# 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)
```

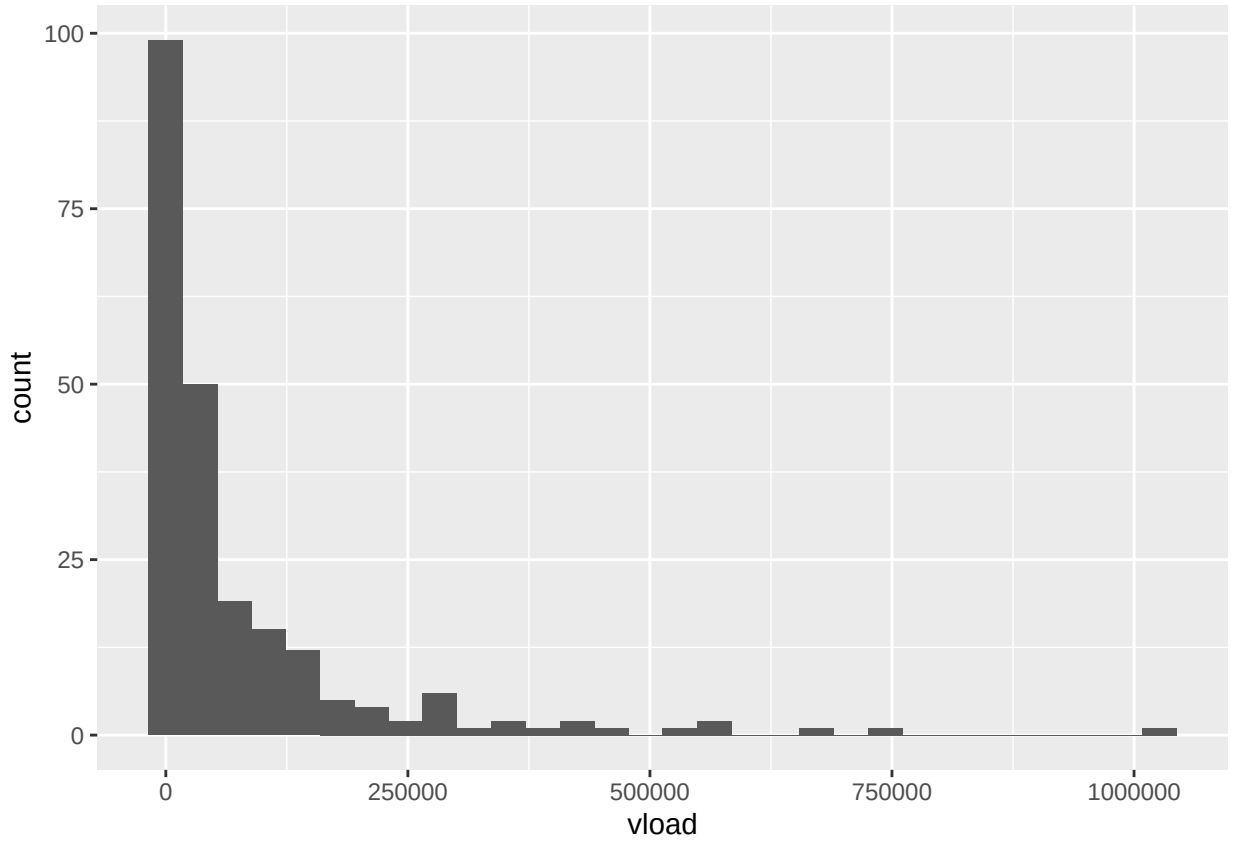
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   2.000   6.000   6.000   6.452   7.000  34.000     225
```

```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
```

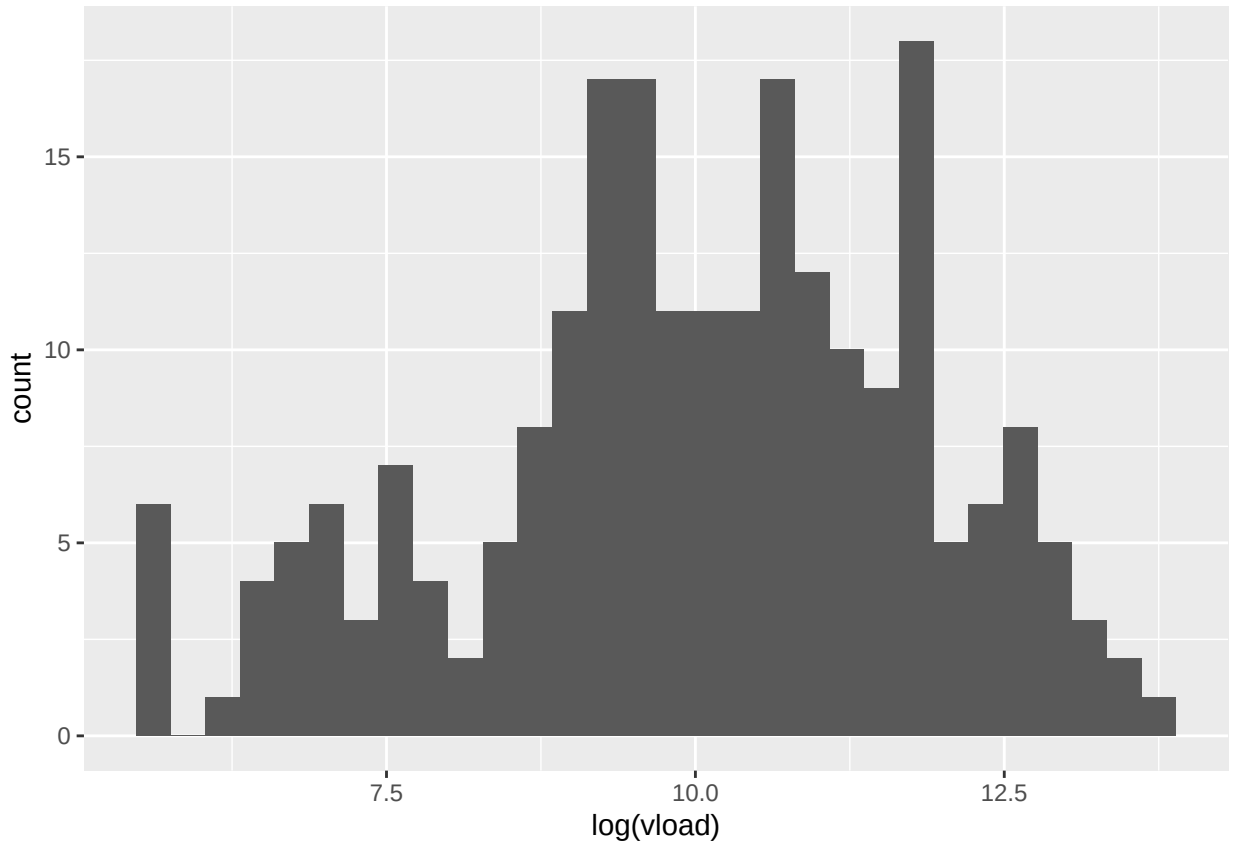
```r
ggplot(vl, aes(x = vload)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```r
ggplot(vl, aes(x = log(vload))) +
  geom_histogram()
```

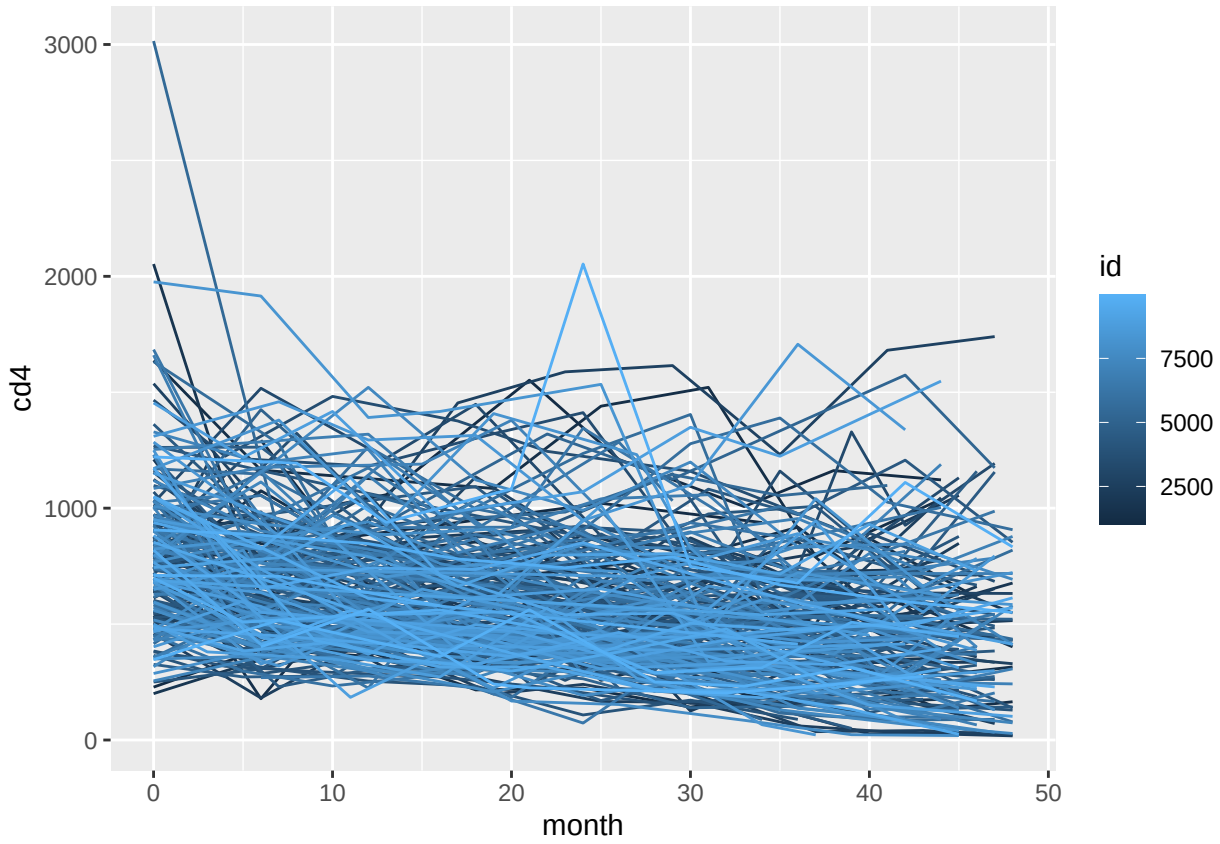
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
# 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
```

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




```
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)
# 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
```

The month variable was mutated into a half-year variable and then I 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)

```
data0 = data |>
  mutate(vload = log(vload))
fit1 = gls(cd4 ~ month*vload, method = "ML", data = data0, corr = corCompSymm(form = ~ 1 | id))
summary(fit1)
```

```
## Generalized least squares fit by maximum likelihood
##   Model: cd4 ~ month * vload
```

```
##   Data: data0
##        AIC      BIC    logLik
##   22928.55 22961.13 -11458.28
##
## Correlation Structure: Compound symmetry
##  Formula: ~1 | id
##  Parameter estimate(s):
##       Rho
## 0.5672909
##
## Coefficients:
##                 Value Std.Error   t-value p-value
## (Intercept) 1108.1012  91.32503 12.133599  0.0000
## month         -3.0861   1.76049 -1.753000  0.0798
## vload        -35.7029   8.99225 -3.970410  0.0001
## month:vload   -0.3806   0.17376 -2.190209  0.0286
##
##  Correlation:
##             (Intr) month  vload
## month       -0.412
## vload       -0.984  0.405
## month:vload  0.405 -0.984 -0.412
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.2033453 -0.7040716 -0.1383408  0.5424951  8.3640477
##
## Residual standard error: 283.383
## Degrees of freedom: 1684 total; 1680 residual
```

```
fit2 = gls(cd4 ~ month*vload, method = "REML", data = data0, corr = corCompSymm(form = ~ 1 | id))
summary(fit2)
```

```
## Generalized least squares fit by REML
##   Model: cd4 ~ month * vload
##   Data: data0
##        AIC      BIC    logLik
##   22917.38 22949.94 -11452.69
##
## Correlation Structure: Compound symmetry
##  Formula: ~1 | id
##  Parameter estimate(s):
##       Rho
## 0.5693312
##
## Coefficients:
##                 Value Std.Error   t-value p-value
## (Intercept) 1108.0965  91.56432 12.101838  0.0000
## month         -3.0868   1.75967 -1.754168  0.0796
## vload        -35.7019   9.01579 -3.959933  0.0001
## month:vload   -0.3805   0.17368 -2.191058  0.0286
##
##  Correlation:
##             (Intr) month  vload
```

```
## month       -0.410
## vload       -0.984  0.404
## month:vload  0.403 -0.984 -0.411
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.1965586 -0.7019290 -0.1379146  0.5408349  8.3385249
##
## Residual standard error: 284.2494
## Degrees of freedom: 1684 total; 1680 residual
```

Both ML and REML give significant estimations of the effect of both baseline virus load on CD4+ cell count and the influence of baseline virus load on the decline rate of cell count. Generally, keeping baseline virus load fixed, with one unit increase in month, the cell count would decrease by -3.08-0.38log(vload); keeping month fixed, with one unit increase in log(virus load), the cell count would decrease by -35.7-0.38month. The p-value of the interaction term is smaller than 0.0286, indicating that under significance level of 0.05, there is a significant association between baseline viral load and the rate of decline in CD4+.

### Question 1(d)

```
vl = data0$vload
summary(vl)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   5.704   8.971  10.109   9.978  11.346  13.842
```

```
min = min(vl)
max = max(vl)
med = median(vl)
mean = mean(vl)
q1 = quantile(vl,0.25)
q3 = quantile(vl,0.75)
breaks = c(min-1, q1, med, q3, max+1)
cats = c("1", "2", "3", "4")

dataj = data0 |>
  mutate(cats = cut(vload, breaks = breaks, labels = cats, right = FALSE))

fit3 = gls(cd4 ~ month*cats, method = "REML", data = dataj, corr = corCompSymm(form = ~ 1 | id))
summary(fit3)
```

```
## Generalized least squares fit by REML
##   Model: cd4 ~ month * cats
##   Data: dataj
##        AIC      BIC    logLik
##   22893.82 22948.06 -11436.91
##
## Correlation Structure: Compound symmetry
##  Formula: ~1 | id
##  Parameter estimate(s):
##      Rho
```

Table 1: Model with ML ($\rho = 0.5673$)

| | Value | Standard error | t-value | p-value |
|---|---|---|---|---|
| Intercept | 1108.10 | 91.33 | 12.1336 | 0.0000 |
| month | -3.0861 | 1.7605 | -1.7530 | 0.0798 |
| vload | -35.70 | 8.99 | -3.9704 | 0.0001 |
| month:vload | -0.38 | 0.17 | -2.1902 | 0.0286 |

Table 2: Model with REML ($\rho = 0.5693$)

| | Value | Standard error | t-value | p-value |
|---|---|---|---|---|
| Intercept | 1108.10 | 91.56 | 12.10 | 0.00 |
| month | -3.09 | 1.76 | -1.7530 | 0.0798 |
| vload | -35.70 | 9.016 | -3.9599 | 0.0001 |
| month:vload | -0.38 | 0.17 | -2.1911 | 0.0286 |

```
## 0.573092
##
## Coefficients:
##                 Value Std.Error   t-value p-value
## (Intercept)  855.4991  33.56261 25.489645  0.0000
## month         -5.4953   0.62991 -8.723916  0.0000
## cats2       -103.9395  46.89458 -2.216451  0.0268
## cats3       -122.2828  47.61295 -2.568269  0.0103
## cats4       -186.2373  46.70951 -3.987139  0.0001
## month:cats2   -1.9486   0.88776 -2.194921  0.0283
## month:cats3   -2.5694   0.88641 -2.898712  0.0038
## month:cats4   -1.0036   0.89912 -1.116229  0.2645
##
##  Correlation:
##             (Intr) month  cats2  cats3  cats4  mnth:2 mnth:3
## month       -0.410
## cats2       -0.716  0.293
## cats3       -0.705  0.289  0.505
## cats4       -0.719  0.295  0.514  0.507
## month:cats2  0.291 -0.710 -0.403 -0.205 -0.209
## month:cats3  0.291 -0.711 -0.209 -0.409 -0.209  0.504
## month:cats4  0.287 -0.701 -0.206 -0.203 -0.409  0.497  0.498
##
## Standardized residuals:
##        Min         Q1        Med         Q3        Max
## -2.0921343 -0.7122694 -0.1435775  0.5327673  8.2249932
##
## Residual standard error: 285.1964
## Degrees of freedom: 1684 total; 1676 residual
```

Table 3: Stratified Model with REML ($\rho = 0.5731$)

| | Value | Standard error | t-value | p-value |
|---|---|---|---|---|
| Intercept | 855.50 | 33.56 | 25.4896 | 0.0000 |
| month | -5.50 | 0.63 | -8.7239 | 0.0000 |
| cats2 | -103.94 | 46.89 | -2.2165 | 0.0268 |
| cats3 | -122.28 | 47.61 | -2.5683 | 0.0103 |
| cats4 | -186.24 | 46.71 | -3.9871 | 0.0001 |
| month:cats2 | -1.95 | 0.89 | -2.1949 | 0.0283 |
| month:cats3 | -2.57 | 0.89 | -2.8987 | 0.0038 |
| month:cats4 | -1.00 | 0.90 | -1.1162 | 0.2645 |