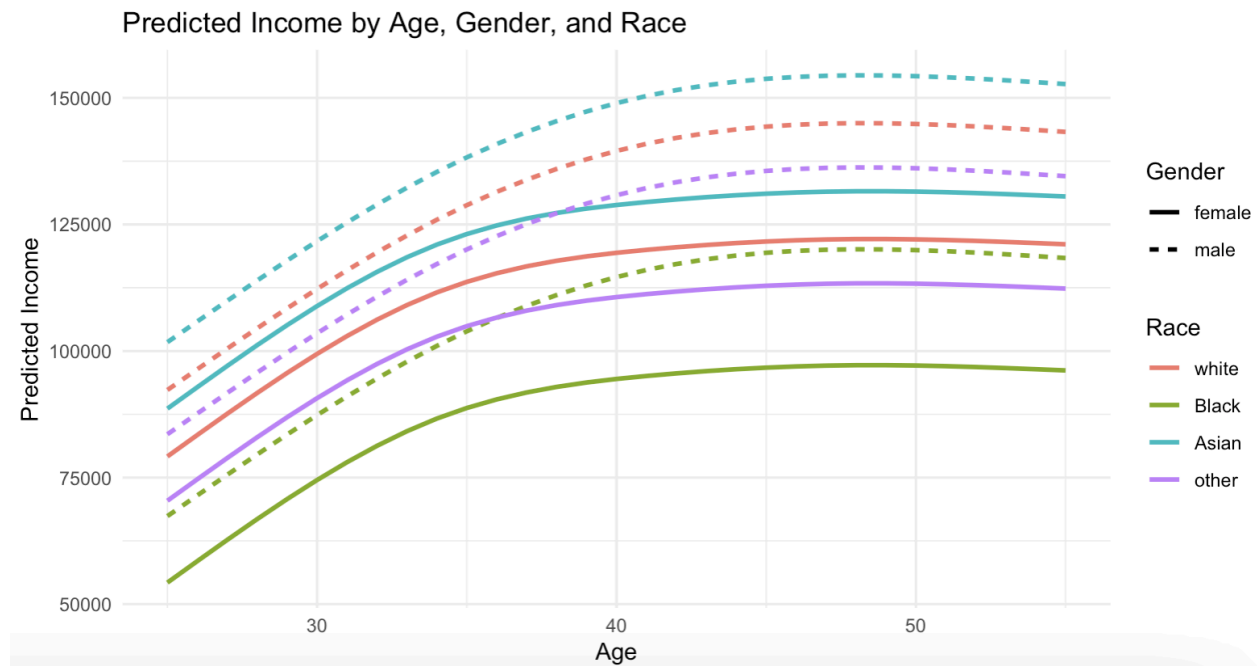Keyla Pereira
Eco B2000
Group Members: Federico Ciandri, Maurice Agonsi, Keyla Pereira

### Predicted Income by Age, Gender, and Race



After using the data to predict income by age, gender and race from ages 25-55, predicted income rises fast from the mid-20s into the late 30s and early 40s and then flattens in the 40s–50s. At every age, men earn more than women and the gap is fairly steady. Age clearly matters overall (joint F = 32.4, p < 0.001), and higher education is associated with higher wages. The female coefficient is -$13,000, t = -13.84, p < 0.001; 95% CI [11,000, 14,000], meaning women earn about $13k less than the men after controlling for age, education, and race. Overall, pay tends to rise with age and education, but a consistent gender gap remains across the prime working years.

```
load("/Users/keylapereira/Downloads/d_HHP2020_24.Rdata")
load("d_HHP2020_24")

prime_age_laborforce_data <- d_HHP2020_24 %>% filter(!is.na(work_kind) &
                              Age >= 25 & Age <= 55)
summary(prime_age_laborforce_data)
model_1 <- lm(income_midpoint ~ Age + Gender + Education + Race + Hispanic,
        data = prime_age_laborforce_data)
summary(model_1)
install.packages("modelsummary")
library(modelsummary)
library(modelsummary)
modelsummary(model_1, stars = TRUE, gof_map = c("nobs", "r.squared"))
install.packages("AER")
```

```r
library(AER)
to_be_predicted1 <- data.frame(Age = 20:55, Gender = "female",
                    Education = "adv degree",
                    Race = "Black",
                    Hispanic = "Hispanic")
to_be_predicted1$yhat <- predict(model_1, newdata = to_be_predicted1)
to_be_predicted1 <- expand.grid(
  Age = 20:55,
  Gender = c("female"),
  Education = "college grad",
  Race = unique(d_HHP2020_24$Race),
  Hispanic = "not Hispanic"
)
to_be_predicted1$yhat <- predict(model_1, newdata = to_be_predicted1)
library(ggplot2)
ggplot(to_be_predicted1, aes(x = Age, y = yhat, color = Gender)) +
  geom_line(size = 1) +
  facet_wrap(~ Race) +
  labs(
    title = "Predicted outcome by Age, Gender, and Race",
    x = "Age",
    y = "Predicted Value (yhat)"
  ) +
  theme_minimal()
to_be_predicted1 <- expand.grid(
  Age = 20:55,
  Gender = c("male"),
  Education = "college grad",
  Race = unique(d_HHP2020_24$Race),
  Hispanic = "not Hispanic"
)
ggplot(to_be_predicted1, aes(x = Age, y = yhat, color = Gender)) +
  geom_line(size = 1) +
  facet_wrap(~ Race) +
  labs(
    title = "Predicted outcome by Age, Gender, and Race",
    x = "Age",
    y = "Predicted Value (yhat)"
  ) +
  theme_minimal()
library(ggplot2)
library(dplyr)
to_be_predicted_f <- expand.grid(
  Age = 20:55,
  Gender = "female",
  Education = "college grad",
  Race = unique(d_HHP2020_24$Race),
  Hispanic = "not Hispanic"
)
```

```r
to_be_predicted_f$yhat <- predict(model_1, newdata = to_be_predicted_f)
to_be_predicted_f$GenderGroup <- "Female"
to_be_predicted_m <- expand.grid(
  Age = 20:55,
  Gender = "male",
  Education = "college grad",
  Race = unique(d_HHP2020_24$Race),
  Hispanic = "not Hispanic"
)
to_be_predicted_m$yhat <- predict(model_1, newdata = to_be_predicted_m)
to_be_predicted_m$GenderGroup <- "Male"
predictions_combined <- bind_rows(to_be_predicted_f, to_be_predicted_m)
ggplot(predictions_combined, aes(x = Age, y = yhat, color = Race, linetype = GenderGroup)) +
  geom_line(size = 1) +
  labs(
    title = "Predicted Values by Age, Gender, and Race",
    x = "Age",
    y = "Predicted Outcome",
    color = "Race",
    linetype = "Gender"
  ) +
  theme_minimal()
prime_age_laborforce_data <- d_HHP2020_24 %>%
  filter(!is.na(work_kind) & Age >= 25 & Age <= 55)
summary(prime_age_laborforce_data)
library(splines)
model_1 <- lm(income_midpoint ~ ns(Age, df = 4) * Gender + Education + Race + Hispanic,
          data = prime_age_laborforce_data)
summary(model_1)
library(modelsummary)
modelsummary(model_1, stars = TRUE, gof_map = c("nobs", "r.squared"))
to_be_predicted <- expand.grid(
  Age = 25:55,
  Gender = c("female", "male"),
  Education = "college grad",
  Race = unique(prime_age_laborforce_data$Race),
  Hispanic = "not Hispanic"
)
to_be_predicted$yhat <- predict(model_1, newdata = to_be_predicted)
ggplot(to_be_predicted, aes(x = Age, y = yhat, color = Race, linetype = Gender)) +
  geom_line(size = 1) +
  labs(
    title = "Predicted Income by Age, Gender, and Race",
    x = "Age",
    y = "Predicted Income",
    color = "Race",
    linetype = "Gender"
  ) +
  theme_minimal()
```