BIOST540HW2

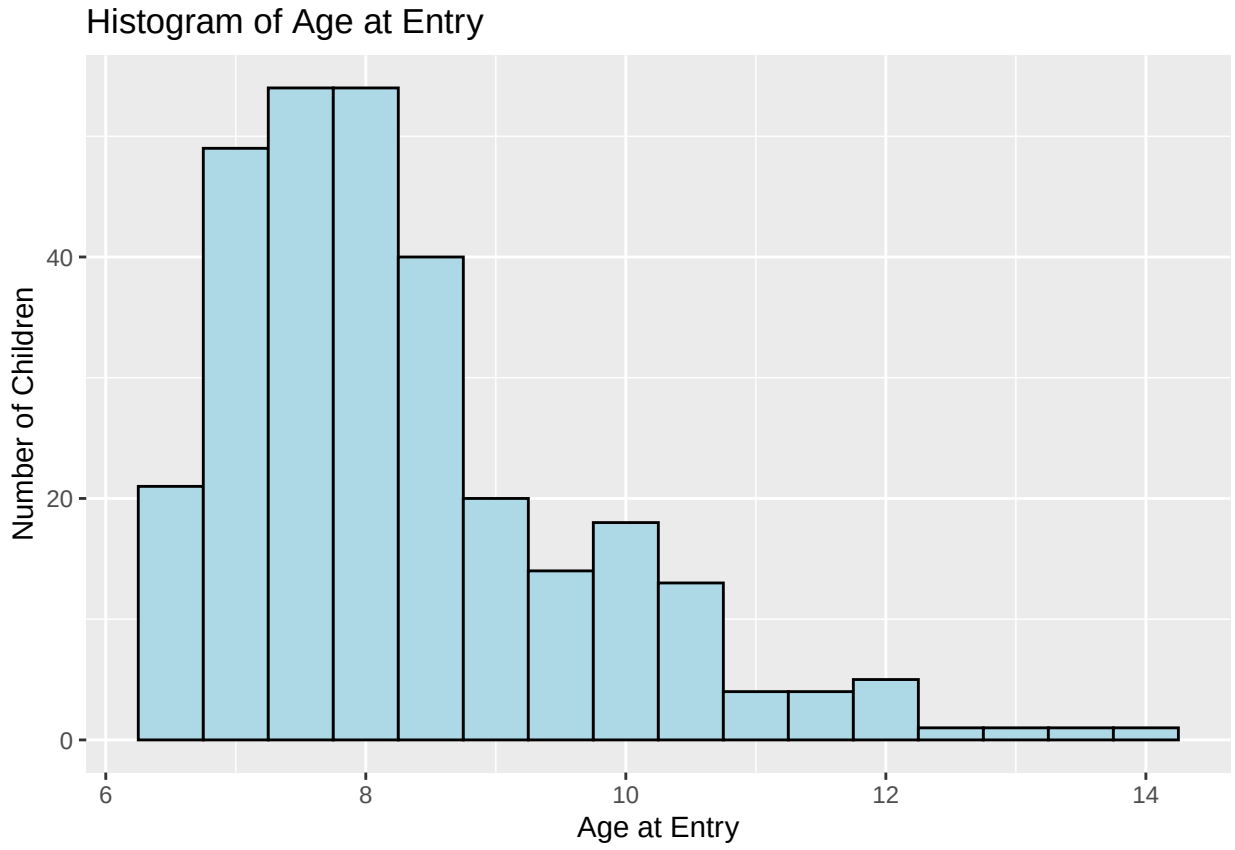
Bryan Ng, 2427348

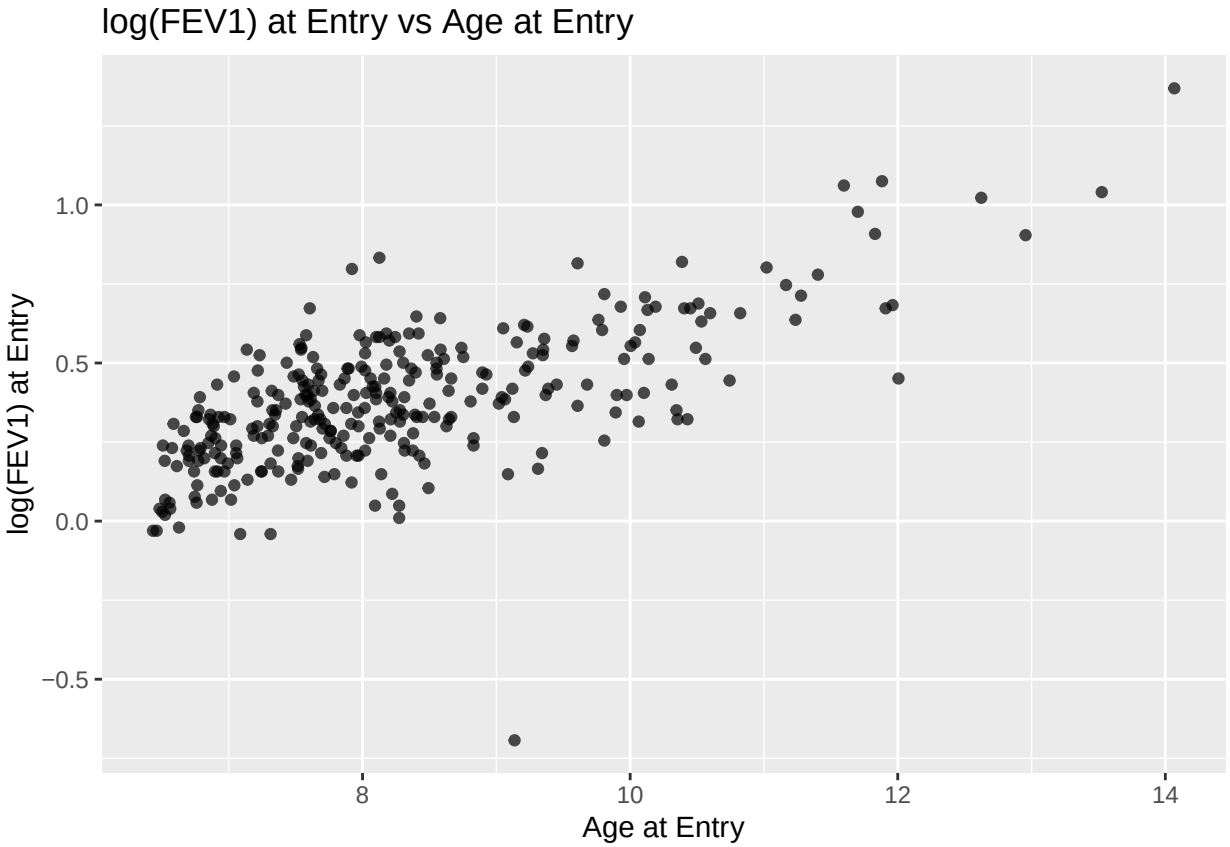
2025-05-09

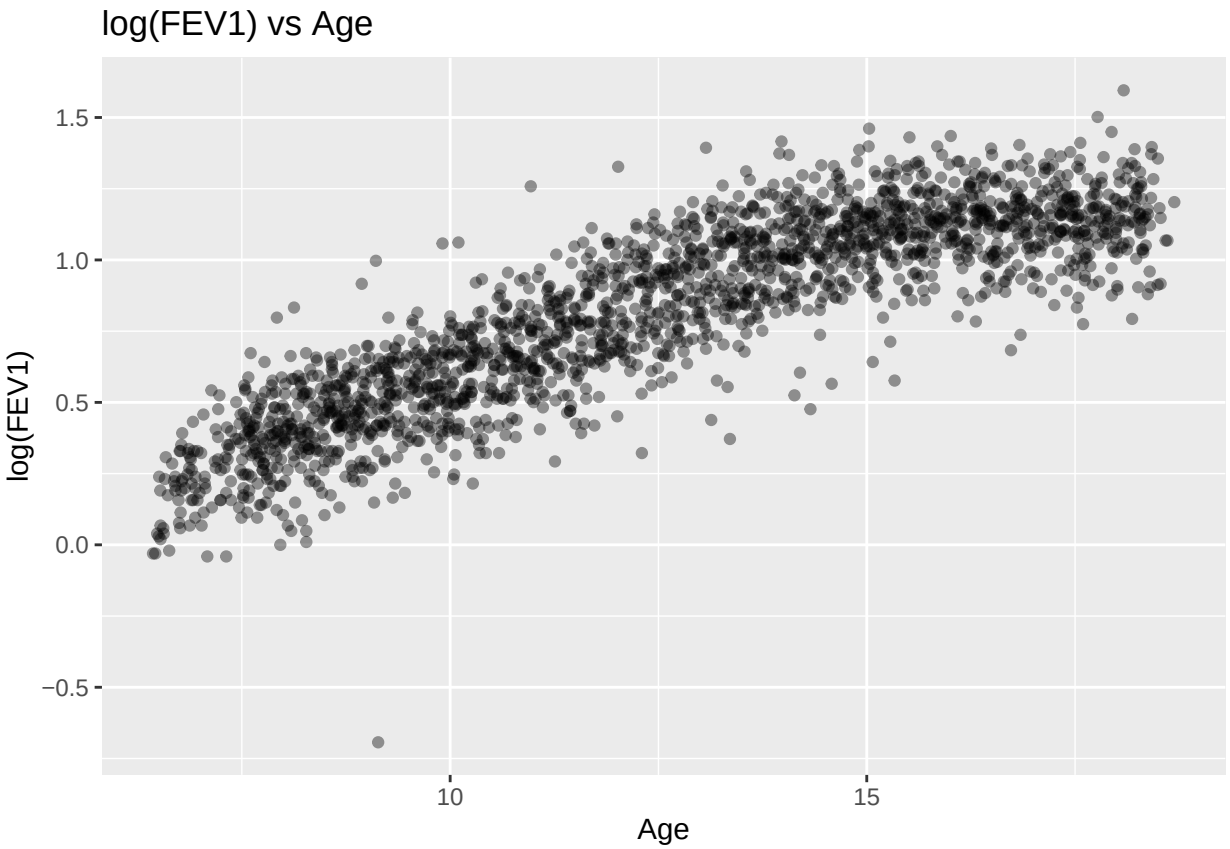
Problem 1

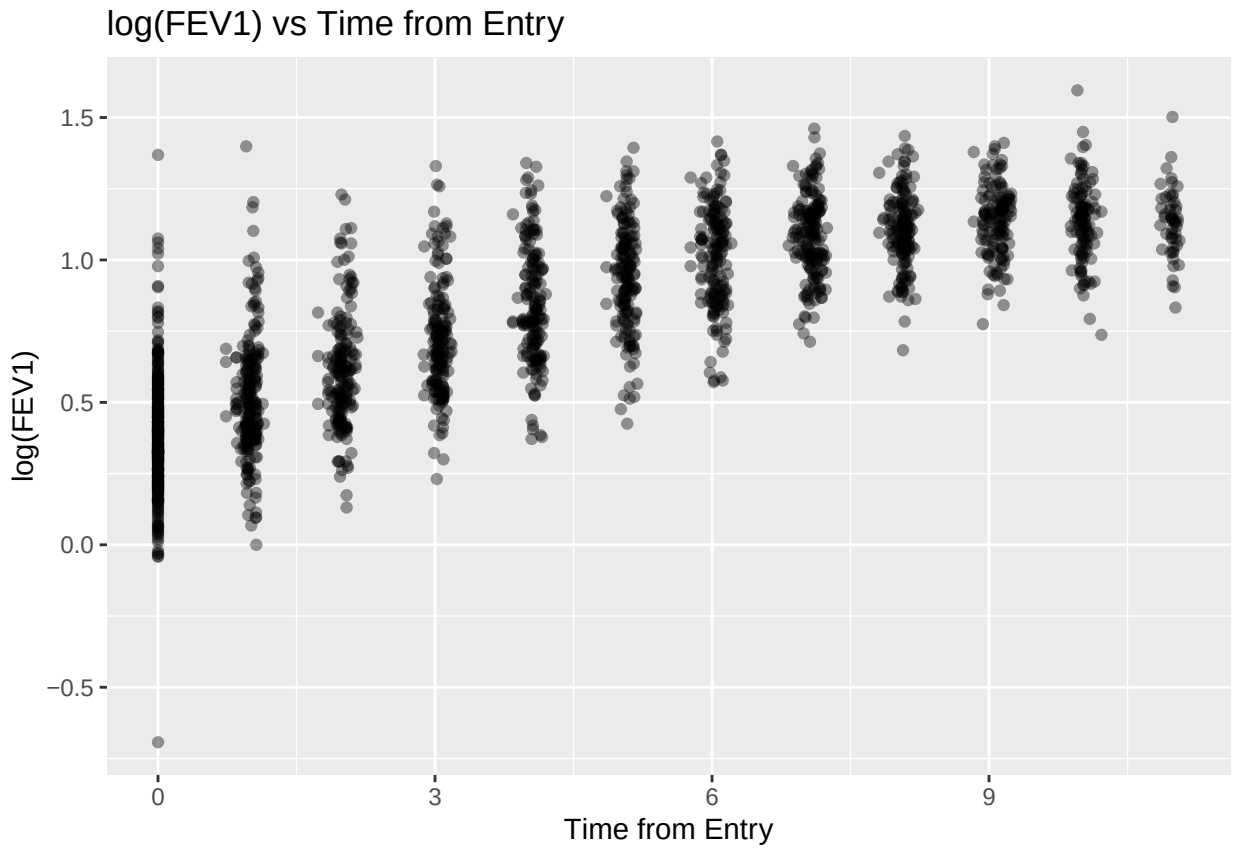
a)

```
## New names:
## Rows: 1994 Columns: 7
## -- Column specification
## -------------------------------------------------------- Delimiter: "," dbl
## (7): ...1, id, height, age, height0, age0, logFEV1
## i Use `spec()` to retrieve the full column specification for this data. i
## Specify the column types or set `show_col_types = FALSE` to quiet this message.
## * `` -> `...1`
```




log(FEV1) at Entry vs Age at Entry
log(FEV1) at Entry
1.0
0.5
0.0
−0.5
8
10
12
14
Age at Entry


log(FEV1) vs Age
log(FEV1)
1.5
1.0
0.5
0.0
−0.5
10
15
Age



b)

The histogram shows most girls entered the study between ages 6 and 9, with only a few older entrants. At baseline, there's a clear positive correlation between age and log(FEV1): six to seven year olds already have higher volumes of air on their first visit. When we incorporate all follow-ups, log(FEV1) continues to rise steadily with both chronological age and time since enrollment, reflecting normal growth and lung-maturation over the school years. We also see increasing spread in log(FEV1) at older ages and later follow-ups, suggesting individual differences in the timing or rate of lung growth become more pronounced over adolescence.

c)

d)

Table 1: Predicted mean log(FEV1) at selected ages

Age	Step	Linear	Quadratic	L-spline	C-spline
9	0.475	0.500	0.491	0.471	0.477
11	0.693	0.671	0.733	0.700	0.695
13	0.929	0.843	0.921	0.935	0.933
15	1.103	1.014	1.056	1.075	1.085

e)

Every model shows a clear positive relationship—older girls have higher lung function on average. Between roughly six and ten years, the slope is steeper, reflecting rapid lung growth during the early school years whereas after about thirteen to fourteen years, the curves begin to plateau, indicating that the rate of increase in the volume of air slows as girls approach teenage years.

Models allowing non-linear: step-function, quadratic, linear spline, cubic spline.

Models allowing non-constant rates of change: quadratic, linear spline, cubic spline.

f) i)

$$E\big[\log FEV1_{ij} \mid age_{ij}, group_i\big] = \beta_0 + \beta_1\, group_i + \beta_2\, age_{ij} + \beta_3\, (age_{ij} - 14)_+$$

We can rewrite it as the function of age_{ij} and group_i.

$$f(\text{age}_{ij}, \text{group}_i) = \beta_0 + \beta_1\, group_i + \beta_2\, age_{ij} + \beta_3\, (age_{ij} - 14)_+$$

For the unexposed group:

Slope for $age < 14$: $\partial f(\text{age}_{ij} < 14, \text{group}_i = 0)/\partial \text{age}_{ij} = \beta_2$
Slope for $age \geq 14$: $\partial f(\text{age}_{ij} \geq 14, \text{group}_i = 0)/\partial \text{age}_{ij} = \beta_2 + \beta_3$

For the exposed group:

Slope for $age < 14$: $\partial f(\text{age}_{ij} < 14, \text{group}_i = 1)/\partial \text{age}_{ij} = \beta_2$
Slope for $age \geq 14$: $\partial f(\text{age}_{ij} \geq 14, \text{group}_i = 1)/\partial \text{age}_{ij} = \beta_2 + \beta_3$

f) ii)

For model:

$$E\big[\log FEV1_{ij} \mid age_{ij}, group_i\big] = \beta_0 + \beta_1\, group_i + \beta_2\, age_{ij} + \beta_3\, (age_{ij} \times group_i) + \beta_4\, (age_{ij} - 14)_+$$

Similarly, we can derive

For the unexposed group: Slope for $age < 14$: β_2
Slope for $age \geq 14$: $\beta_2 + \beta_4$

For the exposed group: Slope for $age < 14$: $\beta_2 + \beta_3$
Slope for $age \geq 14$: $(\beta_2 + \beta_3) + \beta_4$

f) iii)

For model:

$$E\big[\log FEV1_{ij} \mid age_{ij}, group_i\big] = \beta_0 + \beta_1\, group_i + \beta_2\, age_{ij} + \beta_3\, (age_{ij} \times group_i) + \beta_4\, (age_{ij} - 14)_+ + \beta_5\, (age_{ij} - 14)_+ \times group_i$$

Similarly, we can derive

For the unexposed group: Slope for $age < 14$: β_2
Slope for $age \geq 14$: $\beta_2 + \beta_4$

For the exposed group: Slope for $age < 14$: $\beta_2 + \beta_3$
Slope for $age \geq 14$: $(\beta_2 + \beta_3) + (\beta_4 + \beta_5)$

g)

Model i: the group difference is constant over time.

Model ii: the group difference changes linearly with age.

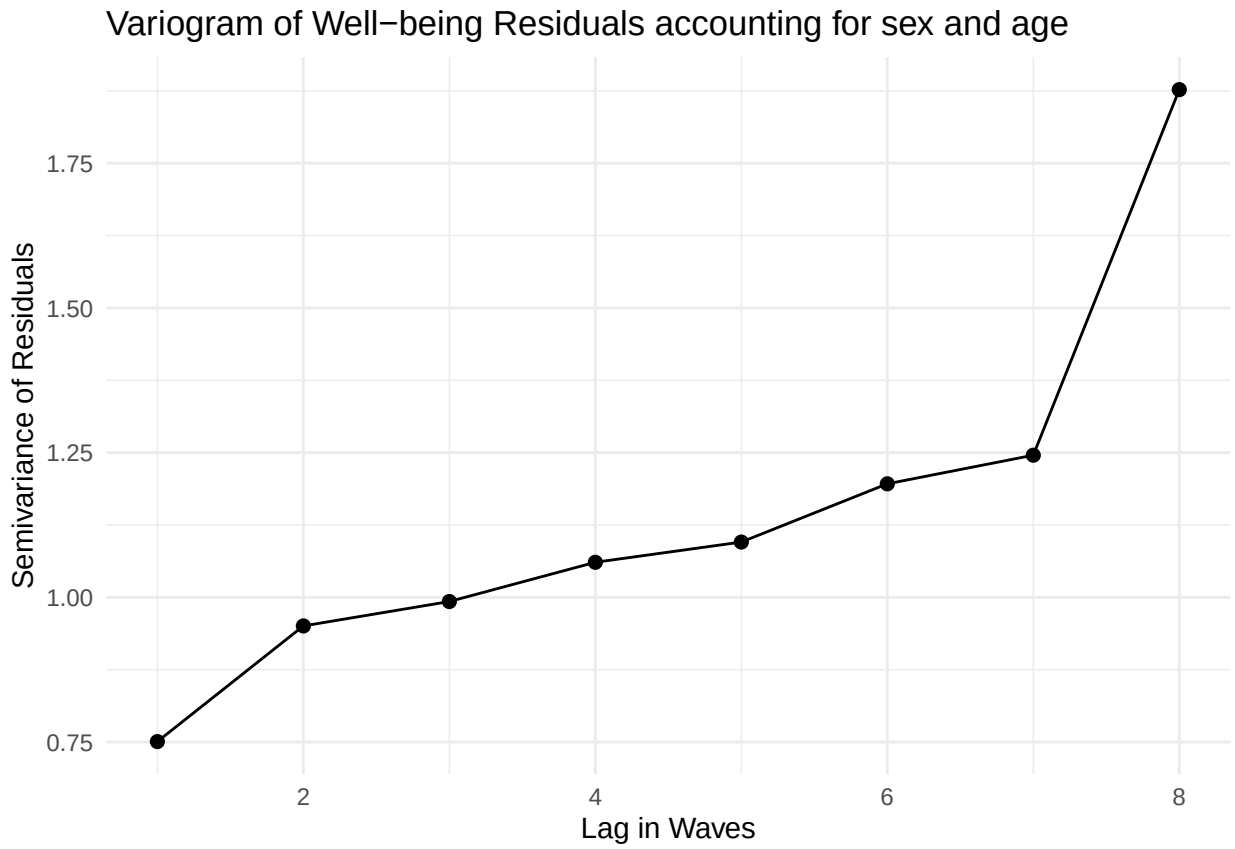
Model iii: the group difference is non-linear.

Problem 2

a)

```
## Rows: 1101 Columns: 5
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): sex_dv
## dbl (4): pidp, wave, age_dv, scghq1_dv
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

b)



c) i

Linear model without accounting for any within-subject correlation.

```
##
## Call:
## lm(formula = fixed_form, data = df, na.action = na.exclude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.011  -3.894  -1.534   1.502  24.314
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         14.05446    0.62544  22.471  < 2e-16 ***
## age_dv              -0.03600    0.01061  -3.395 0.000712 ***
## factor(sex_dv)Male -1.46099    0.35680  -4.095 4.54e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.839 on 1098 degrees of freedom
## Multiple R-squared:  0.02569,    Adjusted R-squared:  0.02391
## F-statistic: 14.47 on 2 and 1098 DF,  p-value: 6.24e-07
```

c) ii

Linear mixed model with a subject-specific random intercept.

```
## Linear mixed-effects model fit by REML
##   Data: df
##        AIC     BIC    logLik
##   6554.943 6579.95 -3272.472
##
## Random effects:
##  Formula: ~1 | pidp
##         (Intercept) Residual
## StdDev:    4.350412 4.009344
##
## Fixed effects:  list(fixed_form)
##                         Value Std.Error  DF   t-value p-value
## (Intercept)        13.521375 1.1207590 912 12.064480  0.0000
## age_dv             -0.021494 0.0190597 912 -1.127719  0.2597
## factor(sex_dv)Male -1.766836 0.7197442 186 -2.454810  0.0150
##  Correlation:
##                    (Intr) age_dv
## age_dv             -0.917
## factor(sex_dv)Male -0.213 -0.038
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -3.5495289 -0.4926014 -0.1234438  0.3428276  4.9056038
##
## Number of Observations: 1101
## Number of Groups: 188
```

c) iii

Linear mixed model with a subject-specific random intercept and serial correlation over data collection waves.

```
## Linear mixed-effects model fit by REML
##   Data: df
##        AIC      BIC    logLik
##   6495.765 6525.772 -3241.882
##
## Random effects:
##  Formula: ~1 | pidp
##         (Intercept) Residual
## StdDev:    4.023875 4.307608
##
## Correlation Structure: ARMA(1,0)
##  Formula: ~wave | pidp
##  Parameter estimate(s):
##      Phi1
## 0.3379215
## Fixed effects:  list(fixed_form)
##                         Value Std.Error  DF   t-value p-value
## (Intercept)        13.422799 1.1425556 912 11.748049  0.0000
```

```
## age_dv             -0.020359 0.0194874 912 -1.044729  0.2964
## factor(sex_dv)Male -1.706848 0.7138426 186 -2.391071  0.0178
##  Correlation:
##                    (Intr) age_dv
## age_dv             -0.921
## factor(sex_dv)Male -0.209 -0.037
##
## Standardized Within-Group Residuals:
##        Min         Q1        Med         Q3        Max
## -3.0878707 -0.4931979 -0.1542192  0.2985998  4.4621208
##
## Number of Observations: 1101
## Number of Groups: 188
```

d)

Table 2: Fixed-effect estimates for age and sex under three covariance structures

Term	OLS	RI	RI + AR(1)
Intercept	14.054 (0.625), <.001	13.521 (1.121), <.001	13.423 (1.143), <.001
age_dv	–0.036 (0.011), .001	–0.021 (0.019), .260	–0.020 (0.019), .296
factor(sex_dv)Male	–1.461 (0.357), <.001	–1.767 (0.720), .015	–1.707 (0.714), .018

We found that the age_dv coefficient appears significant in the OLS model, but it loses significance, and all coefficients' standard errors increase substantially in both the RI and RI + AR(1). This difference arises because OLS treats all 1,100 observations as independent, thereby underestimating uncertainty. By contrast, the mixed-effects models explicitly account for within-subject clustering and serial correlation, which increases the standard errors and yields more reliable inference.

Since we observed an increasing trend in the variogram plotted in (b), we chose the RI + AR(1) model to test whether age or sex is associated with well-being, as it accounts for both serial correlation and within-subject clustering—thereby yielding more reliable inference for our question.

Results from the RI + AR(1) model show that age is not significantly associated with well-being, but there is a statistically significant sex difference: males report lower well-being than females in this cohort.

Appendix: R Code

```
library(tidyverse)
library(nlme)
library(broom.mixed)
library(dplyr)
library(tidyr)
library(ggplot2)
library(splines)
library(knitr)

# Problem 1 a)
df <- read_csv("C:/Users/ncwbr/Desktop/Topeka.csv")
df <- df %>%
  mutate(time = age - age0)

baseline <- df %>%
  filter(abs(time) < 1e-6)

ggplot(baseline, aes(x = age0)) +
  geom_histogram(binwidth = 0.5, color = "black", fill = "lightblue") +
  labs(
    x = "Age at Entry",
    y = "Number of Children",
    title = "Histogram of Age at Entry"
  )

ggplot(baseline, aes(x = age0, y = logFEV1)) +
  geom_point(alpha = 0.7) +
  labs(
    x = "Age at Entry",
    y = "log(FEV1) at Entry",
    title = "log(FEV1) at Entry vs Age at Entry"
  )

ggplot(df, aes(x = age, y = logFEV1)) +
  geom_point(alpha = 0.4) +
  labs(
    x = "Age",
    y = "log(FEV1)",
    title = "log(FEV1) vs Age"
  )

ggplot(df, aes(x = time, y = logFEV1)) +
  geom_point(alpha = 0.4) +
  labs(
    x = "Time from Entry",
    y = "log(FEV1)",
    title = "log(FEV1) vs Time from Entry"
  )

# Problem 1 c)
df$age_dummy <- cut(df$age,
                    breaks = c(-Inf, 8, 10, 12, 14, Inf),
```

```
                  labels = c("<8","8-10","10-12","12-14","14+"))

model1 <- lme(
  fixed = logFEV1 ~ age_dummy,
  random = ~ 1 | id,
  data = df
)

model2 <- lme(
  fixed = logFEV1 ~ age,
  random = ~ 1 | id,
  data = df
)

model3 <- lme(
  fixed = logFEV1 ~ age + I(age^2),
  random = ~ 1 | id,
  data = df
)

df <- df %>%
  mutate(
    a1 = pmax(age - 10, 0),
    a2 = pmax(age - 14, 0)
  )

model4 <- lme(
  fixed = logFEV1 ~ age + a1 + a2,
  random = ~ 1 | id,
  data = df
)

model5 <- lme(
  fixed = logFEV1 ~ bs(age, knots = c(10, 14), degree = 3),
  random = ~ 1 | id,
  data = df
)

age_grid <- seq(min(df$age), max(df$age), length.out = 200)
new_df <- tibble(age = age_grid) %>%

  mutate(
    age_dummy = cut(age,
                    breaks = c(-Inf, 8, 10, 12, 14, Inf),
                    labels = c("<8","8-10","10-12","12-14","14+")),
    a1 = pmax(age - 10, 0),
    a2 = pmax(age - 14, 0)
  )

new_df <- new_df %>%
  mutate(
    step = predict(model1, new_df, level = 0),
    linear = predict(model2, new_df, level = 0),
    quadratic = predict(model3, new_df, level = 0),
```

```
    lspline = predict(model4, new_df, level = 0),
    cspline = predict(model5, new_df, level = 0)
  )

plotdat <- new_df %>%
  pivot_longer(
    cols = step:cspline,
    names_to = "model",
    values_to = "pred_logFEV1"
  )

ggplot() +
  geom_point(
    data = df,
    aes(x = age, y = logFEV1),
    colour = "grey70",
    alpha  = 0.3,
    size   = 0.8
  ) +

  geom_line(
    data = plotdat,
    aes(x = age, y = pred_logFEV1, colour = model),
    size = 1
  ) +
  labs(
    x = "Age",
    y = "log(FEV1)",
    colour = "Fitted Models",
    title = "Fitted Curves"
  ) +
  theme_minimal()

# Problem 1 d)
pred_ages <- tibble(age = c(9, 11, 13, 15)) %>%

  mutate(
    age_dummy = cut(age,
                    breaks = c(-Inf, 8, 10, 12, 14, Inf),
                    labels = c("<8","8-10","10-12","12-14","14+")),
    a1 = pmax(age - 10, 0),
    a2 = pmax(age - 14, 0)
  )

pred_table <- pred_ages %>%
  mutate(
    step  = predict(model1, ., level = 0),
    linear = predict(model2, ., level = 0),
    quadratic = predict(model3, ., level = 0),
    lspline = predict(model4, ., level = 0),
    cspline = predict(model5, ., level = 0)
  ) %>%

  select(age, step:cspline)
```

```
kable(
  pred_table,
  digits = 3,
  col.names = c("Age", "Step", "Linear", "Quadratic", "L-spline", "C-spline"),
  caption = "Predicted mean log(FEV1) at selected ages"
)

# Problem 2 a)
df = read_csv("C:/Users/ncwbr/Desktop/uk_sub.csv")
ggplot(df, aes(x = age_dv,
               y = scghq1_dv,
               group = pidp)) +
  geom_line(alpha = 0.3) +
  labs(
    x = "age",
    y = "scghq1_dv",
    title = "Spaghetti Plot of Well-being by Age"
  ) +
  theme_minimal()

# Problem 2 b)
m0 <- lme(
  fixed = scghq1_dv ~ age_dv + factor(sex_dv),
  random = ~ 1 | pidp,
  data = df,
  na.action = na.exclude
)

vg <- Variogram(
  object  = m0,
  form = ~ wave | pidp,
  resType = "pearson",
  na.rm = TRUE
)

vg_avg <- as_tibble(vg) %>%
  rename(lag = dist) %>%
  group_by(lag) %>%
  summarize(semivar = mean(variog, na.rm = TRUE)) %>%
  ungroup()

ggplot(vg_avg, aes(x = lag, y = semivar)) +
  geom_point(size = 2) +
  geom_line() +
  labs(
    x = "Lag in Waves",
    y = "Semivariance of Residuals",
    title = "Variogram of Well-being Residuals accounting for sex and age"
  ) +
  theme_minimal()

# Problem 2 c)
fixed_form <- scghq1_dv ~ age_dv + factor(sex_dv)
```

```
model1 <- lm(fixed_form, data = df, na.action = na.exclude)

model2 <- lme(fixed = fixed_form,
              random  = ~ 1 | pidp,
              data = df,
              na.action = na.exclude)

model3 <- lme(fixed = fixed_form,
              random = ~ 1 | pidp,
              correlation = corAR1(form = ~ wave | pidp),
              data = df,
              na.action = na.exclude)
summary(model1)
summary(model2)
summary(model3)
```