HW2

1 Question 1

1.1 Question 1(a)

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

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

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

```r
# 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
```

```r
# 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
```

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

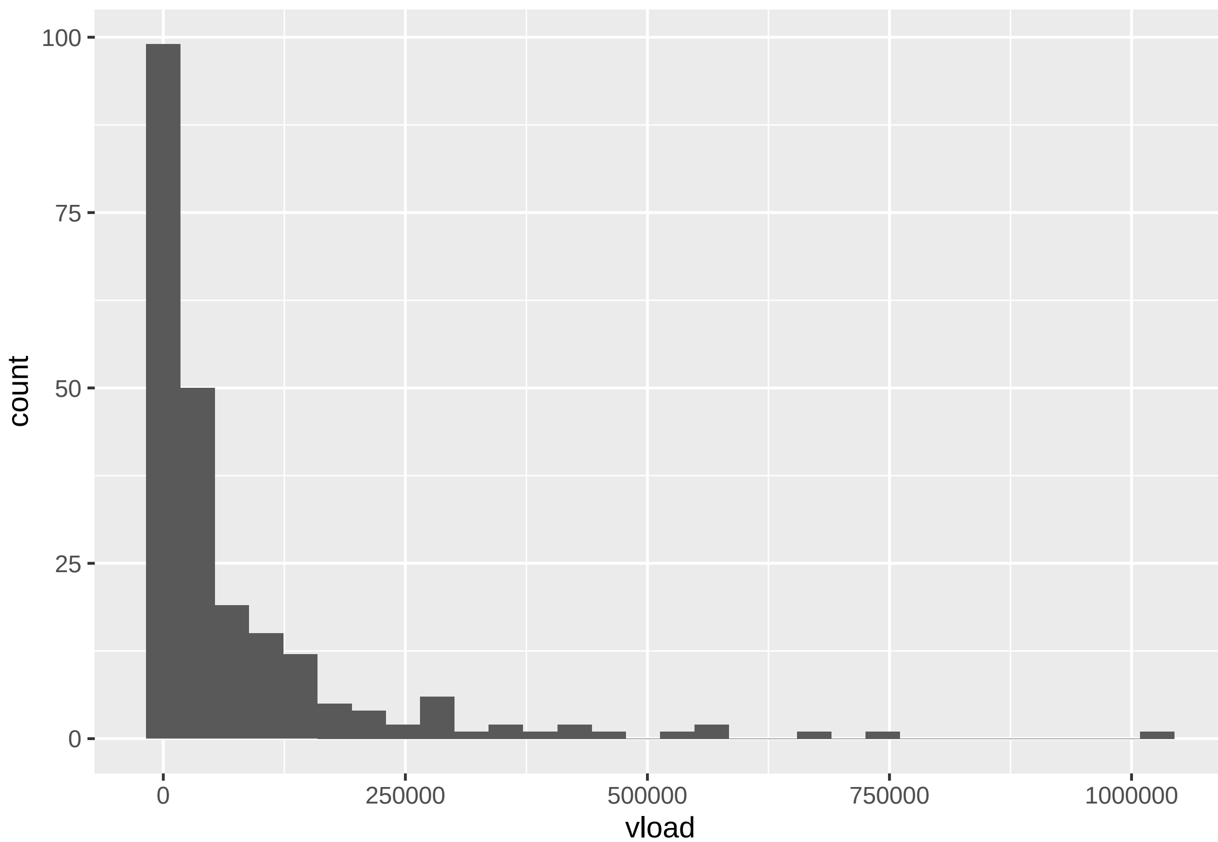
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.000   6.000   6.000   6.452   7.000  34.000
```

```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
```

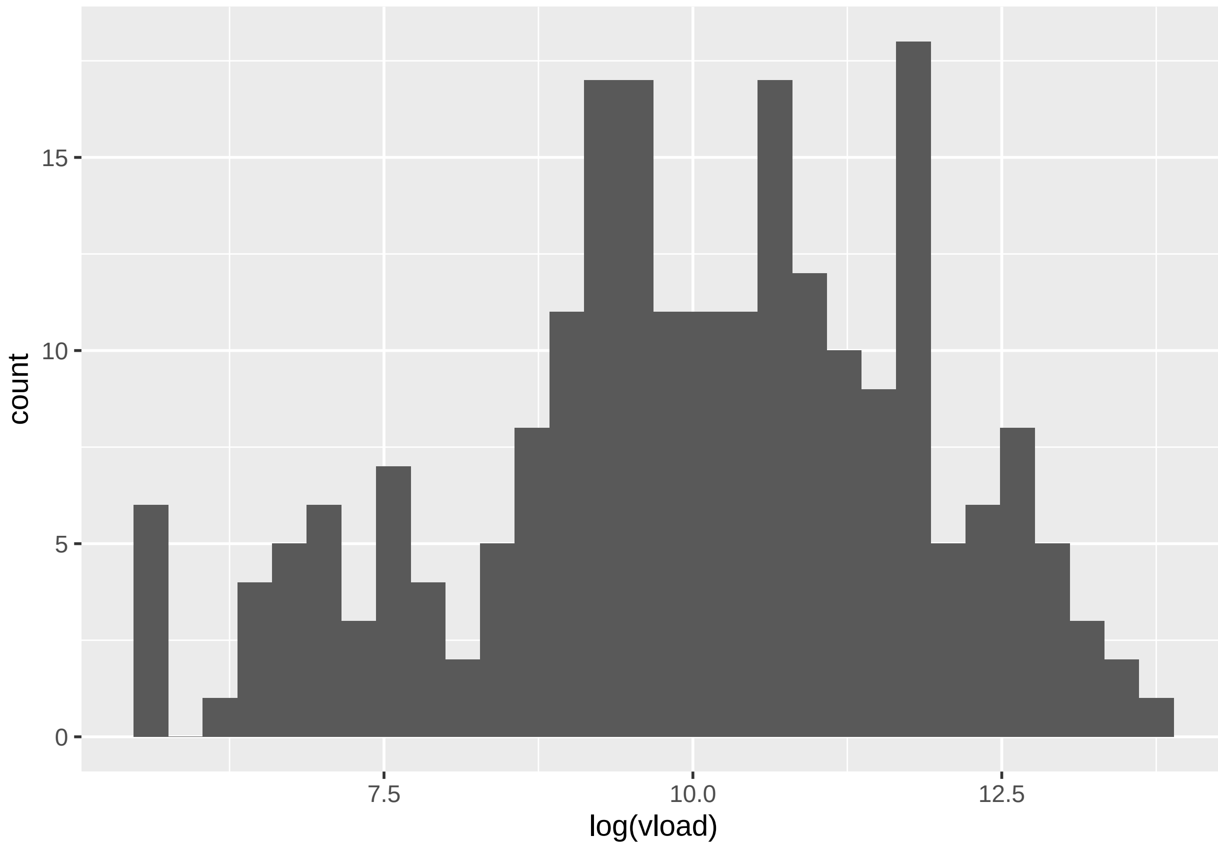
```r
ggplot(vl, aes(x = vload)) +
  geom_histogram()
```

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




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

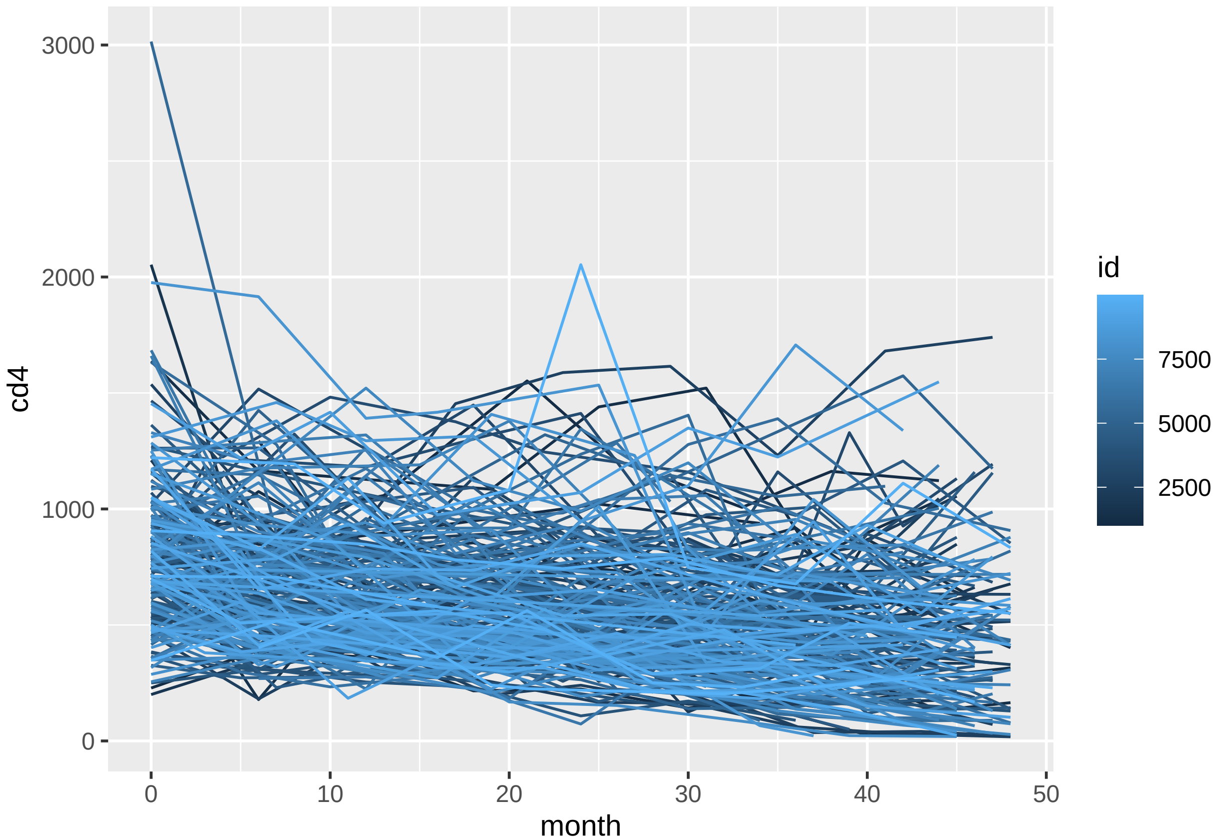
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```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
# 2-stage analysis
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)$coefficients
```

```
##                  Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept) -1.437582e-02 1.368276e-03 -10.506514 3.095558e-21
## vload       -2.026470e-08 8.725514e-09  -2.322465 2.110946e-02
```

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

1.2 Question 1(b)

```r
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. The covariance structure of the data was explored afterwards. 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.

1.3 Question 1(c) – please refer to the last page for model summary tables

```
data0 = data |>
  mutate(vload = log(vload))
fit1 = gls(cd4 ~ month*vload, method = "ML", data = data0, corr = corCompSymm(form = ~ 1 | id))
sum1 = summary(fit1)
fit2 = gls(cd4 ~ month*vload, method = "REML", data = data0, corr = corCompSymm(form = ~ 1 | id))
sum2 = summary(fit2)
sum1$coefficients
```

```
##  (Intercept)        month        vload  month:vload
## 1108.1012484   -3.0861417  -35.7029313   -0.3805617
```

```
sum2$coefficients
```

```
##  (Intercept)        month        vload  month:vload
## 1108.0965067   -3.0867594  -35.7019231   -0.3805324
```

1.4 Question 1(d)

```
vl = data0$vload
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))
sum3 = summary(fit3)
sum3$coefficients
```

```
## (Intercept)       month       cats2       cats3       cats4 month:cats2
##  855.499066   -5.495277 -103.939529 -122.282835 -186.237296   -1.948572
## month:cats3 month:cats4
##   -2.569434   -1.003622
```

Interpretation:

- In the non-categorized data, both ML and REML give significant estimations of the effects 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 expected cell count would decrease by $-3.08 - 0.38log(vload)$;
  - keeping month fixed, with one unit increase in $log(vload)$, the expected cell count would decrease by $-35.7 - 0.38 * month$.

The p-value of the interaction term is 0.0286, indicating that under a significance level of 0.05, there is a significant association between baseline viral load and the rate of decline in CD4+.

- In the vload-categorized data, we categorize $log(vload)$ into 4 categories according to the three quantiles, so that each category has nearly equal number of corresponding measurements. From the result we can see that:

  - for thoese with baseline virus load within the 1^{st} category: expected CD4+ cell count at baseline is 855.5; and with each unit increase in month, the expectation of their cell count would decrease by -5.5;
  - for thoese with baseline virus load within the 2^{nd} category: expected CD4+ cell count at baseline is 751.56; with each unit increase in month, the expectation of their cell count would decrease by -7.45;
  - for thoese with baseline virus load within the 3^{rd} category: expected CD4+ cell count at baseline is 733.22; with each unit increase in month, the expectation of their cell count would decrease by -8.07;
  - for thoese with baseline virus load within the 4^{th} category: expected CD4+ cell count at baseline is 669.26; with each unit increase in month, the expectation of their cell count would decrease by -6.5;

The p-value of all terms except for the month*category4 term is below 0.05, indicating that under a significance level of 0.05, the baseline CD4+ cell count in the four categories differ significantly; while the rate of the decline of CD4+ cell count at least differ significantly in the first 3 categories.

Table 1: Non-Categorized 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: Non-Categorized 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

Table 3: Categorized 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

Table 4: New Coefficients of the Categorized Model with REML

Coefficients	j=2	j=3	j=4
$\beta_0 + \beta_{2,j}$	751.56	733.22	669.26
$\beta_1 + \beta_{3,j}$	-7.45	-8.07	-6.5