University of Michigan

Investigating the Relationship Between Work Cessation in Aging Individuals and Social Demographic Factors:

Insights from the Longitudinal NHATS Survey

Carlos Cristiano Botia

Chia Wen Cheng

Ruiling Kang

SurvMeth 687: Applications of Statistical Modeling

Brady T. West

December 5, 2023

I. Motivation and Research Question

In the expansive domain of labor studies and public health, the cessation of work due to health issues among individuals, especially older adults, remains a pivotal area warranting investigation. Studies by Luth and Prigerson (2022) underscore the need to consider these nuanced factors when examining the association between work cessation in aging individuals and social demographic elements like gender, race, and age. Additionally, research by Sofia von Humboldt highlights negative job-related stereotypes associated with older workers, particularly regarding perceived work adaptability, effectiveness, and workplace age discrimination (von Humboldt et al., 2022). Despite these insights, the specific linkages between work cessation in aging individuals, social demographic factors, and their impact on stable incomes and economic independence remain inadequately explored.

Our paper will utilize longitudinal NHATS Survey data to explore the correlation between work cessation among aging individuals and various social demographic factors in a comprehensive study. National Health and Aging Trends Study is a longitudinal cohort study that captures a nationally representative sample of Medicare beneficiaries aged 65 and older in the United States. It serves as a vital research platform dedicated to reducing disability, fostering health and independence, and enhancing the well-being of older adults. (National Health and Aging Trends Study, 2023).

This study aimed to illuminate the connections between demographic factors and the cessation of work among the elderly due to health concerns, specifically examining trends across race, age, and gender within longitudinal survey data. In other

words, we will answer the research question: How does the probability of experiencing work cessation due to health differ among individuals with distinct demographic attributes, including race, gender, and age, over four years considered?

II. Data Set Introduction

Since 2011, the NHATS has been annually conducting face-to-face interviews with a sampled population of ages over (including) 65 and enrolled in Medicare, aiming to understand the impact of health conditions, caregiving, and physical environments on the quality of life in late adulthood. NHATS survey waves include questions about demographic information, detailed independent functions, medical resource utilization, caregiver and device assistance, and individuals' subjective perspectives (Ankuda et al., 2022). The study employed an initial stratified three-stage equal probability sample design, considering age and race/ethnicity groups relative to the targeted sample sizes. In Round 5, the research was refreshed using the same sample design, providing valuable data for both national-level and individual trajectory studies (Freedman, Schrack, & Skehan, 2023).

For this study, we selected a subset spanning four years, encompassing the fifth round in 2015 through the eighth round in 2018. During the data cleaning phase, we excluded missing values such as "don't know," "inapplicable," and "missing" from each variable incorporated. This step was taken to ensure that our statistical modeling yields results that are more interpretable and avoid complications in explanation. We also excluded individuals having less than 4 responses after filtering out invalid values in our variables, so that every round from waves five to eight had the same sample size of 4,831, with a total of 19324 observations.

A. Dependent Variable

From the extensive set of over 1,000 variables, we chose a binary variable to investigate our outcome of interest: whether the person's health condition prevented them from engaging in paid work in the past month. This binary variable, originally encompassing -9 for “missing values,” -8 for “don't know,” and -1 for “inapplicable” situations, was narrowed down to solely include values of 1 for “yes” and 0 for “no” in our analysis.

B. Independent Variables

We identified 3 independent variables to address our research questions, ensuring both quantitative significance with substantial sample sizes and qualitative relevance to our research inquiries.

Categorical demographic variables included:

i. The Age categories were defined as follows: 1) 65 to 69 years old, 2) 70 to 74 years old, 3) 75 to 79 years old, 4) 80 to 84 years old, 5) 85 to 89 years old, and 6) 90 years old and above.

ii. Gender was represented by two classes: 1) male and 2) female. To align with binary conventions, we converted female to 0 during subsequent analysis.

iii. The Race variable encompassed four categories: 1) White, non-Hispanic; 2) Black, non-Hispanic; 3) Other (American Indian/Asian/Native Hawaiian/Pacific Islander/other specify), non-Hispanic; and 4) Hispanic.

III. Modeling Approach

A. Descriptive Statistics

Prior to constructing our modeling approach of the dependent variable, we conducted a thorough exploration of each variable to identify strong correlations among covariates and ascertain significant differences in the relationships between our dependent variable and each of the three independent variables. Initially, we assessed race-based disparities in the percentages of respondents experiencing work cessation due to health issues.

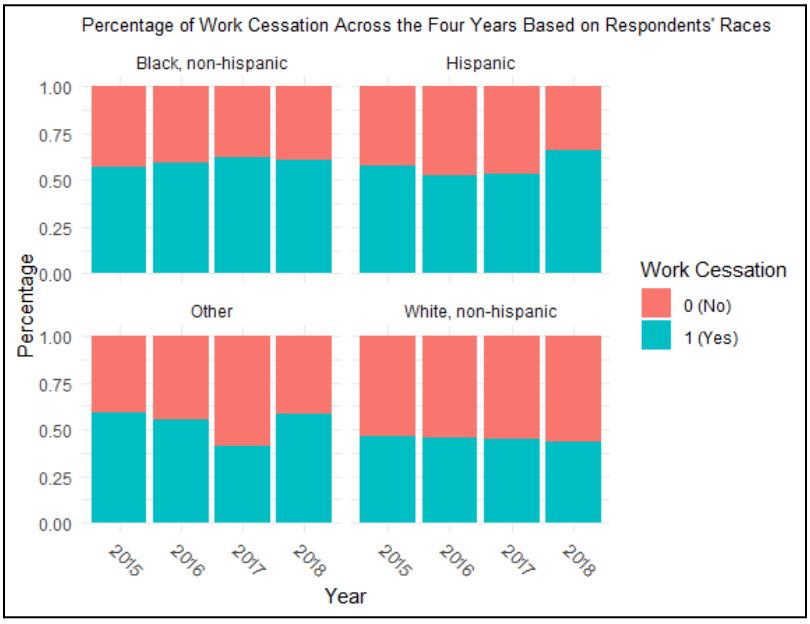

Figure 1. Percentage Bar Plot: Race-based Work Cessation Due to Health

Indicated by Figure 1, non-Hispanic Whites consistently exhibited the lowest percentages of experiencing work termination due to personal health over the years, maintaining a stable trend. In the initial years, a decline in work cessation experiences was observed among Hispanic and non-Hispanic respondents, categorized as "Other," while non-Hispanic Blacks demonstrated an increase in percentages over the years.

However, in the most recent year, 2018, the percentage of work termination was highest among Hispanic respondents, followed by non-Hispanic Blacks, non-Hispanic individuals of other races, and the lowest among non-Hispanic Whites.

We also looked at the differences in groups of genders and ages.

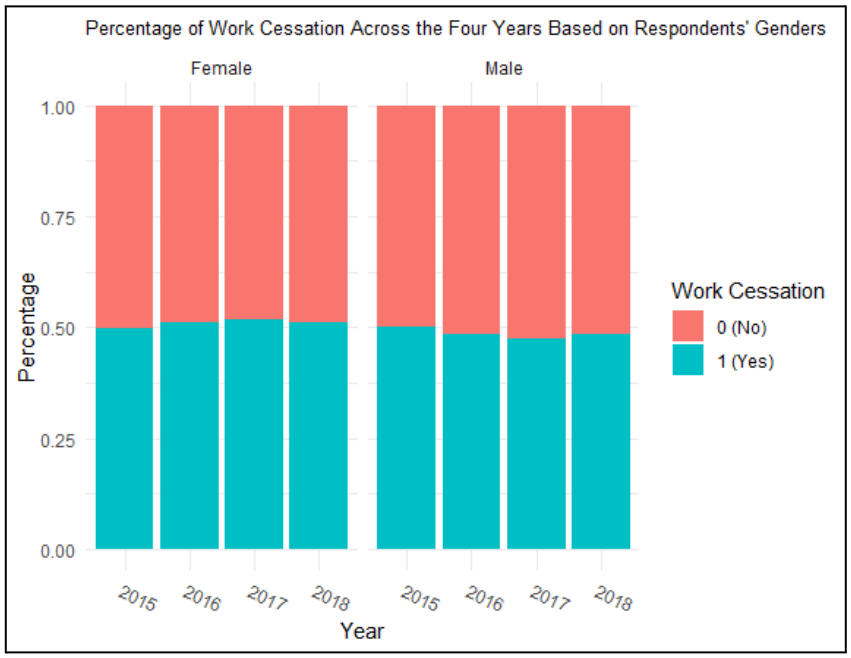

Figure 2. Percentage Bar Plot: Gender-based Work Cessation Due to Health

According to Figure 2, there were comparable percentages of work cessation due to health between men and women. Approximately half of both male and female respondents experienced work termination. Nevertheless, the percentage of women exhibited an upward trend over the years, experiencing a slight decrease from 2017 to 2018. In contrast, the percentage for men declined, with a slight increase noted from 2017 to 2018.

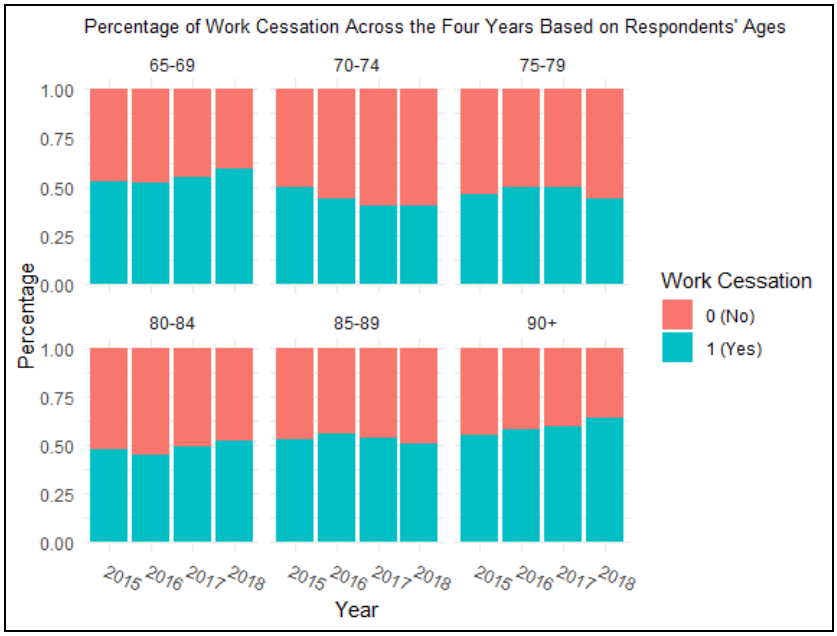

Figure 3. Percentage Bar Plot: Age-based Work Cessation Due to Health

Figure 3 presents indications of variations in work termination due to health conditions based on age groups. Among respondents aged between 65-69 and those aged 90 or above, the percentages increased over the years. Conversely, for individuals within the 70-74 age group, the percentages decreased. Those falling between 75-79 or 85-89 years experienced an initial increase, followed by a decrease in the most recent two years. An opposite trend was observed for respondents aged between 80-84 years.

B. Model Fitting

Given the non-normally distributed nature of the outcome variable, we employed the Generalized Estimating Equations (GEE) approach to construct our logistic regression model. This method was apt for analyzing binary outcome data with repeated measurements or clustered data. Accordingly, we specified a binomial distribution with a logit link function. The model incorporates a time variable, “year,” covering the period from 2015 to 2018, and respondent identifiers, “id,” were used as

observation identifiers for individuals interviewed. The marginal model facilitated an exploration of how these factors influenced the likelihood of work cessation due to health issues and whether these effects exhibited significant changes over time.

To determine an optimal model, we take a set of demographic variables to make our analysis. After employing Stepwise Backward Elimination and Variance Inflation Factor (VIF), we identified the best model among our trials, characterized by the minimum BIC value of 1053. This model excluded gender from our covariates and included two predictors: race, and age.

Covariate	GVIF	d.f.	GVIF^(1/(2*d.f.))
factor(race)	1	3	1
factor(age)	1	5	1

Table 1: GVIFs for the Remaining Predictor Variables in the Optimal Model

We recorded the Generalized Variance Inflation Factors (GVIFs) along with their respective independent variables in Table 1. GVIFs exactly at 1 indicated that the variables are independent, signifying that the variance of the estimated regression coefficient for each variable was not substantially inflated by multicollinearity.

Model	QIC
gee.model.independence	8416.8
gee.model.exchangeable	8418.2
gee.model.AR.1	11478

Table 2: Models and Corresponding QIC Values

In our pursuit of identifying the most fitting model, our criterion led us to select the model characterized by the lowest QIC value in Table 2. Among the models evaluated, the independence variance-covariance matrix emerged as the optimal choice, boasting

a QIC of 8416.8. Following the algorithm output and our decisions of predictors and the outcome variable of interest, the equation is specified mathematically as below:

$$logit(keep\ from\ work_i) = \beta_0 + \beta_1 Race_{Black,\ non-Hispanic;\ i} + \beta_2 Race_{Other,\ non-Hispanic;\ i} + \beta_3 Race_{Hispanic,\ i} + \beta_4 Age_{"70-74,"\ i} + \beta_5 Age_{"75-79,"\ i} + \beta_6 Age_{"80-84,"\ i} + \beta_7 Age_{"85-89,"\ i} + \beta_8 Age_{"90+,"\ i}$$

(where i denotes responses from the i-th individual)

$$\Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

IV. Results

Substituting our findings into the equation, the mathematical representation is as follows:

$$logit(keep\ from\ work_i) = -2.82 + 0.61 Race_{Black,\ non-Hispanic;\ i} + 0.30 Race_{other,\ non-Hispanic;\ i} + 0.62 Race_{Hispanic,\ i} - 0.35 Age_{"70-74,"\ i} - 0.26 Age_{"75-79,"\ i} - 0.19 Age_{"80-84,"\ i} - 0.14 Age_{"85-89,"\ i} + 0.13 Age_{"90+,"\ i}$$

$$\Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

From the results, a baseline log-odds of -2.82 signified the likelihood of a 65-69-year-old non-Hispanic White respondent experiencing paid work cessation in the past month due to their health. On one hand, keeping the age constant at 65-69 years old, the log-odds for the non-Hispanic Blacks was greater by 0.6 compared to

non-Hispanic White respondents. Similarly, for non-Hispanic individuals of other races in the same age bracket, the log-odds increased by 0.3, and for Hispanic respondents, the log-odds increased by 0.62. On the other hand, keeping the racial category constant as non-Hispanic White, the log-odds decreased by 0.35 for respondents aged 70-74. In the subsequent age bracket of 75-79, compared to individuals aged 65-69, the log-odds decreased by 0.26. Similarly, for the age group of 80-84, the log-odds decreased by 0.19, and for the age group of 85-89, the log-odds decreased by 0.14. In contrast, for the remaining age group, 90+, the log-odds increased by 0.13.

The exponentiated intercept value, 0.0597, denotes the expected odds ratio for a non-Hispanic White individual aged 65-69 facing work termination due to personal health. Similarly, individuals of the same age group had expected odds ratios of 1.8463 for non-Hispanic Blacks, 1.3533 for non-Hispanics of other races, and 1.8505 for Hispanics. All these values point to an increased risk of work termination due to health compared to non-Hispanic Whites.

When holding race constant, the odds ratios for being kept from work for non-Hispanic Whites aged 70 to 90 or above relative to the reference group at 65-69 years old were 0.7024, 0.7689, 0.827, 0.8719, and 1.1378, respectively. Notably, the eldest group at 90 or more exhibited odds ratios greater than 1, suggesting a heightened risk of work termination due to health situations compared to the baseline group at 65-69 years old.

V. Conclusions

This study revealed the likelihood of individuals being kept from paid work due to health-related issues in the past month differed by demographic characteristics of race

and age across the sample. The modeling process and the outputs from our logistic marginal model developed using the Generalized Estimating Equations (GEE) methodology suggested the following:

- To estimate the work termination in old ages in a longitudinal study, we were advised by the Stepwise Backward Elimination to consider only the variables race and age as predictors as they offered informative insights into the factors influencing the likelihood of work cessation.
- There were significant differences across the races for people older than 65 years old regarding work cessation due to health issues. Non-Hispanic Whites consistently exhibited the lowest probability of being affected compared to non-Hispanic Blacks, non-Hispanics of other races, and Hispanics.
- When comparing various age groups within the non-Hispanic White population, 90+-year-old respondents with paid jobs faced an increased risk of being affected. However, statistically significant differences in the likelihood of work cessation, in comparison to the 65 to 69-year-old category, were observed only among individuals aged 70 to 74 years.

A further conclusion we would like to draw from our descriptive analysis regarding gender trends in work cessation is:

- While not explicitly detailed, the slight variations over the years raised questions about gender-specific factors influencing work cessation and warrant further investigation.

References

Ankuda, C. K., Covinsky, K., Freedman, V. A., Langa, K., Aldridge, M. D., Yee, C., & Kelley, A. S. (2022). The devil's in the details: Variation in estimates of late-life activity limitations across National Cohort Studies. Journal of the American Geriatrics Society, 71(3), 858–868. https://doi.org/10.1111/jgs.18158.

von Humboldt, S., Miguel, I., Valentim, J., Costa, A., Low, G., & Leal, I. (2023). Psychosocial differences in perceived older workers' work (un)adaptability, effectiveness and workplace age discrimination. *European Psychiatry, 66*(Suppl 1), S237. https://doi.org/10.1192/j.eurpsy.2023.547.

Luth, E. A., & Prigerson, H. G. (2022). Socioeconomic Status, Race/Ethnicity, and Unexpected Variation in Dementia Classification in Longitudinal Survey Data. *The journals of gerontology. Series B, Psychological sciences and social sciences, 77*(12), e234–e246. https://doi.org/10.1093/geronb/gbac128.

National Health and Aging Trends Study (NHATS). (n.d.). https://www.nhats.org/researcher/nhats.

Appendix

R syntaxes

```
# Set Up
## packages
library(haven)
library(dplyr)
```

```
library(readr)
library(ggplot2)
library(lme4)
library(tibble)
library(ggcorrplot)
library(DHARMa)
library(tibble)
library(geepack)
library(car)
## working folder path
folder_path <- "C:/Users/cwcheng/Dropbox (University of
Michigan)/Chia/687/datasets/SP/rounds 5-8"

## get a list of all .dta files in the folder
file_list <- list.files(path = folder_path, pattern = ".dta",
full.names = TRUE)

## initialize an empty data frame to use later on
combined_HCAll = NULL

## select variables we need and iterate through each data set file to
combine the data set
for (file in file_list) {
  tryCatch({
    data <- read_dta(file)
    ## extract the year number from the dresid column
    dresid_column <- grep("dresid$", names(data), value = TRUE)
    HCAll <- data %>%
      select(spid, ends_with(c("dgender",
                               "dracehisp",
                               "d2intvrage",
                               "hlkpfrwrk"))) %>%
      mutate(year_number =  dresid_column[1]) %>%
      data.frame()
    names(HCAll) = c(1:ncol(HCAll))
    # combine with the existing data
    combined_HCAll <- bind_rows(combined_HCAll, HCAll)
  }, error = function(e) {
    cat("Error in processing file:", file, "\n")
  })
}

# Data Cleaning
## change column names
names(combined_HCAll) =  c("id",
```

```
                              "gender",
                              "race",
                              "age",
                              "keep_from_work",
                              "year")


## mutate "year" column using for longitudinal analysis
final <- combined_HCAll %>%
  mutate(year = as.numeric(sub("^r(\\d+)dresid$", "\\1", year))) %>%
  filter(year %in% c(5:8))
## save the data frame to an Rdata in our shared folder
save(final, file = "C:/Users/cwcheng/Dropbox (University of
Michigan)/Chia/687/project/final1.Rdata") # 27,469 obs.

## drop invalid data in columns from our data frame
finalkfw <- final %>%
  subset(select = c(1:6),
         subset = rowSums(final > 0,
                          na.rm = TRUE) == ncol(final)) %>%
  ## sort out missing values
  subset(race != 5) %>% ## sort out don't know or no primary
selection
  subset(race != 6) %>% ## sort out don't know
  mutate(keep_from_work = ifelse(keep_from_work == 2,
                                 0,
                                 ifelse(keep_from_work == 1,
                                        1,
                                        NA))) %>%
  ## change not being kept from work to 0 and being kept from work to
1
  mutate(gender = ifelse(gender == 2,
                         0,
                         ifelse(gender == 1,
                                1,
                                NA)))
  ## change female to 0 while male stays 1

## save this dataframe as an Rdata for my groupmates again
save(finalkfw, file = "C:/Users/cwcheng/Dropbox (University of
Michigan)/Chia/687/project/finalkfw1.Rdata") # 23,984 obs.

# Pre-model Building Descriptive Analysis
## a. for independent variables
## race
```

```
## create a dataset for the plot later on
race_summary <- finalkfw %>%
  group_by(year, keep_from_work, race) %>%
  summarise(percentage = n()) %>%
  ungroup() %>%
  group_by(year, keep_from_work) %>%
  mutate(Cumpercentage = percentage/sum(percentage)) %>%
  ungroup() %>%
  mutate(keep_from_work = factor(ifelse(keep_from_work == 0, "0
(No)", "1 (Yes)"))) %>%
  mutate(race = case_when(
    race == 1 ~ "White, non-hispanic",
    race == 2 ~ "Black, non-hispanic",
    race == 3 ~ "Other",
    race == 4 ~ "Hispanic"
  ),
  year = case_when(
    year == 5~2015,
    year == 6~2016,
    year == 7~2017,
    year == 8~2018,
  ))

## create a percentage bar plot
ggplot(race_summary, aes(x = factor(year),
                         y = Cumpercentage, fill = keep_from_work)) +
  geom_bar(stat = "identity", position = "fill") +
  labs(title = "Percentage of Work Cessation Across the Four Years
Based on Respondents' Races",
       x = "Year",
       y = "Percentage") +
  theme_minimal() +
  guides(fill = guide_legend(title = "Work Cessation")) +
  theme(axis.text.x = element_text(angle = -45, hjust = 0),
        plot.title = element_text(size = 10)) +
  facet_wrap(~ factor(race))

## gender
## create a dataset for the plot later on
gender_summary <- finalkfw %>%
  group_by(year, keep_from_work, gender) %>%
  summarise(percentage = n()) %>%
  ungroup() %>%
  group_by(year, keep_from_work) %>%
  mutate(Cumpercentage = percentage/sum(percentage)) %>%
```

```
  ungroup() %>%
  mutate(keep_from_work = factor(ifelse(keep_from_work == 0, "0
(No)", "1 (Yes)"))) %>%
  mutate(gender = case_when(
    gender == 0 ~ "Female",
    gender == 1 ~ "Male"
  ),
  year = case_when(
    year == 5~2015,
    year == 6~2016,
    year == 7~2017,
    year == 8~2018,
  ))

## create a percentage bar plot
ggplot(gender_summary, aes(x = factor(year),
                           y = Cumpercentage, fill = keep_from_work)) +
  geom_bar(stat = "identity", position = "fill") +
  labs(title = "Percentage of Work Cessation Across the Four Years
Based on Respondents' Genders",
       x = "Year",
       y = "Percentage") +
  theme_minimal() +
  guides(fill = guide_legend(title = "Work Cessation")) +
  theme(axis.text.x = element_text(angle = -25, hjust = 0),
        plot.title = element_text(size = 10)) +
  facet_wrap(~ factor(gender))

## age
## create a dataset for the plot later on
age_summary <- finalkfw %>%
  group_by(year, keep_from_work, age) %>%
  summarise(percentage = n()) %>%
  ungroup() %>%
  group_by(year, keep_from_work) %>%
  mutate(Cumpercentage = percentage/sum(percentage)) %>%
  ungroup() %>%
  mutate(keep_from_work = factor(ifelse(keep_from_work == 0, "0
(No)", "1 (Yes)"))) %>%
  mutate(age = case_when(
    age == 1 ~ "65-69",
    age == 2 ~ "70-74",
    age == 3 ~ "75-79",
    age == 4 ~ "80-84",
    age == 5 ~ "85-89",
```

```
    age == 6 ~ "90+",
  ),
  year = case_when(
    year == 5~2015,
    year == 6~2016,
    year == 7~2017,
    year == 8~2018,
  ))

## create a percentage bar plot
ggplot(age_summary, aes(x = factor(year),
                        y = Cumpercentage, fill = keep_from_work))
+
  geom_bar(stat = "identity", position = "fill") +
  labs(title = "Percentage of Work Cessation Across the Four Years
Based on Respondents' Ages",
       x = "Year",
       y = "Percentage") +
  theme_minimal() +
  guides(fill = guide_legend(title = "Work Cessation")) +
  theme(axis.text.x = element_text(angle = -25, hjust = 0),
        plot.title = element_text(size = 10)) +
  facet_wrap(~ factor(age))

# Model Fitting
## linear regression
full_model <- lm(keep_from_work ~ factor(gender) + factor(race) +
factor(age),
                 data = finalkfw)
## find the best model by stepwise backward elimination
backward_model <- step(full_model, direction = "backward")
summary(backward_model)
## get the gvif of the remaining covariates in the best model
vif(backward_model)
## drop variables with colinearity and get the BIC score
backward_model2 = lm(keep_from_work ~ factor(race) + factor(age),
                     data = finalkfw)
BIC(backward_model2)

# trouble-shooting the dataset because we have the same QIC scores
finalkfw$id <- as.character(finalkfw$id)
class(finalkfw$id)
finalkfw_id <- finalkfw %>%
  group_by(id) %>%
  mutate(id2 = cur_group_id()) %>%
```

```
  mutate(size = n()) %>%
  filter(size == 4) %>%
  ungroup() %>%
  arrange(id2, year)

## marginal model
## independent var-cov matrix
mod_indep = geeglm(keep_from_work ~ factor(race) + factor(age),
                   id = id2,
                   waves = year,
                   family = binomial("logit"),
                   data = finalkfw_id,
                   corstr = 'independence')
summary(mod_indep)
QIC(mod_indep)

## exchangeable var-cov matrix
mod_exch = geeglm(keep_from_work ~ factor(race) + factor(age),
                  id = id2,
                  waves = year,
                  family = binomial("logit"),
                  data = finalkfw_id,
                  corstr = 'exchangeable')
summary(mod_exch)
QIC(mod_exch)

## AR(1) var-cov matrix
mod_ar1 = geeglm(keep_from_work ~ factor(race) + factor(age),
                 id = id,
                 waves = year,
                 family = binomial("logit"),
                 data = finalkfw,
                 corstr = 'ar1')
summary(mod_ar1)
QIC(mod_ar1)

## create a table for the three QICs
qic_tab <- data.frame("gee model-independence" = QIC(mod_indep), "gee
model-exchangeable" = QIC(mod_exch), "gee model-AR(1)" =
QIC(mod_ar1))
qic_tab

# generating a contingency table
final_distinct <- finalkfw %>%
  distinct(id, .keep_all = TRUE)
```

```
counts <- finalkfw %>%
  group_by(race, age, keep_from_work, year) %>%
  summarise(count = n()) %>%
  tibble()
counts[, 1] <- ifelse(counts[, 1] == 1, "1 (White, non-Hispanic)",
ifelse(counts[, 1] == 2, "2 (Black, non-Hispanic)", ifelse(counts[,
1] == 3, "3 (Other)", "4 (Hispanic)")))
counts[, 2] <- ifelse(counts[, 2] == 1, "1 (65-69)", ifelse(counts[,
2] == 2, "2 (70-74)", ifelse(counts[, 2] == 3, "3 (75-79)",
ifelse(counts[, 2] == 4, "4 (80-84)", ifelse(counts[, 2] == 5, "5
(85-89)", "6 (90+)")))))
counts[, 3] <- ifelse(counts[, 3] == 0, "0 (not kept from work)", "1
(kept from work)")
write_csv(counts, "C:/Users/cwcheng/Dropbox (University of
Michigan)/Chia/687/project/counts1.csv")
```