P8157_hw1_rw2844

Ryan Wei

2023-09-18

Question 1

(a)

(i) The variance for the estimates of the treatment effect $\hat{\gamma}_a$ is:

$$\begin{aligned}
\text{Var}(\hat{\gamma}_a) &= \text{Var}(\overline{Y}_2^A - \overline{Y}_2^B) \\
&= \text{Var}(\frac{2}{K}\sum_{l=1}^{\frac{K}{2}} Y_{l2}^A - \frac{2}{K}\sum_{m=1}^{\frac{K}{2}} Y_{m2}^B) \\
&= (\frac{2}{K})^2 \left[\sum_{l=1}^{\frac{K}{2}} \text{Var}(Y_{l2}^A) + \sum_{m=1}^{\frac{K}{2}} \text{Var}(Y_{m2}^B)\right] \\
&= (\frac{2}{K})^2 K\sigma^2 \\
&= \frac{4}{K}\sigma^2
\end{aligned}$$

(ii) The variance for the estimates of the treatment effect $\hat{\gamma}_b$ is:

$$\begin{aligned}
\text{Var}(\hat{\gamma}_b) &= \text{Var}\left[(\overline{Y}_2^A - \overline{Y}_1^A) - (\overline{Y}_2^B - \overline{Y}_1^B)\right] \\
&= \text{Var}\left[\frac{2}{K}\sum_{l=1}^{\frac{K}{2}}(Y_{l2}^A - Y_{l1}^A) - \frac{2}{K}\sum_{m=1}^{\frac{K}{2}}(Y_{m2}^B - Y_{m1}^B)\right] \\
&= (\frac{2}{K})^2 \left[\left(\sum_{l=1}^{\frac{K}{2}} \text{Var}(Y_{l2}^A) + \sum_{l=1}^{\frac{K}{2}} \text{Var}(Y_{l1}^A) - 2\sum_{l=1}^{\frac{K}{2}} \text{Cov}(Y_{l2}^A, Y_{l1}^A)\right)\right. \\
&\left. + \left(\sum_{m=1}^{\frac{K}{2}} \text{Var}(Y_{m2}^B) + \sum_{m=1}^{\frac{K}{2}} \text{Var}(Y_{m1}^B) - 2\sum_{m=1}^{\frac{K}{2}} \text{Cov}(Y_{m2}^B, Y_{m1}^B)\right)\right] \\
&= (\frac{2}{K})^2 [K\sigma^2 - K\rho\sigma^2 + K\sigma^2 - K\rho\sigma^2] \\
&= \frac{8}{K}(1-\rho)\sigma^2
\end{aligned}$$

(iii) The variance for the estimates of the treatment effect $\hat{\gamma}_c$ is:

$$
\begin{aligned}
\mathrm{Var}(\hat{\gamma}_c) &= \mathrm{Var}(\overline{Y}_1^A - \overline{Y}_2^B) \\
&= \mathrm{Var}(\frac{1}{K}\sum_{l=1}^{K} Y_{l1}^A - \frac{1}{K}\sum_{l=1}^{K} Y_{l2}^B) \\
&= (\frac{1}{K})^2 \left[\sum_{l=1}^{K} \mathrm{Var}(Y_{l1}^A) + \sum_{l=1}^{K} \mathrm{Var}(Y_{l2}^B) - 2\sum_{l=1}^{K} \mathrm{Cov}(Y_{l1}^A, Y_{l2}^B)\right] \\
&= (\frac{1}{K})^2 \left[2K(1-\rho)\sigma^2\right] \\
&= \frac{2}{K}(1-\rho)\sigma^2
\end{aligned}
$$

(iv) The variance for the estimates of the treatment effect $\hat{\gamma}_d$ is:

$$
\begin{aligned}
\mathrm{Var}(\hat{\gamma}_d) &= \mathrm{Var}(\overline{\overline{Y}}^A - \overline{\overline{Y}}^B) \\
&= \mathrm{Var}(\frac{1}{K}\sum_{t=1}^{2}\sum_{l=1}^{\frac{K}{2}} {Y_{lt}}^A - \frac{1}{K}\sum_{t=1}^{2}\sum_{m=1}^{\frac{K}{2}} {Y_{mt}}^B) \\
&= (\frac{1}{K})^2 \,\mathrm{Var}(\sum_{t=1}^{2}\sum_{l=1}^{\frac{K}{2}} {Y_{lt}}^A) + (\frac{1}{K})^2 \,\mathrm{Var}(\sum_{t=1}^{2}\sum_{m=1}^{\frac{K}{2}} {Y_{mt}}^B) \\
&= (\frac{1}{K})^2 \left[\mathrm{Var}(\sum_{l=1}^{\frac{K}{2}} {Y_{l1}}^A) + \mathrm{Var}(\sum_{l=1}^{\frac{K}{2}} {Y_{l2}}^A) + 2\,\mathrm{Cov}(\sum_{l=1}^{\frac{K}{2}} {Y_{l1}}^A, \sum_{l=1}^{\frac{K}{2}} {Y_{l2}}^A)\right. \\
&\left. + \mathrm{Var}(\sum_{m=1}^{\frac{K}{2}} {Y_{m1}}^B) + \mathrm{Var}(\sum_{m=1}^{\frac{K}{2}} {Y_{m2}}^B) + 2\,\mathrm{Cov}(\sum_{m=1}^{\frac{K}{2}} {Y_{m1}}^B, \sum_{m=1}^{\frac{K}{2}} {Y_{m2}}^B)\right] \\
&= (\frac{1}{K})^2 2[K(1+\rho)\sigma^2] \\
&= \frac{2}{K}(1+\rho)\sigma^2
\end{aligned}
$$

(b)

- For **Cross-sectional design**, since each participant are only measured once, the number of people we can enrolled under this design is $K = 300000/500 = 600$. The variance of the estimator under this design is $\text{Var}(\hat{\gamma}_a) = \frac{\sigma^2}{150}$. Since the expression of the variance of the estimator $\hat{\gamma}_a$ does not involve ρ, any $\rho \in \{0.2, 0.5, 0.8\}$ can be chosen in order to minimize uncertainty.
- For **Longitudinal comparison of change from baseline**, since each participant are measured twice, the number of people we can enrolled under this design is $K = 300000/(2 \times 500) = 300$. The variance of the estimator under this design is $\text{Var}(\hat{\gamma}_b) = \frac{2}{75}(1-\rho)\sigma^2$. In this case, we choose $\rho = 0.8$ in order to minimize the variance.
- For **Crossover study**, since each participant are measured twice, the number of people we can enrolled under this design is $K = 300000/(2 \times 500) = 300$. The variance of the estimator under this design is $\text{Var}(\hat{\gamma}_c) = \frac{1}{150}(1-\rho)\sigma^2$. In this case, we also choose $\rho = 0.8$ in order to minimize the variance.
- For **Longitudinal comparison of averages**, since each participant are measured twice, the number of people we can enrolled under this design is $K = 300000/(2 \times 500) = 300$. The variance of the estimator under this design is $\text{Var}(\hat{\gamma}_d) = \frac{1}{150}(1+\rho)\sigma^2$. In this case, we choose $\rho = 0.2$ in order to minimize the variance.

(c)

- For **Cross-sectional design**, since each participant are only measured once, the number of people we can enrolled under this design is $K = 300000/(250 + 250) = 600$. The variance of the estimator under this design is $\text{Var}(\hat{\gamma}_a) = \frac{\sigma^2}{150}$. Since the expression of the variance of the estimator $\hat{\gamma}_a$ does not involve ρ, any $\rho \in \{0.2, 0.5, 0.8\}$ can be chosen in order to minimize uncertainty.
- For **Longitudinal comparison of change from baseline**, since each participant are measured twice, the number of people we can enrolled under this design is $K = 300000/(250 + 2 \times 250) = 400$. The variance of the estimator under this design is $\text{Var}(\hat{\gamma}_b) = \frac{1}{50}(1-\rho)\sigma^2$. In this case, we choose $\rho = 0.8$ in order to minimize the variance.
- For **Crossover study**, since each participant are measured twice, the number of people we can enrolled under this design is $K = 300000/(250+2\times250) = 400$. The variance of the estimator under this design is $\text{Var}(\hat{\gamma}_c) = \frac{1}{200}(1-\rho)\sigma^2$. In this case, we also choose $\rho = 0.8$ in order to minimize the variance.
- For **Longitudinal comparison of averages**, since each participant are measured twice, the number of people we can enrolled under this design is $K = 300000/(250 + 2 \times 250) = 400$. The variance of the estimator under this design is $\text{Var}(\hat{\gamma}_d) = \frac{1}{200}(1+\rho)\sigma^2$. In this case, we choose $\rho = 0.2$ in order to minimize the variance.

Question 2

(a)

Mean Model Exploration

First, figure 1 shows the children height against age, with smooth curved fitted with different methods. We can see that there is a clear trend of increasing height as age increases. More specifically, children heights increase rapidly before 12 and then slow down. From this plot we can also see that there is one child with only one observation (in the lower-left part of the plot).

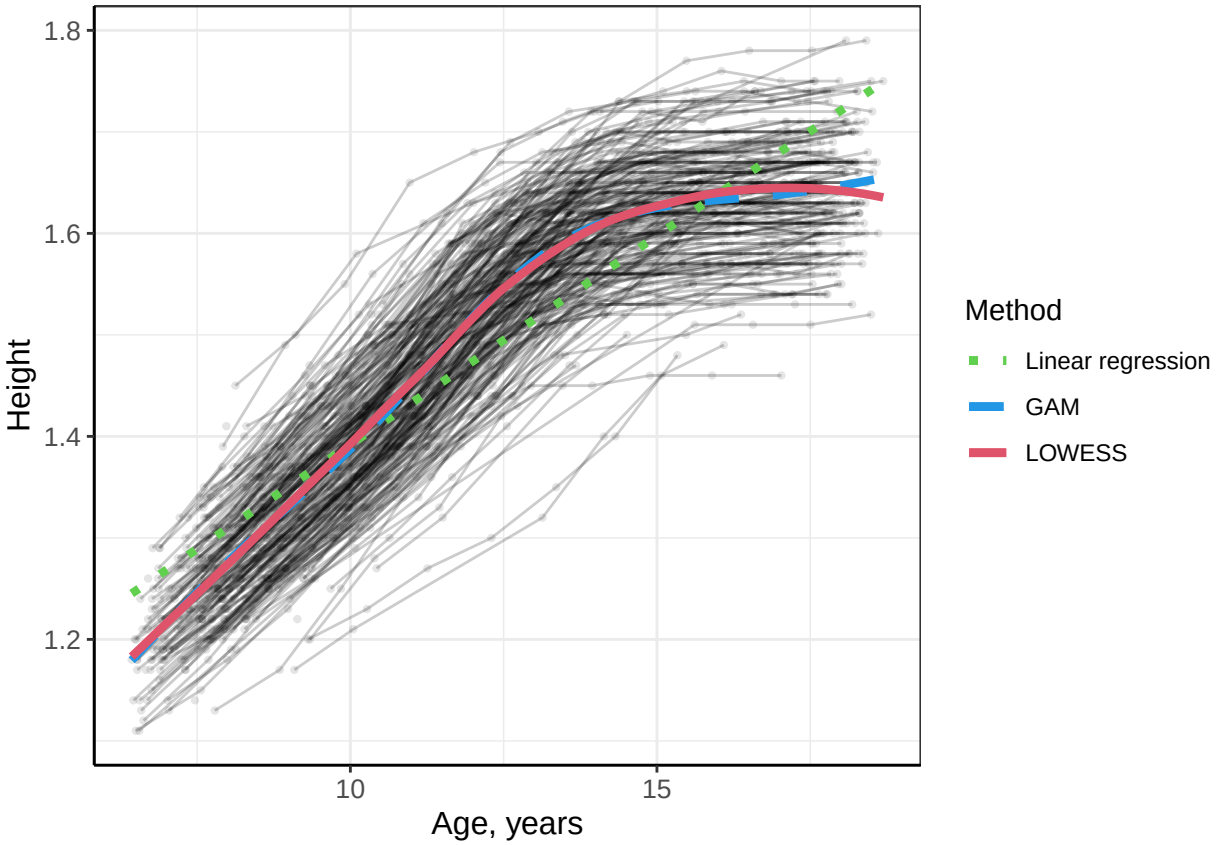

Figure 1: Children's heights with age, with linear, Generalized Additive Model (GAM), and lowess curve.

Secondly, figure 2 shows the children $\log(\text{FEV}_1)$ against age, with smooth curved fitted with different methods. We can see that there is also trend of increasing $\log(\text{FEV}_1)$ as age increases. As we have seen before, children $\log(\text{FEV}_1)$ increase rapidly and constantly before 13 and then slow down. We can not found any obvious outliers in this plot.

Figure 3 and figure 4 are similar to previous two plots, with the x-axis now being the time difference from the beginning of the study. The two plots suggest a population trend that we have seen before. However, by adjusting the time scale, we can now see that there is a single child who has only two observations, with higher height and FEV values at baseline.

Covariance Structure Exploration

The dataset has a continuous time variable, so we divided it into discrete bins (nearest integer time) from 1 to 10 to examine the covariance structure. We then used time as a covariate and $\log(\text{FEV}_1)$ as an outcome to fit a linear marginal model and calculated the mean residual for each individual for each (categorized) time point. The following matrices provide us with the estimated variance and correlation matrix based on

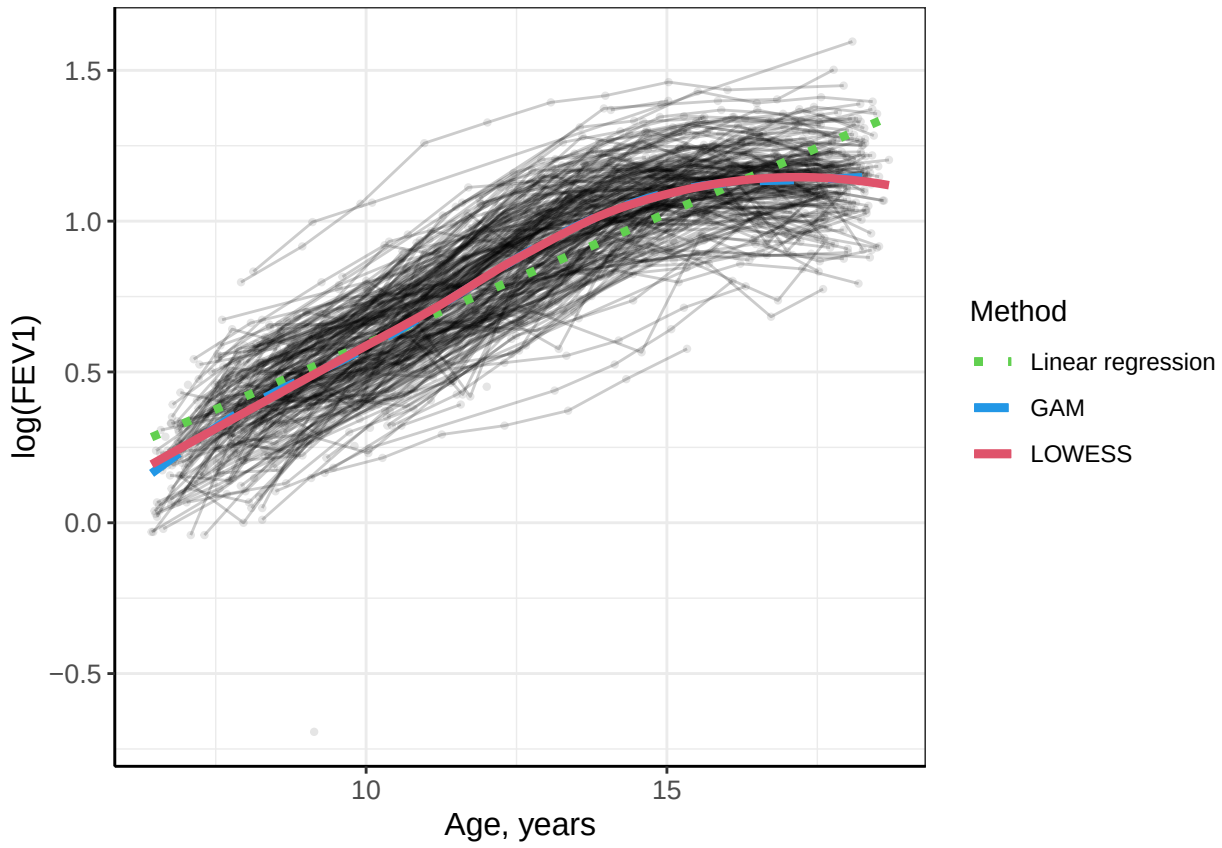

Figure 2: Children's log(FEV_1) with age, with linear, Generalized Additive Model (GAM), and lowess curve.

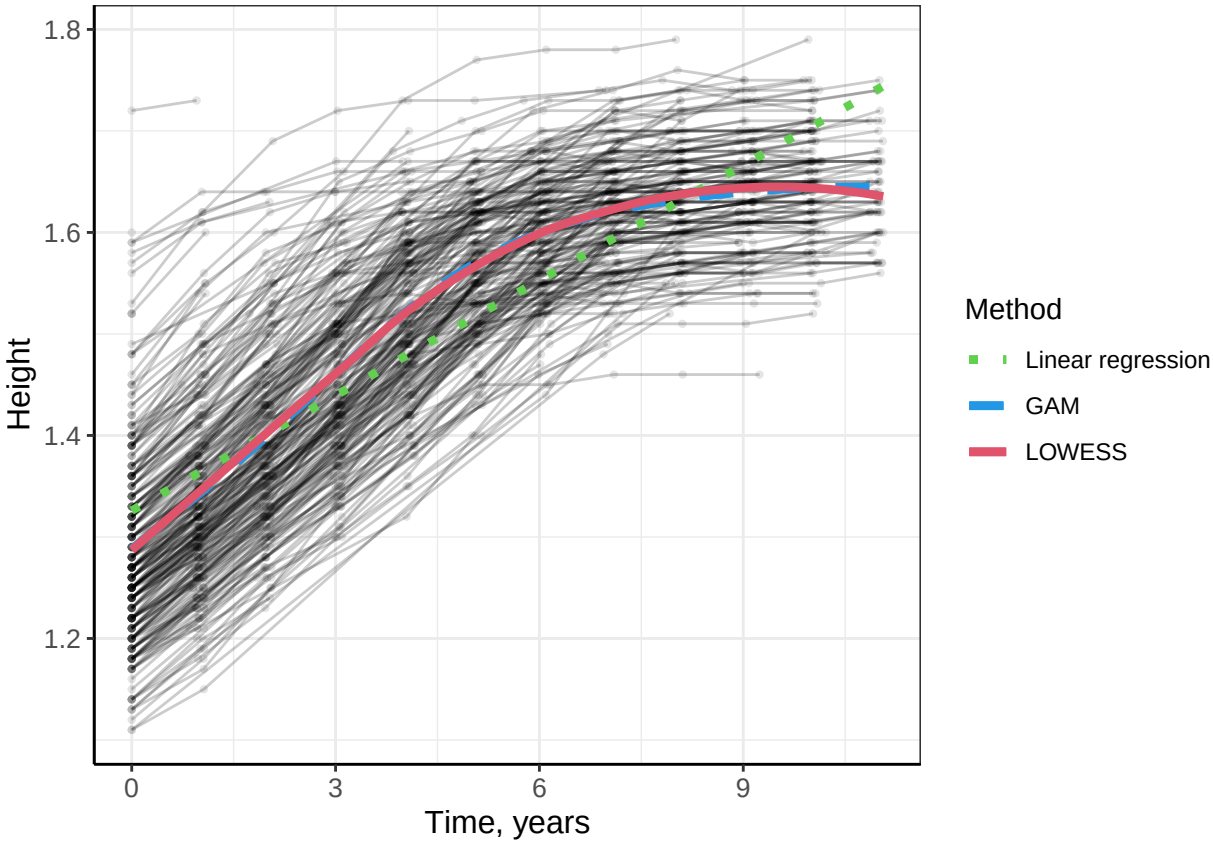

Figure 3: Children's heights over time, with linear, Generalized Additive Model (GAM), and lowess curve.

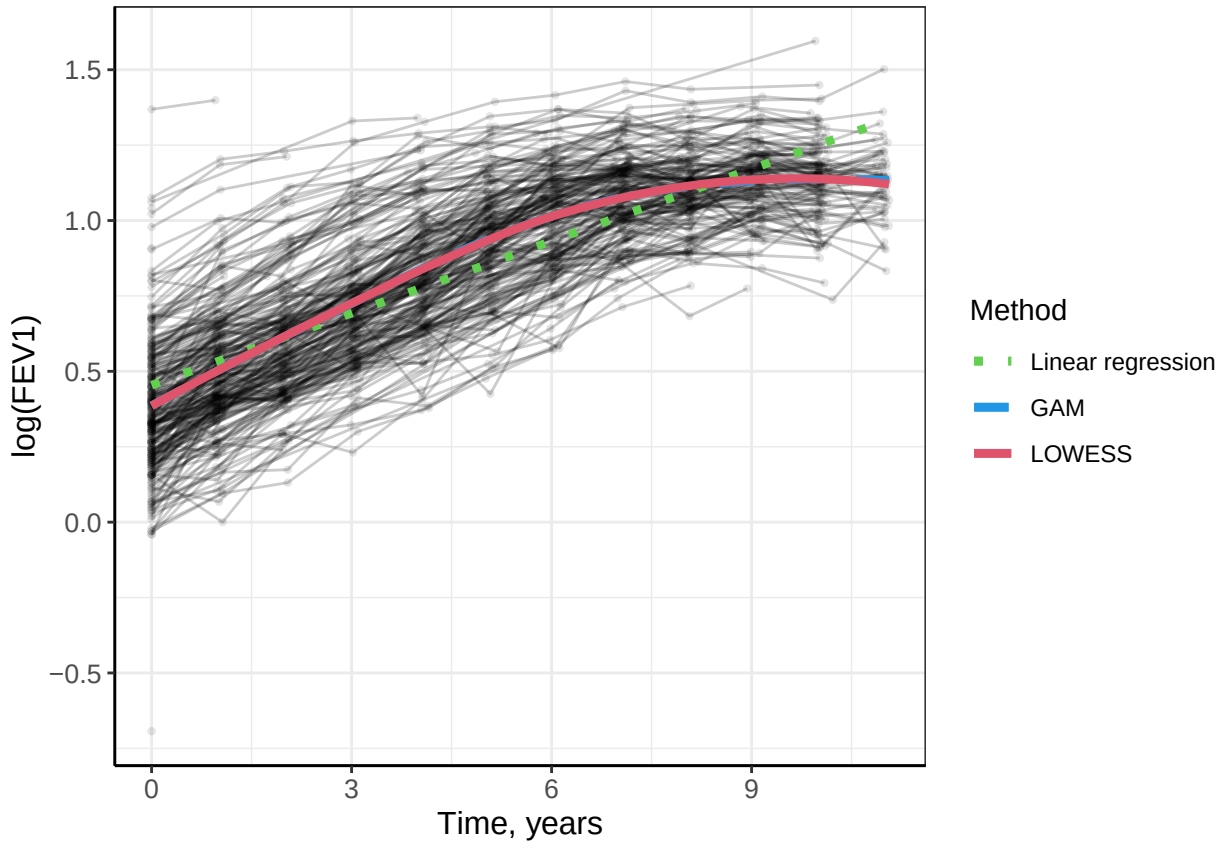

Figure 4: Children's log(FEV_1) over time, with linear, Generalized Additive Model (GAM), and lowess curve.

complete observations. The minimum number of subjects used to estimate the correlation is 83, ensuring that the estimated covariance is reliable.

$$\hat{\sigma}^2 = \begin{bmatrix} 0.219 \\ 0.209 \\ 0.199 \\ 0.198 \\ 0.206 \\ 0.189 \\ 0.177 \\ 0.141 \\ 0.130 \\ 0.121 \\ 0.138 \end{bmatrix},$$

$$\hat{\rho} = \begin{bmatrix} 1 & 0.903 & 0.896 & 0.852 & 0.825 & 0.762 & 0.785 & 0.691 & 0.661 & 0.466 & 0.696 \\ 0.903 & 1 & 0.937 & 0.866 & 0.834 & 0.799 & 0.801 & 0.695 & 0.662 & 0.584 & 0.739 \\ 0.896 & 0.937 & 1 & 0.93 & 0.889 & 0.83 & 0.831 & 0.702 & 0.645 & 0.57 & 0.742 \\ 0.852 & 0.866 & 0.93 & 1 & 0.934 & 0.863 & 0.823 & 0.704 & 0.637 & 0.547 & 0.679 \\ 0.825 & 0.834 & 0.889 & 0.934 & 1 & 0.908 & 0.856 & 0.722 & 0.66 & 0.51 & 0.641 \\ 0.762 & 0.799 & 0.83 & 0.863 & 0.908 & 1 & 0.919 & 0.809 & 0.721 & 0.572 & 0.711 \\ 0.785 & 0.801 & 0.831 & 0.823 & 0.856 & 0.919 & 1 & 0.876 & 0.773 & 0.633 & 0.71 \\ 0.691 & 0.695 & 0.702 & 0.704 & 0.722 & 0.809 & 0.876 & 1 & 0.878 & 0.762 & 0.833 \\ 0.661 & 0.662 & 0.645 & 0.637 & 0.66 & 0.721 & 0.773 & 0.878 & 1 & 0.856 & 0.897 \\ 0.466 & 0.584 & 0.57 & 0.547 & 0.51 & 0.572 & 0.633 & 0.762 & 0.856 & 1 & 0.915 \\ 0.696 & 0.739 & 0.742 & 0.679 & 0.641 & 0.711 & 0.71 & 0.833 & 0.897 & 0.915 & 1 \end{bmatrix}.$$

Since the observation time (age) is continuous and the observed data are not balanced, we also considered the variogram that describes association among repeated observations. Figure 5 shows the variogram for observed measurements, with two components, the total variability in the data (the horizontal dashed line), and the variogram for all time lags in all individuals. We see that the difference between residuals increases as the time difference increases, which means the autocorrelations between two time points decreases as the time difference increases.

Thus far, we have considered displays of the response against time. Next, we consider the relationship between the $\log(\text{FEV}_1)$ and heights other than time. Figure 6 plots the residuals with time trends removed for heights against similar residuals for $\log(\text{FEV}_1)$. We can clearly see that there is a positive correlations between heights and $\log(\text{FEV}_1)$, which suggesting us to investigate more on that.

(b)

Te types of questions that one might be able to address with the Topeka data might be:

1. Estimation of the average $\log(\text{FEV}_1)$ trajectory among all children.
2. Estimation of the average height change trajectory among all children.
3. Testing whether $\log(\text{FEV}_1)$ trajectories are associated with age.
4. Testing whether height change trajectories are associated with age.
5. Prediction of the $\log(\text{FEV}_1)$ or height change trajectory for an individual child.
6. Testing whether height is associated with $\log(\text{FEV}_1)$.

(c)

If we only have a cross-sectional data, for the 5 questions I considered in (b):

1. We can not estimate the average $\log(\text{FEV}_1)$ trajectory since we only have a single observation for each child. However, we can estimate the average $\log(\text{FEV}_1)$ of all children.

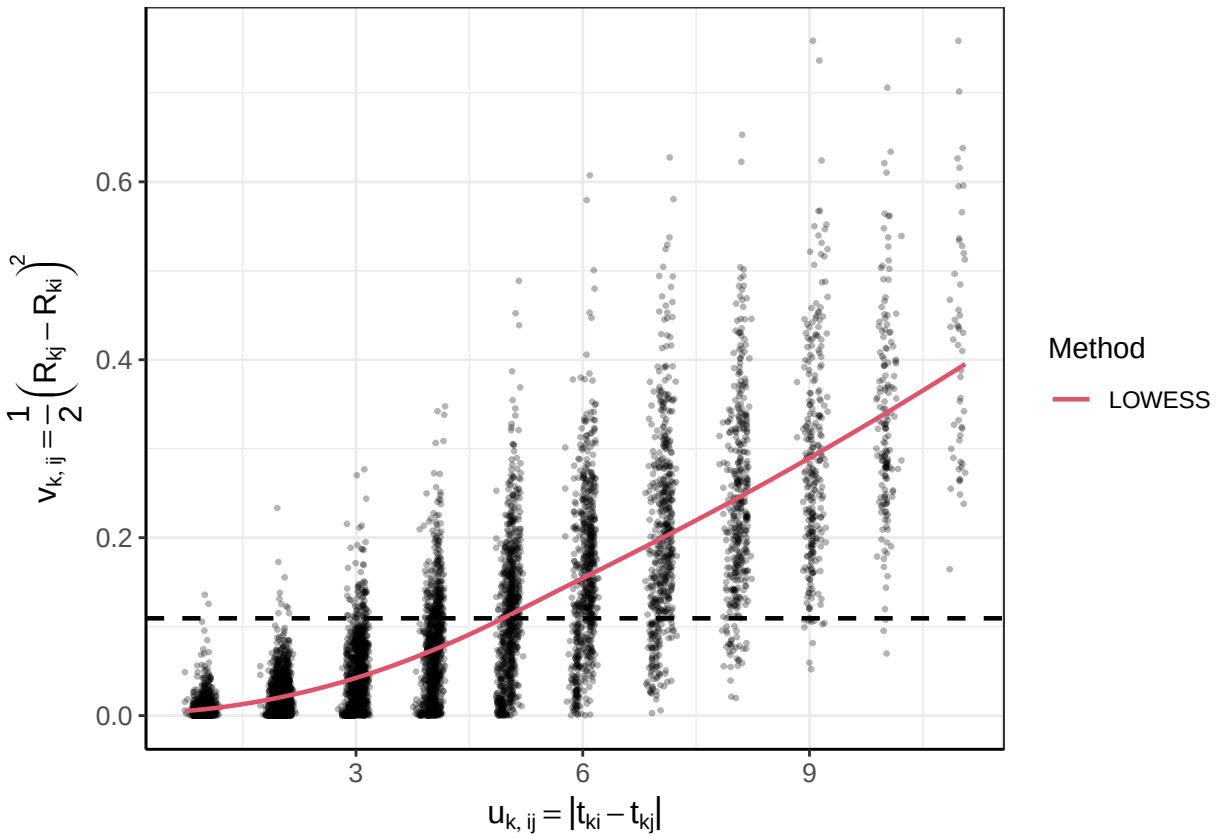

Figure 5: Sample variogram of log(FEV_1) residuals. Horizontal dashed line estimates process variance.

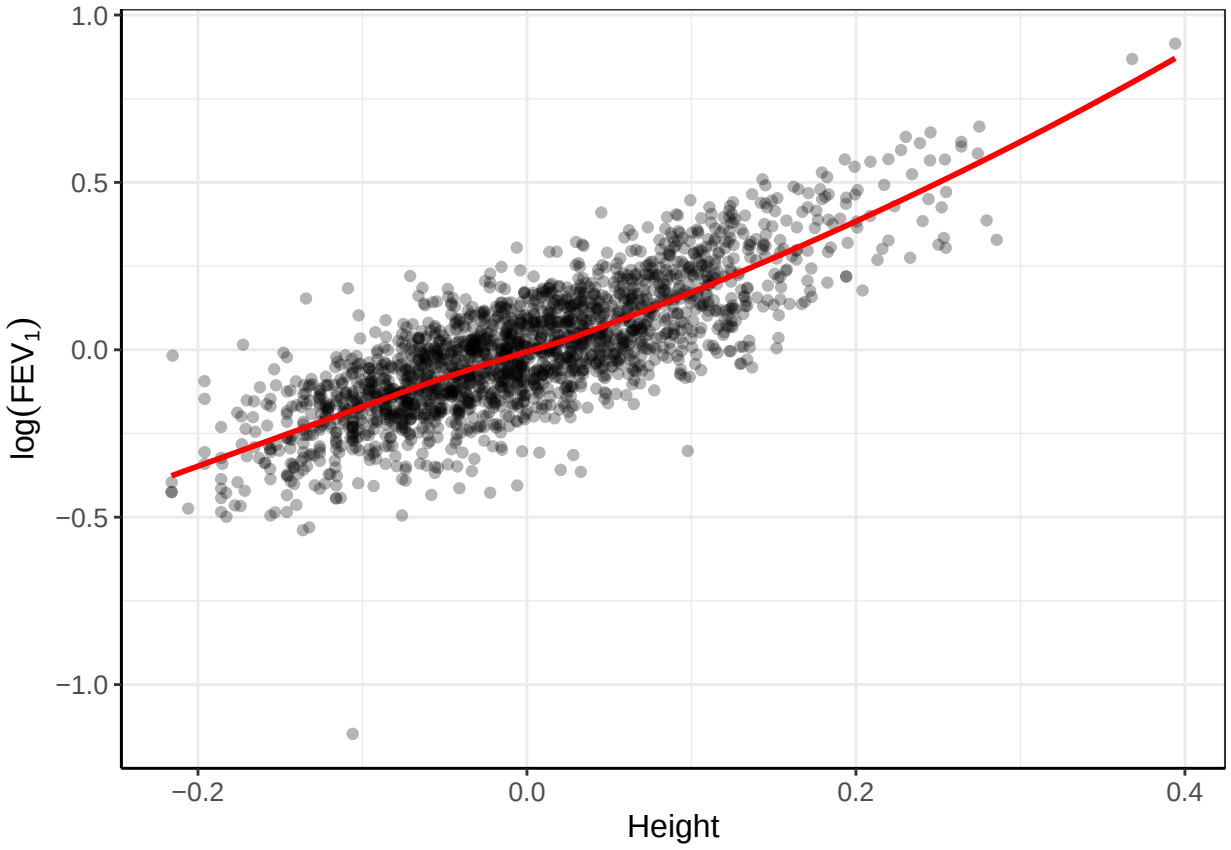

Figure 6: The residuals with time trends removed for height and $\log(FEV_1)$.

2. We cannot estimate the average height change trajectory since we only have one observation per child. However, we can estimate the average height of all children.

3. We can test whether $\log(\text{FEV}_1)$ are associated with age, but not the trajectory, since we only have one observation per child.

4. We can test whether height are associated with age, but not the trajectory, since we only have one observation per child.

5. We can predict the $\log(\text{FEV}_1)$ or height for an individual child, the prediction may solely based on age, since age is the only other variable we can get from a cross-sectional data.

6. We can test whether there is an association between height and $\log(\text{FEV}_1)$, since this test does not involve time variables.

Question 3

(a)

Table 1 shows the descriptive statistics of baseline characteristics of the 266 subjects who has at least one measure before and at least one measure after seroconversion. Notice that the age might be centered,

Table 1: Descriptive statistics of baseline characteristics of the 266 subjects.

	Overall
	(N=266)
Age	
Mean (SD)	2.34 (7.44)
Median [Min, Max]	1.28 [-11.3, 29.1]
Smoking (packs per day)	
Mean (SD)	1.16 (1.52)
Median [Min, Max]	0 [0, 4.00]
Recreational drug use	
No	48 (18.0%)
Yes	218 (82.0%)
Number of sexual partners	
Mean (SD)	6.72 (3.48)
Median [Min, Max]	8.00 [0, 10.0]
CESD scale	
Mean (SD)	10.0 (10.0)
Median [Min, Max]	7.00 [0, 55.0]
CD4+ cell counts	
Mean (SD)	999 (422)
Median [Min, Max]	933 [283, 3180]

(b)

To better understand how CD4+ counts change over time since seroconversion, we first used a spaghetti plot to discover this trend. Figure 7 shows how the CD4+ counts varies over time. From the plot, we can see that the change of CD4+ counts is not linearly in time, and the natural splines regression (with degree of freedom = 3) seems to be a good fit of the trend. Therefore, we decided to use the natural splines regression with degree of freedom equals 3 for the first stage least square analysis (3 coefficients for 3 splines basis).

Table 2 to table 5 show the stage-2 least square analysis on the stage-1 coefficients, $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, respectively. From the tables we can see that the smoking packs per day and recreational drug use is significantly associated with the $\hat{\beta}_0$. Number of partners is significantly associated with the $\hat{\beta}_1$ and $\hat{\beta}_2$. CESD scale is significantly associated with the $\hat{\beta}_2$. Notice that the coefficients from stage-1 analysis do not have a direct interpretation since we used splines regression.

Appendix: Code for this report

```
knitr::opts_chunk$set(echo = FALSE, message = F, warning = F, fig.pos = "H")
library(tidyverse)
library(caret)
library(latex2exp)
library(gstat)
```

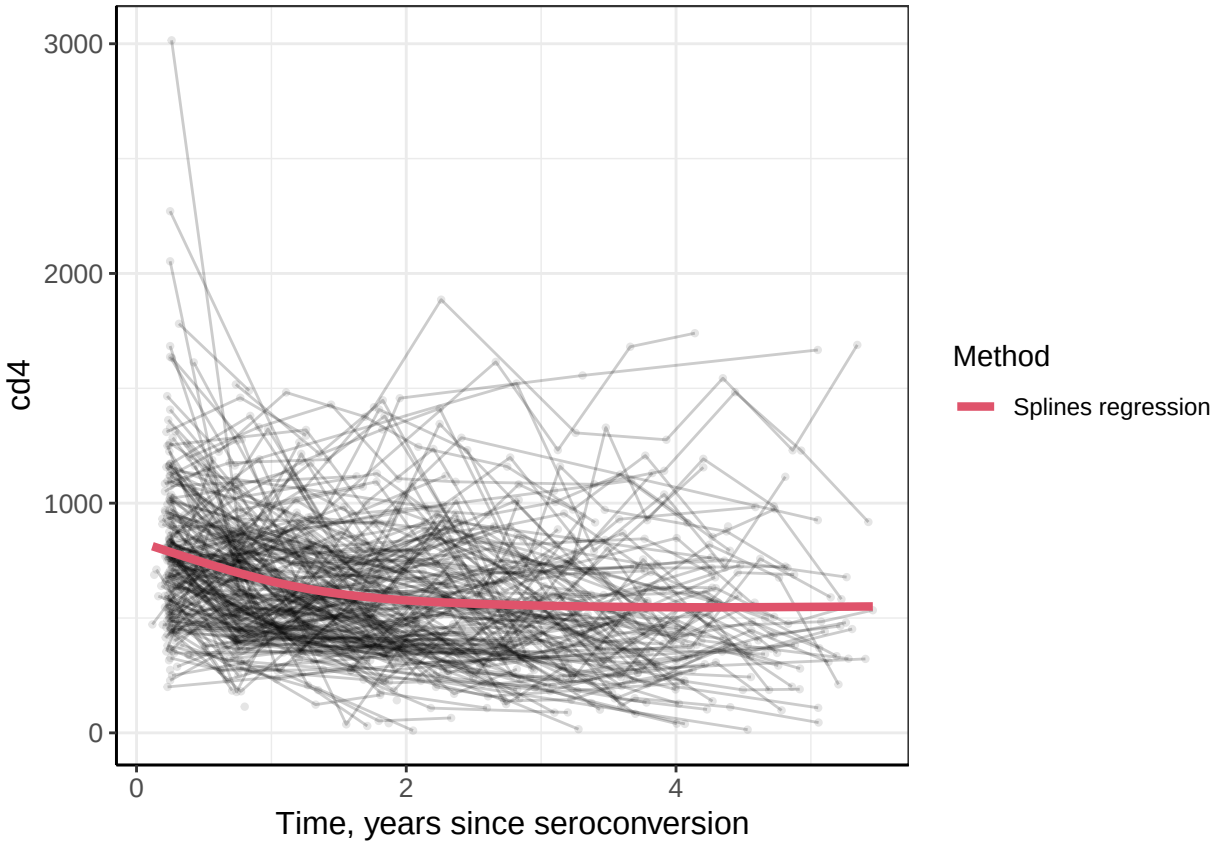

Figure 7: CD4+ counts against time since seroconversion.

Table 2: Stage-2 least square analysis on stage-1 coefficient $\hat{\beta}_0$.

Term	Estimate	Standard Error	P-value
Intercept	860.1278	71.3663	0.0000
Age	0.2467	3.1929	0.9385
Smoking(packs per day)	92.4505	15.3645	0.0000
Recreational drug use	90.7566	61.7815	0.1430
Number of partners	-2.0625	6.6956	0.7583
CESD scale	-5.1296	2.3225	0.0281

Table 3: Stage-2 least square analysis on stage-1 coefficient $\hat{\beta}_1$.

Term	Estimate	Standard Error	P-value
Intercept	731.8375	475.4291	0.1249
Age	-2.3840	21.2703	0.9108
Smoking(packs per day)	187.6186	102.3552	0.0679
Recreational drug use	65.9747	411.5772	0.8728
Number of partners	-104.5020	44.6047	0.0199
CESD scale	18.5617	15.4723	0.2314

Table 4: Stage-2 least square analysis on stage-1 coefficient $\hat{\beta}_2$.

Term	Estimate	Standard Error	P-value
Intercept	-460.7150	224.3198	0.0411
Age	0.5887	9.7034	0.9517
Smoking(packs per day)	-61.9865	47.0782	0.1893
Recreational drug use	12.8548	190.7450	0.9463
Number of partners	-49.5082	20.9908	0.0192
CESD scale	19.9636	7.0708	0.0052

Table 5: Stage-2 least square analysis on stage-1 coefficient $\hat{\beta}_3$.

Term	Estimate	Standard Error	P-value
Intercept	-97.4707	68.3048	0.1552
Age	-1.3970	3.0338	0.6457
Smoking(packs per day)	-20.1239	14.7639	0.1745
Recreational drug use	-87.4603	58.0636	0.1336
Number of partners	-12.4070	6.5546	0.0599
CESD scale	2.9218	2.2573	0.1971

```
library(sp)
library(nlme)
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")
topeka_df = topeka %>% janitor::clean_names() %>%
  mutate(time = age - age_init) %>%
  mutate(delta_height = height - height_init)
baseline_df =
  topeka_df %>%
  group_by(id) %>%
  filter(row_number()==1) %>%
  select(id, height, age, log_fev1)

spaghetti_height <- ggplot(topeka_df, aes(x=age, y=height)) +
  geom_line(alpha = 0.2, aes(group = factor(id))) + geom_point(alpha = 0.1, size = 0.8) +
  geom_smooth(method = "lm", se = FALSE, aes(linetype = "Linear regression", color = "Linear regression
  geom_smooth(method = "gam", se = FALSE, aes(linetype = "GAM", color = "GAM"), linewidth = 1.5) +
  geom_smooth(method = "loess", se = FALSE, aes(linetype = "LOWESS", color = "LOWESS"), linewidth = 1.5
  theme_bw() +
  scale_linetype_manual(name = "Method", values = c(3,2,1),breaks = c( "Linear regression","GAM","LOWESS
  scale_color_manual(name = "Method", values = c(3,4,2),breaks = c( "Linear regression","GAM","LOWESS")
  xlab("Age, years") +
  ylab("Height") +
  theme(
    plot.title = element_text(size = 16, hjust = 0.5),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    axis.text = element_text(size = 10),
    axis.line = element_line(color = "black", size = 0.5),
  )

spaghetti_height
spaghetti_log_fev1 <- ggplot(topeka_df, aes(x=age, y=log_fev1)) +
  geom_line(alpha = 0.2, aes(group = factor(id))) + geom_point(alpha = 0.1, size = 0.8) +
  geom_smooth(method = "lm", se = FALSE, aes(linetype = "Linear regression", color = "Linear regression
  geom_smooth(method = "gam", se = FALSE, aes(linetype = "GAM", color = "GAM"), linewidth = 1.5) +
  geom_smooth(method = "loess", se = FALSE, aes(linetype = "LOWESS", color = "LOWESS"), linewidth = 1.5
  theme_bw() +
  scale_linetype_manual(name = "Method", values = c(3,2,1),breaks = c( "Linear regression","GAM","LOWESS
  scale_color_manual(name = "Method", values = c(3,4,2),breaks = c( "Linear regression","GAM","LOWESS")
  xlab("Age, years") +
  ylab("log(FEV1)") +
  theme(
    plot.title = element_text(size = 16, hjust = 0.5),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    axis.text = element_text(size = 10),
```

```
    axis.line = element_line(color = "black", size = 0.5),
  )

spaghetti_log_fev1
baseline_df =
  topeka_df %>%
  group_by(id) %>%
  filter(row_number()==1) %>%
  select(id, height, age, log_fev1)

spaghetti_height_time <- ggplot(topeka_df, aes(x=time, y=height)) +
  geom_line(alpha = 0.2, aes(group = factor(id))) + geom_point(alpha = 0.1, size = 0.8) +
  geom_smooth(method = "lm", se = FALSE, aes(linetype = "Linear regression", color = "Linear regression
  geom_smooth(method = "gam", se = FALSE, aes(linetype = "GAM", color = "GAM"), linewidth = 1.5) +
  geom_smooth(method = "loess", se = FALSE, aes(linetype = "LOWESS", color = "LOWESS"), linewidth = 1.5
  theme_bw() +
  scale_linetype_manual(name = "Method", values = c(3,2,1),breaks = c( "Linear regression","GAM","LOWES
  scale_color_manual(name = "Method", values = c(3,4,2),breaks = c( "Linear regression","GAM","LOWESS")
  xlab("Time, years") +
  ylab("Height") +
  theme(
    plot.title = element_text(size = 16, hjust = 0.5),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    axis.text = element_text(size = 10),
    axis.line = element_line(color = "black", size = 0.5),
  )

spaghetti_height_time

spaghetti_log_fev1_time <- ggplot(topeka_df, aes(x=time, y=log_fev1)) +
  geom_line(alpha = 0.2, aes(group = factor(id))) + geom_point(alpha = 0.1, size = 0.8) +
  geom_smooth(method = "lm", se = FALSE, aes(linetype = "Linear regression", color = "Linear regression
  geom_smooth(method = "gam", se = FALSE, aes(linetype = "GAM", color = "GAM"), linewidth = 1.5) +
  geom_smooth(method = "loess", se = FALSE, aes(linetype = "LOWESS", color = "LOWESS"), linewidth = 1.5
  theme_bw() +
  scale_linetype_manual(name = "Method", values = c(3,2,1),breaks = c( "Linear regression","GAM","LOWES
  scale_color_manual(name = "Method", values = c(3,4,2),breaks = c( "Linear regression","GAM","LOWESS")
  xlab("Time, years") +
  ylab("log(FEV1)") +
  theme(
    plot.title = element_text(size = 16, hjust = 0.5),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    axis.text = element_text(size = 10),
    axis.line = element_line(color = "black", size = 0.5),
  )

spaghetti_log_fev1_time
spaghetti_delta_height_time <- ggplot(topeka_df, aes(x=time, y=delta_height)) +
  geom_line(alpha = 0.2, aes(group = factor(id))) + geom_point(alpha = 0.1, size = 0.8) +
  geom_smooth(method = "lm", se = FALSE, aes(linetype = "Linear regression", color = "Linear regression
  geom_smooth(method = "gam", se = FALSE, aes(linetype = "GAM", color = "GAM"), linewidth = 1.5) +
```

```
  geom_smooth(method = "loess", se = FALSE, aes(linetype = "LOWESS", color = "LOWESS"), linewidth = 1.5)
  theme_bw() +
  scale_linetype_manual(name = "Method", values = c(3,2,1),breaks = c( "Linear regression","GAM","LOWESS
  scale_color_manual(name = "Method", values = c(3,4,2),breaks = c( "Linear regression","GAM","LOWESS")
  xlab("Time") +
  ylab("Height") +
  theme(
    plot.title = element_text(size = 16, hjust = 0.5),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    axis.text = element_text(size = 10),
    axis.line = element_line(color = "black", size = 0.5),
  )

# spaghetti_delta_height_time
# not showing this plot in the report
baseline_age_plot =
  ggplot(aes(x = age), data = baseline_df) +
  geom_histogram() +
  geom_boxplot(position = "dodge2", outlier.color = "red")+
  theme_bw()+
  theme(
    plot.title = element_text(size = 16, hjust = 0.5),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    axis.text = element_text(size = 10),
    axis.line = element_line(color = "black", size = 0.5),
  )

#baseline_age_plot

baseline_height_plot =
  ggplot(aes(x = height), data = baseline_df) +
  geom_histogram() +
  geom_boxplot(position = "dodge2", outlier.color = "red")+
  theme_bw()+
  theme(
    plot.title = element_text(size = 16, hjust = 0.5),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    axis.text = element_text(size = 10),
    axis.line = element_line(color = "black", size = 0.5),
  )

#baseline_height_plot

baseline_log_fev1_plot =
  ggplot(aes(x = log_fev1), data = baseline_df) +
  geom_histogram() +
  geom_boxplot(position = "dodge2", outlier.color = "red") +
  theme_bw()+
  theme(
    plot.title = element_text(size = 16, hjust = 0.5),
```

```
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    axis.text = element_text(size = 10),
    axis.line = element_line(color = "black", size = 0.5),
  )

#baseline_log_fev1_plot

# not showing these histograms since they are not that informative
tcat <- round(topeka_df$time)
#table(tcat)
tcat <- ifelse(tcat == 11, 10, tcat)
#table(tcat)
fit.fev1.time <- lm(log_fev1 ~ time, data = topeka_df)
resMat.fev1 <- tapply(residuals(fit.fev1.time), list(topeka_df$id, tcat), FUN=mean)
#round(resMat.fev1, 3)
res.Var <- sqrt(diag(cov(resMat.fev1, use="pairwise.complete.obs")))
res.Cov <- cor(resMat.fev1, use="pairwise.complete.obs")
nS <- matrix(NA, nrow=11, ncol=11)
for(i in 1:11){
  for(j in 1:11) nS[i,j] <- nrow(na.omit(resMat.fev1[,c(i,j)]))
}
#nS %>% as.data.frame(row.names = c(1:10))

#write_matex(as.matrix(round(res.Var,3)))
#write_matex(round(res.Cov,3))
tcat <- round(topeka_df$time)
#table(tcat)
tcat <- ifelse(tcat == 11, 10, tcat)
#table(tcat)
fit.height.time <- lm(height ~ time, data = topeka_df)
resMat.height <- tapply(residuals(fit.height.time), list(topeka_df$id, tcat), FUN=mean)
#round(resMat.height, 3)
sqrt(diag(cov(resMat.height, use="pairwise.complete.obs")))
cor(resMat.height, use="pairwise.complete.obs")
library(joineR)
vgm <- variogram(indv=topeka_df$id, time=topeka_df$age, Y=topeka_df$log_fev1)
#plot(vgm, smooth = TRUE, xlab="X-axis label", ylab="y-axix label")

ggplot(aes(x = vt, y = vv), data = as.data.frame(vgm[["svar"]])) + geom_point(size = 0.5,alpha = 0.3) +
  geom_smooth(method = "loess", se = FALSE, aes(linetype = "LOWESS", color = "LOWESS"), linewidth = 0.8
  geom_hline(yintercept = vgm$sigma2, color = "black", linewidth = 0.8, linetype = 2)+
  xlab(TeX("$u_{k,ij} = |t_{ki}-t_{kj}|$")) +
  ylab(TeX("$v_{k,ij} = \\frac{1}{2}\\left(R_{kj} - R_{ki}\\right)^2$")) +
  scale_linetype_manual(name = "Method", values = c(1),breaks = c("LOWESS"))+
  scale_color_manual(name = "Method", values = c(2),breaks = c( "LOWESS"))+
  theme_bw()+
  theme(
    plot.title = element_text(size = 16, hjust = 0.5),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    axis.text = element_text(size = 10),
    axis.line = element_line(color = "black", size = 0.5),
```

```
  )
# exposures other than time
fit.height <- lm(height ~ age, data = topeka_df)
fit.fev1 <- lm(log_fev1 ~ age, data = topeka_df)

res.height <- residuals(fit.height)
res.fev1 <- residuals(fit.fev1)

res.age.plot <-
  ggplot() +
  geom_point(aes(x = res.height, y = res.fev1), alpha = 0.3) +
  geom_smooth(aes(x = res.height, y = res.fev1),method = "loess",se = FALSE, formula = y ~ x, color = "
  theme_bw() + theme(legend.position="none") +
  xlab("Height") +
  ylab(TeX("$\\log($FEV$_1)$")) +
  theme(
    plot.title = element_text(size = 16, hjust = 0.5),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    axis.text = element_text(size = 10),
    axis.line = element_line(color = "black", size = 0.5),
  )

#residuals(fit.height)
#residuals(fit.fev1)
fit.height.time <- lm(height ~ time, data = topeka_df)
fit.fev1.time <- lm(log_fev1 ~ time, data = topeka_df)

res.height.time <- residuals(fit.height.time)
res.fev1.time <- residuals(fit.fev1.time)

topeka_df_residual <-
  topeka_df %>%
  mutate(res.height = res.height.time, res.fev1 = res.fev1.time)

res.time.plot <-
  ggplot() +
  geom_point(aes(x = res.height.time, y = res.fev1.time), alpha = 0.3) +
  geom_smooth(aes(x = res.height.time, y = res.fev1.time),method = "loess", formula = y ~ x,se = FALSE,
  theme_bw() + theme(legend.position="none") +
  xlab("Height") +
  ylab(TeX("$\\log($FEV$_1)$")) +
  theme(
    plot.title = element_text(size = 16, hjust = 0.5),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    axis.text = element_text(size = 10),
    axis.line = element_line(color = "black", size = 0.5),
  )
res.time.plot
load("../../datasets/MACS/MACS.RData")
load("../../datasets/MACS/MACS-VL.RData")
```

```r
#n_distinct(macs$id) # 369
#n_distinct(macsVL$id)
# target number of subject K* = 266
macs_pre_id <-
  macs %>%
  group_by(id) %>%
  mutate(pre_sero = ifelse(time < 0 & time >= -0.5, 1, 0)) %>%
  ungroup() %>%
  filter(pre_sero == 1) %>%
  select(id)


macs_post_id <- macs %>%
  group_by(id) %>%
  mutate(post_sero= ifelse(time >= 0 , 1, 0)) %>%
  ungroup() %>%
  filter(post_sero == 1) %>%
  select(id) %>%
  unique()

macs_pre_post <- macs %>%
  filter(id %in% macs_pre_id$id & id %in% macs_post_id$id) %>%
  filter(time >= -0.5) %>%
  mutate(baseline = ifelse(time <= 0, T, F))

#n_distinct(macs_pre_post$id)
library(table1)
library(kableExtra)
macs_pre_post_tbl = macs_pre_post
label(macs_pre_post_tbl$age) = "Age"
label(macs_pre_post_tbl$packs) =  "Smoking (packs per day)"
macs_pre_post_tbl$drug = as.factor(ifelse(macs_pre_post_tbl$drug == 1, "Yes", "No"))
label(macs_pre_post_tbl$drug) = "Recreational drug use"
label(macs_pre_post_tbl$partners) = "Number of sexual partners"
label(macs_pre_post_tbl$cesd) = "CESD scale"
label(macs_pre_post_tbl$cd4) = "CD4+ cell counts"
table1(~ age + packs + drug + partners + cesd + cd4, data = macs_pre_post_tbl %>% filter(baseline == TRU
spaghetti_cd4 <- ggplot(macs_pre_post %>% filter(baseline == FALSE), aes(x=time, y=cd4)) +
  geom_line(alpha = 0.2, aes(group = factor(id))) + geom_point(alpha = 0.1, size = 0.8) +
  geom_smooth(method = "lm", formula = y ~ splines::ns(x,3),se = FALSE, aes(linetype = "Splines regressi
  theme_bw() +
  scale_linetype_manual(name = "Method", values = c(1),breaks = c( "Splines regression"))+
  scale_color_manual(name = "Method", values = c(2),breaks = c( "Splines regression")) +
  xlab("Time, years since seroconversion") +
  ylab("cd4") +
  theme(
    plot.title = element_text(size = 16, hjust = 0.5),
    axis.title.x = element_text(size = 12),
    axis.title.y = element_text(size = 12),
    axis.text = element_text(size = 10),
    axis.line = element_line(color = "black", size = 0.5),
  )
spaghetti_cd4
```

```r
# K is the number of subjects
K = n_distinct(macs_pre_post$id)
betaMat <- data.frame(age=rep(NA, K), packs=rep(NA, K), drug=rep(NA, K), partners=rep(NA, K), cesd=rep(

macs_pre_post$id = as.numeric(macs_pre_post$id)

for(k in 1:K){
  id.k = unique(macs_pre_post$id)[k]
  temp.k <- macs_pre_post[macs_pre_post$id == id.k,]
  temp.k.pre <- temp.k[temp.k$time <= 0,]
  temp.k.post <- temp.k
  temp.k.post$time <- ifelse(temp.k.post$time<=0,0,temp.k.post$time)
  fit.k <- lm(cd4 ~ splines::ns(time,3), data=temp.k.post)
  betaMat[k,] <- c(temp.k.pre$age, temp.k.pre$packs, temp.k.pre$drug, temp.k.pre$partners, temp.k.pre$c
}

library(gtsummary)
fit.beta0 <-lm(beta0 ~ age + packs + drug + partners + cesd , data=betaMat)
fit.beta1 <- lm(beta1 ~ age + packs + drug + partners + cesd, data=betaMat)
fit.beta2 <- lm(beta2 ~ age + packs + drug + partners + cesd, data=betaMat)
fit.beta3 <- lm(beta3 ~ age + packs + drug + partners + cesd, data=betaMat)
#tbl_regression(fit.beta0) %>% knitr::kable(col.names =  c("Covariates"))
broom::tidy(fit.beta0) %>% mutate(term = c("Intercept", "Age", "Smoking(packs per day)", "Recreational
broom::tidy(fit.beta1) %>% mutate(term = c("Intercept", "Age", "Smoking(packs per day)", "Recreational
broom::tidy(fit.beta2) %>% mutate(term = c("Intercept", "Age", "Smoking(packs per day)", "Recreational
broom::tidy(fit.beta3) %>% mutate(term = c("Intercept", "Age", "Smoking(packs per day)", "Recreational
```