MID

Zirui Zhang

2023-10-25

```r
library(tidyverse)
library(ggplot2)
library(patchwork)
library(nlme)
library(lme4)
```

```r
load("~/Documents/2023Fall/P8157/P8157/Six Cities.RData")
data = topeka |>  group_by(id) |> filter(n() >= 5) |> ungroup()
length(unique(data$id))
```

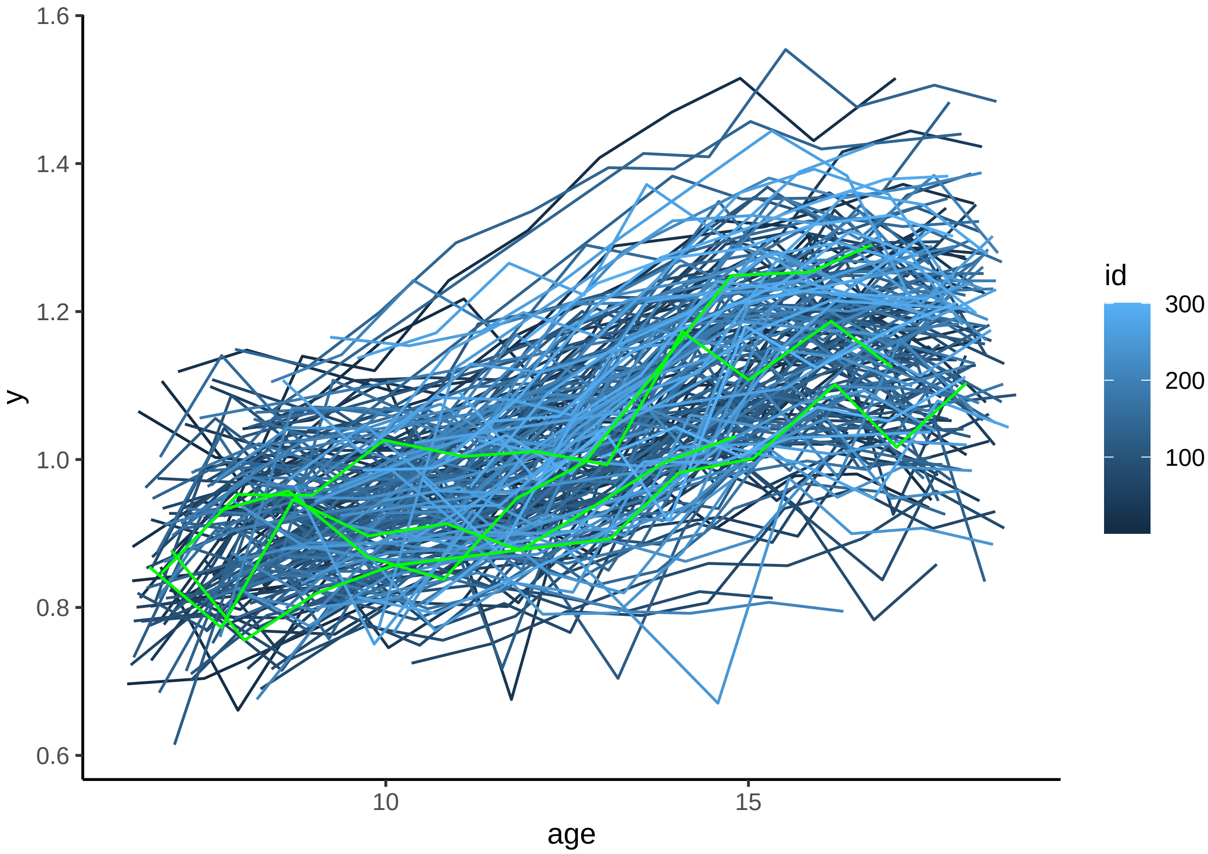
```
## [1] 196
```

```r
data = data |>
  mutate(y = exp(log.FEV1)/(height^2),
         age.2 = age^2,
         age.3 = age^3)
```

Question (a)

Produce a figure of the response, Yki as a function of age. On the figure indicate the individual trajectories for a random sample of 4 girls.

```r
set.seed(200324)
sample = data  |>
  filter(id %in% sample(unique(data$id), 4))
ggplot(data, aes(x = age, y = y, group = id, color = id)) +
  geom_line() +
  geom_line(data = sample, color = "green") +
  theme_classic()
```



Question (b)

```
# 1 naivee
fit1.ML = glm(y ~ age + age.2 + age.3, data, family=gaussian)

# 2 randon intercept + independent error
fit2.ML = lme(fixed=y ~ age + age.2 + age.3, random=reStruct(~ 1 | id), data, method="ML")

# 3 random intercept/slope + independent error
fit3.ML = lme(fixed=y ~ age + age.2 + age.3, random=reStruct(~ age | id, pdClass="pdDiag"), data, metho

# 4. random intercept + auto_regressive error
fit4.ML = lme(fixed=y ~ age + age.2 + age.3, random=reStruct(~ 1 | id), correlation=corAR1(form= ~ age|

# 5 random intercept + exponential spatial error
fit5.ML = lme(fixed=y ~ age + age.2 + age.3, random=reStruct(~ 1 | id), correlation=corExp(form= ~ age|

# 6 random intercept + exponential spatial error + independent homo error
fit6.ML = lme(fixed=y ~ age + age.2 + age.3, random=reStruct(~ 1 | id), correlation=corExp(form= ~ age|

# 7 random intercept + independent hetero error
data_cat = data |>
  dplyr::mutate(age.cat = floor(age/2))
fit7.ML = lme(fixed=y ~ age + age.2 + age.3, random=reStruct(~ 1 | id), weights=varIdent(form= ~1 | age
```

```r
# 8 random intercept/slope + independent hetero error
fit8.ML = lme(fixed=y ~ age + age.2 + age.3, random=reStruct(~ age | id), weights=varIdent(form= ~1 | ag
```

```r
sum = (data.frame(
  logLik = c(logLik(fit1.ML), logLik(fit2.ML), logLik(fit3.ML),logLik(fit4.ML),
             logLik(fit5.ML), logLik(fit6.ML), logLik(fit7.ML), logLik(fit8.ML)),
  AIC = c(AIC(fit1.ML),AIC(fit2.ML),AIC(fit3.ML),AIC(fit4.ML),
          AIC(fit5.ML),AIC(fit6.ML),AIC(fit7.ML),AIC(fit8.ML))
))

colnames(sum) = c("log-Like", "AIC")
rownames(sum) = c("0. Independence", "1. Random intercept + inde. errors",
                  "2. Random intercept/slope + inde. errors", "3. Random intercept + AR errors",
                  "4. Random intercept + ES errors", "5. Random intercept + ES with a 'nugget'",
                  "6. Random intercept + heteroske inde. errors",
                  "7. Random intercept/slope + heteroske inde. errors")
knitr::kable(sum, format = "markdown")
```

	log-Like	AIC
0. Independence	1291.696	-2573.393
1. Random intercept + inde. errors	2073.425	-4134.850
2. Random intercept/slope + inde. errors	2138.977	-4263.954
3. Random intercept + AR errors	2156.186	-4298.373
4. Random intercept + ES errors	2168.018	-4322.035
5. Random intercept + ES with a 'nugget'	2175.572	-4335.145
6. Random intercept + heteroske inde. errors	2092.402	-4160.803
7. Random intercept/slope + heteroske inde. errors	2162.089	-4296.178

Model 4 and 5 give the largest loglikelihood and lowest AIC, provide best fits of the data.

- Model 4:

$$Y_{ki} = \beta_0 + \beta_1 \cdot Age_{ki} + \beta_2 \cdot Age_{ki}^2 + \beta_3 \cdot Age_{ki}^3 + \gamma_{0k} + W_k(T_{ki}) + \epsilon_{ki}^*$$
$$Cov[W_k(T_{ki}), W_k(T_{kj})] = \sigma_W^2 exp\{-U_{k,ij}/range\}$$

where $U_{k,ij} = |T_{ki} - T_{kj}|$

- Model 5:

$$Y_{ki} = \beta_0 + \beta_1 \cdot Age_{ki} + \beta_2 \cdot Age_{ki}^2 + \beta_3 \cdot Age_{ki}^3 + \gamma_{0k} + W_k(T_{ki}) + \epsilon_{ki}^*$$
$$Cov[W_k(T_{ki}), W_k(T_{kj})] = \sigma_W^2 (1-n) exp\{-U_{k,ij}/range\}$$

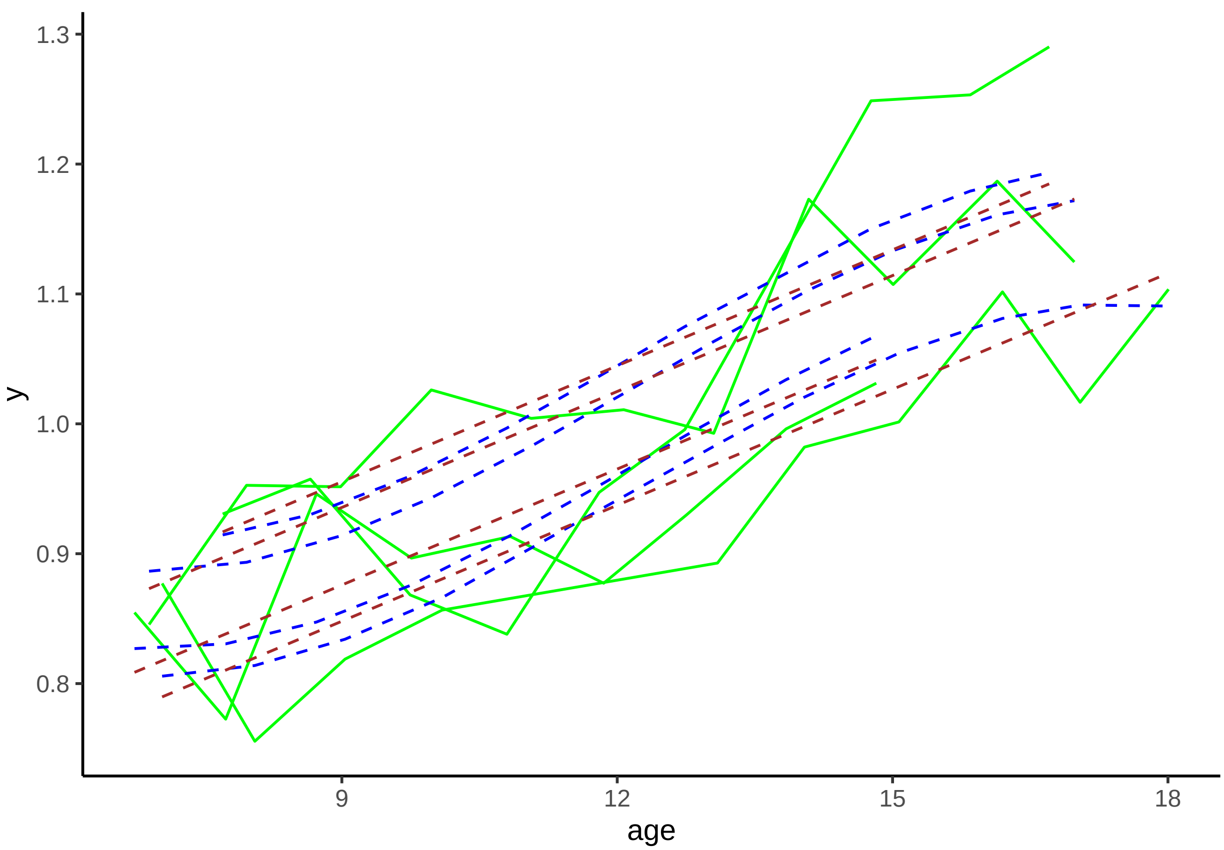
where n denotes the nugget effect.

```r
coef = t((data.frame(
  fit5 = c(summary(fit5.ML)$coefficients$fixed, sqrt(diag(summary(fit5.ML)$varFix))),
  fit6  = c(summary(fit6.ML)$coefficients$fixed, sqrt(diag(summary(fit6.ML)$varFix)))
)))
rownames(coef) = c("Model.4", "Model.5")
colnames(coef) = c("b0", "b1", "b2", "b3",
                   "sd(b0)", "sd(b1)", "sd(b2)", "sd(b3)")
knitr::kable(coef, format = "markdown")
```

	b0	b1	b2	b3	sd(b0)	sd(b1)	sd(b2)	sd(b3)
Model.4	1.400998	-0.1741504	0.0175949	-0.0004801	0.1088313	0.0276627	0.0022521	5.89e-05
Model.5	1.434293	-0.1825535	0.0182797	-0.0004980	0.1051375	0.0266974	0.0021707	5.67e-05

Question (c)

```
# fit6.ML = lme(fixed=y ~ age + age.2 + age.3, random=reStruct(~ 1 | id), correlation=corExp(form= ~ ag
fit9.ML = lme(fixed=y ~ age, random=reStruct(~ 1 | id), correlation=corExp(form= ~ age| id, nugget=TRUE)
predict.1 = predict(fit6.ML, sample)
predict.2 = predict(fit9.ML, sample)
ggplot(data, aes(x = age, y = y, group = id, color = id)) +
  geom_line(data = sample, color = "green") +
  geom_line(data = sample, aes(y = predict.1), color = "blue", linetype = "dashed") +
  geom_line(data = sample, aes(y = predict.2), color = "brown", linetype = "dashed") +
  theme_classic()
```



Question (d)

no inter for coef, fit curve visually

Question (e)

```r
anova(fit6.ML, fit9.ML)
```

```
##         Model df       AIC       BIC   logLik   Test  L.Ratio p-value
## fit6.ML     1  8 -4335.145 -4291.230 2175.573
## fit9.ML     2  6 -4255.874 -4222.937 2133.937 1 vs 2 83.27109  <.0001
```

I did an ANOVA test on the two models, the p-vlaue was smaller than 0.0001. Thus the null was rejected, suggesting that model(3) with more variables does provide better fit of the data.

Question (f)

complexity of the model…