HW2

Question 1

Question 1(a)

```
load("~/Documents/2023Fall/P8157/P8157/MACS-VL.RData")
data = macsVL
macs = data |>
  group_by(id)  |>
  mutate(idd = group_indices())  |>
  ungroup()
```

```
# number of clusters
length(unique(data$id))
```

```
## [1] 225
```

```
# number of measurements within each cluster
obs = data |> group_by(id) |> summarize(n_obs = n())
summary(obs$n_obs)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   7.000   8.000   7.484   9.000  10.000
```

```
# follow-up period
fl = data |> group_by(id) |> mutate(max_mon = max(month)) |>
  filter(month == max_mon)
summary(fl$max_mon)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   42.00   45.00   42.22   47.00   48.00
```

```
# time interval between measurements within each cluster
int = data |>
  group_by(id) |>
  mutate(delta_mon = month - lag(month))
summary(int$delta_mon)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   2.000   6.000   6.000   6.452   7.000  34.000     225
```

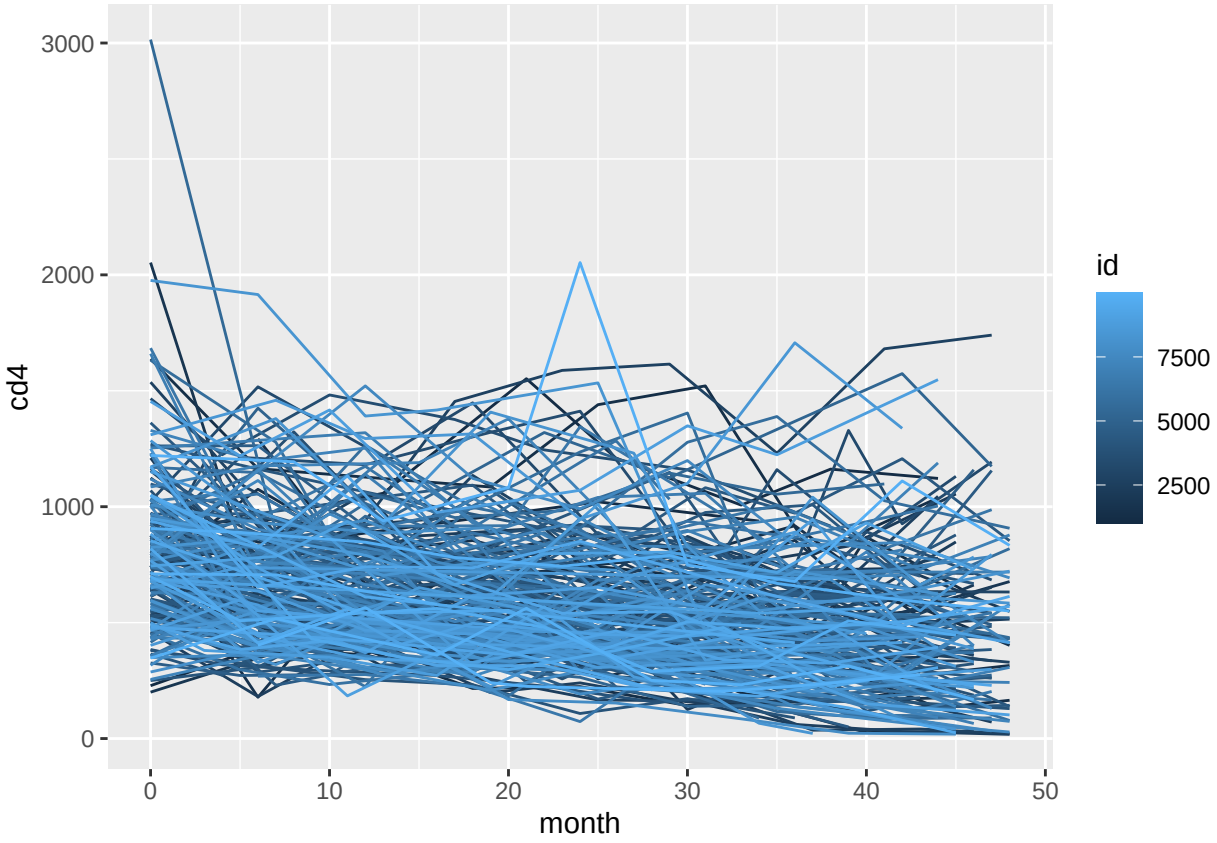
```r
# baseline vload
vl = data |> group_by(id) |> summarize(vload = first(vload))
summary(vl$vload)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     300    7928   24573   78348   91195 1026656
```

```r
# cd4+ count
c4 = data |> group_by(id) |> summarize(base_cd4 = first(cd4), last_cd4 = last(cd4)) |>
  mutate(loss_cd4 = base_cd4 - last_cd4)
summary(c4$loss_cd4)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -452.0   115.0   283.0   316.4   467.0  1917.0
```

```r
# spaghetti plot
ggplot(data, aes(x = month, y = cd4, group = id, color = id)) +
  geom_line()
```




```r
K = 225
# Stage 1
betaMat = data.frame(beta0=rep(NA, K), beta.time=rep(NA, K))
for(k in 1:K) {
  temp.k = macs[macs$idd == k,]
  fit.k = lm(log(cd4) ~ month, data = temp.k)
```

```
  betaMat[k, 1:2] = c(fit.k$coef)
}

# Stage 2
data_2 = cbind(vl, betaMat)
model_time = lm(beta.time ~ vload, data = data_2)
summary(model_time)
```

```
##
## Call:
## lm(formula = beta.time ~ vload, data = data_2)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.078360 -0.006548  0.003285  0.011558  0.033420
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.438e-02  1.368e-03 -10.507   <2e-16 ***
## vload       -2.026e-08  8.726e-09  -2.322   0.0211 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01778 on 223 degrees of freedom
## Multiple R-squared:  0.02362,    Adjusted R-squared:  0.01924
## F-statistic: 5.394 on 1 and 223 DF,  p-value: 0.02111
```

The modeling result indicates that vload is certainly a significant modifier of the rate of decline of CD4+ cell count.

Question 1(b)

```
data_1 = data |>
  mutate(halfyr = round(month/6))
fitf = lm(cd4 ~ halfyr, data = data_1)
resMat = matrix(residuals(fitf), ncol=8, byrow=TRUE)
```

```
## Warning in matrix(residuals(fitf), ncol = 8, byrow = TRUE): data length [1684]
## is not a sub-multiple or multiple of the number of rows [211]
```

```
# covariance matrix diagonal
sd = round(sqrt(diag(cov(resMat))), 2)
sd = c(266.63, 323.47, 312.31, 299.70, 272.13, 315.27, 286.79, 274.45, 332.57)
sd = c(330.30, 264.27, 272.81, 320.29, 338.98, 288.09, 279.74, 292.83)
# correlation
comat = round(cor(resMat), 2)
# sd and corr matrix:
diag(comat) = sd
comat
```

```
##        [,1]   [,2]   [,3]   [,4]   [,5]   [,6]   [,7]   [,8]
## [1,] 330.30   0.60   0.48   0.45   0.28   0.27   0.19   0.13
## [2,]   0.60 264.27   0.67   0.51   0.35   0.30   0.23   0.18
## [3,]   0.48   0.67 272.81   0.57   0.44   0.40   0.30   0.26
## [4,]   0.45   0.51   0.57 320.29   0.47   0.38   0.34   0.26
## [5,]   0.28   0.35   0.44   0.47 338.98   0.53   0.49   0.39
## [6,]   0.27   0.30   0.40   0.38   0.53 288.09   0.63   0.53
## [7,]   0.19   0.23   0.30   0.34   0.49   0.63 279.74   0.68
## [8,]   0.13   0.18   0.26   0.26   0.39   0.53   0.68 292.83
```

I mutate the month variable into a half-year variable and explored the covariance structure of the data. There isn't evident trend whether the variances change with time, but the correlation does seem to be decaying as a function of time between observations. Thus the auto-regressive correlation structure seems most appropriate here.

Question 1(c)

Question 1(d)