

Longitudinal Data Analysis, MACS Data (partial solution)

SISCER: Generalized Estimating Equations and Mixed-Effects Models for LDA

Contents

About This Document	1
Data Recap	1
R packages	2
Scientific Question	2
Data Preparation	2
Inferential Analyses	3
Linear Mixed Models	3
GEE	7
References	9

About This Document

This file provides an analysis in R of longitudinal data collected from the Multicenter Aids Cohort Study (MACS), a study that aims to characterize the time course of CD4 cell depletion (Kaslow et al. 1987; see also Fitzmaurice et al. 2018). Some components of the analysis are left for you to complete (look for **Your turn:**). If you'd rather work from complete code, see the “full” version of this document. An overview of the dataset and some exploratory analyses are available in the Day 1 activity.

Data Recap

Variables: (columns of the data file)

- ID = subject ID
- MONTHS = months since seroconversion (detection of HIV)
- AGE = age of subject
- CD4-COUNT = # of CD4 positive cells (helper cells) per mm^3
- CD8-COUNT = # of CD8 positive cells (suppressor cells) per mm^3
- VLOAD0 = viral load at baseline (copies per ml)
- AIDSCASE = 1 if no AIDS observed; 2 if AIDS observed; 3 if died prior to AIDS
- VTIME = calendar time of study visit in months since January 1984
- SCTIME = calendar time of seroconversion (detection of HIV) in months since 1/1984
- ATIME = calendar time of AIDS diagnosis in months since 1/1984
- DTIME = calendar time of death in months since 1/1984, or follow-up time
- IDEATH = indicator of death at DTIME (1=death, 0=censored)

Note 1: ATIME is missing (NA) if the time was not observed during study follow-up (i.e., subjects remained AIDS free and alive).

Note 2: There is a lower limit of detection for viral load and thus measurements at 300 reflect this detection limit.

Note 3: The ability to measure viral loads actually became available many years after the study was started, and for many subjects this measurement needed to be obtained from stored samples. Thus, not all subjects have a viral load at baseline (perhaps due to limited blood samples).

For the rest of the analysis, we will consider the baseline viral load categorized into the following groups (see Chapter 18 of van Belle et al.):

- Low viral load: baseline value less than 15×10^3
- Medium viral load: baseline value between 15×10^3 and 46×10^3
- High viral load: baseline value greater than 46×10^3

R packages

We first load the packages we will need for this activity.

```
library(tidyverse)
library(conflicted) # will tell you if two packages have the same function
library(lmtest)
library(sandwich)
library(car)
library(nlme) # run mixed model
library(multcomp)
library(geepack) # run gee model
library(doBy) #
library(broom) # inferential results (CI) for gee
library(broom.mixed) # inferential results (CI) for mixed models
```

If you need to install any of the listed packages (e.g., if you do not have the `geepack` package installed, you might get an error message “Error in library(geepack) : there is no package called geepack”. In this case, you can install the missing package by

```
install.packages('geepack')
```

Scientific Question

Is baseline viral load associated with rate of decline in CD4+ cells among men who are HIV+?

Data Preparation

We will load and prepare the dataset like we did yesterday:

```
# Load dataset
macs <- read.csv("./macs.csv", row.names = 1)

# Create categorical baseline viral load variable
# This time, leave missing as NA (R will exclude by default)
macs <- macs %>%
  mutate(vload0_cat = cut(vload0, breaks = c(0, 15000, 46000, Inf),
                          labels = c("Low", "Medium", "High"), right=FALSE))

# View dataset
head(macs)

##      id months age cd4 cd8 vload0 aidsstage vtime sctime atime dtime ideath
```

```
## 1 1022      6 27 391 300 70737      3 18 12 NA 66 1
## 2 1022     12 27 361 596 70737      3 24 12 NA 66 1
## 3 1022     16 28 288 845 70737      3 28 12 NA 66 1
## 4 1022     27 29 378 774 70737      3 39 12 NA 66 1
## 5 1022     33 29 197 868 70737      3 45 12 NA 66 1
## 6 1022     46 30 39 362 70737      3 58 12 NA 66 1
##  vload0_cat
## 1      High
## 2      High
## 3      High
## 4      High
## 5      High
## 6      High
```

Inferential Analyses

Recall our scientific question of interest: is there an association between **baseline viral load** and **rate of decline** of CD4+ cell depletion?

Note: for the remainder of our analysis, we will consider what's known as the available data analysis; that is, we will simply ignore the observations with missing baseline viral loads.

Let Y_{ij} be the CD4+ cell counts for measurement j on individual i , t_{ij} be the month for individual i 's j th measurement since sero-conversion, and $vload0_i$ be the baseline viral load category for individual i . We will consider the following model:

$$E[Y_{ij}|t_{ij}, vload0_i] = \beta_0 + \beta_1 t_{ij} + \beta_2 I(vload0_i = \text{medium}) + \beta_3 I(vload0_i = \text{high}) + \quad (1)$$

$$\beta_4 t_{ij} \times I(vload0_i = \text{medium}) + \beta_5 t_{ij} \times I(vload0_i = \text{high}), \quad (2)$$

where our reference category is “low” baseline viral load and $I(A)$ is an indicator function that equals 1 if event A holds, and 0 otherwise.

Note that in the model above we include two interaction terms (corresponding to β_4 and β_5) because we are interested in understanding whether the **rate of decline** of CD4 is different for individuals who have medium (or high) viral load at baseline when compared to individuals with low viral load at baseline.

Your turn: To help us understand the model better, consider the following questions:

- In terms of the regression coefficients β , what is the rate of change/decline in CD4 cells for the group with low viral load at baseline?
- In terms of the regression coefficients β , what is the rate of change/decline in CD4 cells for the group with medium viral load at baseline?
- In terms of the regression coefficients β , what is the rate of change/decline in CD4 cells for the group with high viral load at baseline?
- Given the above, what is the null hypothesis corresponding to the scientific question of interest?

Linear Mixed Models

In this section, we will demonstrate how to analyze the data using linear mixed models (LMM), which *explicitly* distinguishes between *between-subject* and *within-subject* sources of variability. In this case, our results from an available data analysis are valid provided that (i) data are missing at random (MAR); and (ii) the likelihood function of LMM is correctly specified.

As covered in lectures, there are many choices to make when fitting a linear mixed effects model:

- Random intercepts? Random slopes? Both?

- Maximum likelihood? REML?

For demonstration purposes we will use REML, which is also the default in the `lme` function in **R**. **Note:** recall that with REML, we *cannot* use the likelihood ratio test to compare two models with different fixed effects.

In terms of random intercepts vs slopes, recall that random intercepts allows each subject to have their own level, but the rate of change is the same. Random slopes allows each subject to have their own rate of change.

A reasonable starting point to model the data here might be to allow each participant to have their own level of CD4 and rate of change (i.e., we are fitting a model with random intercepts and slopes).

We will fit the model using categorized baseline viral load, as discussed earlier:

$$E[Y_{ij}|t_{ij}, \text{vload0}_i, b_{0i}, b_{1i}] = \beta_0 + b_{0i} + (\beta_1 + b_{1i})t_{ij} + \beta_2 I(\text{vload0}_i = \text{medium}) + \beta_3 I(\text{vload0}_i = \text{high}) + \beta_4 t_{ij} \times I(\text{vload0}_i = \text{medium}) + \beta_5 t_{ij} \times I(\text{vload0}_i = \text{high}),$$

where

$$\begin{pmatrix} b_{0i} \\ b_{1i} \end{pmatrix} \stackrel{i.i.d.}{\sim} \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix}\right). \quad (3)$$

The following code fits the model and generates inferential summaries (e.g., CI).

```
model_lmm <- lme(fixed=cd4 ~ months*vload0_cat,
                 method="REML",
                 random=reStruct(~1 + months| id, REML=TRUE),
                 data=macs,
                 na.action=na.omit)
summary(model_lmm)
```

```
## Linear mixed-effects model fit by REML
##   Data: macs
##       AIC      BIC    logLik
##  19691.47 19744.42 -9835.733
##
## Random effects:
## Formula: ~1 + months | id
## Structure: General positive-definite, Log-Cholesky parametrization
##              StdDev   Corr
## (Intercept) 247.13352 (Intr)
## months       5.71936 -0.439
## Residual    142.23084
##
## Fixed effects:  cd4 ~ months * vload0_cat
##
##              Value Std.Error   DF   t-value p-value
## (Intercept)    800.0181  32.19763 1250  24.847114  0.0000
## months        -5.2886   0.87136 1250  -6.069411  0.0000
## vload0_catMedium -120.3072  45.20558  223  -2.661335  0.0083
## vload0_catHigh  -132.0101  45.16795  223  -2.922649  0.0038
## months:vload0_catMedium  0.2614   1.21698 1250   0.214769  0.8300
## months:vload0_catHigh  -2.3827   1.21438 1250  -1.962042  0.0500
## Correlation:
##              (Intr) months vld0_M vld0_H mn:0_M
## months          -0.548
## vload0_catMedium -0.712  0.390
## vload0_catHigh  -0.713  0.391  0.508
```

```
## months:vload0_catMedium 0.392 -0.716 -0.548 -0.280
## months:vload0_catHigh 0.393 -0.718 -0.280 -0.549 0.514
##
## Standardized Within-Group Residuals:
##      Min      Q1      Med      Q3      Max
## -4.03856716 -0.51978478 -0.05957684 0.42249730 6.86257347
##
## Number of Observations: 1479
## Number of Groups: 226
```

```
tidy(model_lmm, conf.int = TRUE)
```

```
## # A tibble: 10 x 10
##   effect group term estimate std.e~1 df stati~2 p.value conf.low
##   <chr> <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 fixed <NA> (Interc~ 800. 32.2 1250 24.8 4.38e-111 737.
## 2 fixed <NA> months -5.29 0.871 1250 -6.07 1.70e- 9 -7.00
## 3 fixed <NA> vload0_~ -120. 45.2 223 -2.66 8.35e- 3 -209.
## 4 fixed <NA> vload0_~ -132. 45.2 223 -2.92 3.83e- 3 -221.
## 5 fixed <NA> months:~ 0.261 1.22 1250 0.215 8.30e- 1 -2.13
## 6 fixed <NA> months:~ -2.38 1.21 1250 -1.96 5.00e- 2 -4.77
## 7 ran_pars id sd_(Int~ 247. NA NA NA NA 220.
## 8 ran_pars id cor_mon~ -0.439 NA NA NA NA -0.572
## 9 ran_pars id sd_mont~ 5.72 NA NA NA NA 4.91
## 10 ran_pars Residual sd_Obse~ 142. NA NA NA NA NA
## # ... with 1 more variable: conf.high <dbl>, and abbreviated variable names
## # 1: std.error, 2: statistic
```

From this model we can easily see what the estimated association between CD4 and time is for the low baseline viral load category: $\hat{\beta}_1$. To obtain estimates for the “Medium” and “High” categories, one option is to use the multcomp package:

```
model_lmm_med <- glht(model_lmm, linfct=c("months + months:vload0_catMedium = 0"))
confint(model_lmm_med)
```

```
##
## Simultaneous Confidence Intervals
##
## Fit: lme.formula(fixed = cd4 ~ months * vload0_cat, data = macs, random = reStruct(~1 +
## months | id, REML = TRUE), method = "REML", na.action = na.omit)
##
## Quantile = 1.96
## 95% family-wise confidence level
##
## Linear Hypotheses:
##              Estimate lwr      upr
## months + months:vload0_catMedium == 0 -5.0273 -6.6924 -3.3621
```

Your turn: Repeat the `glht()` command for the High baseline viral load group, and then summarize your findings (estimation and inference) with respect to the rate of change in CD4 cell count in each group.

We can also test for a statistically significant association between baseline viral load (category) and rate of change in CD4.

Your turn: Write down the null hypothesis corresponding to this test.

Here is the code for the hypothesis test:

```
print(anova(model_lmm)) # Wald test
```

```
##                numDF denDF    F-value p-value
## (Intercept)         1  1250 1482.1092 <.0001
## months              1  1250  148.7498 <.0001
## vload0_cat          2   223  11.7130 <.0001
## months:vload0_cat    2  1250   2.9405  0.0532
```

Your turn: Write down your p-value and conclusion from this hypothesis test.

Alternatively, if we fitted our models using ML (as opposed to REML), we can also apply a LRT. Adapt the following code so that it runs, and then summarize your findings:

```
model_lmm_ML <- lme(fixed=cd4 ~??,
  method="ML",
  random=??,
  data=macs,
  na.action=na.omit)
model_lmm_ML_reduced <- lme(fixed=cd4 ~ ??,
  method="ML",
  random=??,
  data=macs,
  na.action=na.omit)

print(anova(model_lmm_ML, model_lmm_ML_reduced)) # LRT
```

Note: if you force anova to compare two REML-fitted objects with different fixed effects, you will get a warning message stating that “REML comparisons are not meaningful”!

Sensitivity Analysis: A sensitivity analysis could be done where we compare models with the same systematic trend, but different covariance structures. Then, using the “best fitting” model, see how our inference changes.

Here are some models that we may consider:

1. Random intercepts only
2. Random intercepts + random slopes, uncorrelated
3. Random intercepts + random slopes, correlated (`model_lmm`)

We will use AIC to determine the best fitting model.

Your turn: Adapt the following code so that it runs (to fit models 1 and 2 above), and then summarize your findings – which model is “best”? Do your inferential results depend on the choice of covariance structure?

```
model_lmm_1 <- lme(fixed=??,
  method="REML",
  random=??,
  data=macs,
  na.action=na.omit)
model_lmm_2 <- lme(fixed=??,
  method="REML",
  random=??,
  data=macs,
  na.action=na.omit)
print(anova(model_lmm_1, model_lmm_2, model_lmm))
```

GEE

As an alternative to LME, we could consider modeling our data using GEE. The idea of GEE is that we treat the correlation structure as a nuisance – i.e., something that exists in the data but not of primary interest, and we need to acknowledge, but don't want to make assumptions about it. Our default will be to use empirical (also known as sandwich/robust) standard errors. Note that there is also a different approach to model fitting. GEE does *not* use a likelihood so there is no distinction of REML or ML, and likelihood ratio tests are *not* applicable.

Recall that we need to specify a *working covariance model* for GEE, which does not need to match the true covariance model for the regression coefficient estimate to be correct (in large samples). But if the working covariance is close to the true covariance, we can get efficiency gains (i.e., higher power). Recall that the parameter estimates and estimated standard errors do depend on the working covariance matrix (so you will get different, but valid results, depending on the working covariance model you choose). There is a stricter assumption regarding the missing data — for valid inference, data need to be missing completely at random (MCAR).

So what working covariance model should you use? As with pretty much every modeling choice, it depends on the application. If you have knowledge of the true correlation structure, it makes sense to include it. If not, working independence is *usually* OK for many practical purposes.

```
# geeglm assumes that the input data is organized by subject id and then time
macs <- macs %>%
  arrange(id, months)
```

Next we will fit the model using GEE with working independence model after removing the subjects with missing baseline viral data.

```
mod_gee_ind <- macs %>%
  dplyr::select(cd4, months, vload0_cat, id) %>% # select variables for analysis
  na.omit() %>% # exclude cases with missing data on those variables
  geeglm(cd4 ~ months*vload0_cat, id = id, data = ., corstr = "independence")
print(anova(mod_gee_ind))
```

```
## Analysis of 'Wald statistic' Table
## Model: gaussian, link: identity
## Response: cd4
## Terms added sequentially (first to last)
##
##              Df      X2 P(>|Chi|)
## months          1 108.346 < 2.2e-16 ***
## vload0_cat       2  19.549 5.688e-05 ***
## months:vload0_cat 2   4.405  0.1106
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Your turn: What do we conclude – do we have evidence for an association between baseline viral load and rate of change in CD4 counts?

As we did with the linear mixed model formulation we can use the `glht()` function to get estimates and confidence intervals for linear combinations of the parameters (that way we can get estimates for the rate of change in mean CD4 in the medium and high groups):

```
tidy(mod_gee_ind, conf.int=T)
```

```
## # A tibble: 6 x 7
##   term                estimate std.error statistic  p.value conf.~1 conf.~2
##   <chr>              <dbl>    <dbl>    <dbl>    <dbl>    <dbl>    <dbl>
## 1 (Intercept)        791.      36.8    462.      0        719.     864.
```

```
## 2 months          -4.84      1.13  18.3    0.0000193   -7.06  -2.62
## 3 vload0_catMedium -123.      46.4   7.04    0.00797   -214.  -32.2
## 4 vload0_catHigh   -142.      46.3   9.34    0.00224   -232.  -50.8
## 5 months:vload0_catMedium  0.122    1.40   0.00759  0.931     -2.62   2.86
## 6 months:vload0_catHigh   -1.93    1.33   2.12    0.146     -4.54   0.671
## # ... with abbreviated variable names 1: conf.low, 2: conf.high
```

```
mod_gee_ind_med <- glht(mod_gee_ind, linfct=c("months + months:vload0_catMedium = 0"))
confint(mod_gee_ind_med)
```

```
##
## Simultaneous Confidence Intervals
##
## Fit: geeglm(formula = cd4 ~ months * vload0_cat, data = ., id = id,
## corstr = "independence")
##
## Quantile = 1.96
## 95% family-wise confidence level
##
## Linear Hypotheses:
##                                Estimate lwr      upr
## months + months:vload0_catMedium == 0 -4.7213  -6.3283  -3.1144
```

```
mod_gee_ind_high <- glht(mod_gee_ind, linfct=c("months + months:vload0_catHigh = 0"))
confint(mod_gee_ind_high)
```

```
##
## Simultaneous Confidence Intervals
##
## Fit: geeglm(formula = cd4 ~ months * vload0_cat, data = ., id = id,
## corstr = "independence")
##
## Quantile = 1.96
## 95% family-wise confidence level
##
## Linear Hypotheses:
##                                Estimate lwr      upr
## months + months:vload0_catHigh == 0 -6.7754  -8.1326  -5.4182
```

Comparison of results: For categorized baseline viral load, the estimated rates of change in CD4 counts over time are included as below:

	LMM (rand int&slope)	GEE (working ind)
Low	-5.3 (-7.0, -3.6)	-4.8 (-7.1, -2.6)
Medium	-5.0 (-6.7, -3.4)	-4.7 (-6.3, -3.1)
High	-7.7 (-9.3, -6.0)	-6.8 (-8.1, -5.4)

As a sensitivity analysis, we will fit the same model using GEE with working exchangeable model.

Your turn: Complete the code below to fit GEE with a working exchangeable covariance structure, then summarize your conclusions.


```
mod_gee_exch <- geeglm(???)
print(anova(mod_gee_exch))
```

With `summary()` and a non-independence working correlation model, we can also extract the estimated correlation between any two observations within a subject: 0.686 (given by the `alpha` estimate under “Estimated Correlation Parameters”). This indicates that the observations within the same subject are quite correlated empirically.

```
summary(mod_gee_exch)
```

Note: Working independence or exchangeable correlation models are quite easy to apply to most study designs. In this case, because our data are unbalanced (i.e., measurement times for each individual are not the same) and there are many observations within each subject (most individuals have 4-8 observations), we cannot directly use other working covariance models such as the AR-1 model or the unstructured model.

Bonus: Fill in the same table but for the models we considered in the sensitivity analysis: **Fill in your answers**

	LMM (rand int) G	EE (exchangeable)
Low		
Medium		
High		

References

- Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2018), Applied Longitudinal Analysis, Wiley & Sons, Limited, John.
- Kaslow, R. A., Ostrow, D. G., Detels, R., Phair, J. P., Polk, B. F., and Rinaldo, C. R., Jr (1987), “The Multicenter AIDS Cohort Study: rationale, organization, and selected characteristics of the participants,” American Journal of Epidemiology, 126, 310–318. <https://doi.org/10.1093/aje/126.2.310>.
- Van Belle, G., Fisher, L. D., Heagerty, P. J., & Lumley, T. (2004), Biostatistics: a methodology for the health sciences, John Wiley & Sons.