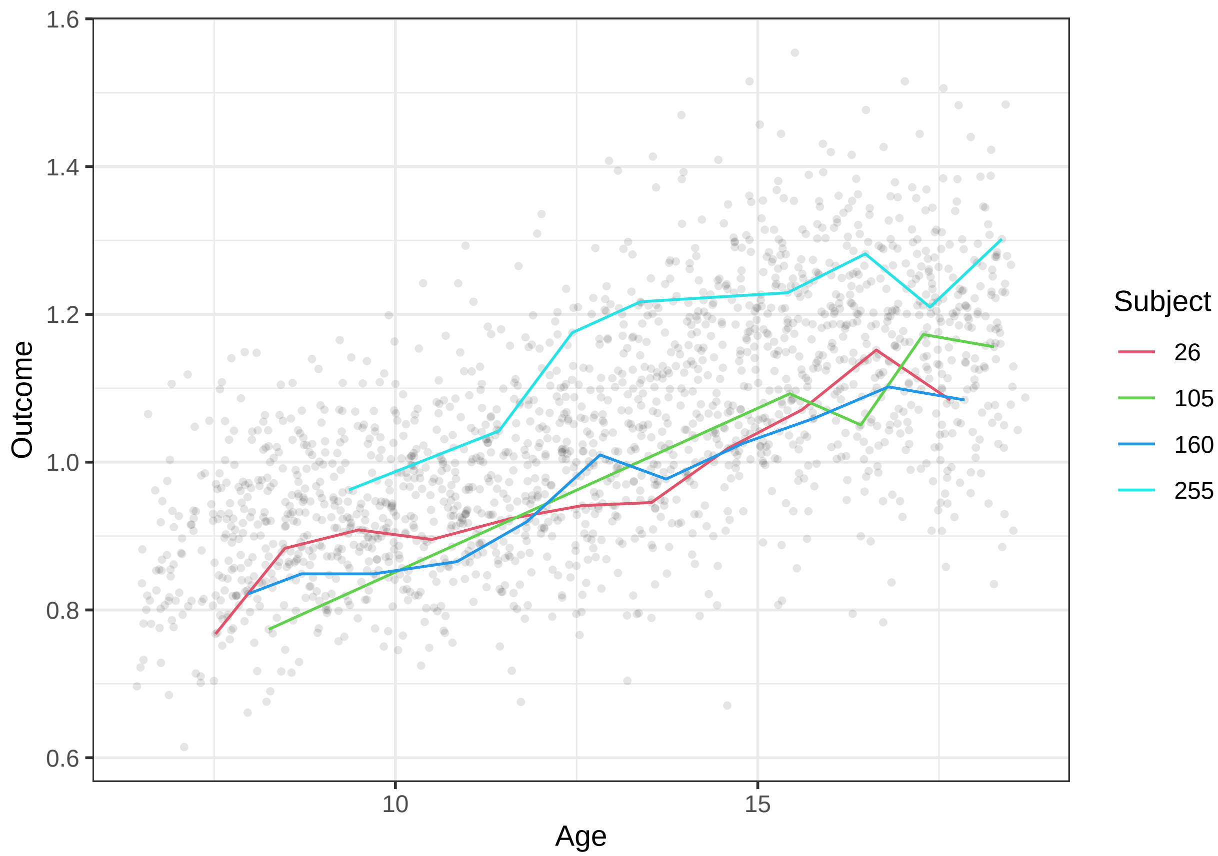

Figure 1: Scatter plot of response vs. age, with individual trajectories for a random sample of 4 girls.

(b)

The marginal mean model is

$$E[Y_{ki} \mid X_{ki}] = \beta_0 + \beta_1 \text{Age}_{ki} + \beta_2 \text{Age}_{ki}^2 + \beta_3 \text{Age}_{ki}^3.$$

Based on this mean model, we are going to fit the following models:

- (i) independent, homoskedastic errors
- (ii) random intercepts plus independent, homoskedastic errors
- (iii) random intercepts/slopes plus independent, homoskedastic errors
- (iv) random intercepts plus auto-regressive errors
- (v) random intercepts plus exponential spatial errors
- (vi) random intercepts plus exponential spatial errors and independent, homoskedastic errors
- (vii) random intercepts plus independent, heteroskedastic errors
- (viii) random intercepts/slopes plus independent, heteroskedastic errors

Table 1 shows the log-likeihood and AIC of the different model fits, using REML. From the table we can see that the models 5 (random intercepts plus exponential spatial errors) and 6 (random intercepts plus exponential spatial errors and independent, homoskedastic errors) are the best two fits of the data, based on AIC.

Table 1: Dependence model with corresponding log-likelihood and AIC, based on model (3).

Number	Dependence Model	log-likelihood	AIC
1	Independent + homoskedastic errors	1291.696	-2573.393
2	Random intercepts + inde.[a], homoskedastic errors	2045.909	-4079.818
3	Random intercepts/slopes + inde., homoskedastic errors	2120.486	-4224.973
4	Random intercepts + AR[b] errors	2129.402	-4244.804
5	Random intercepts + ES[c] errors	2141.178	-4268.357
6	Random intercepts + ES errors with a 'nugget'	2148.787	-4281.574
7	Random intercepts + inde., heteroskedastic errors (age)	2065.037	-4106.074
8	Random intercepts/slopes + inde., heteroskedastic errors (age)	2134.846	-4241.691

[a] Independent [b] Auto-regressive [c] Exponential spatial

For model 5, the full model can be written as

$$Y_{ki} = \beta_0 + \beta_1 \text{Age}_{ki} + \beta_2 \text{Age}_{ki}^2 + \beta_3 \text{Age}_{ki}^3 + \gamma_{0k} + W_k(T_{ki}) + \epsilon_{ki}^*,$$

where γ_{0k} is a cluster-specific random intercept with $\mathrm{E}[\gamma_{0k}] = 0$ and $\mathrm{V}[\gamma_{0k}] = \sigma_\gamma^2$. $W_k(T_{ki})$ is a serial dependence term, we assume the stochastic process $W_k(\cdot)$ is mean zero and is characterized by its covariance function

$$\mathrm{Cov}\left(W_k(T_{ki}), W_k(T_{kj})\right) = \sigma_W^2 \rho(U_{k,ij}),$$

where $U_{k,ij} = |T_{ki} - T_{kj}|$, $\rho(U_{k,ij}) = \exp\{U_{k,ij}/\text{range}\}$.

For model 6, the full model is similar to model 5, which can be written as

$$Y_{ki} = \beta_0 + \beta_1 \text{Age}_{ki} + \beta_2 \text{Age}_{ki}^2 + \beta_3 \text{Age}_{ki}^3 + \gamma_{0k} + W_k(T_{ki}) + \epsilon_{ki}^*,$$

where $\rho(U_{k,ij}) = (1 - \text{nugget}) \exp\{U_{k,ij}/\text{range}\}$ in this setting.

Table 2 shows the coefficients estimates and standard error estimates for these two models.

Table 2: Coefficients estimates and standard error estimates from the best two fits.

Term	Intercept	age	age^2	age^3
Model 5				
Coefficient estimates	1.4004	-0.1740	0.0176	-0.0005
Standard error	0.1090	0.0277	0.0023	0.0001
t-value	12.8437	-6.2788	7.7938	-8.1291
p-value	0.0000	0.0000	0.0000	0.0000
Model 6				
Coefficient estimates	1.4339	-0.1825	0.0183	-0.0005
Standard error	0.1053	0.0267	0.0022	0.0001
t-value	13.6176	-6.8245	8.4059	-8.7670
p-value	0.0000	0.0000	0.0000	0.0000

(c)

Using the following mean model

$$E[Y_{ki} \mid X_{ki}] = \beta_0 + \beta_1 \text{Age}_{ki},$$

we re-fit the models with different dependency sturctures. Table 3 shows the log-likeihood and AIC of the different model fits, using REML. From the table we can see that the models 5 (random intercepts plus exponential spatial errors) and 6 (random intercepts plus exponential spatial errors and independent, homoskedastic errors) are again the best two fits of the data, based on AIC.

Table 3: Dependence model with corresponding log-likelihood and AIC, based on model (2).

Number	Dependence Model	log-likelihood	AIC
1	Independent + homoskedastic errors	1275.095	-2544.190
2	Random intercepts + inde.[a], homoskedastic errors	2009.976	-4011.952
3	Random intercepts/slopes + inde., homoskedastic errors	2076.117	-4140.233
4	Random intercepts + AR[b] errors	2108.761	-4207.522
5	Random intercepts + ES[c] errors	2120.045	-4230.091
6	Random intercepts + ES errors with a 'nugget'	2123.624	-4281.574
7	Random intercepts + inde., heteroskedastic errors (age)	2024.662	-4029.324
8	Random intercepts/slopes + inde., heteroskedastic errors (age)	2087.067	-4150.134

[a] Independent [b] Auto-regressive [c] Exponential spatial

Figure 2 shows the relationship between the outcome $Y_{ki} = \frac{\text{FEV}_{ki}}{\text{height}_{ki}^2}$ and age, with fitted regression curves using the same best-fitting dependence structure with two different mean models.

(d)

The plotted curve based on model (3) represents the average change in outcome with respect to age, along with its polynomial terms up to the third order, for the study population. The curve shows that the outcome tends to increase as age increases, but the rate of increase speeds up at the beginning ($\leq$ 8 years old), then stays constant, and gradually slows down as age progresses. Moreover, the curve demonstrates that the average change in outcome appears to be constant (slightly decrease) after 16 years of age, which indicates that the average outcome does not vary much beyond this age.

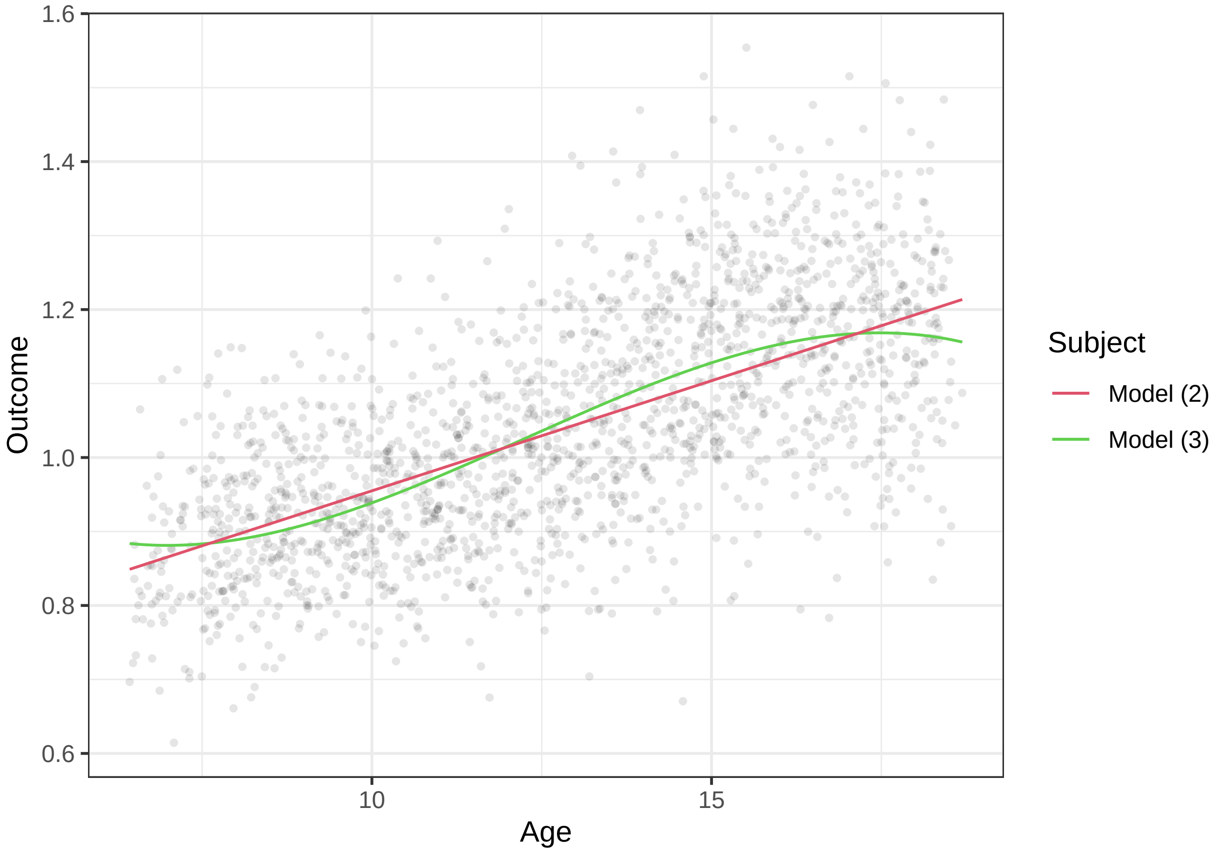

Figure 2: Scatter plot of response vs. age, with fitted regression curves.

(e)

The difference between model (3) with dependency structure 6 and model (2) with dependency structure 6 is that model (2) does not have second- and third-order terms in the mean model. We can perform a likelihood ratio test to determine if adding these higher-order terms improves the model fit. Specifically, the null hypothesis and alternative hypothesis for this test are

$$H_0 : \beta_2 = \beta_3 = 0, \quad H_1 : \beta_2 \neq 0, \beta_3 \neq 0,$$

and the likelihood ratio test has degree of freedom 2.

Table 4: Likelihood ratio test results for two models with same dependency structure with different mean models.

Model	log-likelihood	log-likelihood ratio	p-value
Model (3)	2148.787		
Model (2)	2123.624	50.32555	0

Table 4 shows the likelihood ratio test results. From the table, we can see that the p-value is less than 0.05. Therefore, we reject the null hypothesis at 0.05 siginificance level and conclude that using mean model (3) improves the model fit, which means model (2) might not be adopted in favor of model (3).

(f)

Personally, I would recommend model (2). The first reason is that it is simple and easy to interprate. For instance, we can easily interpret the regression coefficient β_1 as the average change in outcome when the age is increased by one unit. On the other hand, the same coefficient is difficult to interpret in model (3) due to its higher-order terms. Secondly, although model (3) provides a better fit for the data, it may not generalize well. As we can see from figure 2, there is a slight decrease at the right end of the graph that may not be biologically reasonable. Extrapolating beyond our observed data (e.g. subject with age 20) could lead to highly unreliable predictions. In contrast, a linear model (2) has good generalizability due to its simplicity.

Appendix: Code for this report

```
knitr::opts_chunk$set(echo = FALSE, message = F, warning = F)
options(knitr.kable.NA = '')
library(tidyverse)
library(caret)
library(latex2exp)
library(gstat)
library(sp)
library(nlme)
library(kableExtra)
write_matex <- function(x) {
  begin <- "$$\\begin{bmatrix}"
  end <- "\\end{bmatrix}$$"
  X <-
    apply(x, 1, function(x) {
      paste(
        paste(x, collapse = "&"),
        "\\\\"
      )
    })
  writeLines(c(begin, X, end))
}
load("../datasets/Six_Cities/Six Cities.RData")
df = as.tibble(topeka) %>%
  mutate(Y = exp(`log.FEV1`)/(height^2), age_cat = round(age), age_cat2 = cut(age, breaks = seq(6,20,2)
  group_by(id) %>%
  filter(n() >= 5) %>%
  ungroup()
set.seed(8157)
sampled_4_girls = sample(unique(df$id), 4)

spaghetti_topeka =
  ggplot() +
  geom_point(aes(x=df$age, y=df$Y), alpha = 0.1, size = 0.8) +
  geom_line(aes(x = df$age[which(df$id %in% sampled_4_girls)], y = df$Y[which(df$id %in% sampled_4_girl
  scale_color_manual(name = "Subject", values = c(2,3,4,5))+

  xlab("Age") +
  ylab(TeX("Outcome")) +
  theme_bw()

spaghetti_topeka
#(i) independent, homoskedastic errors
fit1.ML = glm(Y ~ age + I(age^2) + I(age^3), data = df, family = gaussian)
#summary(fit1.ML)
#(ii) random intercepts plus independent, homoskedastic errors
fit2.ML = lme(fixed = Y ~ age + I(age^2) + I(age^3),
              random = reStruct(~ 1 | id),
              data = df,
              method = "REML")
#summary(fit2.ML)
#(iii) random intercepts/slopes plus independent, homoskedastic errors
```

```r
fit3.ML = lme(fixed = Y ~ age + I(age^2) + I(age^3),
              random = reStruct(~ age | id),
              data = df,
              method = "REML")
#summary(fit3.ML)
#(iv) random intercepts plus auto-regressive errors
fit4.ML = lme(fixed = Y ~ age + I(age^2) + I(age^3),
              random = reStruct(~ 1 | id),
              correlation = corAR1(form = ~ age | id),
              data = df,
              method = "REML")
#summary(fit4.ML)
#(v) random intercepts plus exponential spatial errors
fit5.ML = lme(fixed = Y ~ age + I(age^2) + I(age^3),
              random = reStruct(~ 1 | id),
              correlation = corExp(form = ~ age | id),
              data = df,
              method = "REML")
#summary(fit5.ML)
#(vi) random intercepts plus exponential spatial errors and independent, homoskedastic errors
fit6.ML = lme(fixed = Y ~ age + I(age^2) + I(age^3),
              random = reStruct(~ 1 | id),
              correlation = corExp(form = ~ age | id, nugget = TRUE),
              data = df,
              method = "REML")
#summary(fit6.ML)
#(vii) random intercepts plus independent, heteroskedastic errors
fit7.ML = lme(fixed = Y ~ age + I(age^2) + I(age^3),
              random = reStruct(~ 1 | id),
              weights = varIdent(form = ~1 | age_cat2),
              data = df,
              method = "REML")
#summary(fit7.ML)
#(viii) random intercepts/slopes plus independent, heteroskedastic errors

fit8.ML = lme(fixed = Y ~ age + I(age^2) + I(age^3),
              random = reStruct(~ age | id),
              weights = varIdent(form = ~1 | age_cat2),
              data = df,
              method = "REML")
#summary(fit8.ML)
loglik_list = c(logLik(fit1.ML), logLik(fit2.ML), logLik(fit3.ML), logLik(fit4.ML), logLik(fit5.ML), log
aic_list = c(AIC(fit1.ML), AIC(fit2.ML), AIC(fit3.ML), AIC(fit4.ML), AIC(fit5.ML), AIC(fit6.ML), AIC(fi
model_list = c(
  "Independent + homoskedastic errors",
  paste0("Random intercepts + inde.",footnote_marker_alphabet(1), ", homoskedastic errors"),
  "Random intercepts/slopes + inde., homoskedastic errors",
  paste0("Random intercepts + AR",footnote_marker_alphabet(2), " errors"),
  paste0("Random intercepts + ES",footnote_marker_alphabet(3)," errors"),
  "Random intercepts + ES errors with a 'nugget'",
  "Random intercepts + inde., heteroskedastic errors (age)",
  "Random intercepts/slopes + inde., heteroskedastic errors (age)"
)
```

```
mdl_comp_tbl = tibble(
  model_number = c(1:8),
  model = model_list,
  loglik = loglik_list,
  aic = aic_list
)

mdl_comp_tbl %>% knitr::kable(booktab = T, escape = F, digits = 3, col.names = c("Number", "Dependence I

fit5.ML.sum = summary(fit5.ML)
#fit5.ML.sum$tTable

fit6.ML.sum = summary(fit6.ML)
#fit6.ML.sum$tTable

coef.sum.tab = rbind(t(fit5.ML.sum$tTable), t(fit6.ML.sum$tTable))
coef.sum.tab = (coef.sum.tab) %>% as.data.frame()
coef.sum.tab$item = rownames(coef.sum.tab)
colnames(coef.sum.tab) = c("intercept", "age", "age^2", "age^3","item")
coef.sum.tab$item = c("Coefficient estimates", "Standard error", "DF", "t-value", "p-value", "Coefficien
coef.sum.tab %>% filter(item != "DF") %>% select(item, everything()) %>% knitr::kable(digits = 4, bookta
  pack_rows("Model 6", 5, 8)
#(i) independent, homoskedastic errors
fit1.ML.2 = glm(Y ~ age, data = df, family = gaussian)
#summary(fit1.ML)
#(ii) random intercepts plus independent, homoskedastic errors
fit2.ML.2 = lme(fixed = Y ~ age,
              random = reStruct(~ 1 | id),
              data = df,
              method = "REML")
#summary(fit2.ML)
#(iii) random intercepts/slopes plus independent, homoskedastic errors
fit3.ML.2 = lme(fixed = Y ~ age,
              random = reStruct(~ age | id),
              data = df,
              method = "REML")
#summary(fit3.ML)
#(iv) random intercepts plus auto-regressive errors
fit4.ML.2 = lme(fixed = Y ~ age,
              random = reStruct(~ 1 | id),
              correlation = corAR1(form = ~ age | id),
              data = df,
              method = "REML")
#summary(fit4.ML)
#(v) random intercepts plus exponential spatial errors
fit5.ML.2 = lme(fixed = Y ~ age,
              random = reStruct(~ 1 | id),
              correlation = corExp(form = ~ age | id),
              data = df,
              method = "REML")
#summary(fit5.ML)
#(vi) random intercepts plus exponential spatial errors and independent, homoskedastic errors
fit6.ML.2 = lme(fixed = Y ~ age,
```

```
              random = reStruct(~ 1 | id),
              correlation = corExp(form = ~ age | id, nugget = TRUE),
              data = df,
              method = "REML")
#summary(fit6.ML)
#(vii) random intercepts plus independent, heteroskedastic errors
fit7.ML.2 = lme(fixed = Y ~ age,
              random = reStruct(~ 1 | id),
              weights = varIdent(form = ~1 | age_cat2),
              data = df,
              method = "REML")
#summary(fit7.ML)
#(viii) random intercepts/slopes plus independent, heteroskedastic errors

fit8.ML.2 = lme(fixed = Y ~ age,
              random = reStruct(~ age | id),
              weights = varIdent(form = ~1 | age_cat2),
              data = df,
              method = "REML")
#summary(fit8.ML)

loglik_list.2 = c(logLik(fit1.ML.2), logLik(fit2.ML.2), logLik(fit3.ML.2), logLik(fit4.ML.2), logLik(fi
aic_list.2 = c(AIC(fit1.ML.2), AIC(fit2.ML.2), AIC(fit3.ML.2), AIC(fit4.ML.2), AIC(fit5.ML.2), AIC(fit6

mdl_comp_tbl.2 = tibble(
  model_number = c(1:8),
  model = model_list,
  loglik = loglik_list.2,
  aic = aic_list.2
)
mdl_comp_tbl.2 %>% knitr::kable(booktab = T, escape = F, digits = 3, col.names = c("Number", "Dependenc
df$pred_3 = predict(fit6.ML, newdata = df)
fit6.ML.coef= fit6.ML$coefficients$fixed
df$pred_2 = predict(fit6.ML.2, newdata = df)
fit6.ML.2.coef= fit6.ML.2$coefficients$fixed

curve_topeka =
  ggplot() +
  geom_point(aes(x=df$age, y=df$Y), alpha = 0.1, size = 0.8) +
  geom_line() +
  scale_color_manual(name = "Subject", values = c(2,3,4,5))+
  geom_function(fun = function(x) fit6.ML.coef[[1]] + x*fit6.ML.coef[[2]] + x^2 * fit6.ML.coef[[3]] + x
  geom_function(fun = function(x) fit6.ML.2.coef[[1]] + x*fit6.ML.2.coef[[2]], aes(color = "Model (2)")
  xlab("Age") +
  ylab(TeX("Outcome")) +
  theme_bw()
curve_topeka
lrt_res = anova.lme(fit6.ML, fit6.ML.2) %>% as.data.frame() %>% select(-call) %>% select(Model,logLik, 

lrt_res$Model = c("Model (3)", "Model (2)")

lrt_res %>% select(-Test) %>% knitr::kable(booktab = T, escape = F, digits = 5, col.names = c("Model",
```

```r
#lrt = abs(as.numeric(2 * (logLik(fit6.ML) - logLik(fit6.ML.2))))

#pchisq(lrt,df = 2, lower.tail = F)
```