HW1

Question 1

Question 1(b)

```
k1 = 600
k2 = 300
rho_1 = 0.2
rho_2 = 0.5
rho_3 = 0.8
rho = c(rho_1, rho_2, rho_3)
V_a = c(1, 1, 1)*4/k1
V_b = c(1, 1, 1)*8*(1-rho)/k2
V_c = c(1, 1, 1)*2*(1-rho)/k2
V_d = c(1, 1, 1)*2*(rho+1)/k2
b  = cbind(rho, V_a, V_b, V_c, V_d) %>% as.data.frame()
options(digits = 2)
print(b)
```

```
##   rho    V_a    V_b    V_c   V_d
## 1 0.2 0.0067 0.0213 0.0053 0.008
## 2 0.5 0.0067 0.0133 0.0033 0.010
## 3 0.8 0.0067 0.0053 0.0013 0.012
```

For minimum variance, for any ρ, I would always choose Crossover study to minimize uncertainty.

Question 1(c)

```
k1 = 600
k2 = 400
rho_1 = 0.2
rho_2 = 0.5
rho_3 = 0.8
rho = c(rho_1, rho_2, rho_3)
V_a = c(1, 1, 1)*4/k1
V_b = c(1, 1, 1)*8*(1-rho)/k2
V_c = c(1, 1, 1)*2*(1-rho)/k2
V_d = c(1, 1, 1)*2*(rho+1)/k2
c  = cbind(rho, V_a, V_b, V_c, V_d) %>% as.data.frame()
options(digits = 2)
print(c)
```

```
##   rho    V_a   V_b    V_c    V_d
## 1 0.2 0.0067 0.016 0.0040 0.0060
## 2 0.5 0.0067 0.010 0.0025 0.0075
## 3 0.8 0.0067 0.004 0.0010 0.0090
```

For minimum variance, for any ρ, I would always choose Crossover study to minimize uncertainty.

Question 2

Question 2(a)

```
load("~/Documents/2023Fall/P8157/P8157/Six Cities.RData")
# data summmary
nrow(topeka) # number of total observations
```

```
## [1] 1994
```

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

```
## [1] 300
```

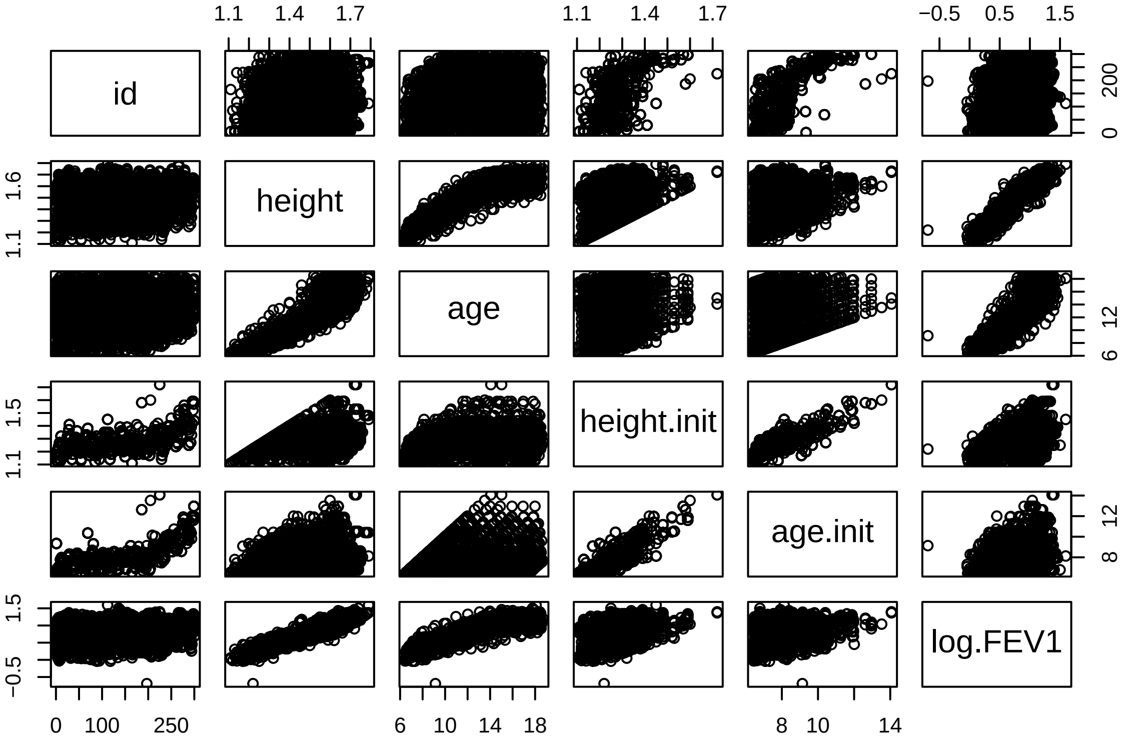
```
visit = topeka %>% group_by(id) %>%tally()
summary(visit$n) # summary of observations in one cluster
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0     3.0     7.0     6.6    10.0    12.0
```

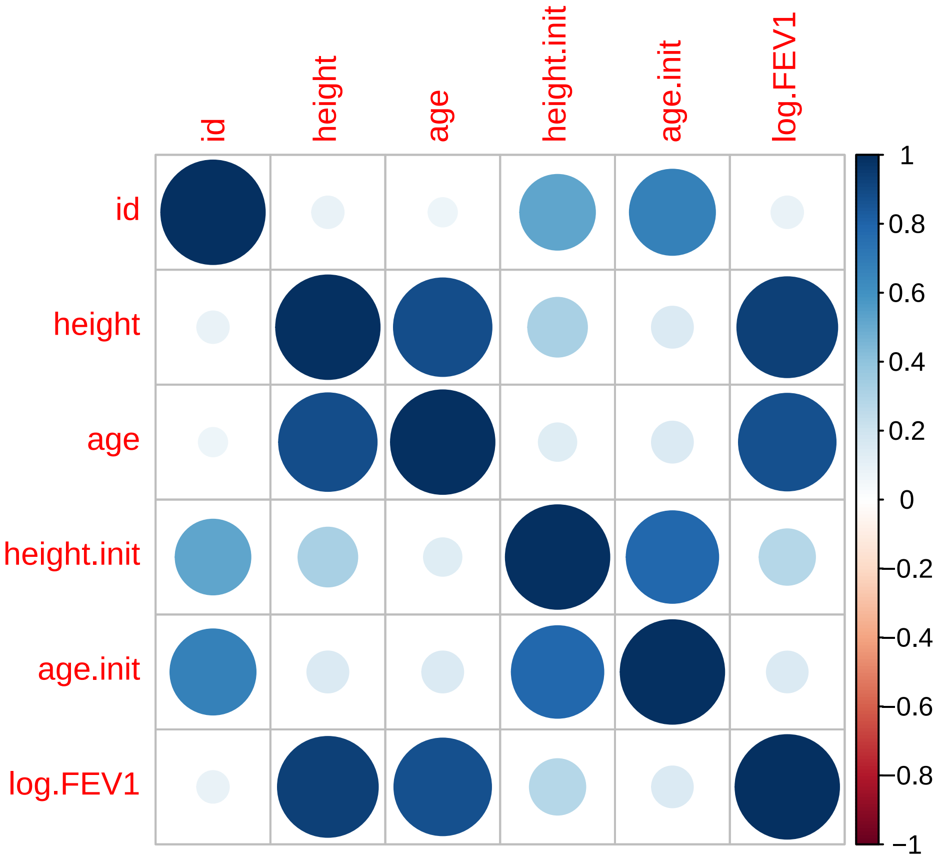
```
skimr::skim(topeka) %>% tibble::as_tibble() # summary of variables
```

```
## # A tibble: 6 x 12
##   skim_type skim_variable n_missing complete_rate numeric.mean numeric.sd
##   <chr>     <chr>             <int>         <dbl>        <dbl>      <dbl>
## 1 numeric   id                    0             1      136.       82.5
## 2 numeric   height                0             1        1.50      0.154
## 3 numeric   age                   0             1       12.6       3.32
## 4 numeric   height.init           0             1        1.28      0.0846
## 5 numeric   age.init              0             1        8.03      1.21
## 6 numeric   log.FEV1              0             1        0.815     0.331
## # i 6 more variables: numeric.p0 <dbl>, numeric.p25 <dbl>, numeric.p50 <dbl>,
## #   numeric.p75 <dbl>, numeric.p100 <dbl>, numeric.hist <chr>
```

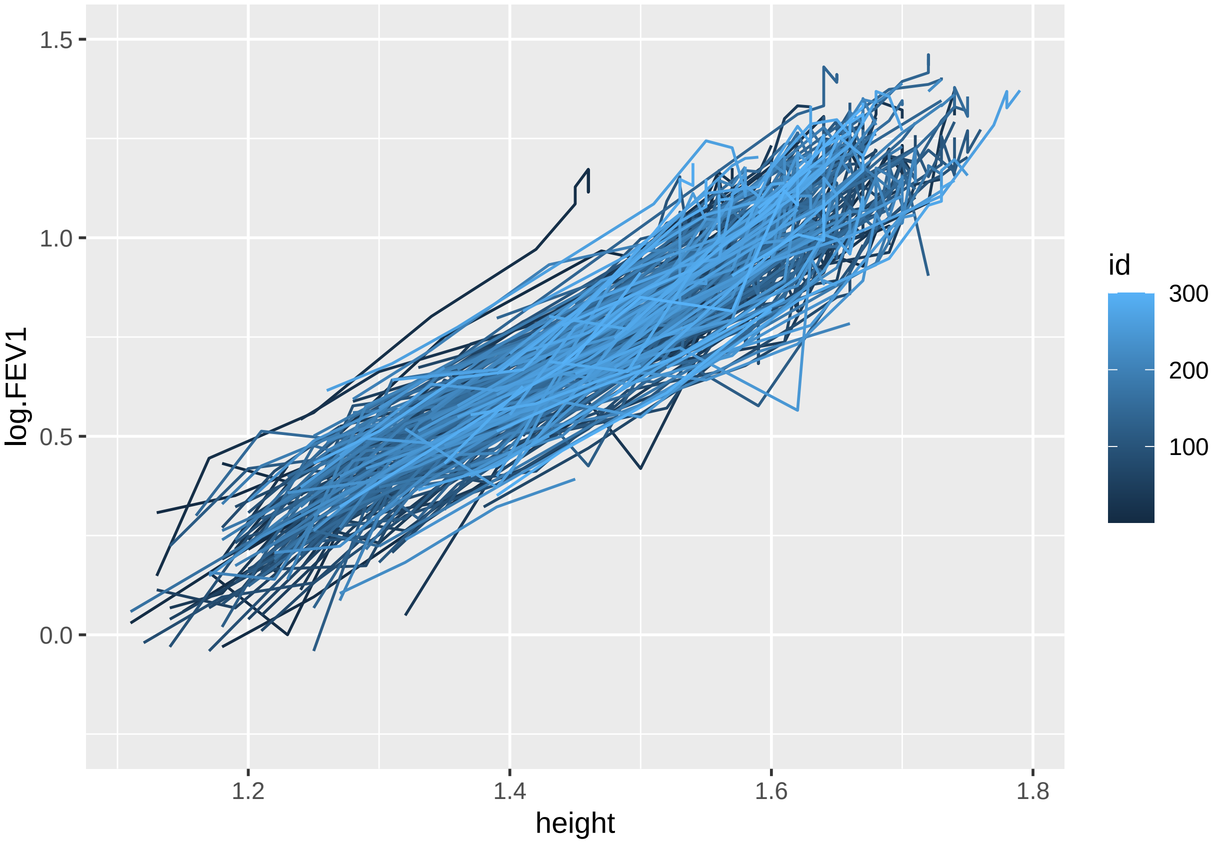
```
# outlier
box = topeka %>% ggplot(aes(x = log.FEV1)) + geom_boxplot()
# dependence
pair = pairs(topeka)
```



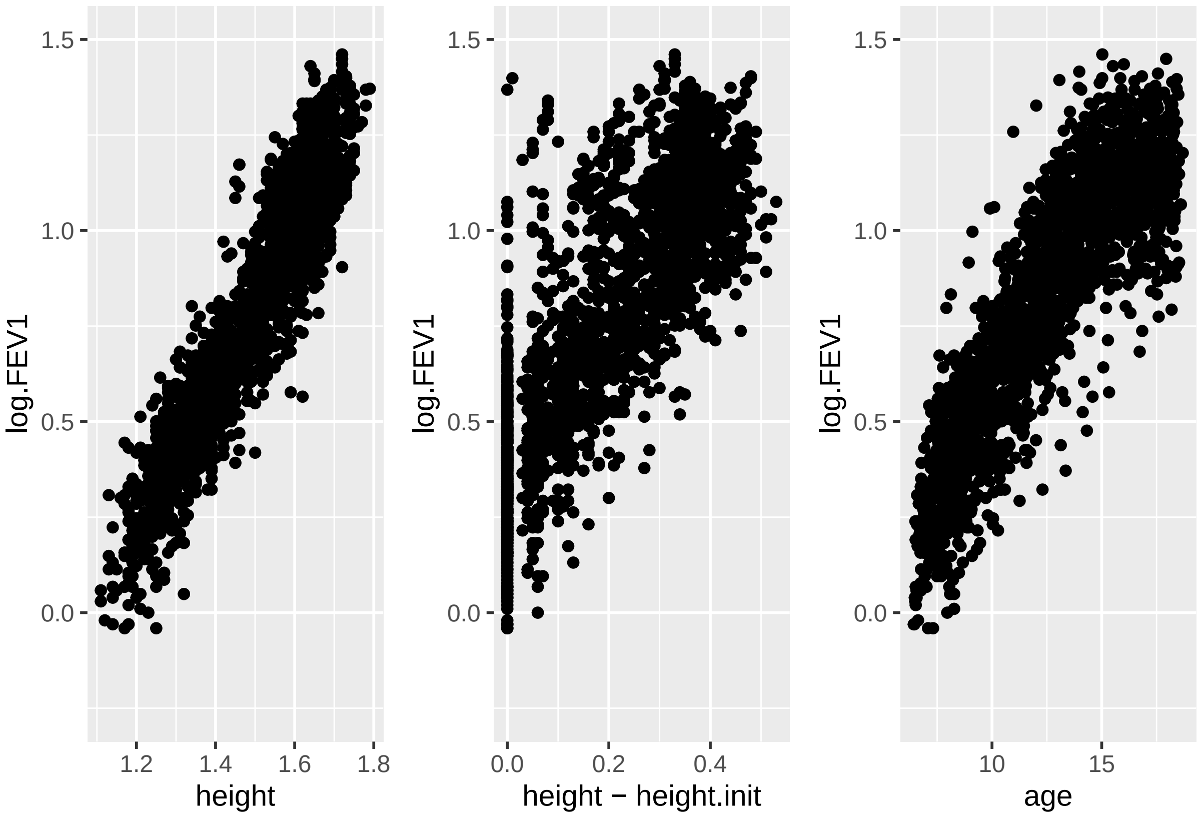
```
cor = corrplot::corrplot(cor(topeka[]))
```



```r
# spaghetti plot
ggplot(topeka, aes(x = height, y = log.FEV1, group = id, color = id)) +
  geom_line() +
  scale_y_continuous(limits = c(-0.25, 1.5))
```



```r
# exploratory
g1 = ggplot(topeka, aes(x = height, y = log.FEV1)) +
  geom_point() +
  scale_y_continuous(limits = c(-0.25, 1.5))
g2 = ggplot(topeka, aes(x = height - height.init, y = log.FEV1)) +
  geom_point() +
  scale_y_continuous(limits = c(-0.25, 1.5))
g3 = ggplot(topeka, aes(x = age, y = log.FEV1)) +
  geom_point() +
  scale_y_continuous(limits = c(-0.25, 1.5))
g1 + g2+ g3
```



There are 5 variables and 1 outcome in the topeka dataset, which consists of 300 clusters with 1994 measurements. Number of measurements within each cluster ranges from 1 to 12, with mean being 6.6 and median being 7. The mean age of the subjects upon their recruitment was 8.03, while the height was 1.28m. For simplicity, the age and height variables could be standardized when conducting the longitudal analysis. One outlier of the outcome log.FEV1 was observed which should be further investigated. In the correlation plot one can see that the covariates age and height are highly correlated, indicating they might be dependent of each other.

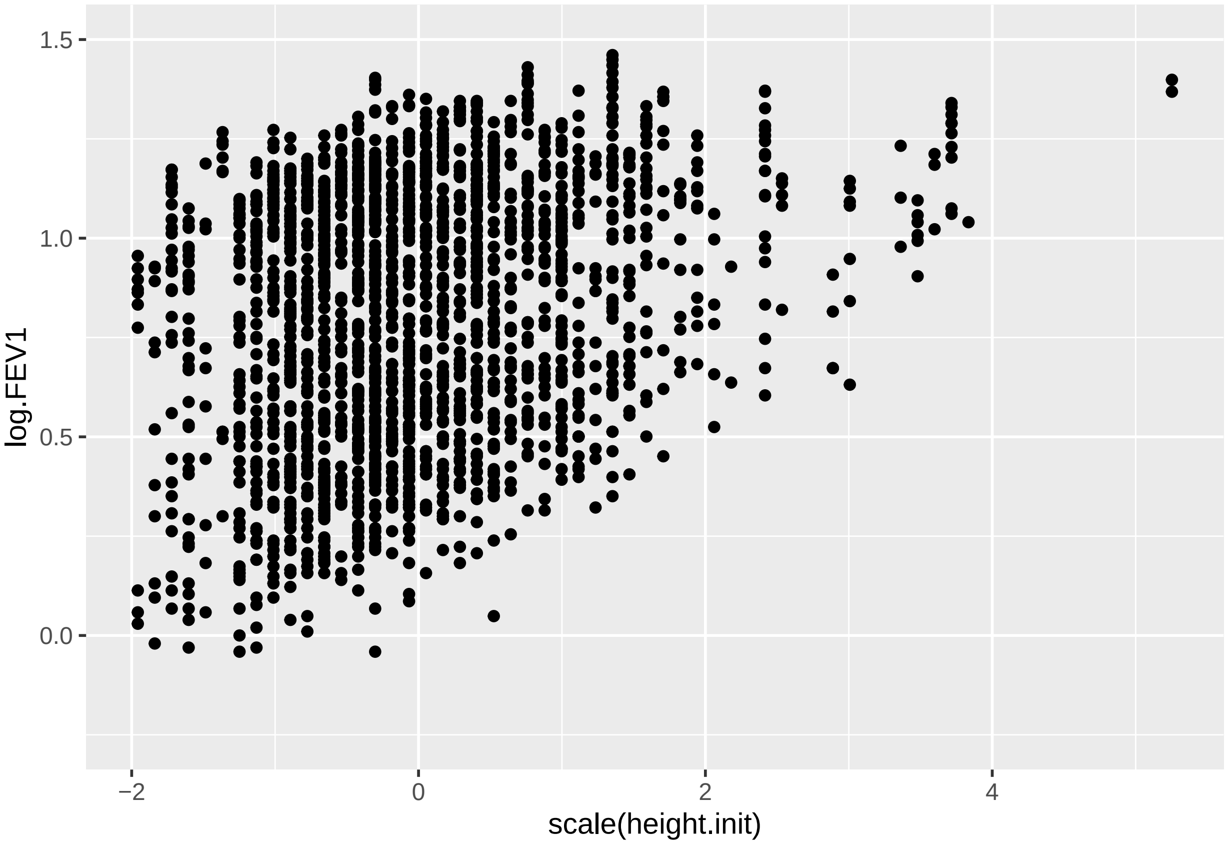
From the scatterplot we can tell that there exists association between age, height, height change and log.FEV1. log.FEV1 seems to be positive related to all the three variables.

Question 2(b)

If we take log.FEV1 as the outcome, we can test if the increase in log.FEV1 is associated with the increase in height. As we can see from the scatterplot that larger increase in height might indicate higher log.FEV1.

Question 2(c)

```
ggplot(topeka, aes(x = scale(height.init), y = log.FEV1)) +
  geom_point() +
  scale_y_continuous(limits = c(-0.25, 1.5))
```



Using cross-sectional data at baseline, we could explore the relationship between initial height and log.FEV1. Assumptions could be made from the scatterplot that those with higher initial height are unlikely to have relative low log.FEV1.

Question 3

```
load("~/Documents/2023Fall/P8157/P8157/MACS.RData")
dmm = macs %>%
  filter(time >= -0.5) %>%
  group_by(id) %>%
  filter(any(-0.5 <= time & time < 0) && any(time > 0)) %>%
  mutate(visit = row_number() - 1) %>%
  ungroup()
```

Question 3(a)

Here's Table 1 summarizing the covariates at baseline:

```
sum = skimr::skim(dmm) %>% tibble::as_tibble()
sum
```

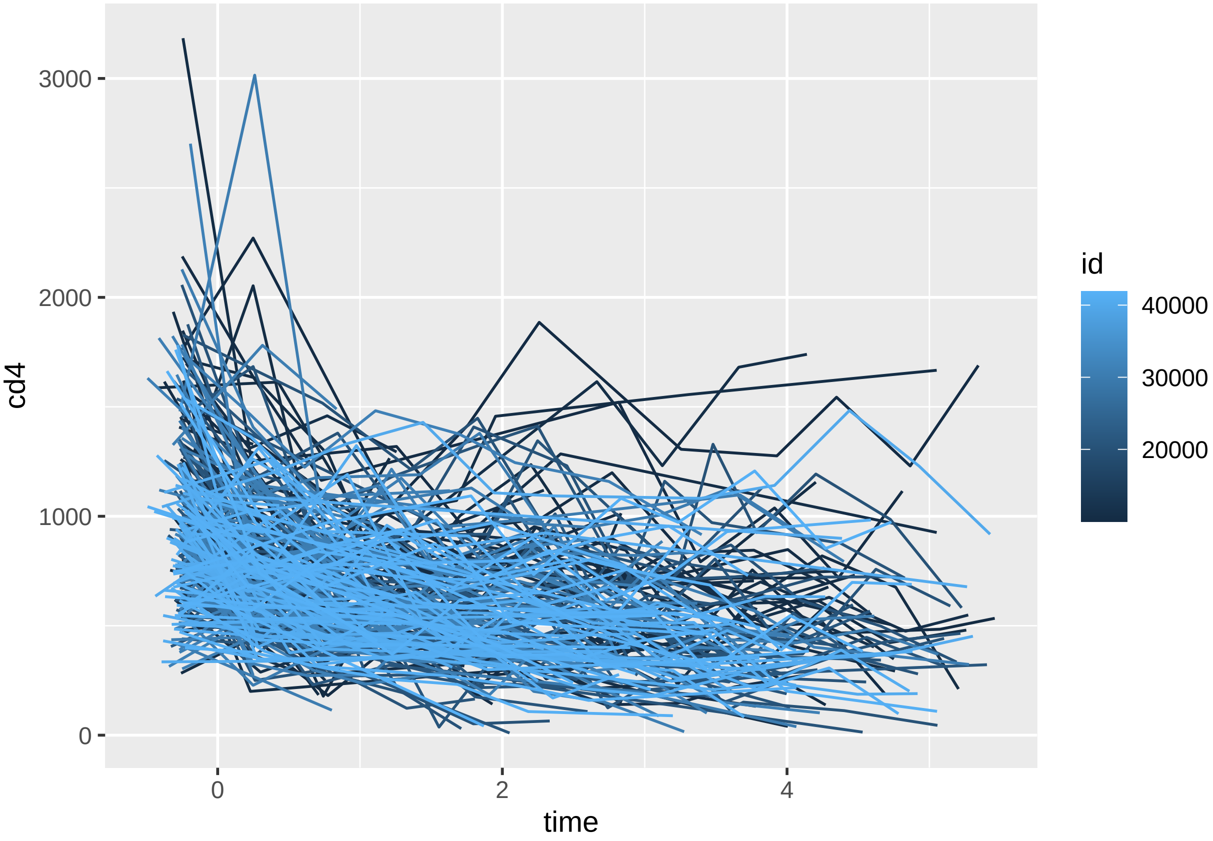
```
## # A tibble: 9 x 12
##   skim_type skim_variable n_missing complete_rate numeric.mean numeric.sd
##   <chr>     <chr>             <int>         <dbl>        <dbl>      <dbl>
## 1 numeric   id                    0             1    25822.     11602.
## 2 numeric   time                  0             1        1.56       1.50
## 3 numeric   age                   0             1        2.57       7.65
## 4 numeric   packs                 0             1        0.980      1.43
## 5 numeric   drug                  0             1        0.756      0.430
## 6 numeric   partners              0             1        4.60       3.56
## 7 numeric   cesd                  0             1        9.57       9.52
## 8 numeric   cd4                   0             1      694.       368.
## 9 numeric   visit                 0             1        3.00       2.52
## # i 6 more variables: numeric.p0 <dbl>, numeric.p25 <dbl>, numeric.p50 <dbl>,
## #   numeric.p75 <dbl>, numeric.p100 <dbl>, numeric.hist <chr>
```

```
dmm = dmm %>%
  group_by(id) %>%
  mutate(idd = group_indices()) %>%
  ungroup()
```

Question 3(b)

Here's the spaghetti plot of the CD4+ cell count progression across time since seroconversion:

```
K = 266
# spaghetti plot
ggplot(dmm, aes(x = time, y = cd4, group = id, color = id)) +
  geom_line()
```



```r
# Stage 1
betaMat = data.frame(beta0=rep(NA, K), beta.time=rep(NA, K))
for(k in 1:K) {
  temp.k = dmm[dmm$idd == k,]
  fit.k = lm(log(cd4) ~ time , data = temp.k)
  betaMat[k, 1:2] = c(fit.k$coef)
}
head(betaMat)
```

Stage 1: Coefficients table (first 6 rows) for time (LogNormal model):

```
##   beta0 beta.time
## 1   6.6    -0.626
## 2   6.3    -0.522
## 3   6.7    -0.053
## 4   6.9    -1.304
## 5   6.7    -0.419
## 6   6.2    -0.995
```

```r
# Stage 2
dmm_base = dmm %>%
  filter(visit == 0)
dmm_base$beta.time = betaMat$beta.time
summary(lm(beta.time ~ age + packs + drug + partners + cesd, data=dmm_base))
```

Stage 2: Explain variance across subject-specific baseline covariates:

```
##
## Call:
## lm(formula = beta.time ~ age + packs + drug + partners + cesd,
##     data = dmm_base)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5765 -0.0821  0.1052  0.2141  0.6733
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.272997   0.074831   -3.65  0.00032 ***
## age          0.001632   0.003348    0.49  0.62636
## packs       -0.000678   0.016110   -0.04  0.96646
## drug        -0.036547   0.064781   -0.56  0.57313
## partners    -0.008908   0.007021   -1.27  0.20563
## cesd         0.004104   0.002435    1.69  0.09316 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.39 on 260 degrees of freedom
## Multiple R-squared:  0.0172, Adjusted R-squared:  -0.00173
## F-statistic: 0.909 on 5 and 260 DF,  p-value: 0.476
```

```
# In addition, provide a brief summary of the results using language that would be suitable for a non-b
```

As time progresses since seroconversion, the CD4+ cell counts of our patients decrease. The decrease rate could be possibly related to the baseline age; packs of cigarettes per day; whether on drug or not; number of partners and cesd depression level of the patients. If our significance level was set higher, say, 0.1, we would find that the more depressed our patients are, the more CD4+ cells he would lose as time progresses. However, there seems to be no significant influence of the other factors on the cell counts.