WELCOME TO THE NDACAN SUMMER TRAINING SERIES!

National Data Archive on Child Abuse and Neglect

Duke University, Cornell University, University of California San Francisco, & Mathematica

NDACAN

Children's Bureau
An Office of the Administration for Children & Families

SUMMER TRAINING SERIES SCHEDULE

- **July 2nd, 2025**
  - Developing a research question & exploring the data
- **July 9th, 2025**
  - Data management
- **July 16th, 2025**
  - Linking data
- **July 23rd, 2025**
  - Exploratory Analysis
- **July 20th, 2025**
  - Visualization and finalizing the analysis

LIFECYCLE OF AN NDACAN RESEARCH PROJECT

This session is being recorded.

Please submit questions to the Q&A box.

See ZOOM Help Center for connection issues: https://support.zoom.us/hc/en-us

If issues persist and solutions cannot be found through Zoom, please contact Andres Arroyo at aa17@cornell.edu.

SESSION AGENDA

- STS recap
- Exploratory analysis
- Demonstration in R

STS RECAP

LINKING DATA

- Record-level linkage possible internally with NDACAN administrative data
- Aggregate linkage possible with other NDACAN data, external data
- Linkage requires clean, well-formatted data files with shared variables
- Linkage is a useful tool for building large datasets, dealing with data limitations, and enabling powerful research designs
- Linkage can create and/or amplify data problems if data limitations are not understood and addressed

RESEARCH QUESTION

What is the relationship between lifetime incidence of removal and full-time employment among youth three years after aging out of foster care?

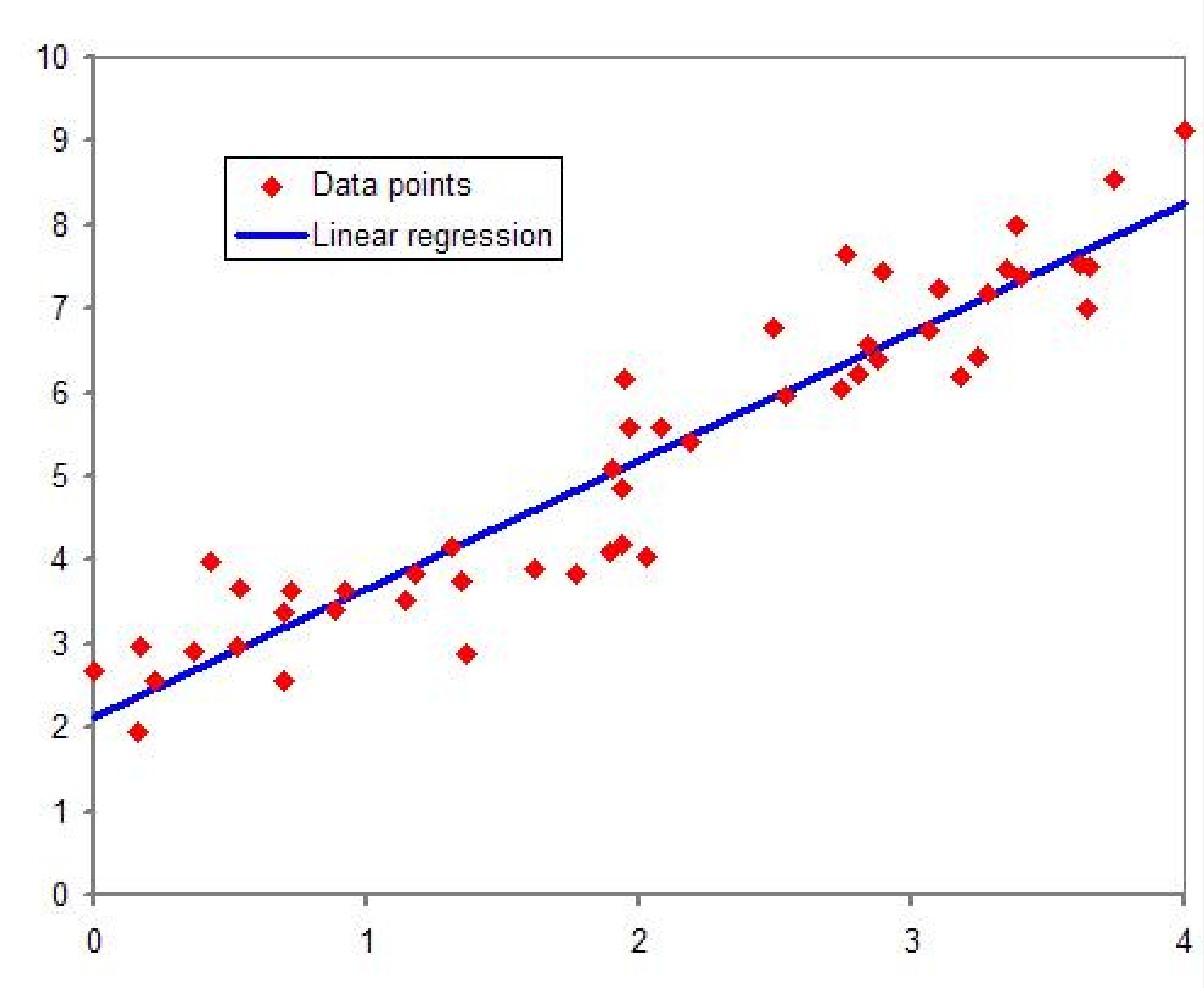
EXPLORATORY ANALYSIS

INTRODUCTION TO LINEAR REGRESSION

- Regression analysis is a statistical method for estimating the relationship between two (or more) random variables:
  - An outcome (or dependent variable)
  - One or more predictors (or independent variables)
- Linear regression is a powerful, flexible class of regression models that assume a linear relationship between the outcome and predictors

ESTIMATING LINEAR REGRESSION MODELS

- Linear regression models find the line (or hyperplane) of best fit representing the relationship between two (or more) random variables
- The most common method for estimating regression models is ordinary least squares (OLS)
- OLS minimizes the sum of the squares of the differences (residuals) between predicted values (blue line) and observed values (red points)



FUNDAMENTAL COMPONENTS OF LINEAR REGRESSION MODELS

- Consider the following bivariate regression:

$$\mathbf{y} = \beta_0 + \mathbf{x}\beta_1 + \boldsymbol{\varepsilon}$$

$\mathbf{y}$ is a $N \times 1$ vector of outcomes, where N is the number of observations in our data
$\mathbf{x}$ is a $N \times 1$ vector of predictors
β_0 is the main intercept (the predicted value of $\mathbf{y}$ when $\mathbf{x} = 0$)
β_1 is the coefficient (or parameter) of interest

- β_1 represents the slope of the line of best fit
- It is the main goal of regression analysis to estimate coefficients of interest validly (without bias) and efficiently (with precision)

$\boldsymbol{\varepsilon}$ is the error term, a $N \times 1$ vector of residuals (distances between red dots and blue line)

BASELINE MODEL

Instead of using matrix notation, we can represent the model using indexing:

$$y_i = \beta_0 + x_i\beta_1 + \varepsilon_i$$

In the case of our research design, the regression model takes the form:

$$CurrFTE\_3_i = \beta_0 + TOTREM_i\beta_1 + \varepsilon_i$$

- Because our outcome is binary, this model is known as a linear probability model.
  - In Presentation 5 we'll explore other models for binary and other categorical outcomes.

CORRELATION AND CAUSALITY

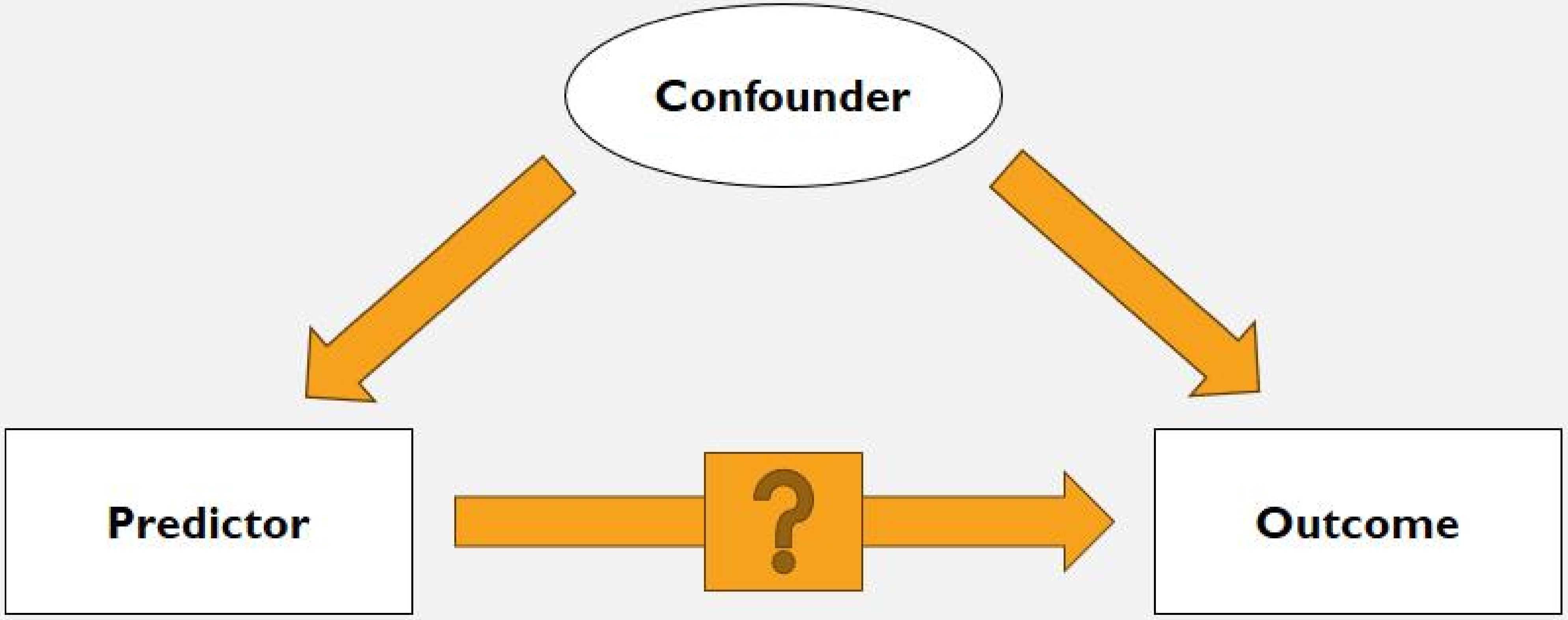
- Recall our research question:

What is the relationship between lifetime incidence of removal and full-time employment among youth three years after aging out of foster care?

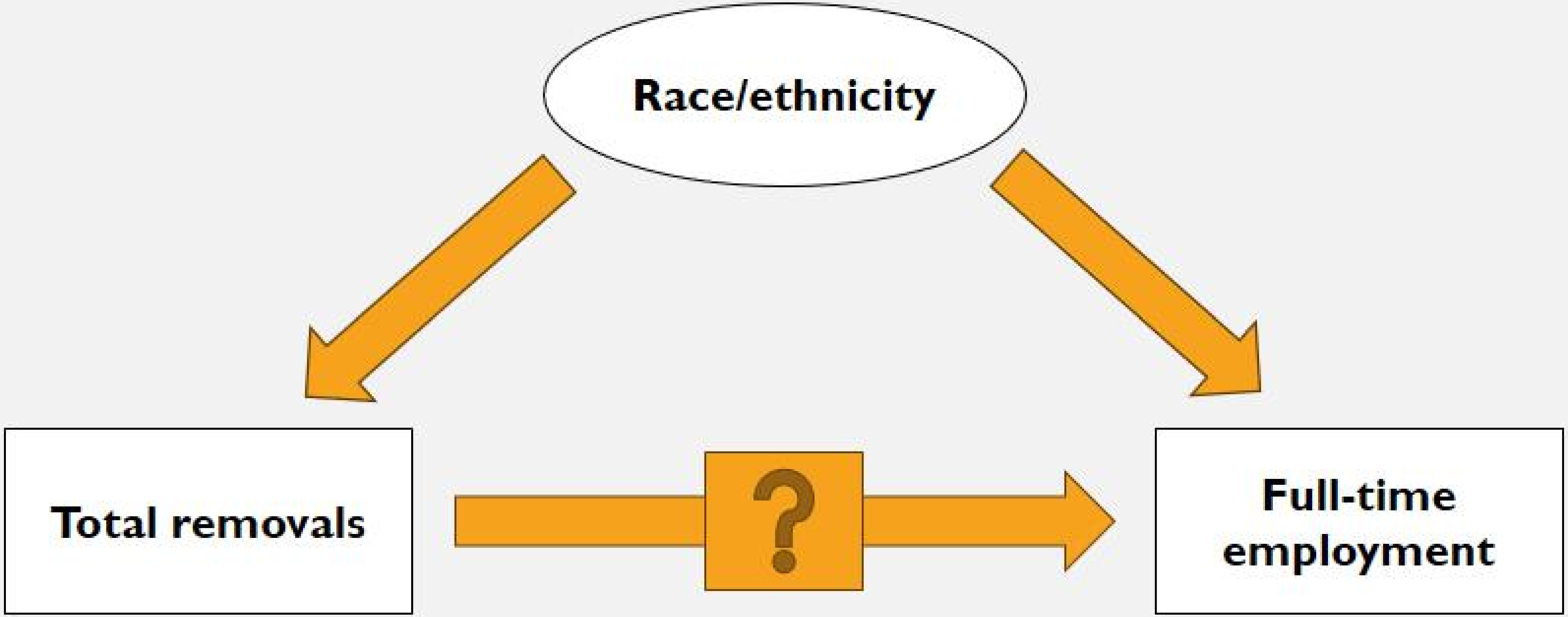
- What if we want to strengthen it to something like:

What is the effect of lifetime incidence of removal on full-time employment among youth three years after aging out of foster care?

OMITTED-VARIABLE BIAS



OMITTED-VARIABLE BIAS: EXAMPLE



CONTROLLING FOR OBSERVABLE CONFOUNDERS

We can deal with observed confounders by incorporating them into our model as additional predictors (or covariates).

Note that adding a predictor with C categories introduces $C-1$ parameters, which measure the difference in outcome for each category relative to a reference category (here, $White_i$)

$$CurrFTE\_3_i = \beta_0 + TOTREM_i\beta_1 + Black_i\beta_2 + AIAN_i\beta_3 + Asian_i\beta_4 + NHPI_i\beta_5 + Multi_i\beta_6 + Hisp_i\beta_7 + \varepsilon_i$$

CONTROLLING FOR UNOBSERVABLE CONFOUNDERS

- There are many, many potential confounders that are not observed or even observable
  - For example, the relationship between foster placement and employment may be confounded by (intangible) features of child welfare policy and practice
- One simple strategy for addressing such unobserved confounders: introduce group-specific intercepts, or fixed effects
  - For example, if CPS systems vary across states but are stable within them, including state intercepts will control for them

EXAMPLE: STATE FIXED EFFECTS

$$\boldsymbol{y} = \beta_0 + \boldsymbol{x}\beta_1 + \boldsymbol{S\gamma} + \boldsymbol{\varepsilon}$$

S is an $N \times (G-1)$ matrix of indicator variables, where G is the number of US states

$\boldsymbol{\gamma}$ is a $(G-1) \times 1$ vector of coefficients, or state fixed effects

$$\boldsymbol{CurrTFE\_3 = \beta_0 + TOTREM\beta_1 + RaceEthn\delta + \ State\gamma + \varepsilon}$$

STRATIFICATION

- Perhaps there's reason to think the answer to your research question will be different for different populations.
  - For example, the relationship between removal incidence and full-time employment may be different for people who are and are not currently enrolled in school
- Stratification allows our model estimates to vary across the values of a stratum variable
  - For example, we could estimate our model separately on currently enrolled and not currently enrolled populations
  - Or we could interact the enrollment variable (CurrEnroll) with all other model parameters

RELAXING PARAMETRIC ASSUMPTIONS

- By default, linear models assume linear relationships between predictors and outcomes
- We can relax this constraint in at least two ways:
  - Adding quasi-linear parameters like quadratic terms or splines
  - Including separate parameters for each level of a variable
- The next presentation will explore models for non-continuous outcomes

EXTENSION: DEALING WITH MISSING DATA

- Statistical software (including most R packages) will almost always listwise-delete records with missing values of modeled variables
- Listwise deletion is rarely advisable, particularly if large amounts of data are missing
- Always:
  - Examine the degree of missingness in your data
  - Consider the mechanisms that generated the missing data
  - Implement a defensible approach to dealing with missing data

DEMONSTRATION IN R

QUESTIONS?

ALEX ROEHRKASSE
AROEHRKASSE@BUTLER.EDU

NOAH WON
NOAH.WON@DUKE.EDU

PAIGE LOGAN PRATER
PAIGE.LOGANPRATER@UCSF.EDU

NEXT WEEK…

Date: July 30th, 2025

Topic: Visualization and Finalizing the Analysis

Instructor: Noah Won

R CODE PAGE 1 OF 5

```
###########
# NOTES #
###########

# This program file demonstrates strategies discussed in
# session 4 of the 2025 NDACAN Summer Training Series
# "Data Management."

# For questions, contact the presenter
# Noah Won (noah.won@duke.edu).

# Note that because of the process used to anonymize data,
# all unique observations include partially fabricated data
# that prevent the identification of respondents.
# As a result, all descriptive and model-based results are fabricated.
# Results from this and all NDACAN presentations are for training purposes only
# and should never be understood or cited as analysis of NDACAN data.


#########################
# TABLE OF CONTENTS #
#########################

# 0. SETUP
# 1. Simple Linear Regression
# 2. Multiple Regression
# 3. Stratified Multiple Regression
```

R CODE PAGE 2 OF 5

```
################
# 0. SETUP #
################

# Clear environment
rm(list=ls())

# Installs packages if necessary, loads packages
if (!requireNamespace("pacman", quietly = TRUE)){
  install.packages("pacman")
}
pacman::p_load(data.table, tidyverse, mice)

# Defines filepaths working directory
project <- "C:/Users/nhwn1/Downloads/STS5/data"
data <- "C:/Users/nhwn1/Downloads/STS5/data"

# Set working directory
setwd(project)

# Set seed
set.seed(1013)
```

R CODE PAGE 3 OF 5

```
#######################
# 2. Simple Linear Regression #
#######################

# Let's read in our cleaned, anonymized
# versions of the 2020 AFCARS files
afcars <- fread(paste0(data,'/afcars_clean_anonymized_linear.csv'))
head(afcars, 20)

# Running frequency tables of predictors of interest
table(afcars$SEX)
table(afcars$RaceEthn)
table(afcars2$FCMntPay)

# Creating Dummy Variables and Age Variables for Predictors
# Also filtering out those older than 30
afcars2 <- afcars %>%
        mutate(SEX_d = case_when(
          SEX == "Male" ~ 1,
          SEX == "Female" ~ 0),
        Hispanic = case_when(
          RaceEthn == "Hispanic" ~ 1,
          TRUE ~ 0),
        age = as.numeric(difftime(Sys.Date(), DOB, units = "days")) / 365.25
        ) %>%
        filter(age <= 30)

# Checking new derived variables
table(afcars2$SEX_d)
table(afcars2$Hispanic)
table(afcars2$age)
```

R CODE PAGE 4 OF 5

```
# Let's run a linear regression using age as a predictor and fcmntpay as an outcome
model <- lm(FCMntPay ~ age, data = afcars2)
summary(model)

#Let's visualize this model
ggplot(afcars2, aes(x = age, y = FCMntPay)) +
  geom_point() +
  geom_smooth(method = "lm", col = "red") +
  labs(title = "Linear Regression: FCMntPay ~ age",
       x = "age",
       y = "FCMntpay")

# A written form of our model is as follows: FCMntPay = 85.594 * age - 139.5495

###########################################################
# 3. Multiple Linear Regression #
# Our model seems to describe a positive relationship between age and FCMntPay but what about
# Hispanic status as a confounder?
model2 <- lm(FCMntPay ~ age + Hispanic, data = afcars2)
summary(model2)

# It seems that Hispanic status has a negative effect on FCMntPay
# Keep in mind that the beta values for age and the intercept have changed
# A written form of our model is as follows: FCMntPay = 85.594 * age + -105.2086 * Hispanic -
119.0115
```

R CODE PAGE 5 OF 5

```
#########################################################
# 4. Stratified Multiple Linear Regression #
#########################################################
# Stratified regression models fit different models based on the stratifications of a provided variable
# Adding a dummy variable and using a stratified regression model can be used to address confounding variables
# Stratified models are helpful when a variable violates linearity or homoscedasticity assumptions and cannot
# be used in a linear model
afcars3 <- afcars2 %>%
  filter(!is.na(SEX_d))

model3 <- afcars3 %>%
  group_by(SEX_d) %>%
  do(model4 = lm(FCMntPay ~ Hispanic + age, data = .))

model3 %>%
  do({
    model_summary <- summary(.$model)
    data.frame(
      SEX_d = unique(.$SEX_d),
      Intercept = coef(model_summary)[1, 1],
      Hispanic_coef = coef(model_summary)[2, 1],
      Age_coef = coef(model_summary)[3, 1]
    )
  })
# A written form of our model is as follows:
# Women - FCMntPay = 98.5 * age + -118 * Hispanic - 206
# Men - FCMntPay = 73.2 * age + -85.7 * Hispanic - 40.1
# It seems that women have a larger increase in FCMntPay compared to men as they age, but Hispanic
# women have less FCMntPay than Hispanic men
```