Biost 540: Homework 2

Department of Biostatistics @ University of Washington

Alejandro Hernandez

Due May 2, April 2024

Problem 1

In the National Cooperative Gallstone Study (NCGS), one of the major interests was to study the safety of the drug chenodiol for the treatment of cholesterol gallstones. In this study, patients were randomly assigned to high-dose (750 mg per day), low-dose (375 mg per day), or placebo. We focus on a subset of data on patients who had floating gallstones and who were assigned to the high-dose and placebo groups. In the NCGS it was suggested that chenodiol would dissolve gallstones but in doing so might increase levels of serum cholesterol. As a result, serum cholesterol (mg/dL) was measured at baseline and at 6, 12, 20 and 24 months of follow-up. Many cholesterol measurements are missing because of missed visits, laboratory specimens that were lost or inadequate, or patient follow-up that was terminated.

(a)

Produce a spaghetti plot of the evolution of cholesterol over time faceted by treatment group. Please label the x-axis using study time (weeks since baseline). Also label the facets according to treatment name.

(b)

Produce a plot or table which characterizes the completeness in cholesterol measurements over time by treatment group. If plotting the data, be sure to include axis labels and text briefly describing what the plot is showing. If a tabular summary, please provide text or a caption describing what data the table is summarizing.

Table 1: Proportion missing in cholesterol measures over time by treatment

group	week0	week6	week12	week20	week24
High dosage	0	0	0.07	0.17	0.23
Placebo	0	0	0.03	0.06	0.10

(c)

What are the possible sources of correlation/clustering in these data? Produce a variogram plot summarizing the dependence in the data as a function of study time.

Repeated measures of an individual's cholesterol are likely to be correlated. Within a given treatment group, the trend of serum cholesterol over time will likely be related. Between treatment groups, so long as the treatment has effect, cholesterol levels may be most similar at baseline.

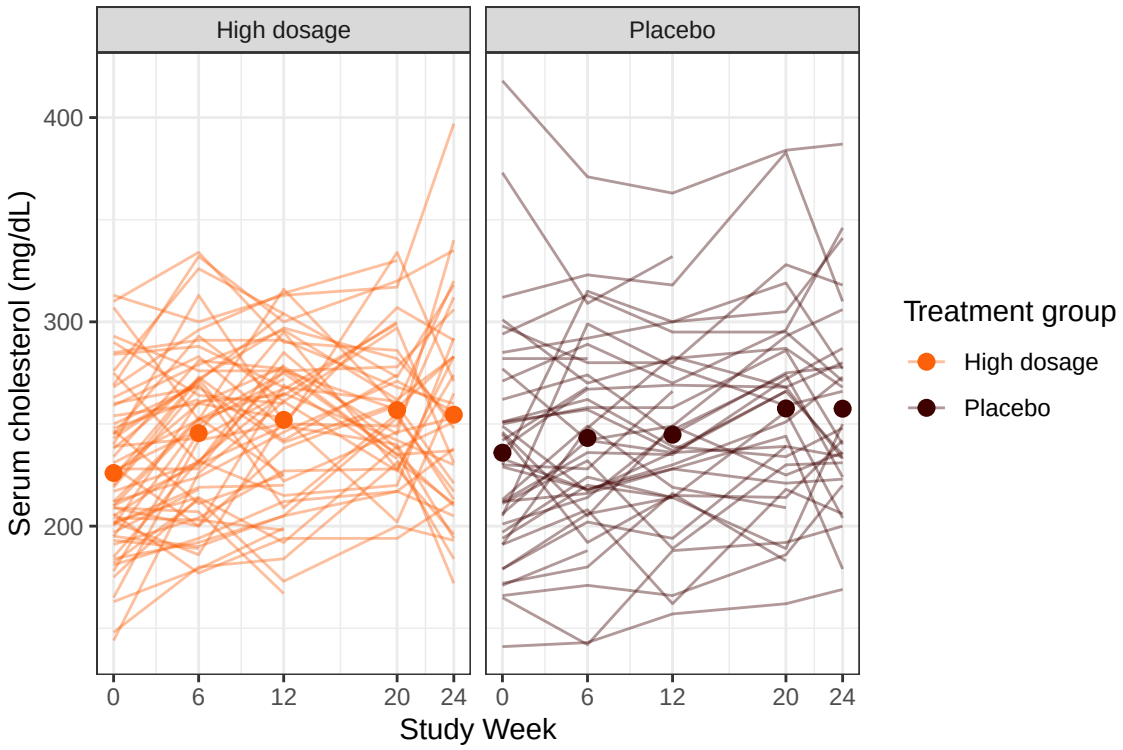


Figure 1: Cholesterol levels and averages over time by treatment

(d)

Incomplete

(e)

Incomplete

Problem 2

The Framingham study is one of the well known long term follow-up study to identify the relationship between various risk factors and diseases and to characterize the natural history of the chronic circulatory disease process. The data on various aspects have been and continue to be collected every two years on a cohort of individuals. It began in 1948 in Framingham, located 21 miles west of Boston, with limited goals of investigating the serum cholesterol, smoking and elevated blood pressure as the risk factors of coronary heart disease. Over the years its goal has been greatly expanded to aid in understanding the numerous etiological factors of various diseases. The data `framingham.dat` is a subset of a large data base collected in the Framingham study over years. There are 12 columns in the data file. The 1st column gives the age of the individual when they entered the study. The 2nd column provides the sex of the individual (1-male, 2-female); the 3rd and 4th columns provide body mass index (BMI) at the baseline and at 10 years from the baseline respectively; the 5th column provides the number of cigarettes per day the individual smoked at the baseline. The columns 6 – 11 provide serum cholesterol levels at the baseline(enrollment) and then every two years through year 10. The column 12 indicates whether the individual is alive(0) or dead(1) at the end of 30 years since enrollment. That is, the data set excludes those who died during the 10 year data

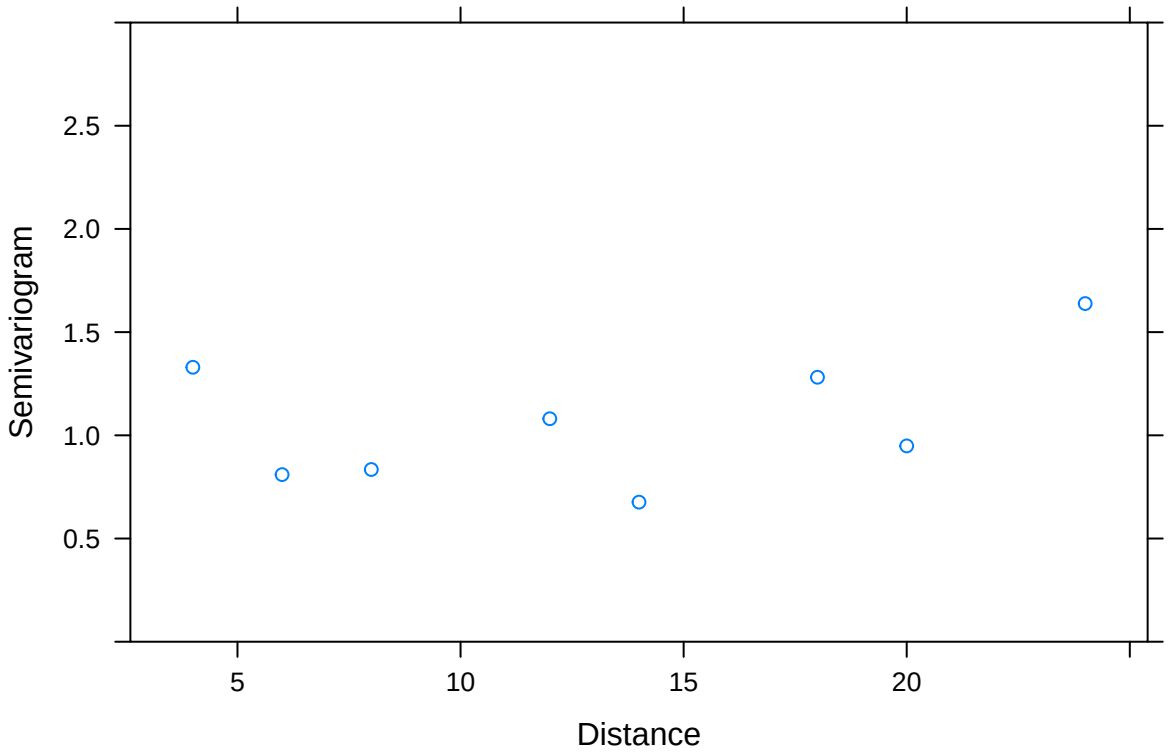


Figure 2: Variogram Plot

collection period. -9 indicates the missing data. Note that you have to convert -9 to NA or empty entries before you do any analysis.

(a)

Please read in the data and name each column by "age0", "gender", "bmi0", "bmi10", "cigarette", "chol_0", "chol_2", "chol_4", "chol_6", "chol_8","chol_10", "death". Use row number to create a new column called "id". Covert the data from wide table into long table. Please convert -9 to NA in your dataframe and remove all observations containing NA. Use head() function to print the first 6 rows to show your result.

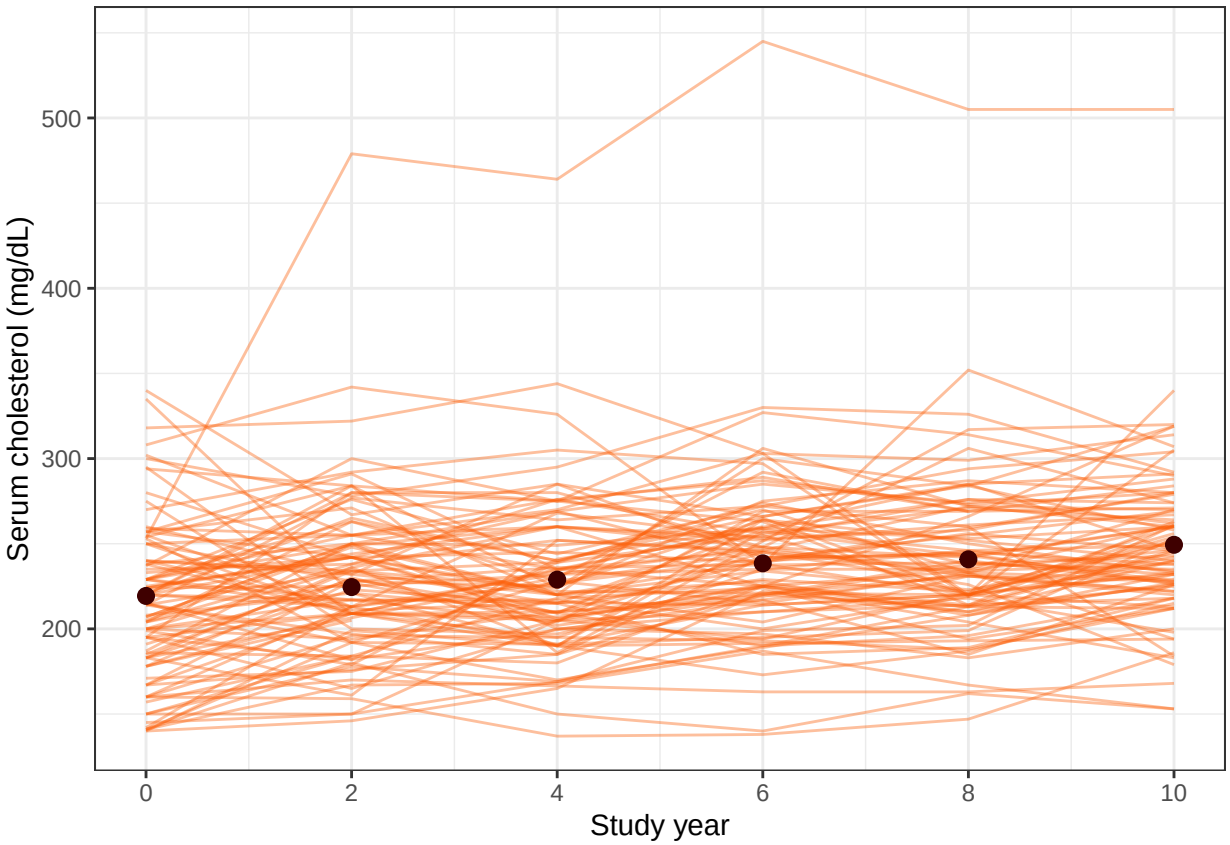
Table 2: Head of long-format Framingham data

id	age0	sex	bmi0	bmi10	cigarette	death	year	chl
1	45	2	22	22	0	0	0	220
1	45	2	22	22	0	0	2	217
1	45	2	22	22	0	0	4	217
1	45	2	22	22	0	0	6	200
1	45	2	22	22	0	0	8	219
1	45	2	22	22	0	0	10	240

(b)

Please make a Spaghetti plot of the cholesterol level over time for first 100 subjects. Please label the x-axis using study time.(eg: Baseline, Year2, Year4, etc) What conclusion could you draw from the plot?

The plot below suggests that each subject's trend of serum cholesterol level over time is similar, aside from a single extreme outlier.



(c)

Suppose we are interested in how cholesterol level changes over time and how it is related to baseline age, gender, and baseline BMI (without interaction between the variables). Please fit three models with different correlation structure: (i) with random intercept only, (ii) with random intercept and slope (with correlated random effects) and (iii) with random intercepts, exponential correlation with measurement error model. For each model, please provide output estimate, se, p value for coefficients of age0, gender, bmi0 and time. Please keep the result to three decimal places.

Table 3: Random intercept model

	Value	Std.Error	DF	t-value	p-value
(Intercept)	139.764	5.731	11023	24.389	0.000
year	2.946	0.057	11023	51.532	0.000
age0	1.235	0.090	2622	13.661	0.000
sex	2.468	1.475	2622	1.674	0.094
bmi0	0.881	0.175	2622	5.038	0.000

Table 4: Random intercepts and slopes model, correlated

	Value	Std.Error	DF	t-value	p-value
(Intercept)	139.764	5.731	11023	24.389	0.000
year	2.946	0.057	11023	51.532	0.000
age0	1.235	0.090	2622	13.661	0.000
sex	2.468	1.475	2622	1.674	0.094

	Value	Std.Error	DF	t-value	p-value
bmi0	0.881	0.175	2622	5.038	0.000

Table 5: Random intercepts and slopes model, exponentially correlated

	Value	Std.Error	DF	t-value	p-value
(Intercept)	139.474	5.732	11023	24.331	0.000
year	2.944	0.068	11023	43.025	0.000
age0	1.235	0.090	2622	13.660	0.000
sex	2.710	1.474	2622	1.838	0.066
bmi0	0.883	0.175	2622	5.047	0.000

(d)

Incomplete

Code Appendix

```
# setup
knitr::opts_chunk$set(echo = F, message = F, warning = 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(ggplot2)  # plotting
library(dplyr)    # data frame manipulation
library(tidyr)    # data frame manipulation
# library(corrplot) # correlation plotting
# library(rigr)     # regression
# library(modelr)   # regression
library(knitr)    # pretty printing data frames
library(nlme)     # (non)linear mixed effects modeling

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

# read in data (wide format)
NCGS_wide <- read.csv("data/cholesterol.csv")

NCGS_wide <- NCGS_wide %>%
  dplyr::mutate(group = ifelse(group == 1, "High dosage", "Placebo"))

# View(NCGS_wide)
### -------------------------------------------------------------
### (1a)

# reformat data to long format
NCGS_long <- NCGS_wide %>%
  tidyr::pivot_longer(cols = starts_with("y"),
                      names_to = "index",
                      names_prefix = "y",
                      values_to = "chl") %>%
  # define variable that matches measure index to time since baseline
  dplyr::mutate(week = c(0,6,12,20,24)[match(index, seq(1,5))])

# produce a spaghetti plot of cholesterol over time by treatment group
NCGS_long %>%
  ggplot(aes(x = week, y = chl, color = group)) +
  geom_line(aes(group = id), alpha = 0.4) +
  geom_point(data = NCGS_long %>%
               group_by(group, week) %>%
               summarize(mean = mean(chl, na.rm = T)),
```

```
            aes(x = week, y = mean), size = 2.5) +
  xlab("Study Week") + ylab("Serum cholesterol (mg/dL)") +
  labs(color = "Treatment group") +
  scale_x_continuous(breaks = c(0,6,12,20,24)) +
  scale_color_manual(values = colors[c(1,3)]) +
  theme_bw() +
  facet_wrap(vars(group))
### ------------------------------------------------------------
### (1b)

# construct a table of percent missing for each variable
NCGS_long %>%
  group_by(group, week) %>%
  summarize(na.prop = sum(is.na(chl)) / nrow(NCGS_wide)) %>%
  pivot_wider(names_from = week,
              names_prefix = "week",
              values_from = na.prop) %>%
  kable(digits = 2, caption = "Proportion missing in cholesterol measures over
        time by treatment")
### ------------------------------------------------------------
### (1c)

NCGS_long <- NCGS_long %>% tidyr::drop_na()

mod1 <- nlme::lme(chl ~ week, method = "ML", data = NCGS_long,
                  random = reStruct( ~ 1 | id, pdClass="pdDiag", REML=F))

plot(nlme::Variogram(mod1, form = ~ week | id, resType="normalized"),
     ylim=c(0, 3), smooth = F, trendline = T)
### ------------------------------------------------------------
### (1d)
### ------------------------------------------------------------
### (1e)
### ------------------------------------------------------------
### Question 2

# load in 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"))

### ------------------------------------------------------------
### (2a)

# create new ID column and position it first
FRM_wide <- FRM_wide %>%
  mutate(id = 1:nrow(FRM_wide)) %>%
  select(id, names(FRM_wide))


# reformat data to long format
FRM_long <- FRM_wide %>%
```

```
  tidyr::pivot_longer(cols = starts_with("chol_"),
                      names_to = "year",
                      names_prefix = "chol_",
                      values_to = "chl") %>%
  # correct `year` to numeric
  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()

head(FRM_long) %>% kable(caption = "Head of long-format Framingham data")
### ------------------------------------------------------------
### (2b)

# produce a spaghetti plot of cholesterol over time for first 100 subjects
FRM_long %>%
  filter(id <= 100) %>%
  ggplot(aes(x = year, y = chl)) +
    geom_line(aes(group = id), alpha= 0.4, 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()
### ------------------------------------------------------------
### (2c)

## random intercepts only
mod1 <- lme(chl ~ year + age0 + sex + bmi0, data = FRM_long, method = "ML",
            random = reStruct( ~ 1 | id, pdClass="pdDiag", REML=F))
mod1 %>% summary %>% coef %>%
  kable(digits = 3, caption = "Random intercept model")

## random intercepts + slopes, correlated
mod2 <- lme(chl ~ year + age0 + sex + bmi0, data = FRM_long, method = "ML",
            random = reStruct( ~ 1 | id, pdClass="pdSymm", REML=F))
mod2 %>% summary %>% coef %>%
  kable(digits = 3, caption = "Random intercepts and slopes model, correlated")

## random intercepts, exponentially correlated
mod3 <- lme(chl ~ year + age0 + sex + bmi0, data = FRM_long, method = "ML",
            random = reStruct(~ 1 | id, pdClass="pdSymm", REML=F),
            correlation = corExp(form =~ year | id, nugget = T))
mod3 %>% summary %>% coef %>%
  kable(digits = 3, caption = "Random intercepts and slopes model, exponentially
        correlated")
### ------------------------------------------------------------
### (2d)
```

End of document.