Code Appendix

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

```r
# load data and data preparation
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)
```

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

```r
# 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
# 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
# model summary and comparison
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")
```

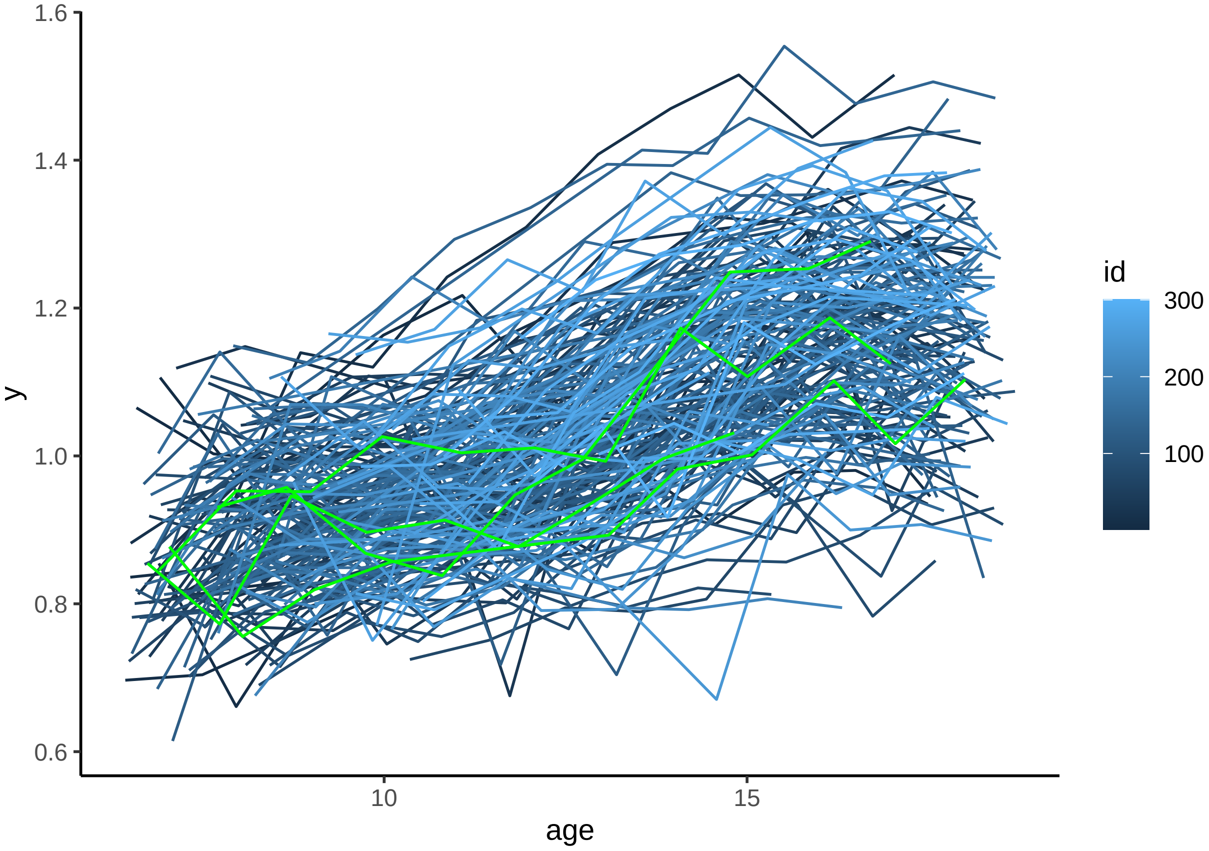
```
# coefficient summary for model4&5
coef = t((data.frame(
  fit5 = c(summary(fit5.ML)$coefficients$fixed, sqrt(diag(summary(fit5.ML)$varFix))),
  fit6  = c(summary(fit6.ML)$coefficients$fixed, sqrt(diag(summary(fit6.ML)$varFix)))
)))
rownames(coef) = c("Model.4", "Model.5")
colnames(coef) = c("b0", "b1", "b2", "b3",
                   "sd(b0)", "sd(b1)", "sd(b2)", "sd(b3)")
```

```
# 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)
p2 = 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()
data$m3 = predict(fit6.ML, data)
data$m2 = predict(fit9.ML, data)
p3 = ggplot(data, aes(x=age, y=y), color = id) +
  geom_point() +
  geom_smooth(aes(y=m3), se = F, color = "blue") +
  geom_smooth(aes(y=m2), se = F, color = "brown") +
  theme_classic()
```

0.1 Problem (a)

Here's the figure for the whole population and a random sample of 4:

```
p
```



0.2 Problem (b)

Here's the model summaries:

```r
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.

Here's the coefficient summaries of the two models:

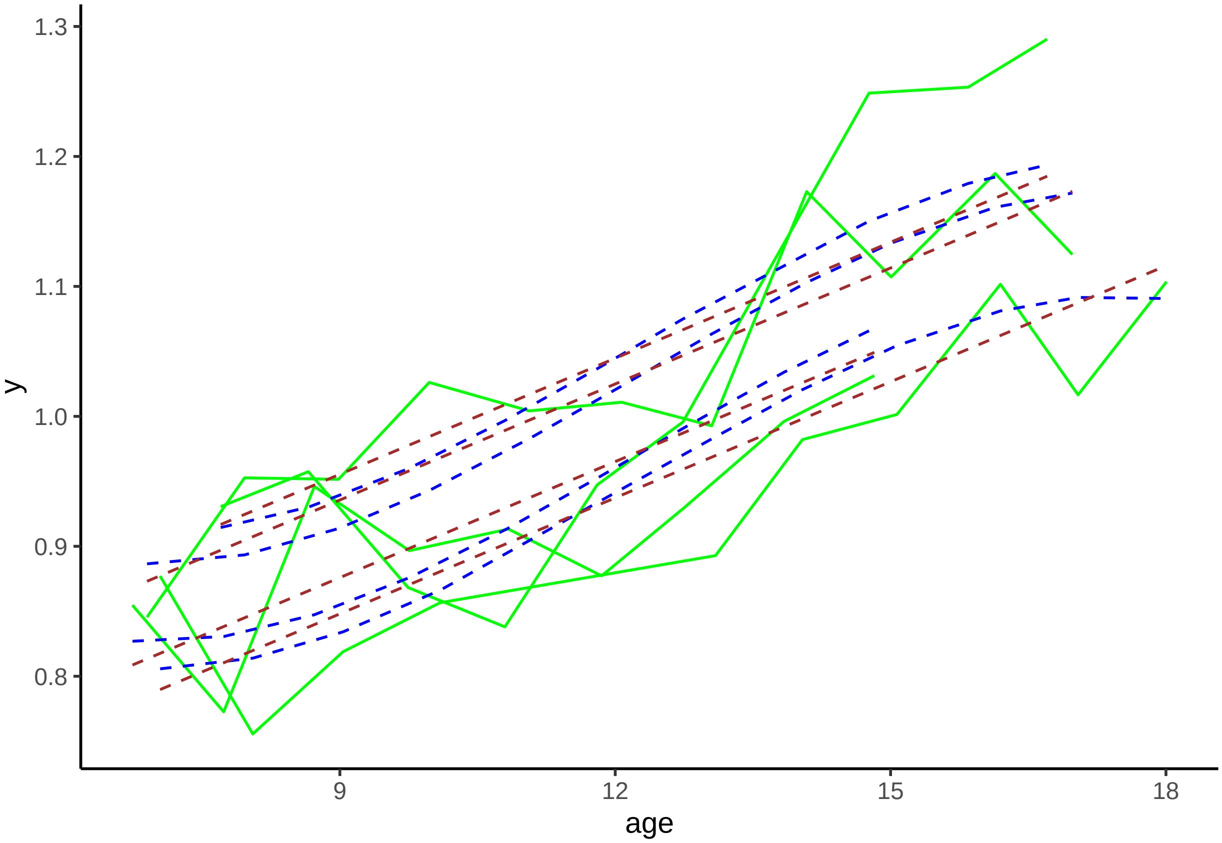
```r
knitr::kable(coef, format = "markdown")
```

	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

0.3 Problem (c)

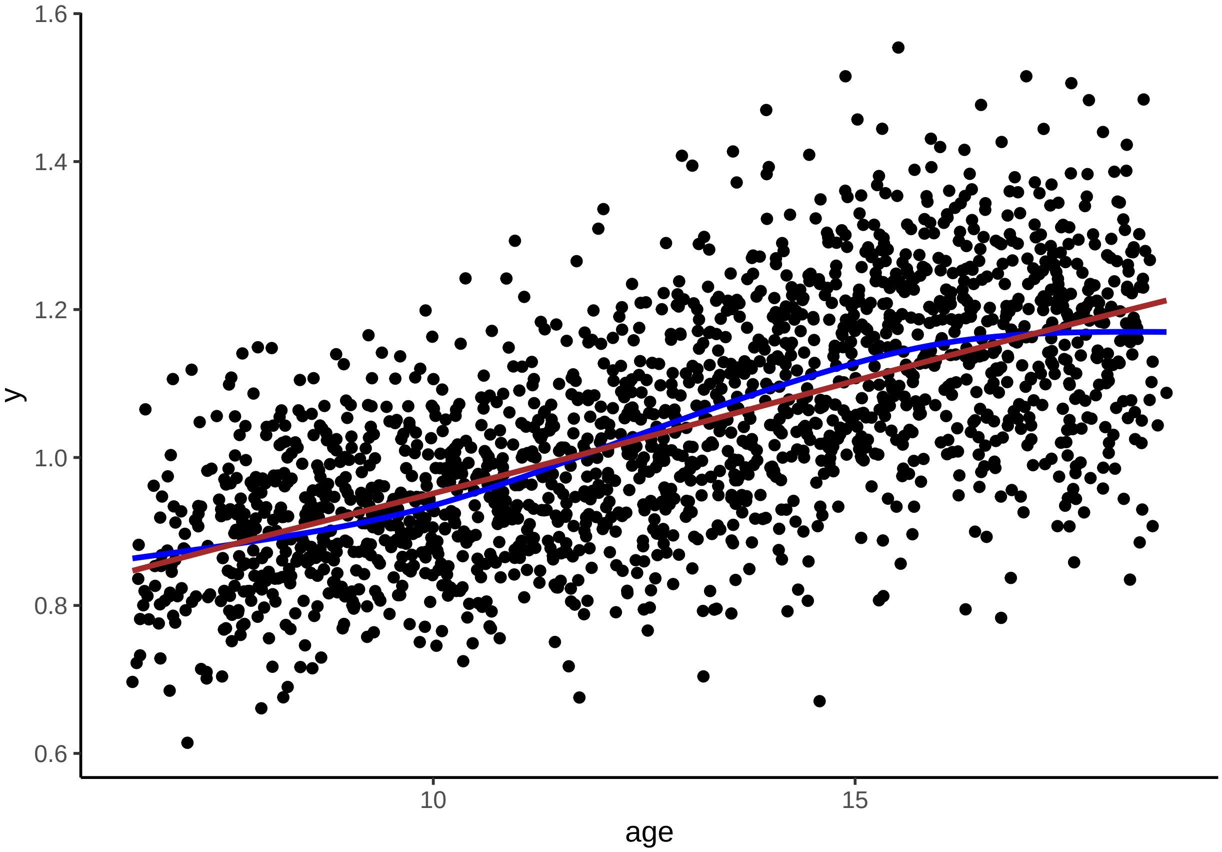
Here's the fitted regression curve for the 4 random samples and the whole population, where green lines are the true Y's, blue lines are the regression curves for model(3) and brown lines for model(2). We can see that model(3) generally provide better fit for the real data.

```
p2
```



```
p3
```

```
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



0.4 Problem (d)

In model(3), the outcome interest of the children increase with the increase of their age. The increasing trend was not obvious when they're young (before 9 yrs) or older (after 16 yrs) but was evident in between. Besides, each children seems to have unique baseline outcome.

0.5 Problem (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, where the null hypotheses is that the smaller model(model(2)) provide better fit of the data. The p-vlaue of the test was smaller than 0.0001, indicating the null should be rejected under a significance level of 0.05, suggesting that model(3) with more variables actually provide better fit of the data.

0.6 Problem (f)

```r
summary(fit6.ML)$tTable
```

```
##                       Value    Std.Error   DF   t-value      p-value
## (Intercept)  1.4342925587 1.052552e-01 1590 13.626806 4.497983e-40
## age         -0.1825534998 2.672730e-02 1590 -6.830227 1.203065e-11
## age.2        0.0182796673 2.173164e-03 1590  8.411544 8.881825e-17
## age.3       -0.0004980445 5.677694e-05 1590 -8.771950 4.430359e-18
```

I would choose model(3).

- It could be seen in part(c) that model(3) provide better fit of the data, with curves detailedly captured the trend at the begining and end of the real data.
- Given the result of model(3), the p-values of the age^2 and age^3 term are both far smaller than 0.05, indicating that they both add significantly better explanation for the data. Abandoning them would be unjustified.
- Given the result of the ANOVA test as above, we shouldn't adopt the parsimonious model.