P8157_hw2_rw2844

Ryan Wei

2023-10-16

Question 1

(a)

First, figure 1 shows the CD4+ cell counts against time, with smooth curved fitted with different methods. We can see that there is a clear trend of decreasing cell counts as time since seroconversion increases.

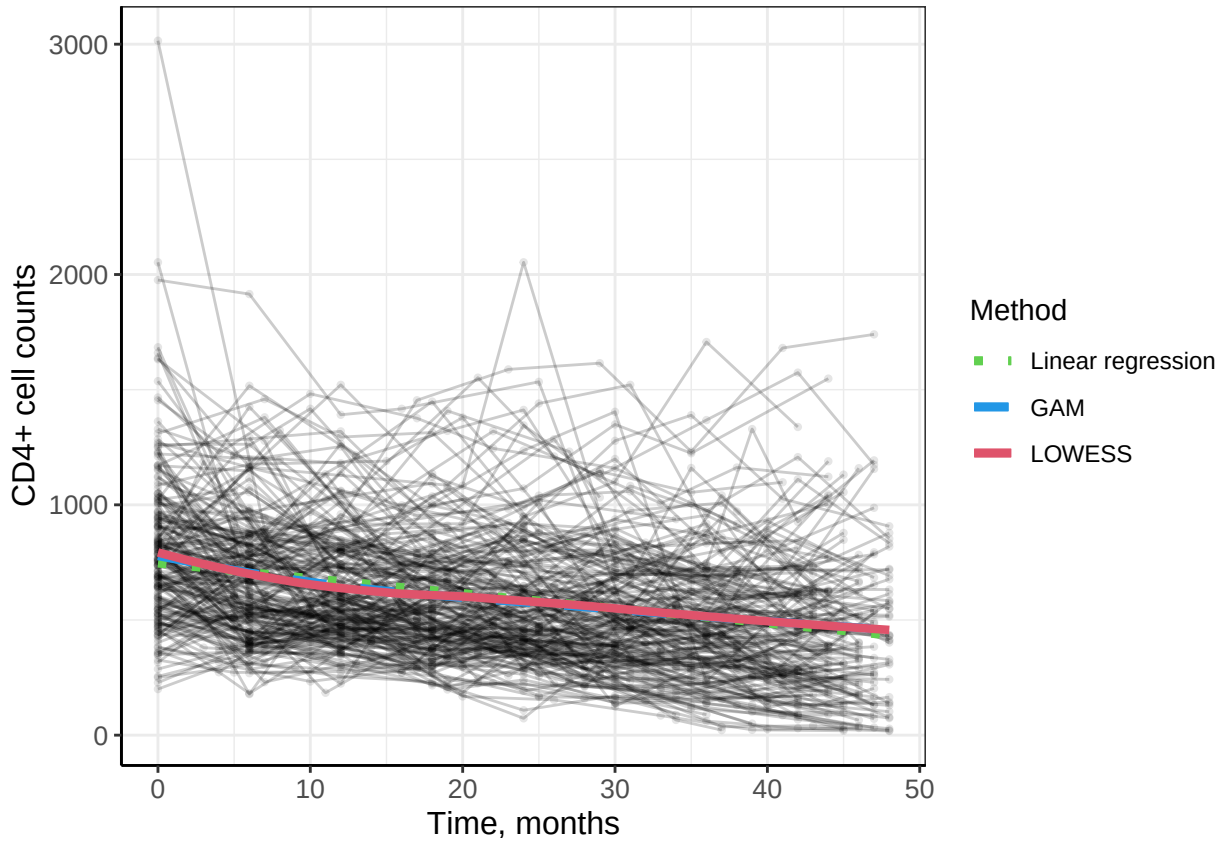
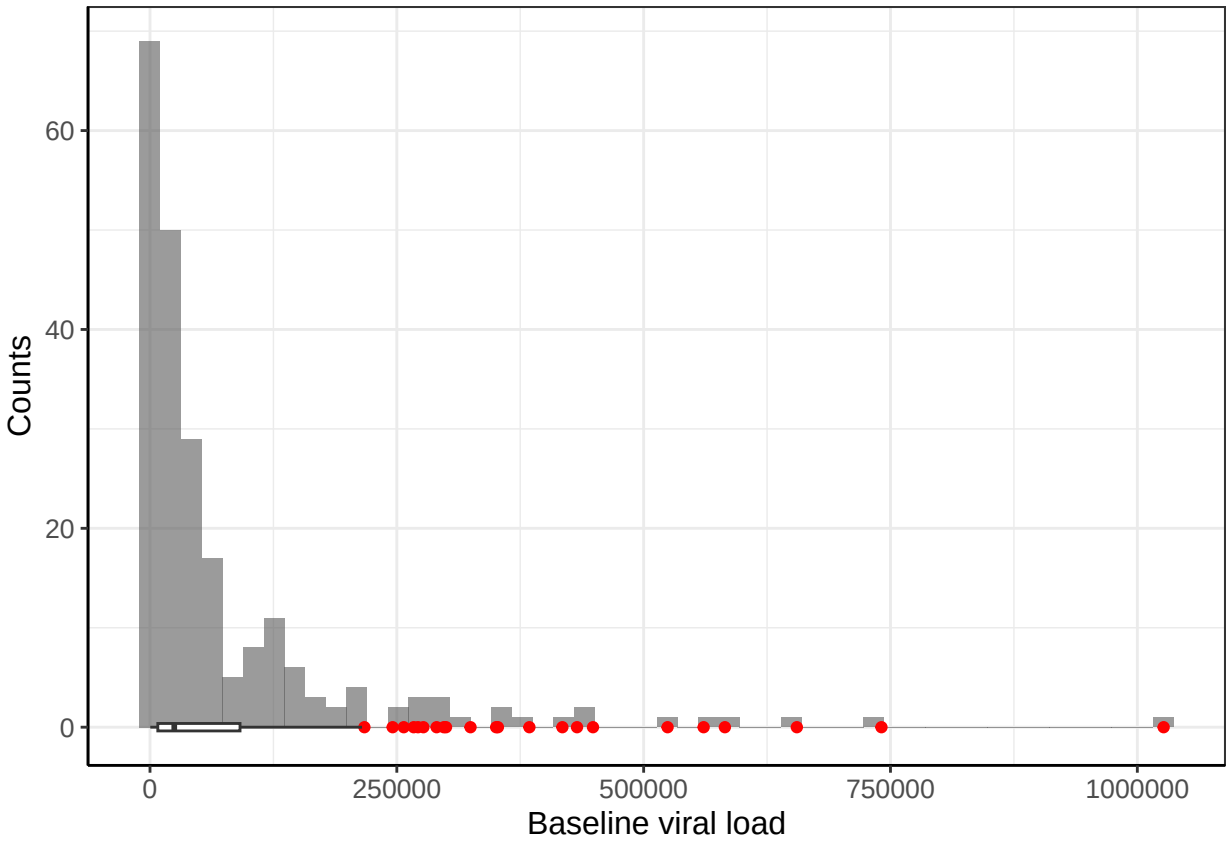


Figure 1: CD4+ cell counts over time, with linear, Generalized Additive Model (GAM), and lowess curve.

Next, to investigate the scientific question of interest, that is, the relationship between baseline viral load and the rate of decline of CD4+ cell counts, I plot the distribution of baseline viral load and the logarithm of baseline viral load in figure **??** and figure **??**, respectively. From the histograms we can see that the distribution of raw baseline viral load is highly right-skewed and the absolute value of it is large with high variability. After log transformation, the distribution of log(baseline viral load) is approximately normal with no outliers. Therefore, I decided to conduct the following analysis based on this transformed value.


Counts
60
40
20
0
0
250000
500000
750000
1000000
Baseline viral load

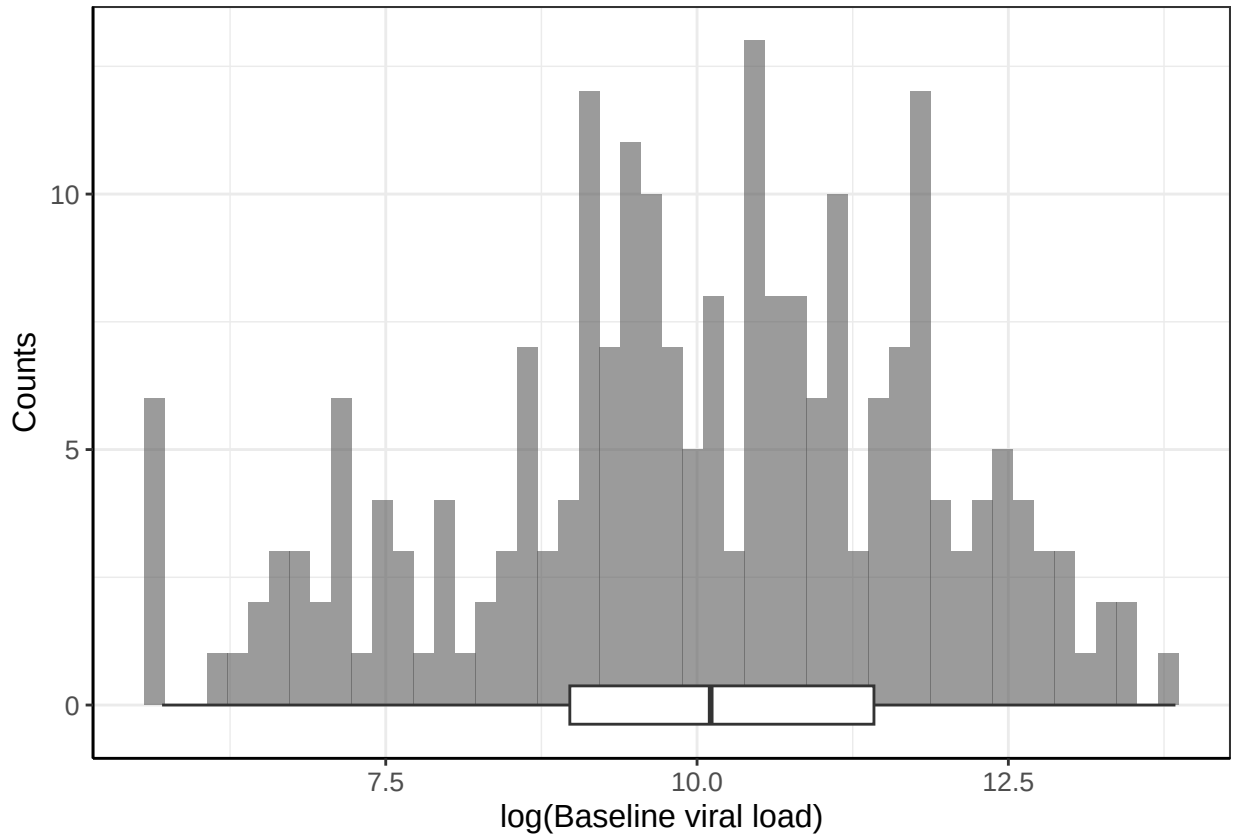


Figure 2 and figure 3 show the scatterplots of the baseline CD4+ cell counts and the average CD4+ cell counts change rate (counts per month) against the baseline viral load. Here, the average CD4+ cell counts change rate is defined as the difference between the CD4+ cell counts at the end of the observation and the CD4+ cell counts at the beginning of the observation, divided by the total observation months. From both plots we can see that there is a trend for CD4+ cell counts to decline with elevated baseline viral loads, which is biologically plausible.

Figure 4 shows the residuals of the CD4+ cell counts and the logarithm of baseline viral load with time trend removed. Again, we can see that there is clearly a trend for CD4+ cell counts to decline with elevated baseline viral loads, which is biologically plausible.

(b)

The dataset has a time variable recorded in monthly intervals ranging from 0 to 48. In order to investigate the covariance structure, I decided to recode the time variable to intervals of six months, resulting in values ranging from 0 to 8. Then I used this recoded time as a covariate and CD4+ cell counts 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 complete observations. The minimum number of subjects used to estimate the correlation is 85, ensuring that the estimated covariance is reliable.

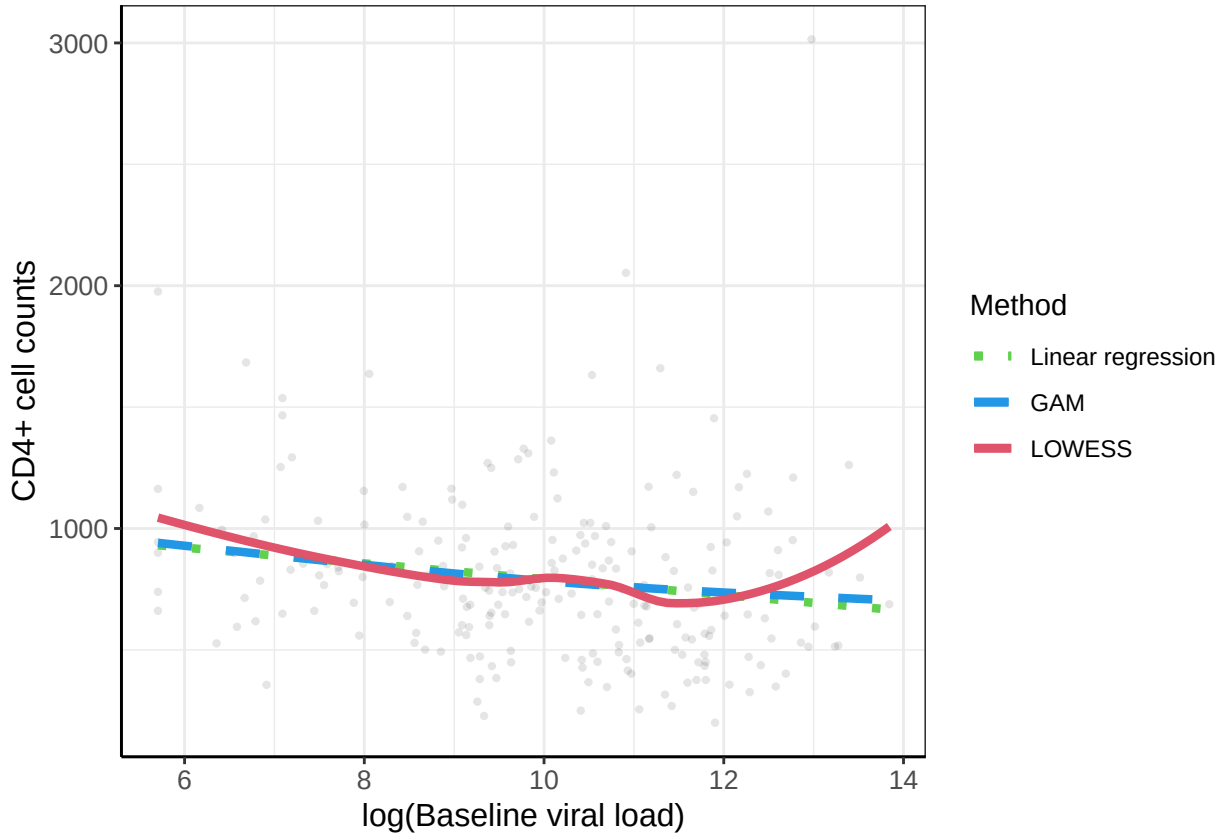


Figure 2: Scatterplot of baseline CD4+ cell counts versus log(baseline viral load), with linear, Generalized Additive Model (GAM), and lowess curve.

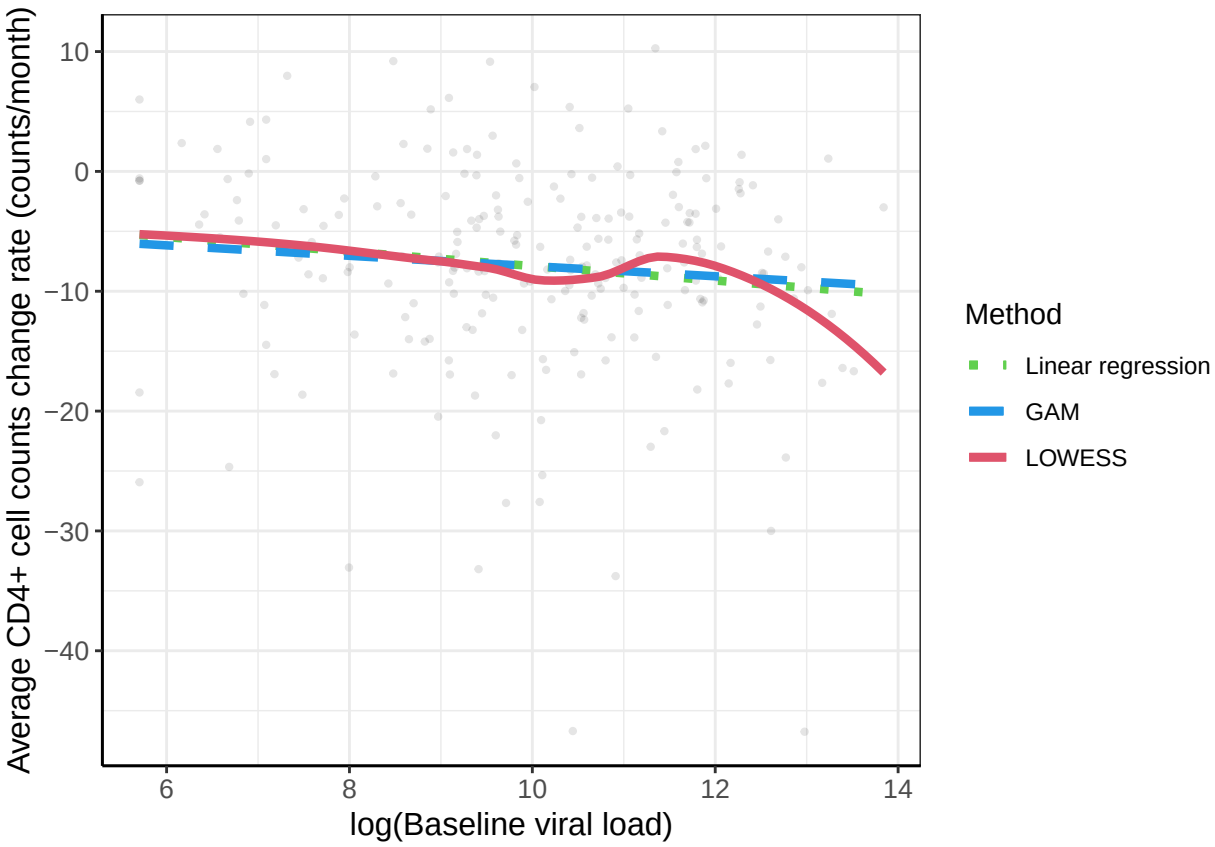


Figure 3: Scatterplot of average CD4+ cell counts change rate versus log(baseline viral load), with linear, Generalized Additive Model (GAM), and lowess curve.

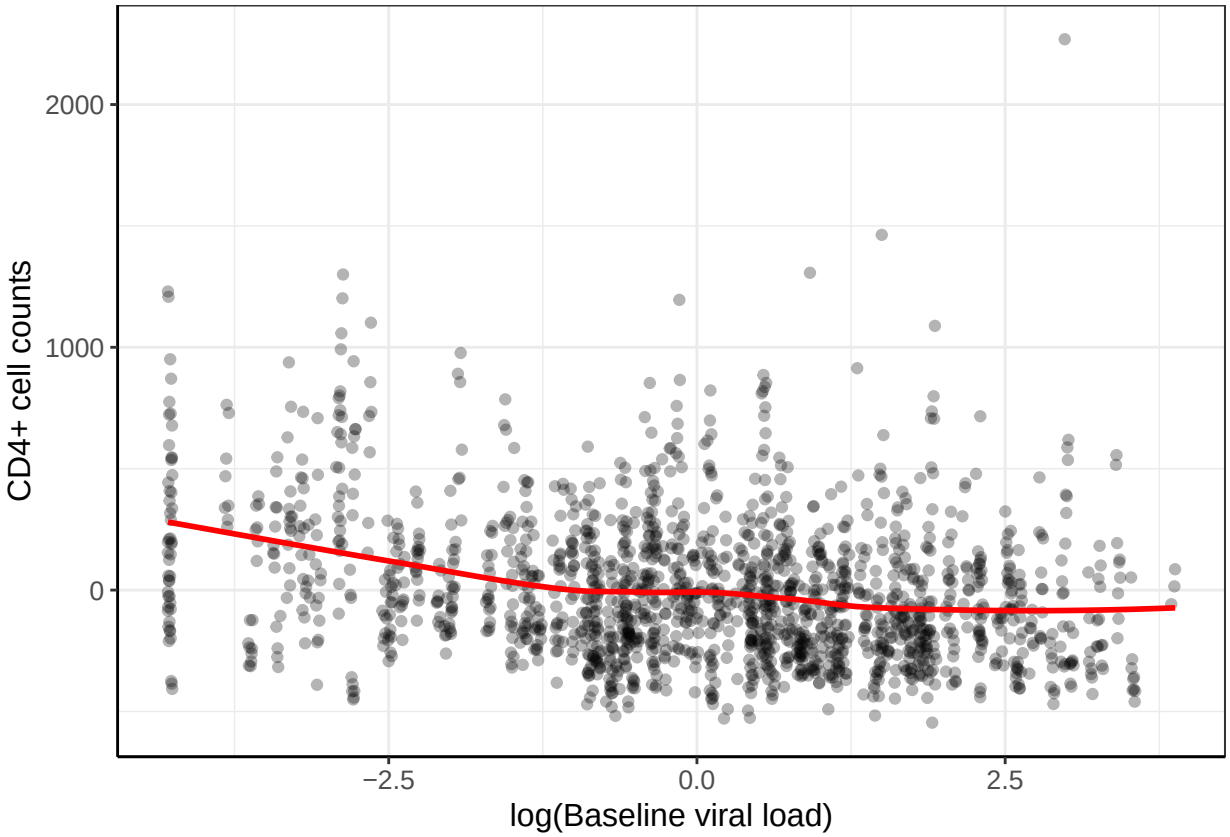


Figure 4: The residuals with time trends removed for CD4+ cell counts and log(Baseline viral load).

$$\hat{\sigma}^2 = \begin{bmatrix} 346.24 \\ 294.026 \\ 259.777 \\ 276.908 \\ 301.191 \\ 292.348 \\ 284.811 \\ 317.881 \\ 313.373 \end{bmatrix}, \hat{\rho} = \begin{bmatrix} 1 & 0.573 & 0.509 & 0.486 & 0.549 & 0.509 & 0.431 & 0.43 & 0.411 \\ 0.573 & 1 & 0.695 & 0.724 & 0.65 & 0.555 & 0.577 & 0.549 & 0.54 \\ 0.509 & 0.695 & 1 & 0.742 & 0.642 & 0.604 & 0.617 & 0.565 & 0.486 \\ 0.486 & 0.724 & 0.742 & 1 & 0.762 & 0.736 & 0.701 & 0.663 & 0.647 \\ 0.549 & 0.65 & 0.642 & 0.762 & 1 & 0.716 & 0.67 & 0.694 & 0.655 \\ 0.509 & 0.555 & 0.604 & 0.736 & 0.716 & 1 & 0.736 & 0.792 & 0.759 \\ 0.431 & 0.577 & 0.617 & 0.701 & 0.67 & 0.736 & 1 & 0.806 & 0.781 \\ 0.43 & 0.549 & 0.565 & 0.663 & 0.694 & 0.792 & 0.806 & 1 & 0.865 \\ 0.411 & 0.54 & 0.486 & 0.647 & 0.655 & 0.759 & 0.781 & 0.865 & 1 \end{bmatrix}.$$

From the matrices above, we can see that there isn't an obvious trend in the change of empirical standard deviations with time. The estimated correlation matrix doesn't show an obvious pattern of the correlations between observations either. Therefore, compond symmetric structure may be appropriate.

Since the observation time (month) 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.

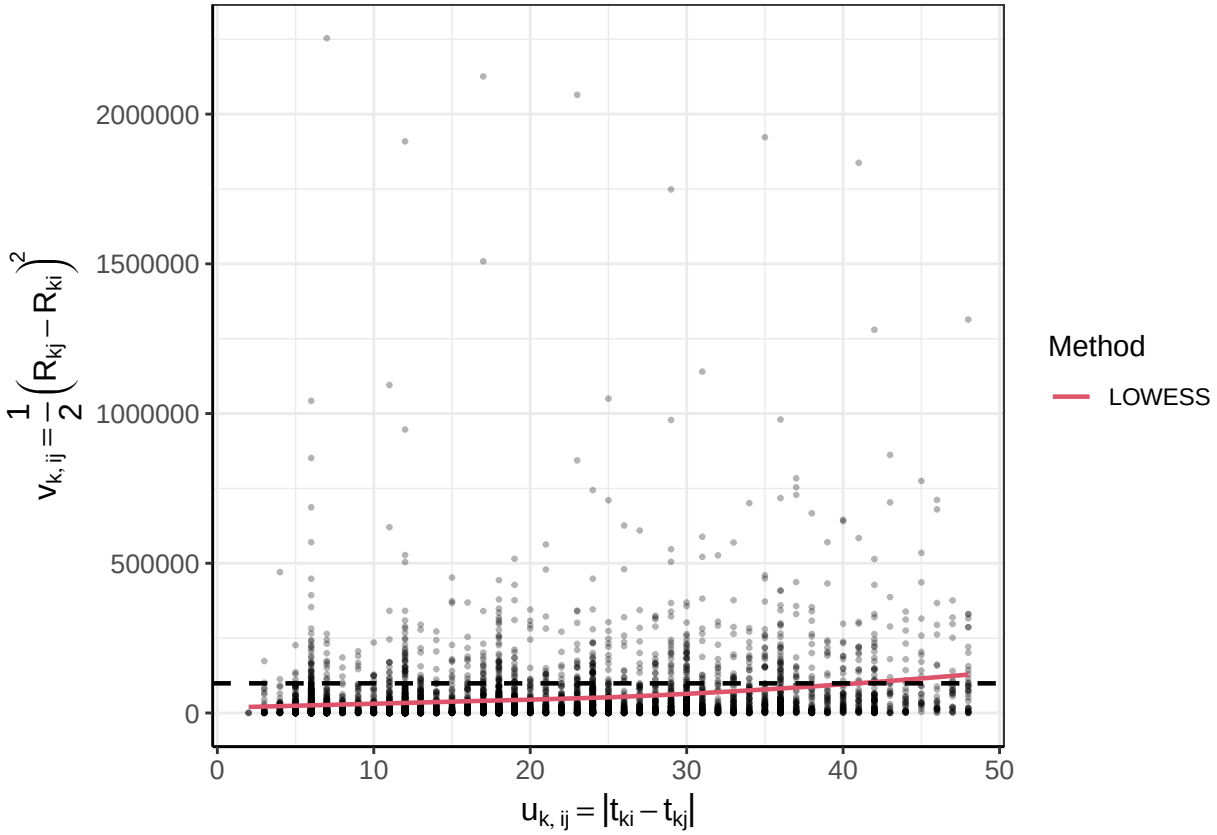


Figure 5: Sample variogram of CD4+ cell counts residuals. Horizontal dashed line estimates process variance.

(c)

For this question, I first applied log transformation on baseline viral load and fitted the following model

$$[Y_{ki}] = \beta_0 + \beta_1 t_{ki} + \beta_2 x_k + \beta_3 t_{ki} x_k E,$$

where x_k is log(Baseline viral load) and t_{ki} is time since seroconversion in months.

Table 1: General least squares coefficient estimates, fit by ML.

Term	Estimate	Standard Error	T-value	P-value
Intercept	1108.101	91.325	12.134	0.000
Month	-3.086	1.760	-1.753	0.080
log(Viral load)	-35.703	8.992	-3.970	0.000
Month×(Viral load)	-0.381	0.174	-2.190	0.029

Table 1 and table 2 show the general least squares coefficient estimates fitted by maximum likelihood (ML) method and restricted maximum likelihood (REML) method. We can see that REML always gives us a estimator with larger standard error. The regression coefficient of month, log(baseline viral load) and their interaction term are both significant. Table 3 shows the goodness of fit statistics of the two method, we can see that REML gives us a lower AIC as well as a lower BIC, which implies a better fit.

Table 2: General least squares coefficient estimates, fit by REML.

Term	Estimate	Standard Error	T-value	P-value
Intercept	1108.097	91.564	12.102	0.000
Month	-3.087	1.760	-1.754	0.080
log(Viral load)	-35.702	9.016	-3.960	0.000
Month×(Viral load)	-0.381	0.174	-2.191	0.029

Table 3: Goodness of fit comparison between ML and REML.

Method	AIC	BIC	logLik
ML	22928.55	22961.12	-11458.28
REML	22917.38	22949.94	-11452.69

For the interpretation of the coefficients, I will pick the results from ML since both method give us a similar coefficient estimates.

Interpretations:

- The CD4+ cell counts decrease by 3.09 for each additional month after seroconversion, holding everything else the same.
- The CD4+ cell counts decrease by 35.70 for one unit increase in log(baseline viral load), holding everything else the same.
- The CD4+ cell counts decrease by 0.38 for each additional month after seroconversion and one unit increase in log(baseline viral load), holding everything else the same.

(d)

The quantiles of the baseline viral load in the population is shown in the table 4. Therefore, I categorized baseline viral load into four categories based on percentile of baseline viral load, corresponding to $0\% - 25\%$,

$25\% - 50\%$, $50\% - 75\%$, and $75\% - 100\%$, respectively, which are Group A: $300 \leq$ baseline viral load < 7928, Group B: $7928 \leq$ baseline viral load < 24573, Group C: $24573 \leq$ baseline viral load < 91195 and Group D: baseline viral load ≥ 91195. Then, I fitted the following model

$$[Y_{ki}] = \beta_0 + \beta_1 t_{ki} + \sum_{j=2}^{4} \beta_{2j} x_k(j) + \sum_{j=2}^{4} \beta_{3j} t_{ki} x_k(j)$$

Table 4: Quantiles of the baseline viral load.

0%	25%	50%	75%	100%
300	7928	24573	91195	1026656

Table 5 shows the coefficient estimates from the general least squares model fitted by REML. From the table we can see that the month is significantly associated with the CD4+ decline rate, as well as baseline viral load groups. The interactions between month and baseline viral load groups are significant except for the interaction of month with group D.

Table 5: General least squares coefficient estimates with categorized baseline viral load, fit by REML.

Term	Estimate	Standard Error	T-value	P-value
Intercept	859.552	33.222	25.873	0.000
Month	-5.656	0.625	-9.052	0.000
Viral load category B	-113.801	46.842	-2.429	0.015
Viral load category C	-133.764	46.918	-2.851	0.004
Viral load category D	-184.821	46.852	-3.945	0.000
Month × Viral load category B	-1.660	0.889	-1.867	0.062
Month × Viral load category C	-2.109	0.875	-2.410	0.016
Month × Viral load category D	-1.121	0.906	-1.238	0.216

Interpretations:

- The CD4+ cell counts decrease by 5.66 on average for each additional month after seroconversion, holding everything else the same.
- The CD4+ cell counts decrease by 113.80 on average for subjects in group B compared to subjects in group A, holding everything else the same.
- The CD4+ cell counts decrease by 133.76 on average for subjects in group C compared to subjects in group A, holding everything else the same.
- The CD4+ cell counts decrease by 184.82 on average for subjects in group D compared to subjects in group A, holding everything else the same.
- The CD4+ cell counts decrease by 1.66 on average for subject in group B compared to group A, for each additional month after seroconversion, holding everything else the same.
- The CD4+ cell counts decrease by 2.11 on average for subject in group C compared to group A, for each additional month after seroconversion, holding everything else the same.
- The CD4+ cell counts decrease by 1.21 on average for subject in group D compared to group A, for each additional month after seroconversion, holding everything else the same. This effect is not significant.

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)
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/MACS/MACS-VL.RData")
#length(unique(macsVL$id))
macsVL$log_vload = log(macsVL$vload)
sphaghetti_cd4 =
  ggplot(macsVL, aes(x=month, y=cd4)) +
  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, months") +
  ylab("CD4+ cell counts") +
  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),
  )

sphaghetti_cd4
macsVL_baseline = macsVL %>% group_by(id) %>% slice(1)

macsVL_baseline %>%
  ggplot(aes(x = vload)) +
  geom_histogram(alpha = 0.6, bins = 50) +
  geom_boxplot(position = "dodge2", outlier.color = "red")+
  theme_bw()+
  xlab(TeX("Baseline viral load")) +
```

```
  ylab(TeX("Counts")) +
  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),
  )
macsVL_baseline %>%
  ggplot(aes(x = log(vload))) +
  geom_histogram(alpha = 0.6, bins = 50) +
  geom_boxplot(position = "dodge2", outlier.color = "red")+
  theme_bw()+
  xlab(TeX("log(Baseline viral load)")) +
  ylab(TeX("Counts")) +
  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),
  )
ggplot(aes(x = log(vload), y = cd4), data = macsVL_baseline) + 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(TeX("log(Baseline viral load)")) +
  ylab("CD4+ cell counts") +
  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),
  )
macsVL%>%
  group_by(id) %>%
  arrange(month, .by_group = T) %>%
  mutate(cd4_change = last(cd4) - first(cd4)) %>%
  mutate(cd4_change_rate_avg = cd4_change/(last(month)- first(month))) %>%
  slice(1) %>%
  ungroup() %>%
  ggplot(aes(x = log_vload, y = cd4_change_rate_avg))+
  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(TeX("log(Baseline viral load)")) +
  ylab("Average CD4+ cell counts change rate (counts/month)") +
  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),
  )
fit.cd4.time <- lm(cd4 ~ month, data = macsVL)
fit.log_vload.time <- lm(log_vload ~  month, data = macsVL)

res.cd4.time <- residuals(fit.cd4.time)
res.log_vload.time <- residuals(fit.log_vload.time)

res.time.plot <-
  ggplot() +
  geom_point(aes(x = res.log_vload.time, y = res.cd4.time), alpha = 0.3) +
  geom_smooth(aes(x = res.log_vload.time, y = res.cd4.time),method = "loess", formula = y ~ x,se = FALSE
  theme_bw() + theme(legend.position="none") +
  xlab(TeX("$\\log$(Baseline viral load)")) +
  ylab(TeX("CD4+ cell counts")) +
  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
macsVL$half_year = round(macsVL$month/6)
tcat <- macsVL$half_year
#table(macsVL$half_year)
#quantile(tcat,probs = seq(0, 1, 0.1))
#table(tcat)
fit.cd4.time <- lm(cd4 ~ month, data = macsVL)
resMat.cd4 <- tapply(residuals(fit.cd4.time), list(macsVL$id, tcat), FUN=mean)
#round(resMat.fev1, 3)
res.Var <- sqrt(diag(cov(resMat.cd4, use="pairwise.complete.obs")))
res.Cov <- cor(resMat.cd4, use="pairwise.complete.obs")
nS <- matrix(NA, nrow=9, ncol=9)
for(i in 1:9){
  for(j in 1:9) nS[i,j] <- nrow(na.omit(resMat.cd4[,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))
library(joineR)
vgm <- variogram(indv=macsVL$id, time=macsVL$month, Y=macsVL$cd4)
#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),
  )
fit.gls.ML = gls(cd4 ~ month * log_vload, method="ML", data=macsVL, corr=corCompSymm(form = ~ 1 | id))
fit.gls.REML <- gls(cd4 ~ month * log_vload, method="REML", data=macsVL, corr=corCompSymm(form = ~ 1 | i

#summary(fit.gls.ML)
coef(summary(fit.gls.ML)) %>%
  as.data.frame() %>%
  mutate(term = row.names(.)) %>%
  mutate(term = c("Intercept", "Month", "$\\log$(Viral load)", "Month$\\times$(Viral load)")) %>%
  select(term, Value, `Std.Error`, `t-value`, `p-value`) %>%
  rename("Term" = "term", "Estimate" = "Value", "Standard Error" = "Std.Error", "T-value" = "t-value",
  knitr::kable(format = "latex", digits = 3, escape = F, booktabs = TRUE, row.names = F, caption = "Gen

#summary(fit.gls.REML)
coef(summary(fit.gls.REML)) %>%
  as.data.frame() %>%
  mutate(term = row.names(.)) %>%
  mutate(term = c("Intercept", "Month", "$\\log$(Viral load)", "Month$\\times$(Viral load)")) %>%
  select(term, Value, `Std.Error`, `t-value`, `p-value`) %>%
  rename("Term" = "term", "Estimate" = "Value", "Standard Error" = "Std.Error", "T-value" = "t-value",
  knitr::kable(format = "latex", digits = 3, escape = F, booktabs = TRUE, row.names = F, caption = "Gen
fit.gls.ML.summary = summary(fit.gls.ML)
fit.gls.REML.summary = summary(fit.gls.REML)

model.comparison.tab <-
  tibble(
    Method = c("ML", "REML"),
    AIC = c(fit.gls.ML.summary$AIC, fit.gls.REML.summary$AIC),
    BIC = c(fit.gls.ML.summary$BIC, fit.gls.REML.summary$BIC),
    logLik = c(fit.gls.ML.summary$logLik, fit.gls.REML.summary$logLik),
  )

model.comparison.tab %>%
  knitr::kable(format = "latex", digits = 3, escape = F, booktabs = TRUE, row.names = F, caption = "Goo
t(quantile(macsVL_baseline$vload)) %>%
  as.data.frame() %>%
  knitr::kable(format = "latex", digits = 4, escape = F, booktabs = TRUE, caption = "Quantiles of the ba
  kable_styling(latex_options = "hold_position")
```

```r
macsVL$vload_cat = cut(macsVL$vload, c(300, 7928, 24573, 91195, 1026657), right = FALSE, labels = c("A"
fit.gls.cat.REML = gls(cd4 ~ month * vload_cat, method="REML", data=macsVL, corr=corCompSymm(form = ~ 1
#summary(fit.gls.cat.REML)

coef(summary(fit.gls.cat.REML)) %>%
  as.data.frame() %>%
  mutate(term = row.names(.)) %>%
  mutate(term = c("Intercept", "Month", "Viral load category B", "Viral load category C", "Viral load ca
  select(term, Value, `Std.Error`,`t-value`, `p-value`) %>%
  rename("Term" = "term", "Estimate" = "Value", "Standard Error" = "Std.Error", "T-value" = "t-value",
  knitr::kable(format = "latex", digits = 3, escape = F, booktabs = TRUE, row.names = F, caption = "Gen
```