```
cd4 = read.table("C:/Users/LEGION/OneDrive - Duke University/610/cd4.dat", header = TRUE)
cd4$pid = as.factor(cd4$pid)
```

3(a)

```
library(lme4)
library(dplyr)
library(tibble)
lmem <- lmer(cd4 ~ as.factor(trt) + time + as.factor(trt)*time + (time|pid), data=cd4, REML=FAL
SE)
summary(lmem)
```

```
## Linear mixed model fit by maximum likelihood ['lmerMod']
## Formula: cd4 ~ as.factor(trt) + time + as.factor(trt) * time + (time |
##     pid)
##    Data: cd4
##
##      AIC       BIC    logLik deviance df.resid
##   3119.4    3159.2   -1551.7   3103.4     1064
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -5.0924 -0.3978  0.0061  0.4101  5.0207
##
## Random effects:
##  Groups   Name        Variance Std.Dev. Corr
##  pid      (Intercept) 1.9083   1.3814
##           time        0.3309   0.5753   -0.06
##  Residual             0.5147   0.7174
## Number of obs: 1072, groups:  pid, 250
##
## Fixed effects:
##                     Estimate Std. Error t value
## (Intercept)          4.82628    0.13671  35.303
## as.factor(trt)1     -0.12792    0.18934  -0.676
## time                -0.30623    0.09871  -3.102
## as.factor(trt)1:time -0.08943    0.13496  -0.663
##
## Correlation of Fixed Effects:
##             (Intr) as.()1 time
## as.fctr(t)1 -0.722
## time        -0.238  0.172
## as.fctr()1:  0.174 -0.243 -0.731
```

```
confint(lmem, level = 0.95)
```

```
##                          2.5 %      97.5 %
## .sig01               1.2474912   1.5336904
## .sig02              -0.3015835   0.2245120
## .sig03               0.4227833   0.7290044
## .sigma               0.6800689   0.7581767
## (Intercept)          4.5570316   5.0950726
## as.factor(trt)1     -0.5002298   0.2448533
## time                -0.5022743  -0.1086288
## as.factor(trt)1:time -0.3565953   0.1774465
```

Confident intervals:

- Treatment: [−0.5002298,0.2448533]
- Time: [−0.5022743,−0.1086288]
- Interaction of treatment and time: [−0.3565953,0.1774465]

To find evidence for evidence for a treatment effect, we shall look at each effect separately.

The treatment effect (trt) alone has a negative estimate of -0.12798. However, if we look at the 95% confidence interval, we see that 0 is included in this interval, which implies that the effect is not statistically significant at the 95% confidence level. This means that we do not have a strong evidence that the treatment alone has a significant effect on cd4 values.

Comparing the two methods:

In the last homework, we used linear model and ANOVA test to determine if there is treatment effect. Specifically, the linear model provided significant evidence for a treatment effect based on the ANOVA test. However, this method does not account for individual subject variability (pid). In the current mixed-effects model, we see that while the model accounts for heterogeneity across subjects, the fixed effects for treatment and treatment-time interaction were not statistically significant. This suggests that after properly accounting for subject-level variability, the treatment effect is not as strong or as significant as it appeared in the simpler linear models.

Also, in the last homework, it was shown that treatment and pid were confounded, meaning that each subject was only exposed to one treatment level. This meant that separating the effects of pid and trt was impossible in that fixed-effects-only model. The mixed-effects model, we now handle this by allowing random effects for pid, which separates the treatment effects from the individual variability across subjects.

3(b)

```r
mpar<-function(...){par(mar=c(3,3,1,1),mgp=c(2,.75,0),tck=-.025,...)}
par(mfrow=c(1,2))
mpar()
BETA.OLS <- NULL
DF <- SSE <- 0
for (j in unique(cd4$pid)) {
  yj <- cd4$cd4[cd4$pid == j]
  xj <- cd4$time[cd4$pid == j]
  fitj <- lm(yj ~ xj)
  BETA.OLS <- rbind(BETA.OLS, fitj$coefficients)
  if (length(yj) >= 2) {
    SSE <- SSE + sum(fitj$residuals^2)
    DF <- DF + length(yj) - 2
  }
}
fixed_effects <- fixef(lmem)
random_effects <- ranef(lmem)$pid
subject_data <- cd4 %>%
  group_by(pid) %>%
  summarise(trt = unique(trt))
random_effects <- as.data.frame(random_effects) %>%
  rownames_to_column(var = "pid") %>%
  left_join(subject_data, by = "pid")
random_effects <- random_effects %>%
  mutate(
    # For intercept, if trt = 0 use beta_0 + (Intercept), else beta_0 + beta_1 + (Intercept)
    final_intercept = ifelse(trt == 0, fixed_effects[1] + `(Intercept)`, fixed_effects[1] + fix
ed_effects[2] + `(Intercept)`),
    # For slope, if trt = 0 use beta_2 + time, else beta_2 + beta_3 + time
    final_slope = ifelse(trt == 0, fixed_effects[3] + time, fixed_effects[3] + fixed_effects[4]
+ time)
  )
s2.ols <- SSE / DF
plot(BETA.OLS[,1],random_effects[, 'final_intercept'],xlab="OLS intercept",ylab="LME intercep
t",)
abline(0, 1)
plot(BETA.OLS[,2],random_effects[, 'final_slope'],xlab="OLS slope",ylab="LME slope",)
abline(0, 1)
```
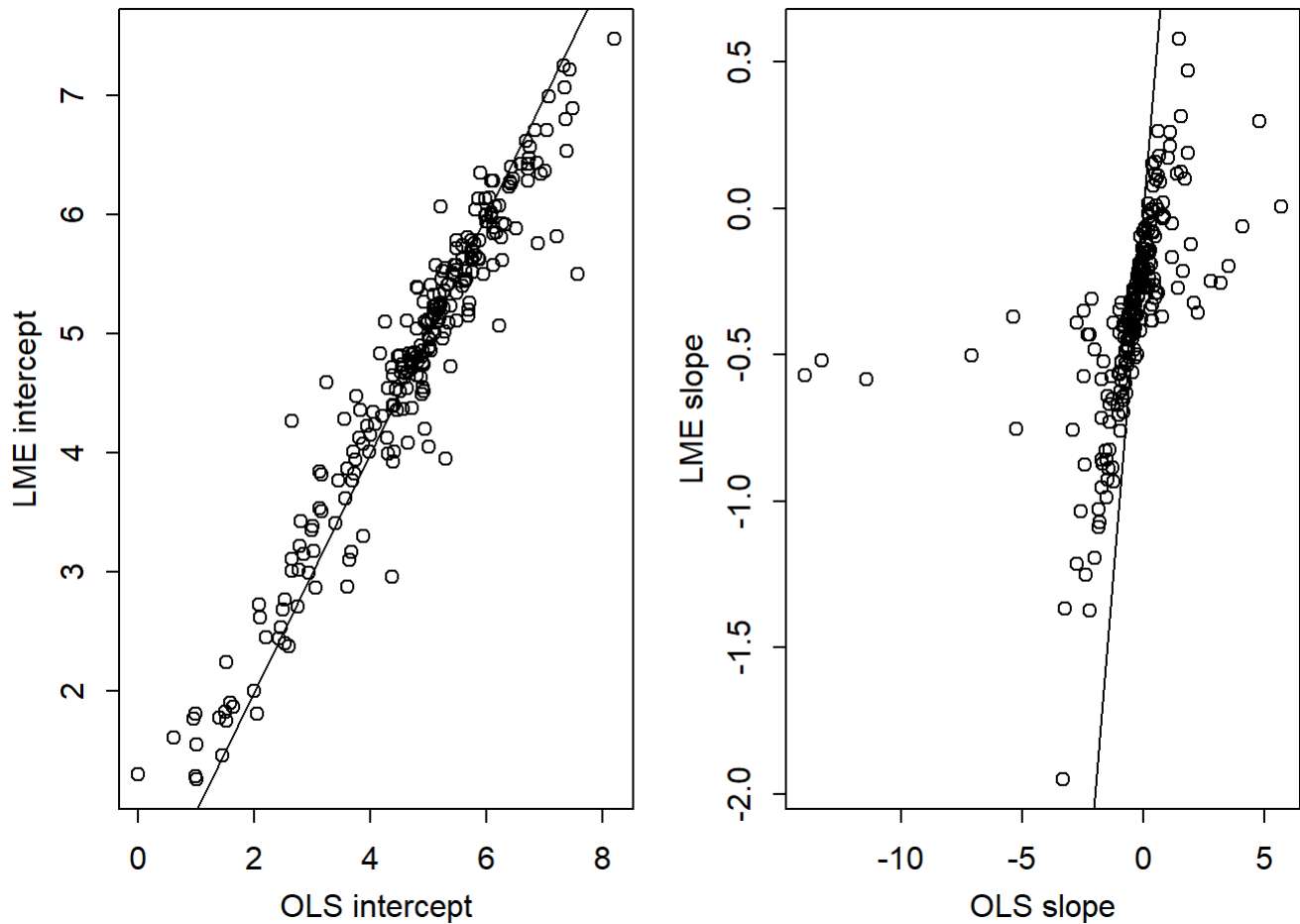
The LME model's intercept estimates are largely consistent with the OLS intercepts, with minimal shrinkage. This indicates that subject-specific intercepts don't differ much from the population average in the LME model.

The LME model significantly shrinks the slope estimates toward zero, which suggests that the model assumes less subject-specific variability in the slopes compared to OLS. This is particularly true for subjects with more extreme OLS slopes, where the LME estimates bring them closer to the population average.

We will now plot the estimate differences for both intercept and slope vs. sample sizes.

```
sample_size <- cd4 %>%
  group_by(pid) %>%
  summarise(sample_size = n())
final_data <- data.frame(
  pid = random_effects$pid,
  OLS_intercept = BETA.OLS[,1],
  LME_intercept = random_effects$final_intercept,
  LME_slope = random_effects$final_slope,
  OLS_slope = BETA.OLS[,2]
)
final_data <- final_data %>%
  left_join(sample_size, by = "pid")
final_data <- final_data %>%
  mutate(difference_intercept = LME_intercept - OLS_intercept,
         difference_slope = LME_slope - OLS_slope)
plot(final_data$sample_size, final_data$difference_intercept,
     xlab = "Sample Size (n)",
     ylab = "Difference (LME - OLS Intercept)",
     main = "Difference in Intercepts vs Sample Size",
     pch = 19, col = "blue")
abline(h = 0, col = "gray", lty = 2)
```


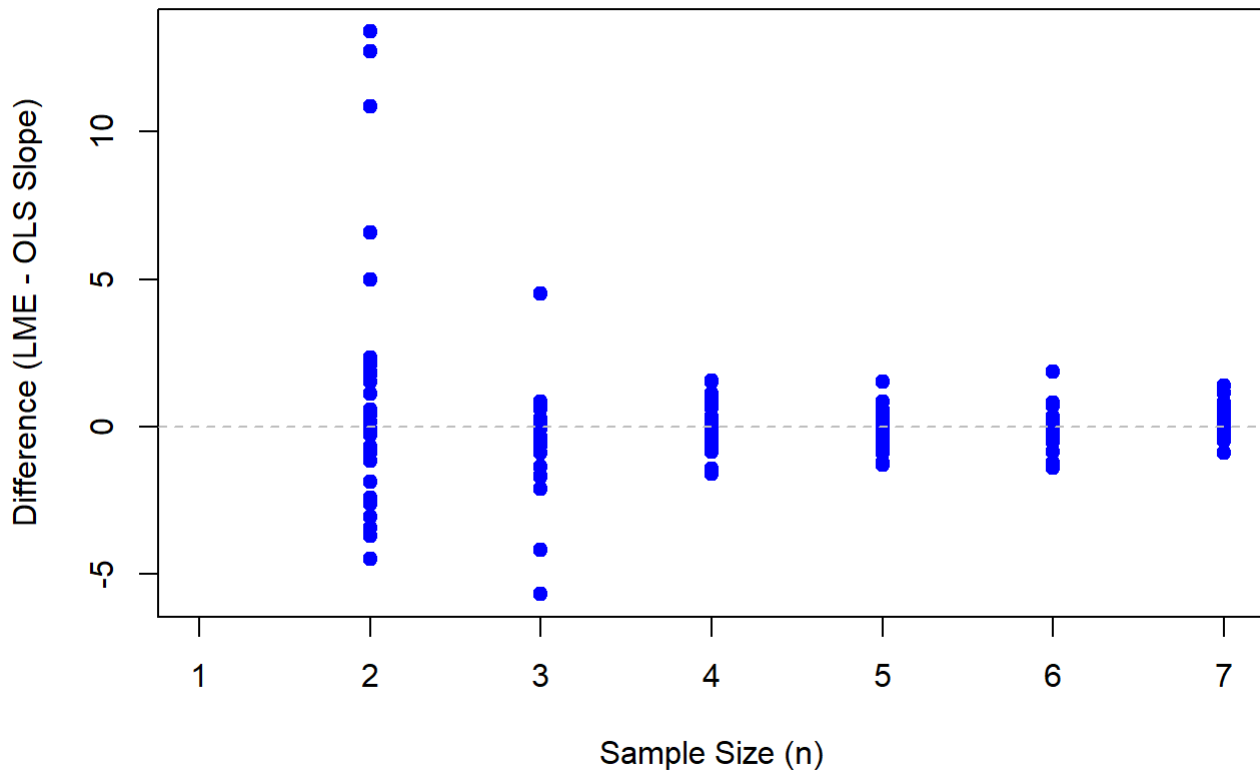**Difference in Intercepts vs Sample Size**
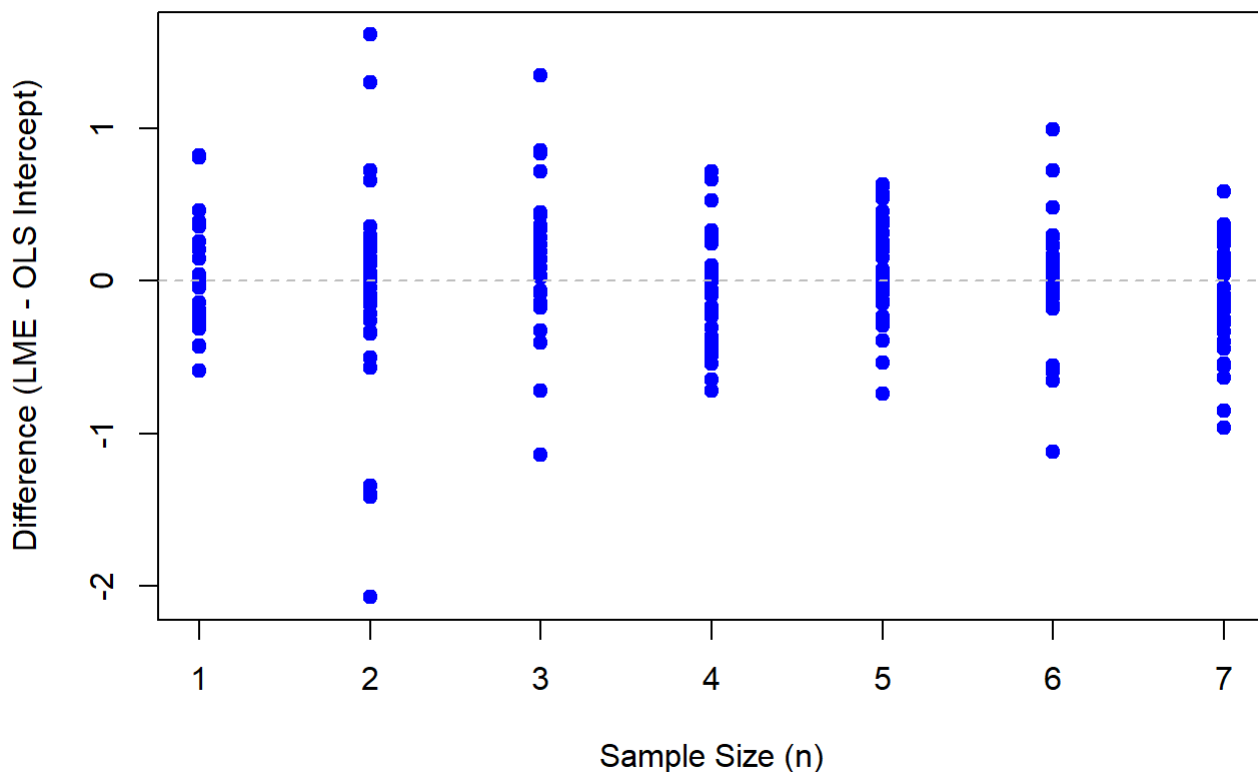
```
plot(final_data$sample_size, final_data$difference_slope,
     xlab = "Sample Size (n)",
     ylab = "Difference (LME - OLS Slope)",
     main = "Difference in Slopes vs Sample Size",
     pch = 19, col = "blue")
abline(h = 0, col = "gray", lty = 2)
```

Difference in Slopes vs Sample Size

We see that for both intercept and slope, the difference tend to converge closer to zero when the sample size increases. As the sample size increases, the estimates from the LME model become more aligned with those from the OLS model, reflecting the greater reliance on subject-specific data when more observations are available. In conclusion, we see that for both intercepts and slopes, the shrinkage effect is most pronounced when the sample size is small, as the LME model pulls estimates toward the population mean.

3(c)

We will plot the residuals of each pid versus time for all pids with more than or equal to 4 sample sizes.

```r
library(lattice)
library(dplyr)

fitted_values <- fitted(lmem)
cd4$residuals <- cd4$cd4 - fitted_values

plotdf <- data.frame(
  pid = cd4$pid,
  time = cd4$time,
  residuals = cd4$residuals,
  fitted_values = fitted_values
)

# Filter for pids with more than 4 observations
filtered_data <- plotdf %>%
  group_by(pid) %>%
  filter(n() > 4) %>%
  ungroup()

xyplot(residuals ~ time | factor(pid),
       data = filtered_data,
       type = "b",
       layout = c(5, 5),
       as.table = TRUE,
       xlab = "Time",
       ylab = "Residuals",
       main = "Residuals vs Time by Subject (pids with > 4 observations)",
       panel = function(x, y, ...) {
         panel.xyplot(x, y, ...)
         panel.abline(h = 0, col = "red", lty = 2)  # Add a red dashed horizontal line at y = 0
       }
)
```
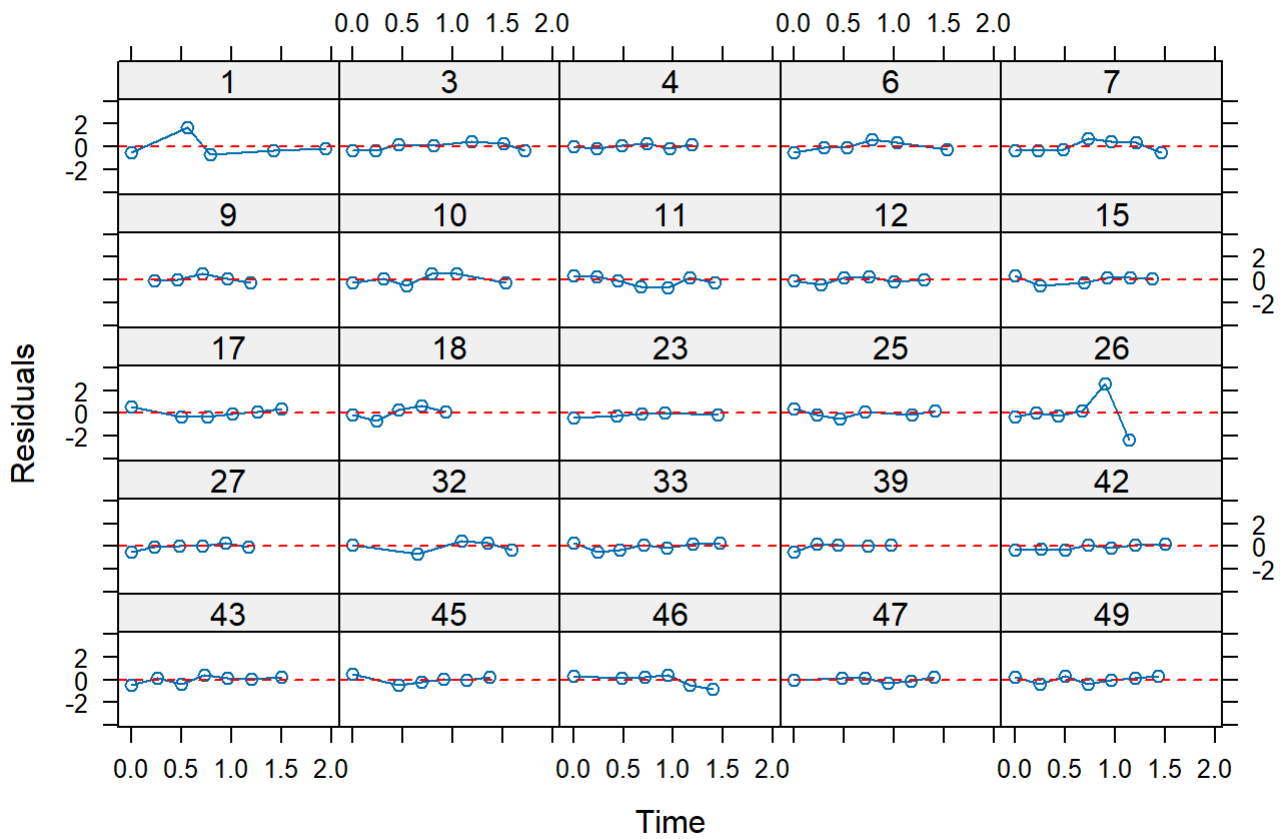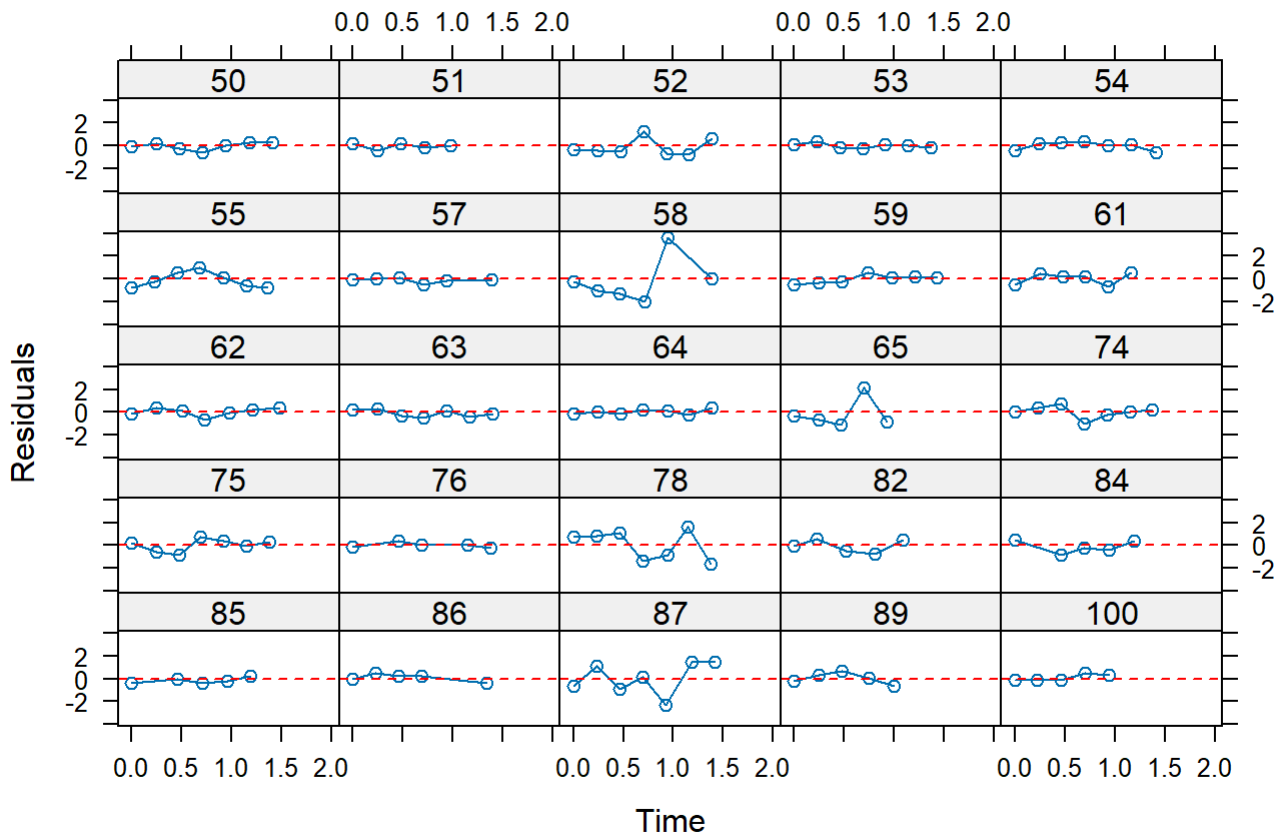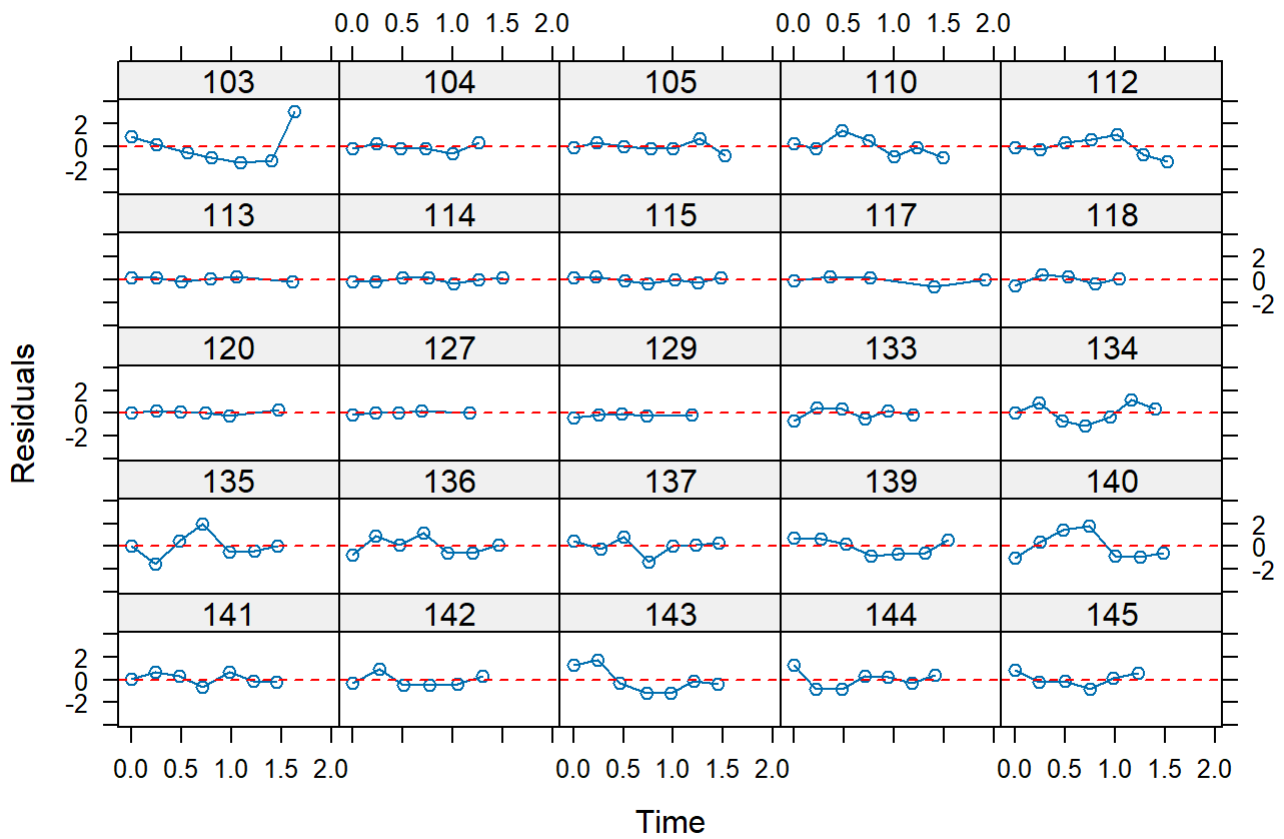
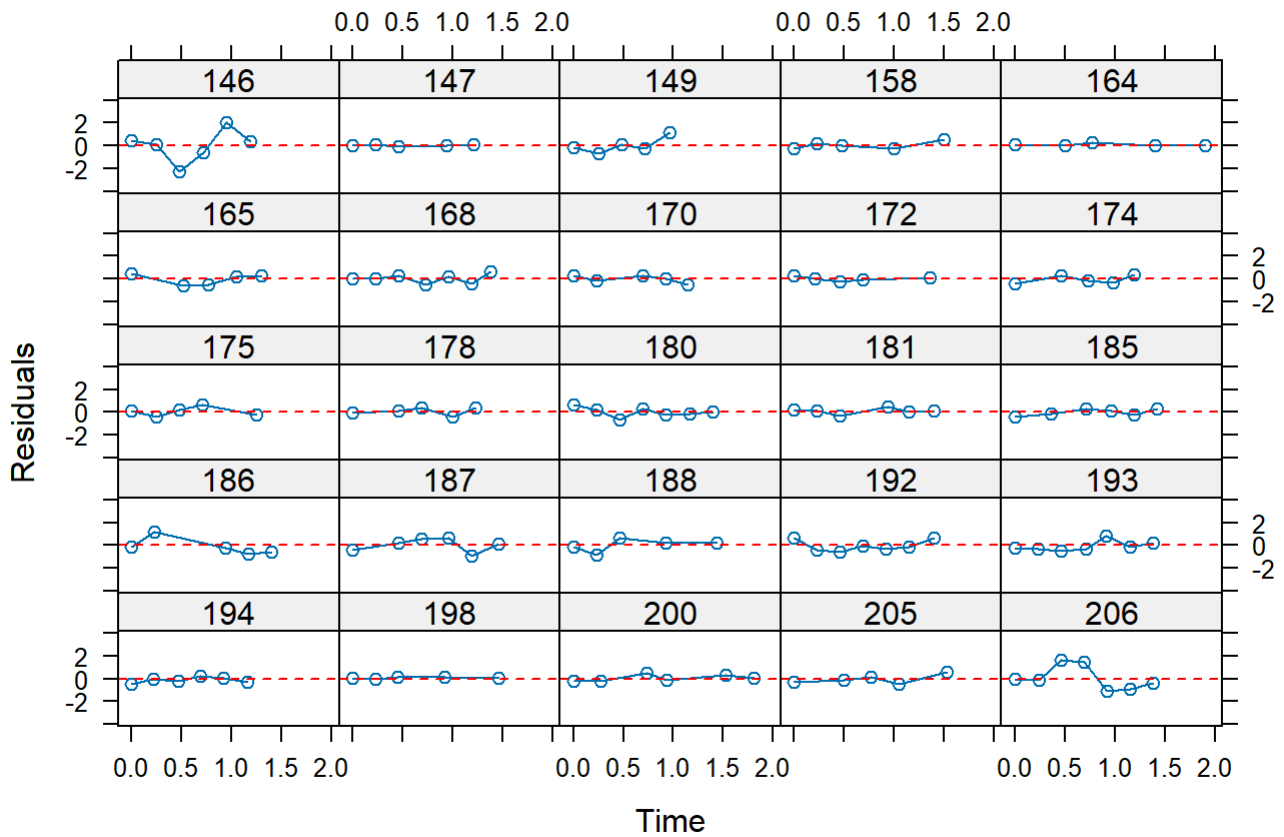Residuals vs Time by Subject (pids with > 4 observations)

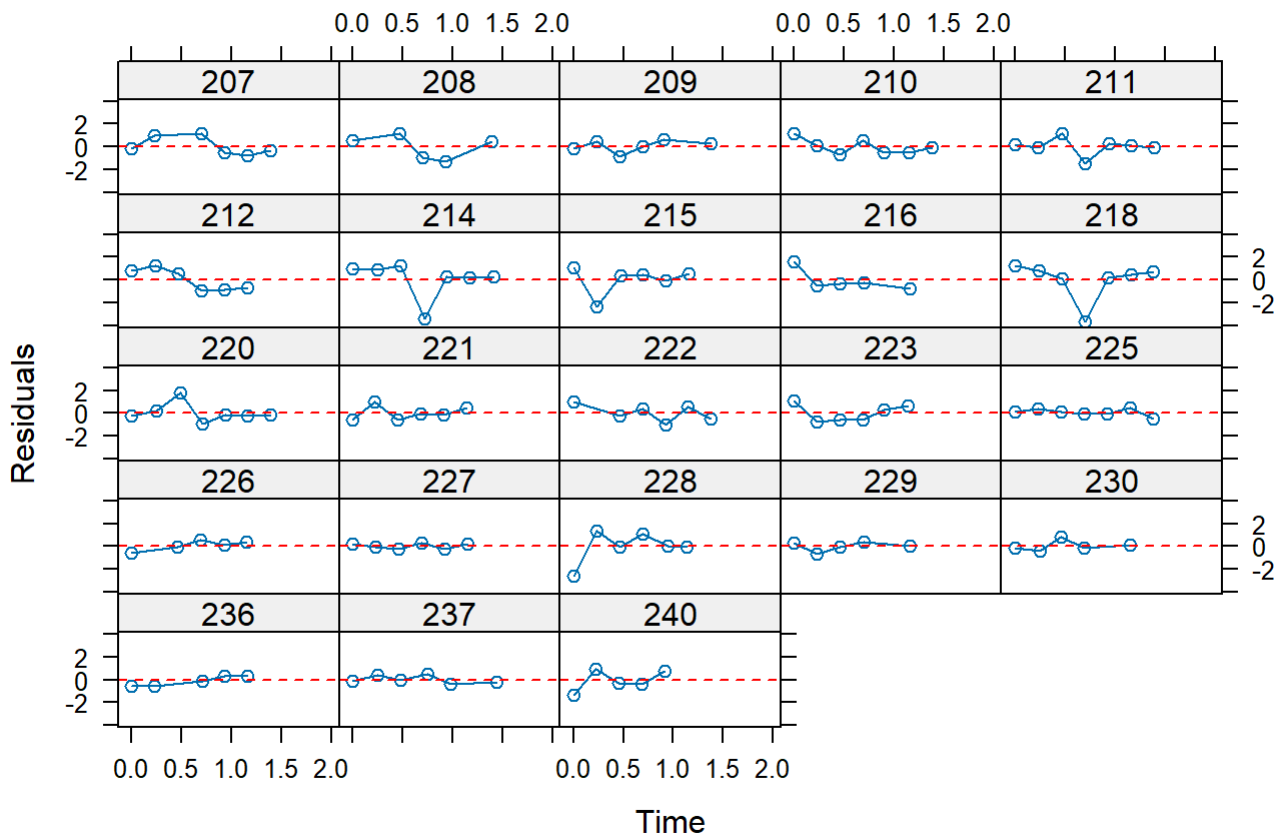Residuals vs Time by Subject (pids with > 4 observations)

Residuals vs Time by Subject (pids with > 4 observations)

Residuals vs Time by Subject (pids with > 4 observations)

Residuals vs Time by Subject (pids with > 4 observations)

From the plots above, we finds out that in most cases, the residuals appear to fluctuate around zero without a noticeable trend. This is a good indication that the model is not systematically biased in overpredicting or underpredicting values for these subjects. For most subjects, there do not appear to be strong patterns over time, such as increasing or decreasing residuals, which would violate the independence assumption. The residuals appear randomly distributed around the red dashed line, which supports the conditional independence assumption. In some subjects, the residuals show small fluctuations, but this may be due to noise rather than a systematic violation of independence.

In conclusion, the assumption of conditional independence of ϵ_{ij} within each subject seems reasonable based on these residual plots. The residuals do not show significant autocorrelation or patterns over time that would suggest otherwise.