lmm\_chop.R

lucp12431

2024-09-30

#---------------------------------------------------  
# FEDERATED DATA ANALYSIS FOR LINEAR [MIXED] MODELS  
# USING CHOP DATA  
#---------------------------------------------------

# Clear memory  
rm(list=ls(all=TRUE))  
  
# Set working directory  
#setwd("G:/My Drive/PhD/FDA")

# Install packages  
#install.packages('medicaldata')  
#install.packages('skimr')  
#install.packages('purrr')  
#install.packages('dplyr')  
  
## I. Load packages and data ------  
library(medicaldata)  
library(skimr)  
library(purrr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

data("covid\_testing")  
skim(covid\_testing) #overview of variables

Data summary

|  |  |
| --- | --- |
| Name | covid\_testing |
| Number of rows | 15524 |
| Number of columns | 17 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| character | 9 |
| numeric | 8 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: character**

| skim\_variable | n\_missing | complete\_rate | min | max | empty | n\_unique | whitespace |
| --- | --- | --- | --- | --- | --- | --- | --- |
| fake\_first\_name | 0 | 1.00 | 3 | 13 | 0 | 832 | 0 |
| fake\_last\_name | 0 | 1.00 | 4 | 11 | 0 | 27 | 0 |
| gender | 0 | 1.00 | 4 | 6 | 0 | 2 | 0 |
| test\_id | 0 | 1.00 | 5 | 5 | 0 | 2 | 0 |
| clinic\_name | 0 | 1.00 | 3 | 32 | 0 | 88 | 0 |
| result | 0 | 1.00 | 7 | 8 | 0 | 3 | 0 |
| demo\_group | 0 | 1.00 | 6 | 12 | 0 | 5 | 0 |
| payor\_group | 7087 | 0.54 | 5 | 18 | 0 | 7 | 0 |
| patient\_class | 7077 | 0.54 | 9 | 23 | 0 | 9 | 0 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 | hist |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| subject\_id | 0 | 1.00 | 5570.65 | 3613.70 | 1.00 | 2329.75 | 5268.0 | 8636.25 | 12346.0 | ▇▆▆▆▅ |
| pan\_day | 0 | 1.00 | 63.21 | 27.22 | 4.00 | 38.00 | 65.0 | 87.00 | 107.0 | ▃▇▆▇▇ |
| age | 0 | 1.00 | 14.19 | 16.47 | 0.00 | 2.00 | 9.0 | 18.00 | 138.0 | ▇▁▁▁▁ |
| drive\_thru\_ind | 0 | 1.00 | 0.51 | 0.50 | 0.00 | 0.00 | 1.0 | 1.00 | 1.0 | ▇▁▁▁▇ |
| ct\_result | 209 | 0.99 | 44.12 | 3.98 | 14.05 | 45.00 | 45.0 | 45.00 | 45.0 | ▁▁▁▁▇ |
| orderset | 0 | 1.00 | 0.70 | 0.46 | 0.00 | 0.00 | 1.0 | 1.00 | 1.0 | ▃▁▁▁▇ |
| col\_rec\_tat | 0 | 1.00 | 7.22 | 492.82 | 0.00 | 0.70 | 1.9 | 3.60 | 61370.2 | ▇▁▁▁▁ |
| rec\_ver\_tat | 0 | 1.00 | 5.64 | 5.09 | -18.60 | 4.00 | 5.0 | 6.20 | 218.2 | ▇▁▁▁▁ |

## II. Preprocess data ------  
# 1. include only valid values for ct\_result  
# 2. select variables for the analysis  
# 3. convert from numeric to factor  
data <- covid\_testing %>%  
 filter(ct\_result > 0, result %in% c('positive', 'negative')) %>%  
 dplyr::select(clinic\_name, result, ct\_result, gender, age, drive\_thru\_ind) %>%  
 mutate(drive\_thru\_ind\_factor = as.factor(drive\_thru\_ind), drive\_thru\_ind = NULL)  
  
## III. Construct summary data table PER GROUP ------  
# Include also summaries for interactions, log transformations, and quadratic terms  
  
# A. Standardize numeric variablesm ------  
data <- data %>%  
 mutate(across(where(is.numeric), ~ scale(.)[, 1], .names = "std\_{.col}"))  
  
# B. Log-transform numeric variables ------  
data <- data %>%  
 mutate(across(where(is.numeric) &  
 !starts\_with("std\_"),  
 ~ log(.x),   
 .names = "log\_{.col}"))  
  
# C. Compute quadratic terms for numeric variables ------  
data <- data %>%  
 mutate(across(where(is.numeric) &  
 !starts\_with("std\_") &  
 !starts\_with("log\_"),  
 ~ (.x)^2,   
 .names = "{.col}\_sq"))  
  
# D. Compute interaction columns ------  
# 1. Construct design matrix (as dataframe) from current dataset  
data\_design\_df <- as.data.frame(model.matrix(~ ., data %>% dplyr::select(-c(clinic\_name))))  
data\_design\_df$clinic\_name <- data$clinic\_name  
# 2. Break up data per group  
group\_data\_design\_df <- data\_design\_df %>%  
 group\_split(clinic\_name) %>%  
 keep(~ nrow(.x) > 1)  
# 3. Extract unique group names  
unique\_group\_names <- do.call(rbind, group\_data\_design\_df) %>%  
 pull(clinic\_name) %>%  
 unique()  
# 4. Assign names to the list elements  
names(group\_data\_design\_df) <- unique\_group\_names  
# 5. Create variable names for interaction terms  
var\_names\_for\_interaction <- combn(names(group\_data\_design\_df[[1]] %>%  
 dplyr::select(-starts\_with('log') & -ends\_with('\_sq') & -'(Intercept)' & -starts\_with('clinic'))), 2)  
var\_names\_of\_interaction <- apply(var\_names\_for\_interaction, 2,   
 function(vars){paste0(vars[1],'X',vars[2])})  
# 6. Include interaction columns in dataframe  
group\_data\_design\_df <- lapply(group\_data\_design\_df,  
 function(grp){  
 interaction\_columns <- do.call(cbind,  
 apply(var\_names\_for\_interaction, 2, function(var\_name) grp[,var\_name[1]]\*grp[,var\_name[2]]))  
 names(interaction\_columns) <- var\_names\_of\_interaction  
 cbind(grp, interaction\_columns)  
 })  
  
# E. Compute sufficient statistics per hospital ------  
summary\_stats <- lapply(group\_data\_design\_df,  
 function(grp){  
 data.frame(  
 variable = names(grp[, c(-1)] %>%   
 dplyr::select(-(starts\_with('clinic')))),  
 n = rep(nrow(grp), length(names(grp[, c(-1)] %>%  
 dplyr::select(-(starts\_with('clinic')))))),  
 mean = apply(grp[, c(-1)] %>%  
 dplyr::select(-(starts\_with('clinic'))),  
 2, mean, na.rm = TRUE),  
 variance = ifelse(is.na(apply(grp[, c(-1)] %>%  
 dplyr::select(-(starts\_with('clinic'))),  
 2, var)), 0, apply(grp[, c(-1)] %>%  
 dplyr::select(-(starts\_with('clinic'))),  
 2, var)),  
 row.names = NULL)  
 })  
var\_cov\_mat <- lapply(group\_data\_design\_df,  
 function(grp) cov(grp[, c(-1)] %>%  
 dplyr::select(-(starts\_with('clinic')))))  
var\_cov\_mat <- lapply(var\_cov\_mat, function(grp) apply(grp, 2, function(column) ifelse(is.na(column),0,column)))  
  
# Merge all dataframes into one  
summary\_stats\_one <- do.call(rbind, summary\_stats)  
row.names(summary\_stats\_one) <- NULL  
summary\_stats\_one$clinic\_name <- rep(names(summary\_stats),   
 each = nrow(summary\_stats[[1]]))  
var\_cov\_mat\_one <- as.data.frame(do.call(rbind, var\_cov\_mat))  
var\_cov\_mat\_one$clinic\_name <- rep(names(var\_cov\_mat),   
 each = nrow(var\_cov\_mat[[1]]))  
write.csv(summary\_stats\_one, file = "summary\_stats\_chop.csv")  
write.csv(var\_cov\_mat\_one, file = "var\_cov\_mat\_chop.csv")

## I. Load packages / functions / [summary] data ------  
#install.packages('lme4')  
library(lme4)

## Loading required package: Matrix

source('pseudo\_data\_gen\_fn.R') #modified mvrnorm  
source('pseudo\_data\_ls\_fn.R')   
#other options are source('pseudo\_data\_gen\_mvrnorm\_option\_fn.R') or source('pseudo\_data\_gen\_chol\_option\_fn.R')  
  
id\_summary\_stats <- "1dw6y5CN1d1Kuh5PMyZ7Nh6\_5kbyrpN4L"  
summary\_stats\_one <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download", id\_summary\_stats))  
summary\_stats <- summary\_stats\_one %>%   
 dplyr::select(-1) %>%   
 split(f = as.factor(.$clinic\_name)) %>%  
 lapply(function(df){  
 df[-ncol(df)]  
 })  
  
id\_var\_cov\_mat <- "1dwuaHNbUAtnzV8FtBiH7H\_isRqkHjT1w"  
var\_cov\_mat\_one <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download", id\_var\_cov\_mat))  
var\_cov\_mat <- var\_cov\_mat\_one %>%   
 dplyr::select(-1) %>%  
 split(f = as.factor(.$clinic\_name)) %>%  
 lapply(function(df) {  
 rownames(df) <- colnames(df)[-ncol(df)]  
 return(as.matrix(df[-ncol(df)]))  
 })  
  
# Preview of summary statistics from 1 data provider  
vars <- c('log\_ct\_result', 'gendermale', 'age', 'drive\_thru\_ind\_factor1', 'gendermaleXage')  
summary\_stats[[26]][summary\_stats[[26]][,'variable'] %in% vars, ]

## variable n mean variance  
## 803 gendermale 208 0.528846154 2.503716e-01  
## 804 age 208 1.372596154 1.050606e+01  
## 805 drive\_thru\_ind\_factor1 208 0.004807692 4.807692e-03  
## 808 log\_ct\_result 208 3.803002662 5.554545e-04  
## 823 gendermaleXage 208 0.778846154 7.085444e+00

round(var\_cov\_mat[[26]][vars,vars], 3)

## log\_ct\_result gendermale age drive\_thru\_ind\_factor1  
## log\_ct\_result 0.001 0.001 0.001 0.000  
## gendermale 0.001 0.250 0.053 0.002  
## age 0.001 0.053 10.506 0.003  
## drive\_thru\_ind\_factor1 0.000 0.002 0.003 0.005  
## gendermaleXage -0.001 0.369 6.621 0.006  
## gendermaleXage  
## log\_ct\_result -0.001  
## gendermale 0.369  
## age 6.621  
## drive\_thru\_ind\_factor1 0.006  
## gendermaleXage 7.085

## II. Generate pseudo-data ------  
# A. Provide model specification ------  
formula <- log\_ct\_result ~ gendermale + std\_age + drive\_thru\_ind\_factor1 + gendermaleXstd\_age   
  
  
  
  
# B. Generate pseudo-data ------  
# (initial random numbers come from a std uniform distribution)  
set.seed(121314)  
pseudo\_data <- lapply(names(summary\_stats), pseudo\_data\_ls\_fn)  
names(pseudo\_data) <- names(summary\_stats)  
  
# C. Combine [list] into one df ------  
pooled\_pseudo\_data <- do.call(rbind, pseudo\_data)  
  
## III. Estimate LMM using pseudo-data  
# A. With random-intercept only ------  
lmm\_pseudo <- lmer(log\_ct\_result ~ gendermale + std\_age + drive\_thru\_ind\_factor1 + gendermaleXstd\_age + (1|clinic\_name), data = pooled\_pseudo\_data)  
summary(lmm\_pseudo)

## Linear mixed model fit by REML ['lmerMod']  
## Formula: log\_ct\_result ~ gendermale + std\_age + drive\_thru\_ind\_factor1 +   
## gendermaleXstd\_age + (1 | clinic\_name)  
## Data: pooled\_pseudo\_data  
##   
## REML criterion at convergence: -20473  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -4.5919 -0.5799 0.0249 0.5762 3.7464   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## clinic\_name (Intercept) 0.0004689 0.02165   
## Residual 0.0149360 0.12221   
## Number of obs: 15068, groups: clinic\_name, 70  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 3.787066 0.003947 959.587  
## gendermale 0.002088 0.001995 1.047  
## std\_age -0.004576 0.001545 -2.961  
## drive\_thru\_ind\_factor1 -0.004276 0.005802 -0.737  
## gendermaleXstd\_age -0.006105 0.001997 -3.058  
##   
## Correlation of Fixed Effects:  
## (Intr) gndrml std\_ag dr\_\_\_1  
## gendermale -0.255   
## std\_age 0.013 -0.005   
## drv\_thr\_n\_1 -0.145 0.013 0.020   
## gndrmlXstd\_ 0.003 0.000 -0.658 -0.003

# Compare with LMM using actual data  
pooled\_actual\_data <- data %>% filter(clinic\_name %in% names(group\_data\_design\_df))  
lmm\_actual <- lmer(log\_ct\_result ~ gender \* std\_age + drive\_thru\_ind\_factor + (1|clinic\_name), data = pooled\_actual\_data)  
summary(lmm\_actual)

## Linear mixed model fit by REML ['lmerMod']  
## Formula: log\_ct\_result ~ gender \* std\_age + drive\_thru\_ind\_factor + (1 |   
## clinic\_name)  
## Data: pooled\_actual\_data  
##   
## REML criterion at convergence: -20473  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -9.3287 0.1482 0.2174 0.2538 1.1855   
##   
## Random effects:  
## Groups Name Variance Std.Dev.  
## clinic\_name (Intercept) 0.0004689 0.02165   
## Residual 0.0149360 0.12221   
## Number of obs: 15068, groups: clinic\_name, 70  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 3.787066 0.003947 959.587  
## gendermale 0.002088 0.001995 1.047  
## std\_age -0.004576 0.001545 -2.961  
## drive\_thru\_ind\_factor1 -0.004276 0.005802 -0.737  
## gendermale:std\_age -0.006105 0.001997 -3.058  
##   
## Correlation of Fixed Effects:  
## (Intr) gndrml std\_ag dr\_\_\_1  
## gendermale -0.255   
## std\_age 0.013 -0.005   
## drv\_thr\_n\_1 -0.145 0.013 0.020   
## gndrml:std\_ 0.003 0.000 -0.658 -0.003

# Compute AIC and confidence intervals (random intercept only)  
AIC(lmm\_pseudo)

## [1] -20459.04

AIC(lmm\_actual)

## [1] -20459.04

BIC(lmm\_pseudo)

## [1] -20405.7

BIC(lmm\_actual)

## [1] -20405.7

confint(lmm\_pseudo)

## Computing profile confidence intervals ...

## 2.5 % 97.5 %  
## .sig01 0.016022295 0.028193030  
## .sigma 0.120828350 0.123591823  
## (Intercept) 3.779332754 3.794900771  
## gendermale -0.001822568 0.005996308  
## std\_age -0.007601562 -0.001547158  
## drive\_thru\_ind\_factor1 -0.015632588 0.007088819  
## gendermaleXstd\_age -0.010024750 -0.002196225

confint(lmm\_actual)

## Computing profile confidence intervals ...

## 2.5 % 97.5 %  
## .sig01 0.016022296 0.028193029  
## .sigma 0.120828350 0.123591823  
## (Intercept) 3.779332754 3.794900771  
## gendermale -0.001822568 0.005996308  
## std\_age -0.007601562 -0.001547158  
## drive\_thru\_ind\_factor1 -0.015632588 0.007088819  
## gendermale:std\_age -0.010024750 -0.002196225

# B. With random intercept+slope) ------  
lmm\_pseudo <- lmer(log\_ct\_result ~ gendermale + std\_age + drive\_thru\_ind\_factor1 + gendermaleXstd\_age + (1+std\_age|clinic\_name), data = pooled\_pseudo\_data)  
summary(lmm\_pseudo)

## Linear mixed model fit by REML ['lmerMod']  
## Formula: log\_ct\_result ~ gendermale + std\_age + drive\_thru\_ind\_factor1 +   
## gendermaleXstd\_age + (1 + std\_age | clinic\_name)  
## Data: pooled\_pseudo\_data  
##   
## REML criterion at convergence: -20513.2  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -4.7364 -0.5806 0.0240 0.5779 3.7682   
##   
## Random effects:  
## Groups Name Variance Std.Dev. Corr   
## clinic\_name (Intercept) 0.0006220 0.02494   
## std\_age 0.0001643 0.01282 -0.10  
## Residual 0.0148649 0.12192   
## Number of obs: 15068, groups: clinic\_name, 70  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 3.7851418 0.0044954 841.996  
## gendermale 0.0020821 0.0019919 1.045  
## std\_age -0.0005126 0.0036856 -0.139  
## drive\_thru\_ind\_factor1 -0.0037530 0.0058529 -0.641  
## gendermaleXstd\_age -0.0052269 0.0020098 -2.601  
##   
## Correlation of Fixed Effects:  
## (Intr) gndrml std\_ag dr\_\_\_1  
## gendermale -0.225   
## std\_age 0.115 -0.007   
## drv\_thr\_n\_1 -0.126 0.013 0.021   
## gndrmlXstd\_ -0.003 0.003 -0.284 -0.002

# Compare with LMM using actual data (random intercept+slope)  
lmm\_actual <- lmer(log\_ct\_result ~ gender \* std\_age + drive\_thru\_ind\_factor + (1+std\_age|clinic\_name), data = pooled\_actual\_data)  
summary(lmm\_actual)

## Linear mixed model fit by REML ['lmerMod']  
## Formula: log\_ct\_result ~ gender \* std\_age + drive\_thru\_ind\_factor + (1 +   
## std\_age | clinic\_name)  
## Data: pooled\_actual\_data  
##   
## REML criterion at convergence: -20513.2  
##   
## Scaled residuals:   
## Min 1Q Median 3Q Max   
## -9.3604 0.1229 0.2030 0.2560 1.8698   
##   
## Random effects:  
## Groups Name Variance Std.Dev. Corr   
## clinic\_name (Intercept) 0.0006220 0.02494   
## std\_age 0.0001643 0.01282 -0.10  
## Residual 0.0148649 0.12192   
## Number of obs: 15068, groups: clinic\_name, 70  
##   
## Fixed effects:  
## Estimate Std. Error t value  
## (Intercept) 3.7851418 0.0044954 841.996  
## gendermale 0.0020821 0.0019919 1.045  
## std\_age -0.0005126 0.0036856 -0.139  
## drive\_thru\_ind\_factor1 -0.0037530 0.0058529 -0.641  
## gendermale:std\_age -0.0052269 0.0020098 -2.601  
##   
## Correlation of Fixed Effects:  
## (Intr) gndrml std\_ag dr\_\_\_1  
## gendermale -0.225   
## std\_age 0.115 -0.007   
## drv\_thr\_n\_1 -0.126 0.013 0.021   
## gndrml:std\_ -0.003 0.003 -0.284 -0.002

# Compute AIC and confidence intervals (random intercept+slope)  
AIC(lmm\_pseudo)

## [1] -20495.15

AIC(lmm\_actual)

## [1] -20495.15

BIC(lmm\_pseudo)

## [1] -20426.57

BIC(lmm\_actual)

## [1] -20426.57

confint(lmm\_pseudo)

## Computing profile confidence intervals ...

## 2.5 % 97.5 %  
## .sig01 0.018517717 0.032316506  
## .sig02 -0.579652278 0.391660240  
## .sig03 0.007933868 0.018998638  
## .sigma 0.120542665 0.123301809  
## (Intercept) 3.776227341 3.794209958  
## gendermale -0.001824739 0.005984991  
## std\_age -0.008058855 0.007360131  
## drive\_thru\_ind\_factor1 -0.015230494 0.007718764  
## gendermaleXstd\_age -0.009198163 -0.001297890

confint(lmm\_actual)

## Computing profile confidence intervals ...

## 2.5 % 97.5 %  
## .sig01 0.018517685 0.032316502  
## .sig02 -0.579664309 0.391678681  
## .sig03 0.007928129 0.018998639  
## .sigma 0.120542665 0.123301809  
## (Intercept) 3.776227341 3.794209958  
## gendermale -0.001824739 0.005984991  
## std\_age -0.008058855 0.007360143  
## drive\_thru\_ind\_factor1 -0.015230494 0.007718764  
## gendermale:std\_age -0.009198163 -0.001297891

# Perform posthoc procedures via lmerTest  
# lmm\_pseudo <- lmerTest::lmer(log\_ct\_result ~ gendermale + age + drive\_thru\_ind\_factor1 + gendermaleXage + (1|clinic\_name), data = pooled\_pseudo\_data)  
# summary(lmm\_pseudo)  
# lmm\_actual <- lmerTest::lmer(log\_ct\_result ~ gender \* std\_age + drive\_thru\_ind\_factor + (1|clinic\_name), data = pooled\_actual\_data)  
# summary(lmm\_actual)  
#   
# anova(lmm\_pseudo, type = 2, ddf = "Kenward-Roger")  
# anova(lmm\_actual, type = 2, ddf = "Kenward-Roger")