Biost 540: Homework 3

Department of Biostatistics @ University of Washington

Alejandro Hernandez

Due May 30, April 2024

Problem 1

In this problem, we will still focus on the Framingham study which we have explored in HW2. Please check the detailed description from HW2. For the pre processing step, please follow the instruction of HW2 Question2 a and b (You can basically use the code in solution). Incorrect pre processing may result in wrong results for this homework.

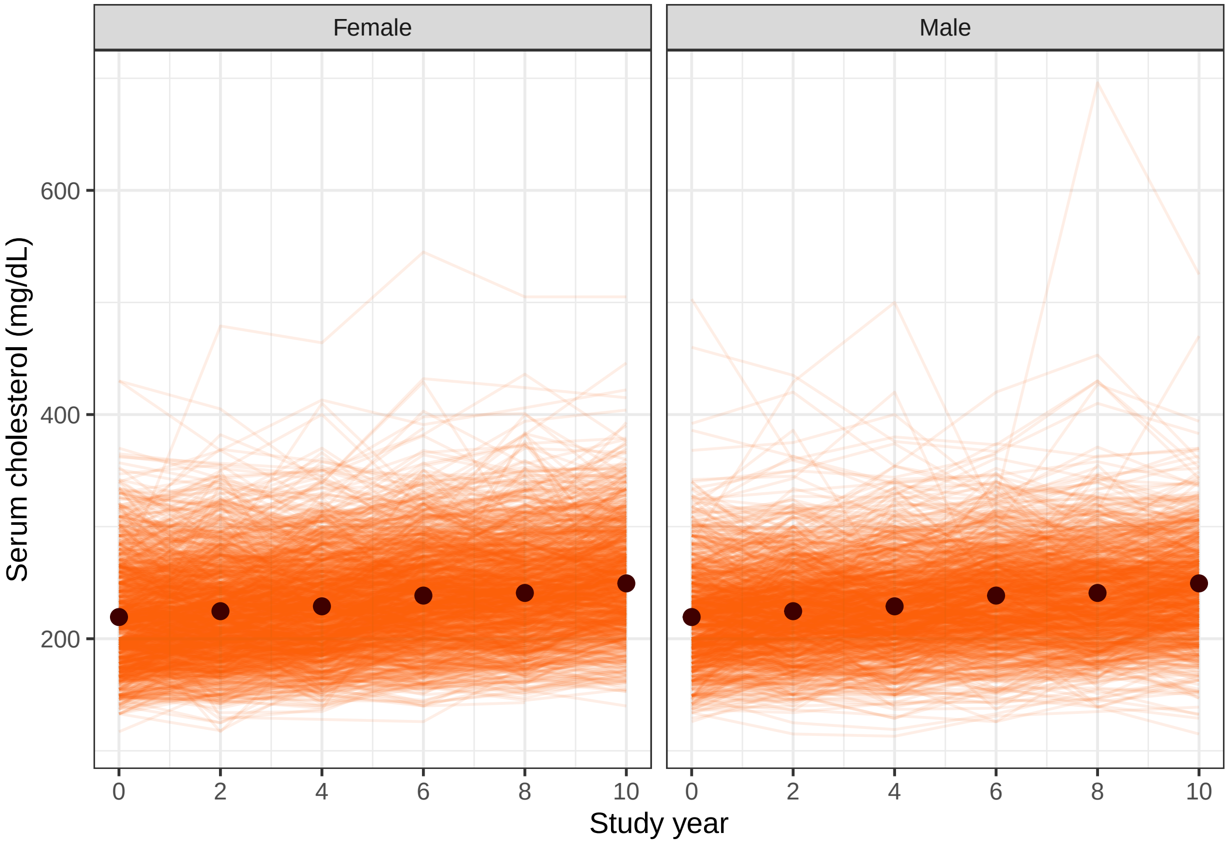


Figure 1: Cholesterol levels and averages over time by sex

(a)

In HW2, we have studied how cholesterol level changes over time and how it is related to baseline age, sex, and baseline BMI without considering interactions. Please provide statistical evidence on whether the relation between cholesterol level and time will change for different sexes. You could just focus on LMM with random intercept and slope. And please treat time as continuous variable instead of categorical variable.

Table 1: Coefficient estimates from linear mixed-effect model

	Estimate	Std. Error	t value
(Intercept)	137.806	5.247	26.262
year	3.447	0.093	37.244
sexMale	2.627	1.527	1.720
age0	1.309	0.090	14.595
bmi0	0.938	0.173	5.407
year:sexMale	-1.106	0.137	-8.092

From a linear mixed effects model, we estimate that cholesterol over time increases, on average, for both female and male participants. The rate of increase between male and females differs on average by 1.11 (95% CI: 1.24, 0.97) units (with the female group having the higher average rate of increase). Therefore, we conclude that the rate of change in cholesterol over time differ significantly between sexes.

(b)

Based on your conclusion from 2a, please fit GEE models using same covariates(i.e. If you think there should be interaction, you should include it in your GEE models.) but with independent, exchangeable, and AR1 working correlation matrix.

(c)

Please make a neat table to compare the results of LMM model in 2a and 2 GEE models (independence, exchangeable) in 2b. You should include estimates and sd of coefficients. What conclusion could you draw from the comparison. Please keep the result to three decimal places.

Table 2: Coefficient estimates and corresponding standard errors

	LMM	SE	GEE.indp	SE	GEE.exch	SE
(Intercept)	137.806	5.247	142.744	5.528	142.254	5.360
year	3.447	0.093	3.493	0.105	3.459	0.096
sexMale	2.627	1.527	2.921	1.538	2.605	1.528
age0	1.309	0.090	1.182	0.096	1.236	0.093
bmi0	0.938	0.173	0.954	0.191	0.885	0.186
year:sexMale	-1.106	0.137	-1.107	0.149	-1.116	0.136

Table 1 shows coefficient estimates and corresponding standard errors for linear mixed-effects model (LMM) and generalized estimating equation (GEE) models with independent and exchangeable working correlations. The effect estimates and standard errors are very similar between each model.

All models agree in their estimation that cholesterol over time increases, on average, across all strata. They also all conclude that the rate of change in cholesterol over time differ significantly between sexes.

Problem 2

Consider the Six City data set which describes Mother's smoking behavior and childhood respiratory disease. Columns are id: Child's ID, resp: binary indicator of respiratory disease, age: standardized age (ages 6-9 minus 8), smok: mother's smoking, aXs: interaction of age and smoke.

(a)

Please fit a GLMM, GEE, and transition model regressing indicator of respiratory disease on age, maternal smoking status, and interactions. Describe the output of each model (you can focus on the smoke variable when printing output). Also provide a brief description of how the assumptions made by each model differ from one another. Hint: see Lecture 3 pg 49, 56, 64 for help with interpretation and assumptions.

Table 3: Estimates and corresponding standard errors

	GLMM	SE	GEE	SE	Transition	SE
(Intercept)	-3.402	0.279	-1.900	0.119	-2.475	0.117
age	-0.217	0.087	-0.141	0.058	-0.187	0.113
smoke	0.478	0.299	0.314	0.188	0.284	0.155
ageXsmoke	0.105	0.139	0.071	0.088	-0.142	0.187
disease.lag	-	-	-	-	2.216	0.188

The GLMM estimates the difference in log odds of having respiratory disease - comparing a child whose mother smokes to a child whose mother does not smoke- is 0.478 (95% CI: 0.179, 0.777), with the later group facing lower risk. Both the GEE and transition models agree with this conclusion, but with smaller effect sizes: 0.314 (95% CI: 0.126, 0.502) and 0.284 (95% CI: 0.129, 0.439), respectively.

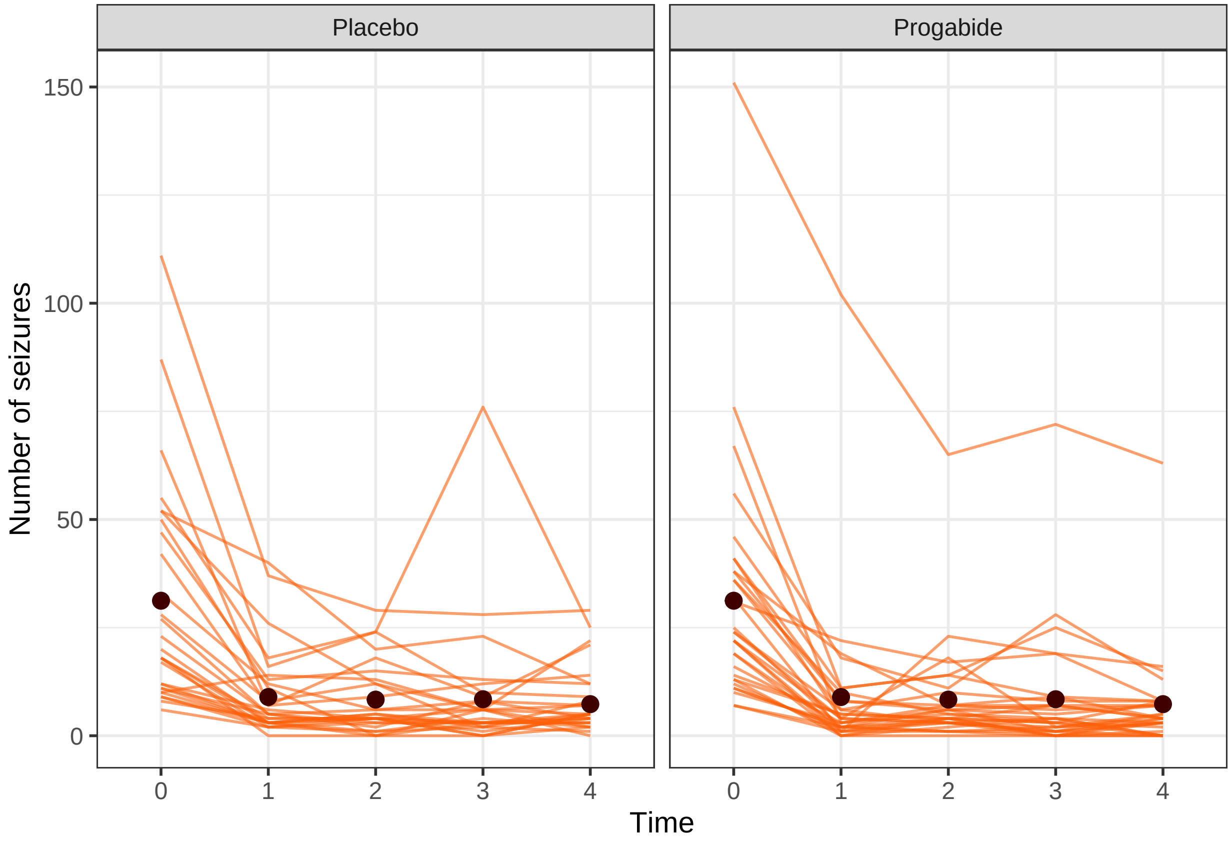
Problem 3

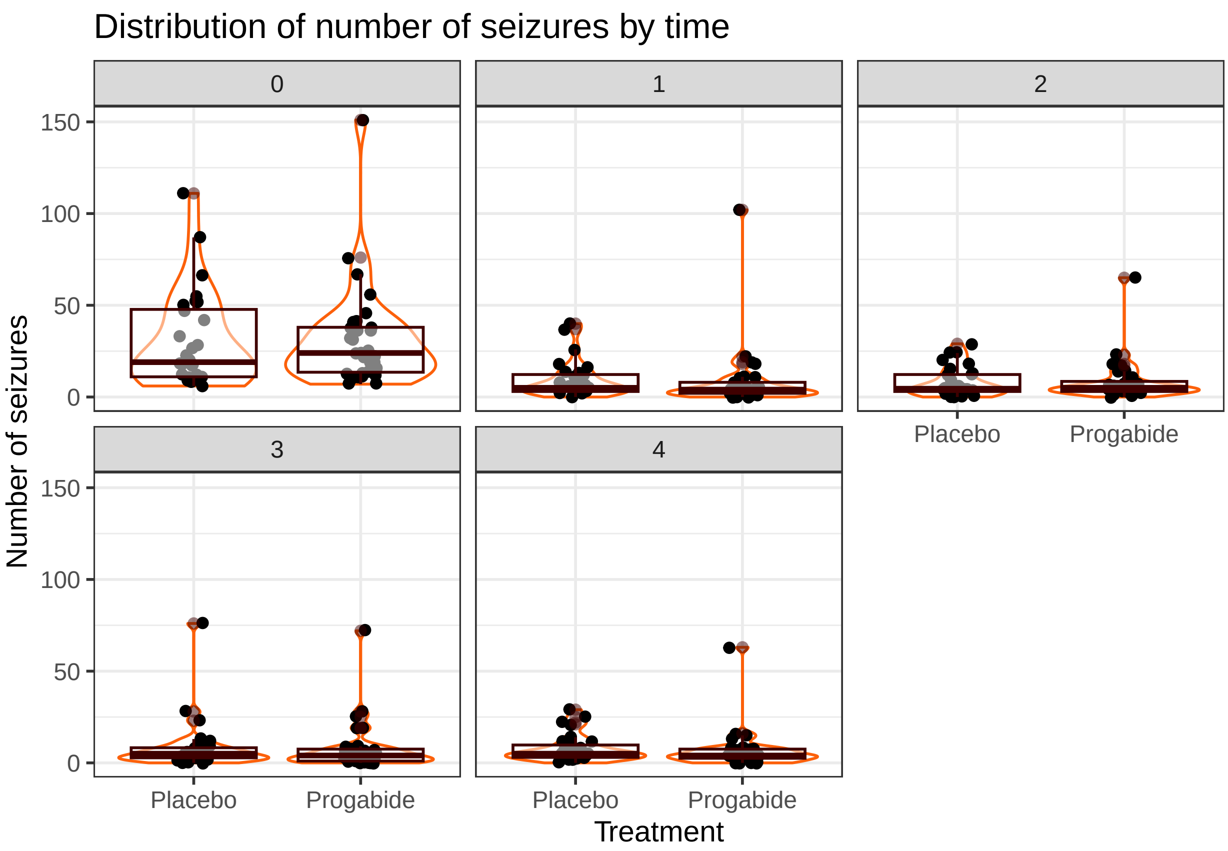
Consider the seizure dataset. The study consisted of 59 patients randomized to the anti-epileptic drug progabide, or to placebo in addition to standard chemotherapy. Over an 8-week period prior to randomization, a "baseline" number of seizures was recorded for each participant. Over four subsequent follow-up time periods the number of seizures in each 2-week period was recorded. Listed below are the variables in the dataset + id: patient id + age: age of the patient + tx: treatment (placebo or progabide) + y0, y1, y2, y3, y4: number of seizures by visit times subjects were randomized. y0 refers to the baseline number of seizures.

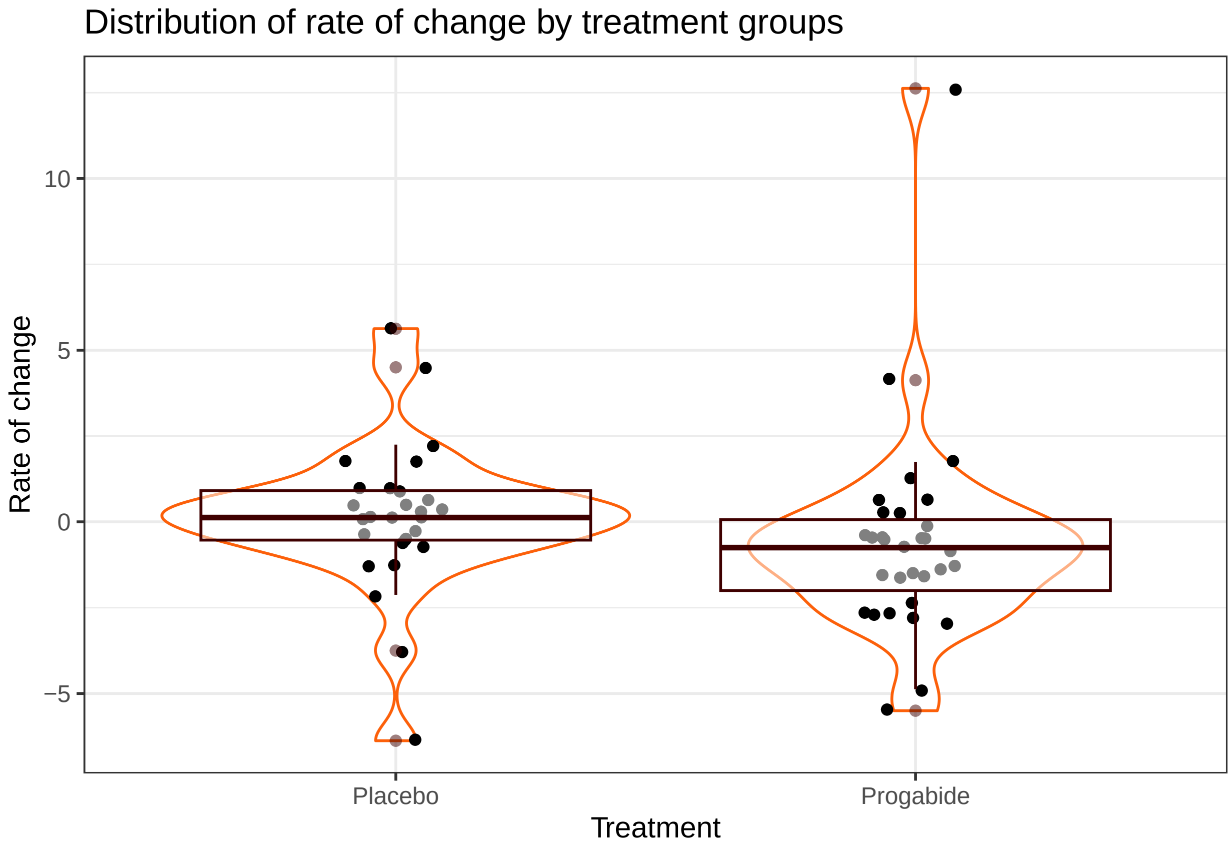
We are interested in evaluating whether the drug progabide is effective at reducing the rate of epileptic seizures.

(a)

Produce a graphical/tabular summary of the rates of seizure incidence at baseline and over the subsequent study visits by treatment arm. Feel free to summarize the individual rates and/or the average rates within each arm. Also pay particular attention to the length of the "at risk" period at each time point and use informative labels for study visits.


Placebo
Progabide
Number of seizures
150
100
50
0
0
1
2
3
4
Time


Distribution of number of seizures by time
0
1
2
3
4
150
100
50
0
Number of seizures
Placebo
Progabide
Treatment



(b)

Describe and fit a Poisson Generalized Estimating Equation (GEE) to the data to evaluate whether treatment has an effect on the rate of seizures. Consider relevant covariate adjustments. Be sure to provide a relevant interpretation of your model: describe the parameter estimate and 95% CI and the conclusion you draw from the hypothesis test. Also, discuss the implications of working correlation specification on the validity of your inference. Hint 1: include an log observation time as an offset in your model to ensure the rate is being modeled. See Lecture 3 Slide 45 for details. Hint 2: the poisson GEE should return coefficients on the log incidence rate ratio scale. You may want to report the exponentiated coefficients, which correspond to the fold changes in incidence rates associated with predictors. Hint 3: Also you may need to use the deltamethod function from the msm package if your model uses an interaction.

```
##
##  Descriptive statistics by group
## group: Placebo
##     vars  n  mean    sd median min max range   se
## age    1 28 29.00  6.00   29.0  19  42    23 1.13
## y0     2 28 30.79 26.10   19.0   6 111   105 4.93
## y1     3 28  9.36 10.14    5.0   0  40    40 1.92
## y2     4 28  8.29  8.16    4.5   0  29    29 1.54
## y3     5 28  8.79 14.67    5.0   0  76    76 2.77
## y4     6 28  7.96  7.63    5.0   0  29    29 1.44
## ------------------------------------------------------------
## group: Progabide
##     vars  n  mean    sd median min max range   se
## age    1 31 27.74  6.60     26  18  41    23 1.19
## y0     2 31 31.61 27.98     24   7 151   144 5.03
## y1     3 31  8.58 18.24      4   0 102   102 3.28
## y2     4 31  8.42 11.86      5   0  65    65 2.13
## y3     5 31  8.13 13.89      4   0  72    72 2.50
## y4     6 31  6.71 11.26      4   0  63    63 2.02
```

```
## seizure_wide$tx: Placebo
##           y0        y1        y2        y3        y4
## y0 1.0000000 0.7442140 0.8313228 0.4931314 0.8180073
## y1 0.7442140 1.0000000 0.7823271 0.5070280 0.6745946
## y2 0.8313228 0.7823271 1.0000000 0.6609345 0.7804471
## y3 0.4931314 0.5070280 0.6609345 1.0000000 0.6756745
## y4 0.8180073 0.6745946 0.7804471 0.6756745 1.0000000
## ------------------------------------------------------------
## seizure_wide$tx: Progabide
##           y0        y1        y2        y3        y4
## y0 1.0000000 0.8542248 0.8463589 0.8350421 0.8749799
## y1 0.8542248 1.0000000 0.9070300 0.9124734 0.9713376
## y2 0.8463589 0.9070300 1.0000000 0.9249302 0.9466338
## y3 0.8350421 0.9124734 0.9249302 1.0000000 0.9522894
## y4 0.8749799 0.9713376 0.9466338 0.9522894 1.0000000
```

```
##                       Estimate
## (Intercept)          3.8482143
## postTRUE             1.1171694
## txProgabide          1.0268692
## postTRUE:txProgabide 0.9015131
```

```
## [1] "Exponentiated confidence interval: 0.822 , 1.282"
```

Because the 95% CI for the coefficient of `txProgabide` contains 1, we do not conclude that Progabide reduces seizure rate.

Code Appendix

```
# setup
knitr::opts_chunk$set(echo = F, message = F, warning = F, cache = F)
options(knitr.kable.NA = '-')
labs = knitr::all_labels()
labs = labs[!labs %in% c("setup", "llm_appendix", "allcode")]

# clear workspace
rm(list = ls())

# load relevant libraries
library(tidyverse)
library(knitr)
library(lme4)
library(geepack)
library(psych)

# color selection
colors <- c("#FC600A", # dark orange
            "#C21460", # dark pink
            "#3F0000") # darker red
#########################
### Question 1
#########################

# read in Framingham data
FRM_wide <- read.table("data/framingham.dat",
                      col.names = c("age0", "sex", "bmi0", "bmi10", "cigarette",
                                "chol_0", "chol_2", "chol_4", "chol_6",
                                "chol_8", "chol_10", "death"))
FRM_wide <- FRM_wide %>%
  # create new ID column and position it first
  dplyr::mutate(id = 1:nrow(FRM_wide)) %>%
  dplyr::select(id, names(FRM_wide)) %>%
  # rename levels of sex
  mutate(sex = ifelse(sex == 1, "Male", "Female"))


# reformat a long format
FRM_long <- FRM_wide %>%
  tidyr::pivot_longer(cols = dplyr::starts_with("chol_"),
                      names_to = "year",
                      names_prefix = "chol_",
                      values_to = "chl") %>%
  # correct `year` to numeric
  dplyr::mutate_at("year", as.integer)

# convert instances of -9 to NA and remove NAs
FRM_long <- replace(FRM_long, FRM_long == -9, NA) %>% tidyr::drop_na()

# produce a spaghetti plot of cholesterol over time for first 100 subjects
FRM_long %>%
  ggplot(aes(x = year, y = chl)) +
    geom_line(aes(group = id), alpha = 0.1, color = colors[1]) +
    geom_point(data = FRM_long %>%
                 group_by(year) %>%
                 summarize(mean = mean(chl)),
               aes(x = year, y = mean), size = 2.5, color = colors[3]) +
  xlab("Study year") + ylab("Serum cholesterol (mg/dL)") +
  scale_x_continuous(breaks = seq(0,10,2)) +
  theme_bw() +
  facet_wrap(vars(sex))
# Fit the linear mixed-effects model
# Random intercept and random slope for each subject
lmm_model <- lme4::lmer(chl ~ year * sex + age0 + bmi0 + (year | id),
                        data = FRM_long)

# Summary of the model to examine the fixed effects
lmm_model %>% summary %>% coef %>%
  knitr::kable(digits = 3,
               caption = "Coefficient estimates from linear mixed-effect model")

# Plot the fitted model to visualize the interaction effect
# FRM_long %>%
#   ggplot(aes(x = year, y = chl, color = sex)) +
#     geom_jitter(width = 0.5, alpha = 0.4) +
#     geom_line(aes(y = predict(lmm_model))) +
#     theme_bw() +
#     scale_x_continuous(breaks = seq(0,10,2)) +
#     xlab("Study year") + ylab("Serum cholesterol (mg/dL)")

# Fit GEE model with independent working correlation structure
gee_independent <- geepack::geeglm(chl ~ year * sex + age0 + bmi0,
                                   id = id, data = FRM_long,
                                   corstr = "independence")

# Fit GEE model with exchangeable working correlation structure
gee_exchangeable <- geeglm(chl ~ year * sex + age0 + bmi0,
                           id = id, data = FRM_long,
                           corstr = "exchangeable")

# Fit GEE model with AR(1) working correlation structure
gee_ar1 <- geeglm(chl ~ year * sex + age0 + bmi0,
                  id = id, data = FRM_long,
                  corstr = "ar1")
# Summary of the models
cbind(
  lmm_model %>% summary %>% coef %>% data.frame %>%
    select(LMM = Estimate, SE = Std..Error),
  gee_independent %>% summary %>% coef %>%
    select(GEE.indp = Estimate, SE = Std.err),
  gee_exchangeable %>% summary %>% coef %>%
    select(GEE.exch = Estimate, SE = Std.err)) %>%
  # pretty print
  kable(digits = 3,
```

```r
        caption = "Coefficient estimates and corresponding standard errors")
#########################
### Question 2
#########################

# read in data
sixcity_long <- read.csv("data/sixcity-1.csv", col.names = c("X", "id",
                        "disease", "age", "smoke", "ageXsmoke")) %>% select(-X)
# head(sixcity_long)
# Fit GLMM
glmm_model <- lme4::glmer(disease ~ age + smoke + ageXsmoke + (1 | id),
                          data = sixcity_long,
                          family = binomial)

# Fit GEE model
gee_model <- geeglm(disease ~ age + smoke + ageXsmoke, id = id,
                    data = sixcity_long,
                    family = binomial, corstr = "exchangeable")

# Fit transition Model
transition_model <- sixcity_long %>%
  # arrange complete cases of id and age
  tidyr::complete(id, age) %>%
  dplyr::arrange(id, age) %>%
  # create a lag variable for disease and remove NA values
  dplyr::group_by(id) %>%
  mutate(disease.lag = dplyr::lag(disease, default=NA)) %>%
  na.omit %>%
  # fit transition model from complete data with lag variable
  geeglm(disease ~ age + smoke + ageXsmoke + disease.lag, id = id, data = .,
         family = binomial, corstr = "independence")

# Summary of the models
cbind(
  # add NA rows for GLMM and GEE columns (they do not have lag variable)
  glmm_model %>% summary %>% coef %>% data.frame %>%
    select(GLMM = Estimate, SE = Std..Error) %>% rbind(disease.lag = rep(NA, 2)),
  gee_model %>% summary %>% coef %>%
    select(GEE = Estimate, SE = Std.err) %>% rbind(rep(NA, 2)),
  transition_model %>% summary %>% coef %>%
    select(Transition = Estimate, SE = Std.err)) %>%
  # pretty print
  kable(digits = 3, caption = "Estimates and corresponding standard errors")
#########################
### Question 3
#########################

# read in data
seizure_wide <- read.csv("data/Seizure.csv")[,-1]
seizure_wide$rate_change <- seizure_wide$y4/2 - seizure_wide$y0/8
# seizure_wide %>% head

# convert to long format
```

```
seizure_long <- seizure_wide %>%
  pivot_longer(cols = starts_with("y"),
                      names_to = "time",
                      names_prefix = "y",
                      values_to = "seizures") %>% arrange(id, time)
seizure_long$tx <- ifelse(seizure_long$tx == "placebo", "Placebo", "Progabide")
seizure_wide$tx <- ifelse(seizure_wide$tx == "placebo", "Placebo", "Progabide")

# produce a spaghetti plot of seizure rates over time
seizure_long %>%
  ggplot(aes(x = time, y = seizures)) +
    geom_line(aes(group = id), alpha = 0.6, color = colors[1]) +
    geom_point(data = seizure_long %>%
                 group_by(time) %>%
                 summarize(mean = mean(seizures)),
               aes(x = time, y = mean), size = 2.5, color = colors[3]) +
  xlab("Time") + ylab("Number of seizures") +
  # scale_x_continuous(breaks = seq(0,10,2)) +
  theme_bw() +
  facet_wrap(vars(tx))

# violin and box plots of number of seizures by treatment
ggplot(seizure_long, aes(x = tx, y = seizures)) +
  geom_violin(color = colors[1]) +
  geom_jitter(width = 0.1) +
  geom_boxplot(alpha = 0.5, color = colors[3]) +
  facet_wrap(~time) +
  theme_bw() +
  labs(title = "Distribution of number of seizures by time",
       x = "Treatment", y = "Number of seizures")

# violin and box plots of rate of change by treatment
ggplot(seizure_wide, aes(x = tx, y = rate_change)) +
  geom_violin(color = colors[1]) +
  geom_jitter(width = 0.1) +
  geom_boxplot(alpha = 0.5, color = colors[3]) +
  theme_bw() +
  labs(title = "Distribution of rate of change by treatment groups",
       x = "Treatment", y = "Rate of change")
describeBy(seizure_wide[c(2,4:8)], seizure_wide$tx, skew=F)

# Correlation
by(seizure_wide[,-c(1,2,3,9)], INDICES = seizure_wide$tx, FUN=cor)


seizure_long$obstime <- ifelse(seizure_long$time > 0, 2, 8)
seizure_long$post <- seizure_long$time > 0

gee.robust <- geeglm(seizures ~ post * tx + offset(log(obstime)),
                     data = seizure_long, id = id,
                     family = poisson(link="log"),
                     corstr = "exchangeable", std.err = "san.se")
```

```r
gee.robust %>% summary %>% coef %>% select(Estimate) %>% exp
paste("Exponentiated confidence interval:", round(exp(0.0265-0.222),3), ",", round(exp(0.0265+0.222),3)
```

End of document.