BIOST 540 Final

Ben Stan | June 8, 2021

Introduction

The Americans' Changing Lives (ACL) survey monitored the health of Black and White Americans from 1986 to 2011. The survey captured the activities and social relationship details for each individual in addition to their health behavior and engagement with health resources. The outcome of interest in this analysis is functional impairment (FI) and the primary covariate is race (defined as White American (W) or African American (AA)). We utilized a subset of the data collected over five measurement times for those who were 50 years or younger in 1986. The first scientific question was to examine the differences in odds of FI over time by race when adjusting for baseline age, sex, socioeconomic status, and past FI status. The second scientific question was similar, but it did not condition on previous FI status and utilized multiple approaches for missing data.

Methods

Descriptive Analysis: To accurately investigate the relationship between FI and race, we compared the sex, age, and socioeconomic status distributions of each group, shown in **Table 1**. The table also shows the trends in FI over the course of the study. As missing data was prevalent, the patterns of missingness were further explored. The rate of FI for each time and race was computed to identify differences for those that have missing measurements compared to those that do not. Outcome values preceding populated and missing records were compared in an attempt to categorize the pattern of missing data.
Confirmatory Analysis: To answer the first scientific question, the past FI status covariate was computed by taking the status of each subject at the previous time. The model (**Model 1**) is as follows: $\text{logit}(\mu_{ij}) = \beta_0 + \beta_1 {*} I_{AA} + \beta_2 {*} \text{year} + \beta_3 {*} \text{age} + \beta_4 {*} I_{Male} + \beta_5 {*} I_{SES\ Middle} + \beta_6 {*} I_{SES\ Upper} + \beta_7 {*} Y_{i,\ j\text{-}1} + \beta_8 {*} I_{AA} {*} \text{year}$, with μ_{ij} being the probability of having FI and year being years since baseline. The null hypothesis for investigating the change in odds of FI over time by race is H_0: $\beta_8 = 0$. We were also interested in whether past FI status is an important predictor of current status, which has a null hypothesis of H_0: $\beta_7 = 0$. To perform this analysis, a logistic regression transition model was used with generalized estimating equations (GEEGLM) on all available data. The GEE method was chosen because it

gives robust standard errors that have valid coverage in large samples even when Markov assumption is incorrect, i.e. even when observations are based on more than just the most recent observation. The point estimates are similar to a generalized linear model with independence. We chose a working independence covariance structure with the GEE because it is robust to the specified covariance structure at large samples. Computation of the regression coefficients also involves a conditional likelihood and provides valid results under missing completely at random (MCAR) and missing at random (MAR) scenarios, given the conditional mean and variance are properly specified. The model for the second scientific question (**Model 2**) is as follows: $\text{logit}(\mu_{ij}) = \beta_0 + \beta_1{*}I_{AA} + \beta_2{*}\text{year} + \beta_3{*}\text{age} + \beta_4{*}I_{Male} + \beta_5{*}I_{SES\ Middle} + \beta_6{*}I_{SES\ Upper} + \beta_7{*}I_{AA}{*}\text{year}$. The first hypothesis test considered if there is a difference in FI at baseline between races; H_0: $\beta_1 = 0$. The second test evaluated H_0: $\beta_7 = 0$ to determine if there is a difference between races in change over time. All tests in both models utilized a Wald test. We considered two methods for handling the missing data; the primary analysis utilized an inverse probability weighting (IPW) GEE, while the sensitivity analysis used an available data GEE. The IPW approach instituted a dropout pattern in the data and compensated for under-representation. The sensitivity approach was chosen due to the robustness of GEE to missing data and was used to validate results.

Results

Descriptive Analysis: Based on **Table 1**, there is no difference in age distribution between the two races and a small difference in the percent of female subjects (65% for AA and 54% for White). In socioeconomic status (SES), however, 62% of White subjects are in the Upper category compared to 36% of AA subjects. Despite having a comparable percentage with FI at baseline, the proportion with FI for AA grows more over the study, reaching 52.9% in 2011, compared to 31.8% for the White population. The pattern of missingness is non monotonic, as shown in **Figure 1**. The two most common patterns after complete data (59.9%) were only missing the fourth measurement (7.4%) and missing all times after baseline (4.6%), a form of dropout. Thus, for the IPW GEE, dropout was implemented and only required discarding one observation time for one of these patterns. The AA population exhibited more missingness than the White population over all times following baseline, but there is no consistent trend over time. This lack of

pattern seems to suggest that the data is MCAR. **Figure 2** shows that the portion of subjects with FI by race over time does not consistently differ between subjects with missing data and those with complete data, further suggesting a lack of pattern. However, the outcome preceding a missing value shows FI 9.2% of the time compared to 12.8% of the time for populated values; this difference in preceding values suggests MAR.
Confirmatory Analysis: By examining the fitted results of **Model 1** in **Table 2**, the estimated odds of FI for both races is 8% higher per year (95% CI (AA): 7%-10% higher; 95% CI (W): 7%-9% higher). We do not have strong evidence that the the trend in odds of FI differ across races when adjusting for baseline age, sex, socioeconomic status, and past FI status (p-value: 0.62). There is, however, strong evidence that past FI status is an important predictor of current status (p-value $< 1.0*10^{-10}$). The estimated odds ratio for a population with FI in the previous time compared to one without is 14.9 (95% CI: 11.5 - 19.3). The results of **Model 2** in **Table 3** show that the estimated odds ratio of FI for AA compared to White subjects at baseline is 1.21 (95% CI: 0.871 - 1.67). There is not strong evidence that the odds of FI differ by race at baseline (p-value = 0.26). For White subjects, the estimated odds of FI is 8% higher per year (95% CI: 7%-10% higher). For AA subjects, the estimated odds of FI is 11% higher per year (95% CI: 9%-12% higher). There is strong evidence of a difference over time across races (p-value = 0.020). All conclusions of the primary analysis were also attained in the sensitivity analysis.

Discussion

When examining the odds of FI, the results of the first model indicate that there is no difference over time between races, but there is a non-zero relationship between previous FI status and current. The second model found that there is no difference by race at baseline but that there is a difference over time between races. The structure of the two models can explain the differing conclusions. The first model uses a conditional response on the previous outcomes, which may explain the change in odds of FI that is captured by the interaction term between race and year in the second model. The imputation method of IPW appeared reasonable given the pattern of missingness but still required assumptions and discarding of data. The use of GEE also required either MCAR data or for the distributional assumptions to hold under an MAR scenario. Further research could examine the mechanisms for missingness to attain greater inference.

Tables and Figures

Race	N	SES	Age	Year	Percent Missing	Percent with functional impairment
African American	546 M: 192 F: 354	Low: 23% Middle: 41% Upper: 36%	Min: 25 Median: 35 Mean: 36.2 Max: 50	1986	0%	7.5%
				1989	23.4%	12.4%
				1994	26.7%	22.5%
				2002	37.9%	34.5%
				2011	18.7%	52.9%
White American	1040 M: 478 F: 562	Low: 12% Middle: 26% Upper: 62%	Min: 25 Median: 35 Mean: 35.8 Max: 50	1986	0%	6.2%
				1989	15.4%	7.5%
				1994	14.6%	10.1%
				2002	20.7%	16.8%
				2011	14.5%	31.8%

Table 1: Descriptive statistics for each race in the study.

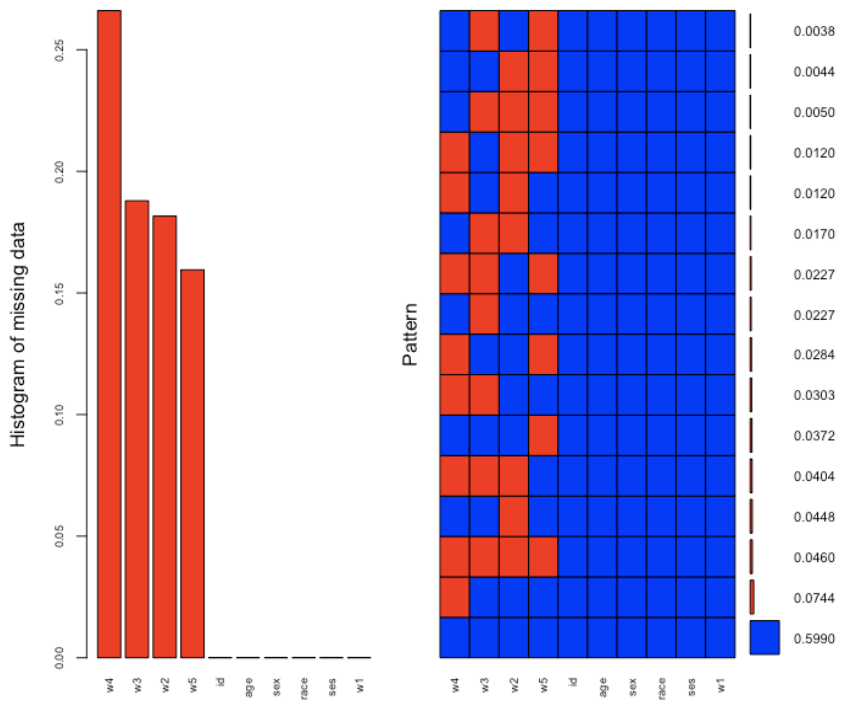

Figure 1: Patterns in missingness in the data: (left) histogram of percent missing observations; (right) illustration of patterns of missingness and their prevalence.

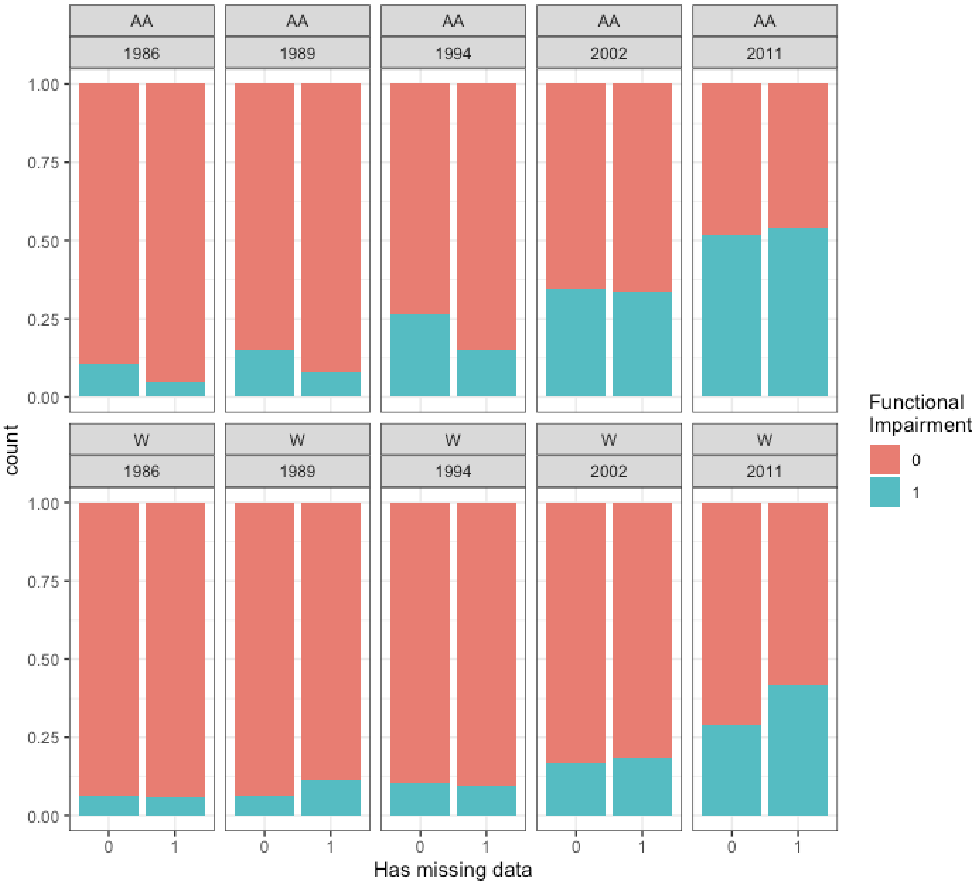

Figure 2: Proportion of subjects with functional impairment by race, year and whether subjects had any missing data.

	Estimate	95% CI Lower	95% CI Upper	P-value
Intercept	0.0189	0.0105	0.0340	$<1.0*10^{-10}$
I_{AA}	1.50	1.03	2.19	0.035
age	1.05	1.03	1.06	$<1.0*10^{-10}$
I_{Male}	0.844	0.694	1.03	0.090
$I_{SES\ Middle}$	0.543	0.416	0.709	$6.8*10^{-6}$
$I_{SES\ Upper}$	0.313	0.241	0.407	$<1.0*10^{-10}$
$Y_{i,\ j-1}$	14.9	11.5	19.3	$<1.0*10^{-10}$
year (W)	1.08	1.07	1.09	$<1.0*10^{-10}$
year (AA)	1.08	1.07	1.10	0.62[a]

Table 2: Exponentiated coefficients, confidence intervals, and p-values for Model 1.

a. p-value corresponds to interaction term for the indicator for being African American and year

	Estimate - Available Data (95% CI)	P-value - Available Data	Estimate - IPW (95% CI)	P-value - IPW
Intercept	0.0205 (0.0133-0.0316)	$<1.0*10^{-10}$	0.0241 (0.0122-0.0475)	$<1.0*10^{-10}$
I_{AA}	1.17 (0.902-1.51)	0.24	1.21 (0.871-1.67)	0.26
age	1.06 (1.05-1.07)	$<1.0*10^{-10}$	1.06 (1.04-1.08)	$<1.0*10^{-10}$
I_{Male}	0.704 (0.608-0.816)	$3.1*10^{-6}$	0.711 (0.555-0.911)	$7.1*10^{-3}$
$I_{SES\ Middle}$	0.515 (0.425-0.624)	$<1.0*10^{-10}$	0.430 (0.312-0.591)	$2.3*10^{-7}$
$I_{SES\ Upper}$	0.234 (0.193-0.283)	$<1.0*10^{-10}$	0.211 (0.154-0.289)	$<1.0*10^{-10}$
year (W)	1.09 (1.08-1.10)	$<1.0*10^{-10}$	1.08 (1.07-1.10)	$<1.0*10^{-10}$
year (AA)	1.11 (1.10-1.13)	0.0071[a]	1.11 (1.09-1.12)	0.020[a]

Table 3: Exponentiated coefficients, confidence intervals, and p-values for Model 2 under two different missing data approaches.

[a] p-value corresponds to interaction term for the indicator for being African American and year

Code

```
rm(list = ls())
library(ggplot2)
library(GGally)
library(reshape2)
library(geepack)
library(lme4)
library(nlme)
library(tidyverse)
library(dplyr)
library(lattice)
library(VIM)
library(mice)
library(broom.mixed)
library(wgeesel)
library(multcomp)

########################
```

```
# Scientific Question 1 #
#########################
## Load data
setwd("/Users/bstan/Documents/UW/Courses/BIOST 540")
dat <- read.csv('datasets/acl_subset.csv')
dat <- dat[,-1]

dat_long <- melt(dat, id=c("id", "sex", "race", "ses", "age"))
dat_long$year <- 1986
dat_long$year[dat_long$variable=="w2"] <- 1989
dat_long$year[dat_long$variable=="w3"] <- 1994
dat_long$year[dat_long$variable=="w4"] <- 2002
dat_long$year[dat_long$variable=="w5"] <- 2011

## Create summary table by race with wide table
summary1 <- dat %>%
  group_by(race) %>%
  summarise(n = n(),
            n_males = sum(sex=="Male", na.rm = T),
            n_females = sum(sex=="Female", na.rm = T),
            min_age = min(age),
            p25_age = quantile(age, 0.25),
            mean_age = mean(age),
            median_age = quantile(age, 0.5),
            p75_age = quantile(age, 0.75),
            max_age = max(age),
            pct_low_ses = sum(ses=="Low", na.rm = T)/n(),
            pct_middle_ses = sum(ses=="Middle", na.rm = T)/n(),
            pct_upper_ses = sum(ses=="Upper", na.rm = T)/n())
summary1

## Create summary table by race and year with long table
summary2 <- dat_long %>%
  group_by(race, year) %>%
  summarise(n = n(),
          pct_missing = sum(is.na(value))/n(),
          pct_func_impairment = sum(value==1, na.rm = T)/sum(!is.na(value))
          )
summary2

## Plot proportion with functional impairment over time by race
ggplot(data=summary2) +
  geom_line(aes(x=year, y=pct_functional_impairment, color=race)) +
  scale_x_continuous(breaks=c(1986,1989,1994,2002,2011)) +
  labs(x='Year', y='Proportion with Functional Impairment', color='Race') +
  theme_bw()

## Plot patterns in missingness
aggr_plot <- aggr(dat,
                col=c('blue','red'),
                numbers=TRUE,
                sortVars=TRUE,
                labels=names(dat),
                cex.axis=.7,
                gap=3,
                ylab=c("Histogram of missing data","Pattern"))

## Plot percent missing by race and year
par(mfrow=c(3,2))
spineMiss(dat[, c("race", "w1")])
spineMiss(dat[, c("race", "w2")])
spineMiss(dat[, c("race", "w3")])
spineMiss(dat[, c("race", "w4")])
spineMiss(dat[, c("race", "w5")])

## Create bar graphs to compare proportions with/without missingness
ids.miss <- unique(dat_long$id[is.na(dat_long$value)])
dat_long$missing <- ifelse(dat_long$id %in% ids.miss, 1, 0)
```

```
smok.labs <- c('Mother smoking: No', 'Mother smoking: Yes')
ggplot(data=dat_long[!is.na(dat_long$value),],
       aes(x=as.factor(missing),
           fill=as.factor(value))) +
  geom_bar(position = 'fill') +
  facet_wrap(~race+year, nrow=2) +
  labs(fill='Functional \nImpairment', x='Has missing data') +
  theme_bw()

## Analysis of proportions of previous measurement
dat_long <- dat_long %>% arrange(id,year) %>% group_by(id) %>%
  mutate(lag1 = dplyr::lag(value, default=NA))
lag_subset <- dat_long[dat_long$year != 1986,]
lag_subset$val_missing <- is.na(lag_subset$value)

lag_subset %>% group_by(val_missing) %>%
  summarise(value0 = sum(lag1=='0', na.rm=T),
            value1 = sum(lag1=='1', na.rm=T),
            prop = value1/(value0+value1))

#########################
# Scientific Question 2 #
#########################
## Use all available, populated data to fit GEE transition model
### Fit model once with W as reference race
dat_long_subset <- dat_long[complete.cases(dat_long),]
dat_long_subset$year <- dat_long_subset$year-1986
transit_gee <- geeglm(value~I(race=="AA")*year+age+sex+ses+lag1,
                      data = dat_long_subset,
                      id = id,
                      family = binomial(link = "logit"),
                      corstr = "independence")
summary(transit_gee)
exp(transit_gee$coefficients)
exp(broom::confint_tidy(transit_gee,conf.int = TRUE))
### Fit model once with AA as reference race
transit_gee <- geeglm(value~I(race=="W")*year+age+sex+ses+lag1,
                data = dat_long_subset,
                id = id,
                family = binomial(link = "logit"),
                corstr = "independence")
summary(transit_gee)
exp(transit_gee$coefficients)
exp(broom::confint_tidy(transit_gee,conf.int = TRUE))

#########################
# Scientific Question 3 #
#########################
## Using available data to fit GEE
### Fit model once with W as reference race
dat_long <- melt(dat, id=c("id", "sex", "race", "ses", "age"))
dat_long$year <- 1986
dat_long$year[dat_long$variable=="w2"] <- 1989
dat_long$year[dat_long$variable=="w3"] <- 1994
dat_long$year[dat_long$variable=="w4"] <- 2002
dat_long$year[dat_long$variable=="w5"] <- 2011
dat_long_subset <- dat_long[complete.cases(dat_long),]
dat_long_subset$year <- dat_long_subset$year-1986
avail_gee <- geeglm(value~I(race=="AA")*year+age+sex+ses,
           data = dat_long_subset,
           id = id,
           family = binomial(link = "logit"),
           corstr = "independence") # by defaults uses robust standard errors
summary(avail_gee)
exp(avail_gee$coefficients)
exp(broom::confint_tidy(avail_gee,conf.int = TRUE))
### Fit model once with AA as reference race
avail_gee <- geeglm(value~I(race=="W")*year+age+sex+ses,
          data = dat_long_subset,
          id = id,
          family = binomial(link = "logit"),
          corstr = "independence") # by defaults uses robust standard errors
summary(avail_gee)
exp(avail_gee$coefficients)
exp(broom::confint_tidy(avail_gee,conf.int = TRUE))

## Make data exhibit dropout for IPW
dat[is.na(dat$w2),8:10] <- NA
dat[is.na(dat$w3),9:10] <- NA
dat[is.na(dat$w4),10] <- NA
aggr_plot <- aggr(dat,
         col=c('blue','red'),
         numbers=TRUE,
         sortVars=TRUE,
         labels=names(dat),
         cex.axis=.7,
         gap=3,
         ylab=c("Histogram of missing data","Pattern"))
dat_long_dropout <- melt(dat, id=c("id", "sex", "race", "ses", "age"))
dat_long_dropout$year <- 1986
dat_long_dropout$year[dat_long_dropout$variable=="w2"] <- 1989
dat_long_dropout$year[dat_long_dropout$variable=="w3"] <- 1994
dat_long_dropout$year[dat_long_dropout$variable=="w4"] <- 2002
dat_long_dropout$year[dat_long_dropout$variable=="w5"] <- 2011
dat_long_dropout$year <- dat_long_dropout$year-1986
dat_long_dropout <- dat_long_dropout %>% arrange(id,year)
dat_long_dropout$R <- 1
dat_long_dropout$R[is.na(dat_long_dropout$value)] <- 0
dat_long_dropout <- dat_long_dropout %>% arrange(id,year) %>% group_by(id) %>%
  mutate(lag1 = dplyr::lag(value, default=NA))
### Run IPW GEE
mod_ipw <- wgee(model = value ~ I(race=="AA")*year+age+sex+ses,
          data = dat_long_dropout, id = dat_long_dropout$id,
          family="binomial", corstr ="independence",
          scale = NULL, mismodel = R ~ race + year + sex + ses + age + lag1)
summary(mod_ipw)
exp(mod_ipw$beta)
### Examine Coefficients
lambda1 <- c(0, 0, 1, 0, 0, 0, 0, 0)
exp(lambda1 %*% mod_ipw$beta)
exp(lambda1 %*% mod_ipw$beta + qnorm(c(0.025, 0.975)) * c(sqrt(lambda1 %*% mod_ipw$var
%*% lambda1)))
lambda2 <- c(0, 0, 1, 0, 0, 0, 0, 1)
exp(lambda2 %*% mod_ipw$beta)
exp(lambda2 %*% mod_ipw$beta + qnorm(c(0.025, 0.975)) * c(sqrt(lambda2 %*% mod_ipw$var
%*% lambda2)))
```