# The Gender Pay Gap : A case study

## Exploring Earnings Differences between Women and Men

**Griffin, Cook**



## Introduction

Our project aims to investigate gender-based earnings disparities in the United States using the CPS-March (CPSMAR) data set. We seek to understand whether there are significant differences in earnings between genders and how these disparities may have changed over time, and if these changes are effected by variables such as: education level and hours worked. We've found evidence suggesting that gender-based wage gaps persist, even after controlling for these factors.

## Data fundamentals

### March 2022 CPS

The standard monthly CPS is a nationally representative survey conducted by the U.S. Census Bureau and the Bureau of Labor Statistics. It collects demographic and labor force data from a selection of around 90,000 households across the U.S. It collects additional socioeconomic information including: income, poverty status, insurance coverage, and is crucial for analyzing economic well-being and income inequality.

```
# Read extract created by cpsmar_t.R
cpsmar_t <- read_csv(here("data", "cpsmar_t.csv"))
```

### March 2022 extract

The script renamed and created several variables for analysis, including indicators for gender, ethnicity, education levels, and employment status. Additionally, it calculated the Labor Force Participation Rate (LFPR) and stored these calculations in a data frame. The script selected a subset of variables from the CPS data for further analysis, including information on age, gender, ethnicity, education, earnings, hours worked, union membership, and insurance coverage, among others. There are 52,097 observations found in the cpsmar_t data set.

---

# Beginning to learn

## From extract to analysis sample

```
cpsmar_a <- cpsmar_t %>%
  filter(
    age >= 23,
    age <= 62,
    earnings>0
  ) %>%
  mutate(
    learnings = log(earnings),
    gender = case_when(
      female == 1 ~ "Female",
      female == 0 ~ "Male"
    )
  )
```

---

The script filters the cpsmar_t dataset to include individuals aged 23 to 62 with positive earnings, and creates a new variable, log_earnings, by taking the natural logarithm of said earnings. Additionally, it creates a gender variable based on the existing female indicator, categorizing individuals as “Female” or “Male.”

---

## The baseline earnings distribution

```
ggplot(cpsmar_a, aes(earnings)) +
  geom_histogram(
    aes(y = after_stat(density)), color=1,
    fill="white", bins=50
    ) +
  stat_function(
    fun = dlnorm,
    args = list(
      meanlog = mean(log(cpsmar_a$earnings)),
      sdlog   = sd(log(cpsmar_a$earnings))
      ),
    colour = "red"
    ) +
  labs(
    title="Figure 1. Earnings Distrubution and Log-Normal Fit ",
    x="Earnings",
    y="Density"
    )
```

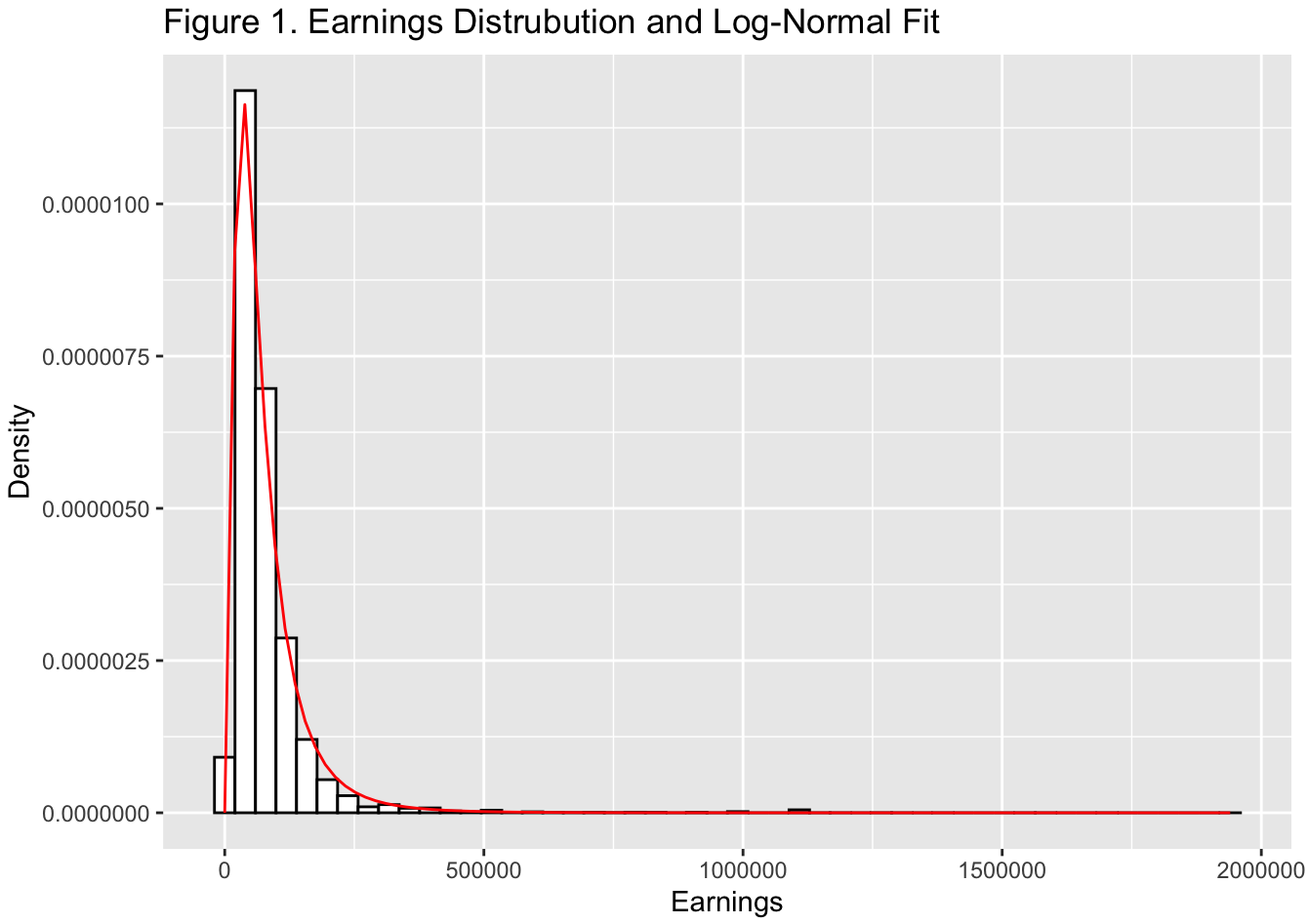



**Figure 1: Earnings Distrubution and Log-Normal Fit**

Figure 1 provides a visual representation of the earnings distribution in the form of a histogram. The right-skewed shape of the histogram suggests that the majority of individuals in the data set have lower earnings, while a smaller proportion of individuals earn higher incomes. The red curve overlaid on the histogram represents a log-normal probability density function fit to the data. It suggests that earnings in the data set follow a log-normal distribution, where most individuals cluster around a central value.

## Summary statistics on earnings, age and gender composition

```
datasummary(
  earnings + age + female ~ N + Mean + Median,
  data=cpsmar_a,
  title="Table 1. Summary of Earnings, Age, and Gender"
  )
```

```
 Table: Table 1. Summary of Earnings, Age, and Gender
```

| | N | Mean | Median |
|---|---|---|---|
| earnings | 46194 | 78540.32 | 58000.00 |
| age | 46194 | 42.28 | 42.00 |
| female | 46194 | 0.43 | 0.00 |

### Table 1: Summary of Earnings, Age, and Gender

Table 1 provides key empirical insights from the data set: it includes 46,194 observations with an average age of 42.28. The average earnings among the observed individuals is 78,540.32, yet they show a significantly lower median earnings of 58,000. Additionally, the data comprises approximately 43% females.

## Earnings distributions for women and men

```
ggplot(cpsmar_a, aes(x = earnings, group = female, fill = female)) +
  geom_density(adjust=1.5, alpha = 0.4) +
  labs(
    title="Figure 2. Earnings Distribution by Gender ",
    x="Earnings",
    y="Density"
    )
```

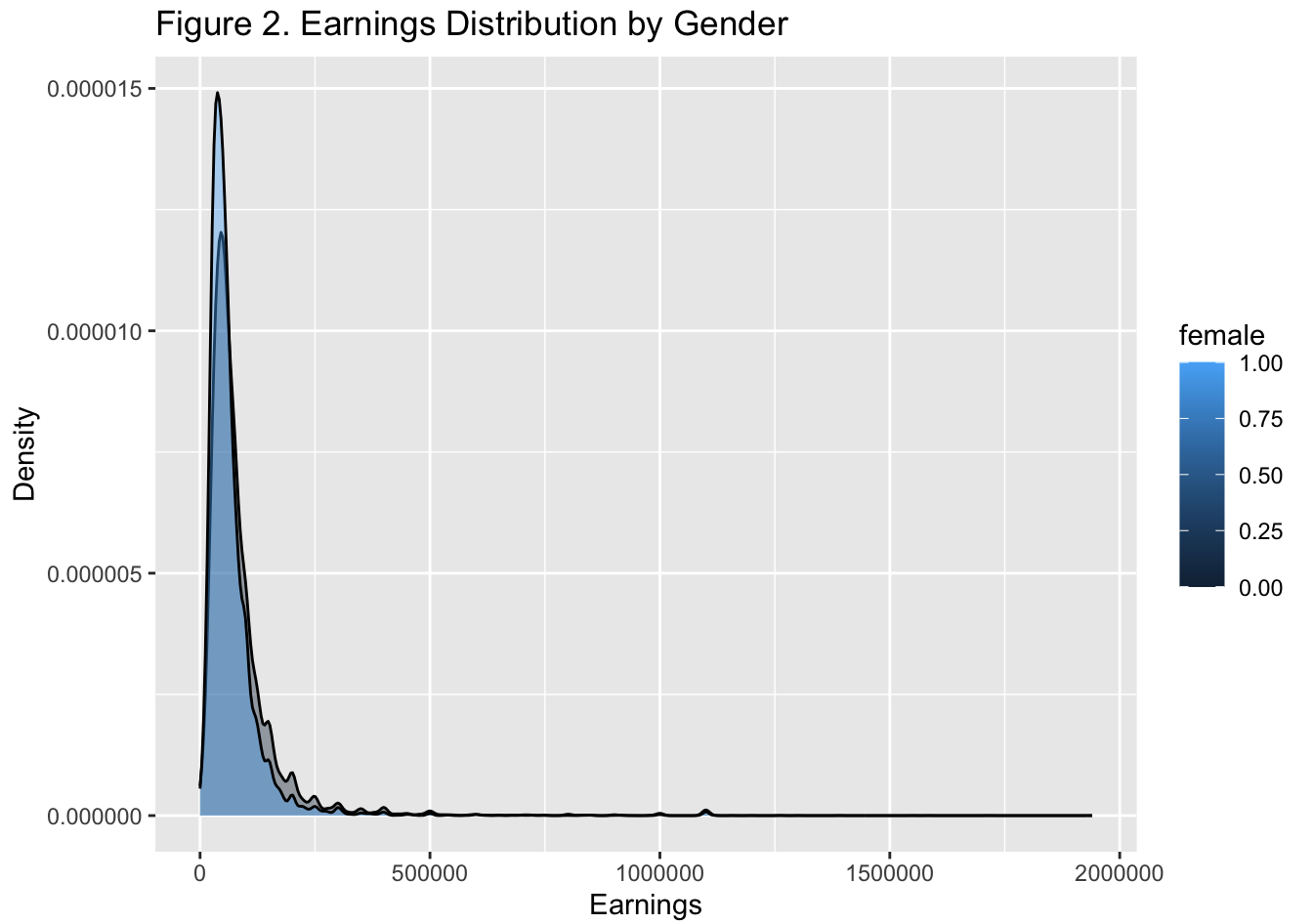



**Figure 2: Earnings Distribution by Gender using Density**

---

Figure 2 illustrates the earnings distribution by gender, showing similar shapes with males having a higher central earnings tendency compared to females. While both genders exhibit comparable earnings variability, this suggests potential gender-based earnings disparities.

---

## Summary statistics on earnings and age by gender

```
datasummary(
  earnings + age + female ~ gender*(N + Mean + Median),
  data=cpsmar_a,
  title="Table 2. Summary Statistics by Gender"
  )
```

```
 Table: Table 2. Summary Statistics by Gender
```

| | Female / N | Female / Mean | Female / Median | Male / N | Male / Mean | Male / Median |
|---|---|---|---|---|---|---|
| earnings | 20030 | 67645.89 | 51000.00 | 26164 | 86880.62 | 62000.00 |
| age | 20030 | 42.36 | 42.00 | 26164 | 42.21 | 42.00 |
| female | 20030 | 1.00 | 1.00 | 26164 | 0.00 | 0.00 |

**Table 2: Summary Statistics by Gender**

Table 2 provides a gender-based breakdown of earnings and age statistics. It reveals that, on average, females have lower earnings than males, with median incomes of 51,000 for females and 62,000 for males, while the ages of both genders are similar, with medians of 42 years.

# Models for exploration

## Estimating the CEF, $E(earnings|age)$

```
cef <- cpsmar_a %>%
  mutate(age=age- 23) %>%
  group_by(age) %>%
  summarise(earnings_bar = mean(earnings))
print(cef[c(1, 5, 10, 15, 20, 30), ])
```

```
## # A tibble: 6 × 2
##     age earnings_bar
##   <dbl>        <dbl>
## 1     0       39685.
## 2     4       56589.
## 3     9       73064.
## 4    14       77242.
## 5    19       89522.
## 6    29       83391.
```

## Assessing the Estimated CEF of Earnings

The estimated CEF of earnings by age reveals a typical career earnings trajectory. Earnings show steady growth at the beginning but then begins to increase more deliberately around 10, signifying accelerated earnings growth during mid-career. The CEF suggests that earnings peak around 30,

marking the prime working years, and then gradually decline as individuals approach retirement. This pattern can be captured by a quadratic model, reflecting initial slow growth, followed by acceleration, and eventually slowing down earnings in later career stages.

## Plotting and modeling the CEF

```
formula <- y~poly(x, 2, raw=TRUE)
ggplot(data = cef, aes(x= age, y= earnings_bar)) +
  geom_line() +
  geom_point() +
  geom_smooth(
    formula= formula,
    method="lm",
    se=FALSE,
    color="red"
    ) +
  stat_poly_eq(
    mapping = use_label(c("eq")),
    formula= , color = "red"
    ) +
  labs(
    title="Figure 3. Estimated Conditional Expectation of Earnings by Age",
    x="Year",
    y="Average earnings"
    )
```

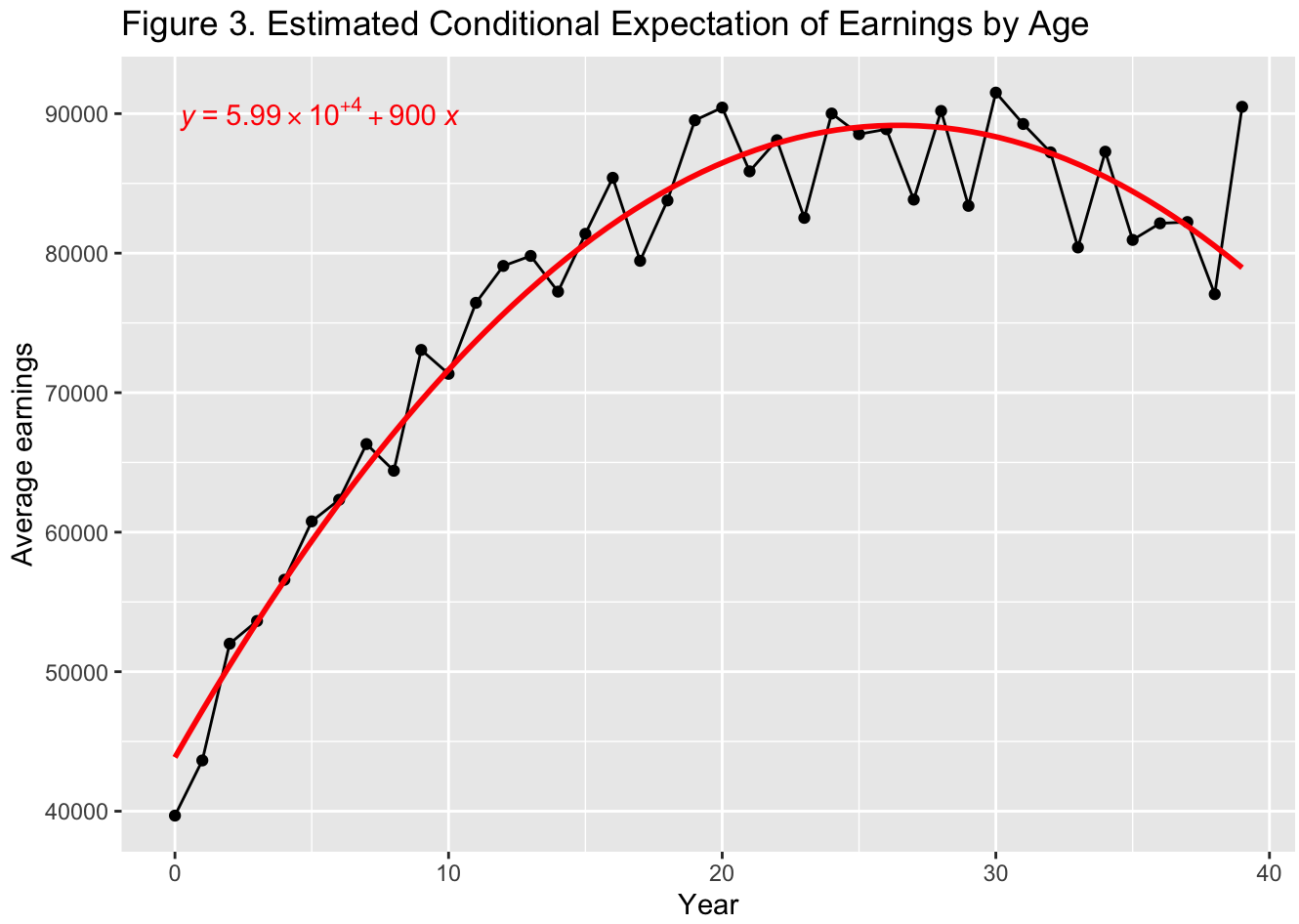



**Figure 3: Fitted Model vs. CEF**

---

The fitted quadratic model's predictions closely align with the CEF for earnings at different ages in a career. At year 0, both the CEF and quadratic model suggest earnings around 40,000. As careers progress, the earnings increase gradually, with both methods indicating around 80,000 at year 14 and approximately 90,000 at year 29. This consistency between the CEF and quadratic model reaffirms the model's effectiveness in capturing the underlying earnings trends across a typical career.

---

## Gender gap in earnings over a career

```
cef_fvm <- cpsmar_a %>%
  mutate(age = age - 23 ) %>%
  group_by(age, gender ) %>%
  summarise(earnings_bar = mean(earnings))

formula <- y~poly(x, 2, raw=TRUE)
ggplot(cef_fvm, aes(x = age ,y = earnings_bar, color = factor(gender))) +
  geom_point() +
  geom_line() +
  geom_smooth(
    formula = formula,
    method = "lm",
    se = FALSE,
    aes(group = gender)
    ) +
  stat_poly_eq(
    aes(label =  paste(stat(eq.label))),
    formula = formula,
    parse = TRUE
    ) +
  labs(
    title="Figure 4. Earnings profiles by gender",
    x="Year",
    y="Average earnings"
    )
```


**Figure 4: A Periodic Assessment**

---

When taking a look at the model, we see that the male earnings curve consistently remains above the female earnings curve at all ages, indicating that, on average, males tend to earn more than females throughout their careers. The gap in earnings between genders also widens over time, this is evident from the increasing separation between the two curves as careers progress. In year 0, the male-female ratio is approximately 1.15 (male: 42,145.27 / female: 36,458.60). In year 14, the male-female ratio increases to around 1.20 (male: 83,410.93 / female: 69,366.65). In year 29, the male-female ratio reaches approximately 1.26 (male: 92,179.52 / female: 72,975.20).

---

## Plotting and fitting the percentage gender gap

```
males    <- filter(cef_fvm, gender == "Male")
females  <- filter(cef_fvm, gender == "Female")
df_ratio <- data.frame(age = males$age,
                       ratio = males$earnings_bar /females$earnings_bar )

ggplot(df_ratio, aes(x = age , y = ratio )) +
  geom_line() +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  stat_poly_eq(mapping = use_label(c("eq")), color = "blue") +
  labs(
    title="Figure 5. Male-female earnings ratio by age",
    x="Year",
    y="Male-female average earnings ratio"
    )
```

Figure 5. Male-female earnings ratio by age

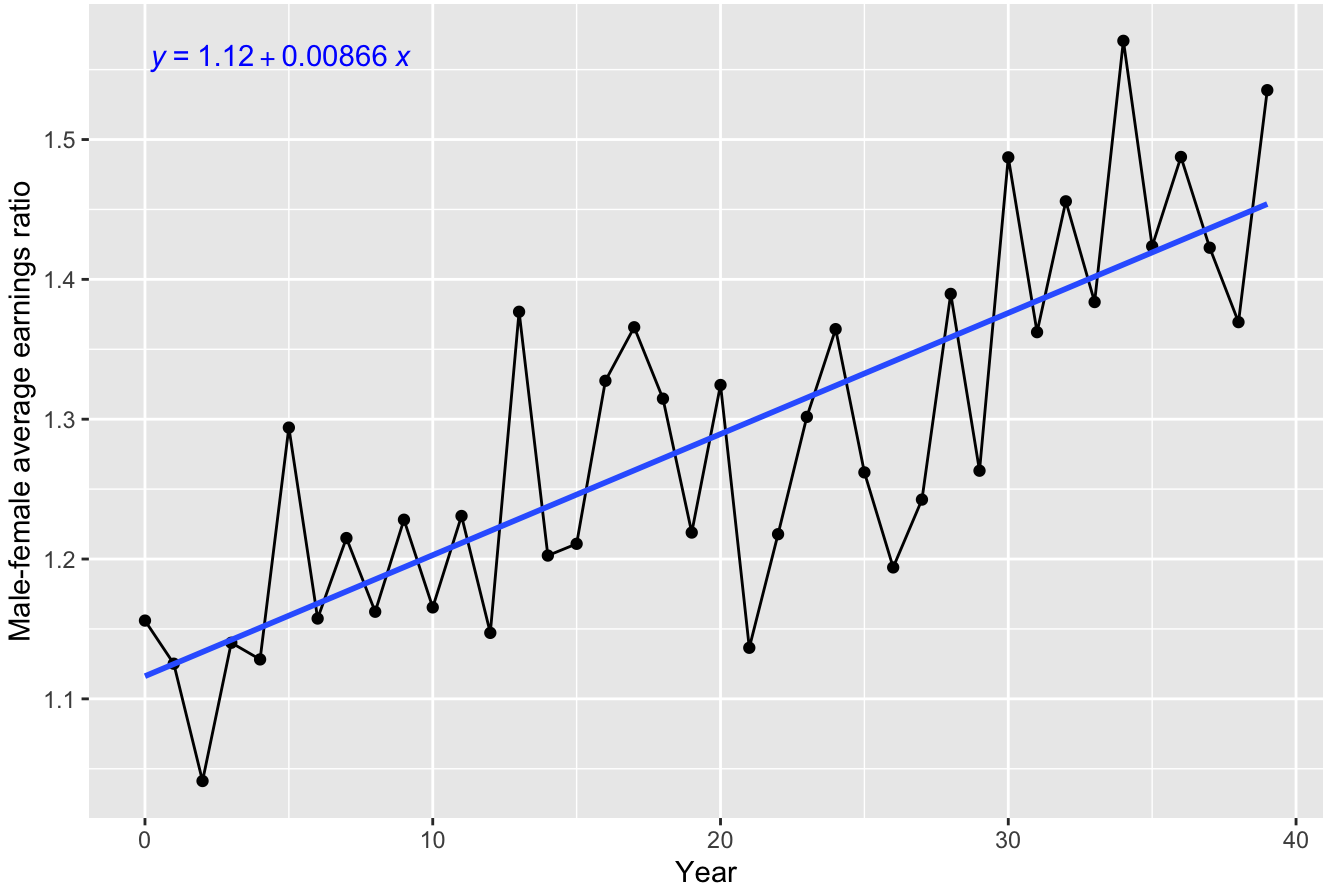



**Figure 5: Predicted Gender Gap by Year**

According to the linear model, the male-female earnings ratio at Year 0 is 1.12, while the peak value for year 0 based on the non-linear model is around 1.15. When looking at the non-linear model, we also see a peak in the gap-ratio around year 34, at a value above 1.55. To predict the gap in each year,

we can use the linear model to find how much the ratio increases each year. Furthermore, based on the linear model, the gap is predicted to increase by 0.00866 each year.

---

# Making inferences

## Earnings, wages and hours

```
cpsmar_a <- cpsmar_a %>%
  filter(earnings>0) %>%
  mutate(
    wage    =  earnings / hours,
    lwage   =  log(wage),
    over40  = case_when(hours > 40 ~"1", TRUE~"0"),
      )

datasummary(
   earnings + wage + age + hours ~ gender*(N + Mean + SD),
  data=cpsmar_a,
  title="Table 3. Summary Statistics for Earnings, Wage, Age, and Hours "
  )
```

```
Table: Table 3. Summary Statistics for Earnings, Wage, Age, and Hours
```

| | Female / N | Female / Mean | Female / SD | Male / N | Male / Mean | Male / SD |
|---|---|---|---|---|---|---|
| earnings | 20030 | 67645.89 | 71615.67 | 26164 | 86880.62 | 97742.32 |
| wage | 20030 | 1590.72 | 1599.38 | 26164 | 1967.43 | 2130.30 |
| age | 20030 | 42.36 | 10.64 | 26164 | 42.21 | 10.65 |
| hours | 20030 | 42.26 | 6.10 | 26164 | 44.03 | 7.77 |

## Table 3: Summary Statistics for Earnings, Wage, Age, and Hours

---

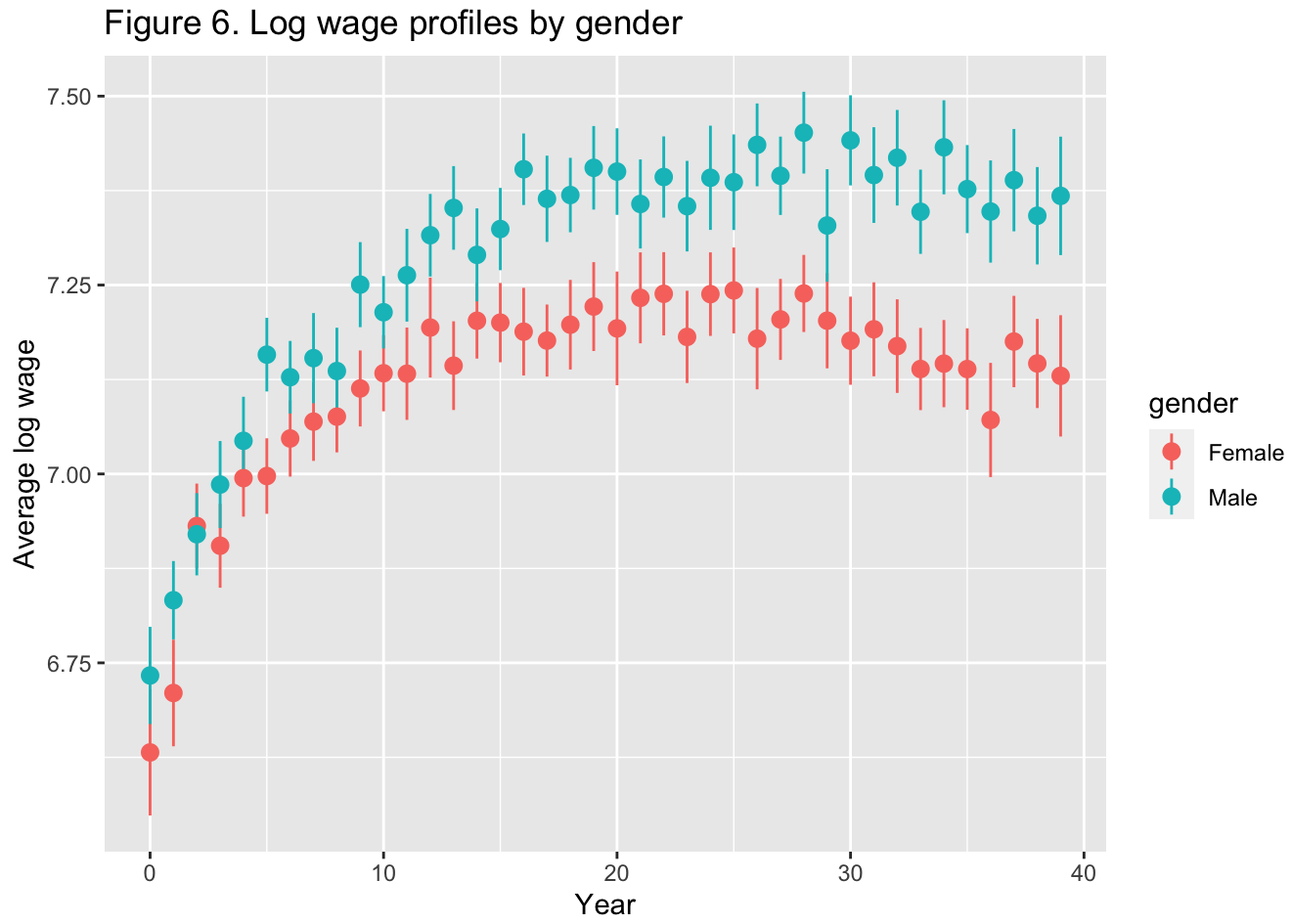
Table 3 reports empirical facts on earnings, wage, age, and hours worked for both Female and Male individuals. Notably, Male individuals generally have higher earnings, wages, and hours worked on average compared to their Female counterparts, while the age distribution appears similar for both genders.

---

## Log wage CEFs by gender

```
cef_fvm_w <- cpsmar_a %>%
  mutate(age=age-23) %>%
  group_by(age, gender) %>%
  summarise(
    lwage_bar =  mean(lwage, na.rm = TRUE),
    lwage_se  =  sd(lwage, na.rm = TRUE)/sqrt(n()),
    upper =  lwage_bar + qnorm(0.975) * lwage_se,
    lower =  lwage_bar - qnorm(0.975) * lwage_se
    )

ggplot(cef_fvm_w, aes(x = age, y = lwage_bar, color=gender)) +
  geom_pointrange(aes(ymin = lower, ymax = upper)) +
  labs(
    title="Figure 6. Log wage profiles by gender",
    x="Year",
    y="Average log wage"
    )
```




## Figure 6: Overlap

The graph of the estimated Conditional Earnings Functions (CEFs) for women and men illustrates that men consistently earn more than women throughout their careers. Notably, there is an initial overlap in earnings distributions during the early career years, before year 5, suggesting gender pay equity in the early stages. However, as careers progress, the earnings gap widens, emphasizing the persistent gender-based earnings disparity in later career stages.

## Gender wage gap by hours of work

```
cpsmar_a1 <- cpsmar_a %>% filter(over40 == "1")

datasummary(
 wage + hours + female~ over40*(N + Mean + SD), data=cpsmar_a1,
  title="Table 4. Summary statistics by over 40 hours per week"
  )
```

```
Table: Table 4. Summary statistics by over 40 hours per week
```

| | N | Mean | SD |
|---|---|---|---|
| wage | 12415 | 1994.27 | 2104.65 |
| hours | 12415 | 52.48 | 8.54 |
| female | 12415 | 0.33 | 0.47 |

## Table 4: Summary statistics by over 40 hours per week

Table 4 presents empirical facts regarding earnings and weekly work hours, stratified by gender and whether individuals worked over or under 40 hours per week. Among females, the mean wage is 1,590.72 with a standard deviation of 1,599.38, while males have a higher mean wage of 1,967.43 with a standard deviation of 2,130.30. In terms of weekly work hours, females worked an average of 42.26 hours with a standard deviation of 6.10, whereas males worked an average of 44.03 hours with a standard deviation of 7.77. This highlights gender-based disparities in earnings and work hours.

```
cef_fvm_w_h <- cpsmar_a %>%
  mutate(age=age-23) %>%
  group_by(age, gender,  over40) %>%
  summarise(lwage_bar = mean(lwage, na.rm = TRUE))

formula <- y~poly(x, 2, raw=TRUE)
ggplot(cef_fvm_w_h, aes(x = age , y = lwage_bar , color = gender)) +
  geom_point() +
  geom_line() +
  geom_smooth(
    method = "lm",
    se = TRUE,
    formula = formula) +
  stat_poly_eq(
    aes(label = paste(stat(eq.label))),
    formula = formula,
    parse = TRUE,
    vstep = .08
    ) +
  stat_poly_eq(
    aes(label =  paste(stat(eq.label))),
    formula = formula,
    parse = TRUE,
    vstep = .08 # To adjust annotation placement
    ) +
  facet_wrap(~ over40, dir = "v") + # dir="v" places figures vertically
  labs(
    title = "Figure 7. Log wage profiles by gender and hours",
    x="Age ",
    y="Average log wage "
    )
```


**Figure 7: Log wage profiles by gender and hours**

---

In Figure 7,when comparing the CEFs for women and men who work at least or more than 40 hours per week, it is evident that men consistently earn more than women throughout their careers. However, there is some overlap in the CEFs during the early stages of their careers, suggesting that wage growth patterns are somewhat similar in the initial years of employment. This overlap is visually reinforced by the standard error bands, which indicate that predicted log wages are more similar in the early-career ages, gradually diverging as careers progress.

---

## Early career gender gap and long hours

```
cpsmar_a_2330 <- cpsmar_a  %>%
  filter(age >= 23, age <= 30, over40 == 1
         )

t.test(lwage~ gender, data=cpsmar_a_2330)
```

```
##
##  Welch Two Sample t-test
##
## data:  lwage by gender
## t = -0.99614, df = 1227.2, p-value = 0.3194
## alternative hypothesis: true difference in means between group Female and
group Male is not equal to 0
## 95 percent confidence interval:
##  -0.10430065  0.03405266
## sample estimates:
## mean in group Female   mean in group Male
##             6.955681             6.990805
```

### Results

The t-test results suggest that there is no statistically significant difference in average log wages between women and men aged 23 to 30 who work more than 40 hours per week, with a p-value of 0.3194.

## Summary

In our analysis of the gender pay gap, we initially observed that men tend to earn higher average earnings and wages than women across various age groups, indicating a persistent gender wage disparity. However, this gap tends to narrow during early career stages, particularly among individuals aged 23 to 30, where some overlap in earnings is evident, as supported by both the estimated conditional earnings functions and log wage profiles. Nevertheless, as careers progress, men consistently earn more than women, which is further confirmed by our statistical tests. Despite adjusting for numerous factors, the result of men making more than woman remains constant. These findings emphasize the need for continued efforts to address gender-based wage inequalities, especially as careers advance.

## Appendix 1

March 2022 CPS extract documentation

The Current Population Survey (CPS) is a monthly survey conducted by the Bureau of the Census of the Bureau of Labor Statistics, encompassing approximately 90,000 U.S. households. It serves as the primary data source for understanding various aspects of the U.S. population's labor force

characteristics, encompassing topics such as employment, earnings, education, income, poverty, health insurance coverage, job history, voting behavior, computer usage, veteran status, and more. Further details can be accessed at www.census.gov/cps and dataferrett.census.gov.

For our analysis, we utilized data from the March 2022 CPS survey, specifically focusing on individuals with complete and non-allocated variables who met the criteria of full-time employment, while excluding those affiliated with the military. This sample comprises 152,732 individuals, and we extracted a total of 22 variables from the CPS dataset for our analysis.

Variable list

- age
- earnings
- hours
- race
- gender

Race

- 1 = White only
- 2 = Black only
- 3 = American Indian, Alaskan Native only (AI)
- 4 = Asian only
- 5 = Hawaiian/Pacific Islander only (HP)
- 6 = White-Black
- 7 = White AI
- 8 = White- AI
- 9 = White- HP
- 10 = Black - AI
- 11 = AI - Asian
- 12 = AI - HP
- 13 = AI-Asian
- 14 = AI-HP
- 15 = Asian-HP
- 16 = White-Black-AI

- 17 = White-Black-Asian
- 18 = White-Black-HP
- 19 = White-AI-Asian
- 20 = White-AI-HP
- 21 = White-Asian-HP
- 22 = Black-AI-Asian
- 23 = White-Black-AI-Asian
- 24 = White-AI-Asian-HP
- 25 = Other 3 race comb.
- 26 = Other 4 or 5 race comb.

Marital Status

- 1 = Married - civilian spouse present
- 2 = Married- AF spouse person
- 3 = Married – spouse absent
- 4 = Widowed
- 5 = Divorced
- 6 = Seperated
- 7 = Never married

Education Level

- HSGrad = 1, if education= 1
- SomeColl= 1, if <=education<= 2
- CollDeg = 1, if education>= 2

Region of residence

- 1 = Northeast
- 2 = Midwest
- 3 = South
- 4 = West

Other

- female = 1 if individual is female

- hisp = 1 if individual is Hispanic
- over40 = 1 if hours > 40
- union = 1 if in a union

---