BUSN 5000 Project

Exploring Pay Differences between Women and Men

Spencer Katzman

Spring 2025
(updated 20 Apr 25)

Academic honesty statement

I have been academically honest in all of my work and will not tolerate academic dishonesty of others, consistent with UGA's Academic Honesty Policy.

Sign the academic honesty statement by typing your name on the **Signature** line.

Signature: Spencer Katzman

We will not accept submissions that omit a signed Academic Honesty statement.

Introduction

Overview

This project examines pay differences between males and females using data from the March 2022 edition of the Current Population Survey (CPS). My analysis focuses on working-age individuals, their demographic and household characteristics, and their earnings. I identify a statistically significant wage gap where men earn more than women on average and show that the gap increases at a decreasing rate over a career. The gap persists when education and demographic controls are used in estimations.

Data

March 2022 CPS

The March 2022 CPS surveyed about 54,000 households. The monthly CPS collects labor force information about households' employment and demographics. The ASEC supplement adds data on income, work experience, poverty, and other variables. Importantly ASEC data is based on the previous year while CPS data is current.

March 2022 CPS Extract

```
cpsmar_e <- read_csv(here("data", "cpsmar_e.csv"))
```

Age, earnings, hours, weeks, race, marital status, education, and job type were extracted from the person file, and region, county, and city from the household file. Rename was used to simplify variable names, and mutate to create indicators for categorical data. Households with children under six were identified using group_by and mutate. After filtering for full-time workers only and merging the files based on household sequence number, 52,097 observations of 20 variables remained.

Analysis sample

```
cpsmar_a <- cpsmar_e %>%
  filter(
    age >= 23,
    age <= 62,
    earnings > 0
  ) %>%
  mutate(
    gender = ifelse(female == 1, "Female", "Male"),
    wage = earnings/(hours*weeks),
    lwage = log(wage),
    Black = case_when(race==2~1, TRUE ~ 0),
    south = case_when(region==3~1, TRUE ~ 0),
    married = case_when((marital==1 | marital==2 | marital==3)~1, TRUE ~ 0),
    age_centered = age - 23
  )
```

The analysis sample includes 46,194 observations of individuals with positive earnings who are between 23 and 62 years old.

Baseline earnings distributions

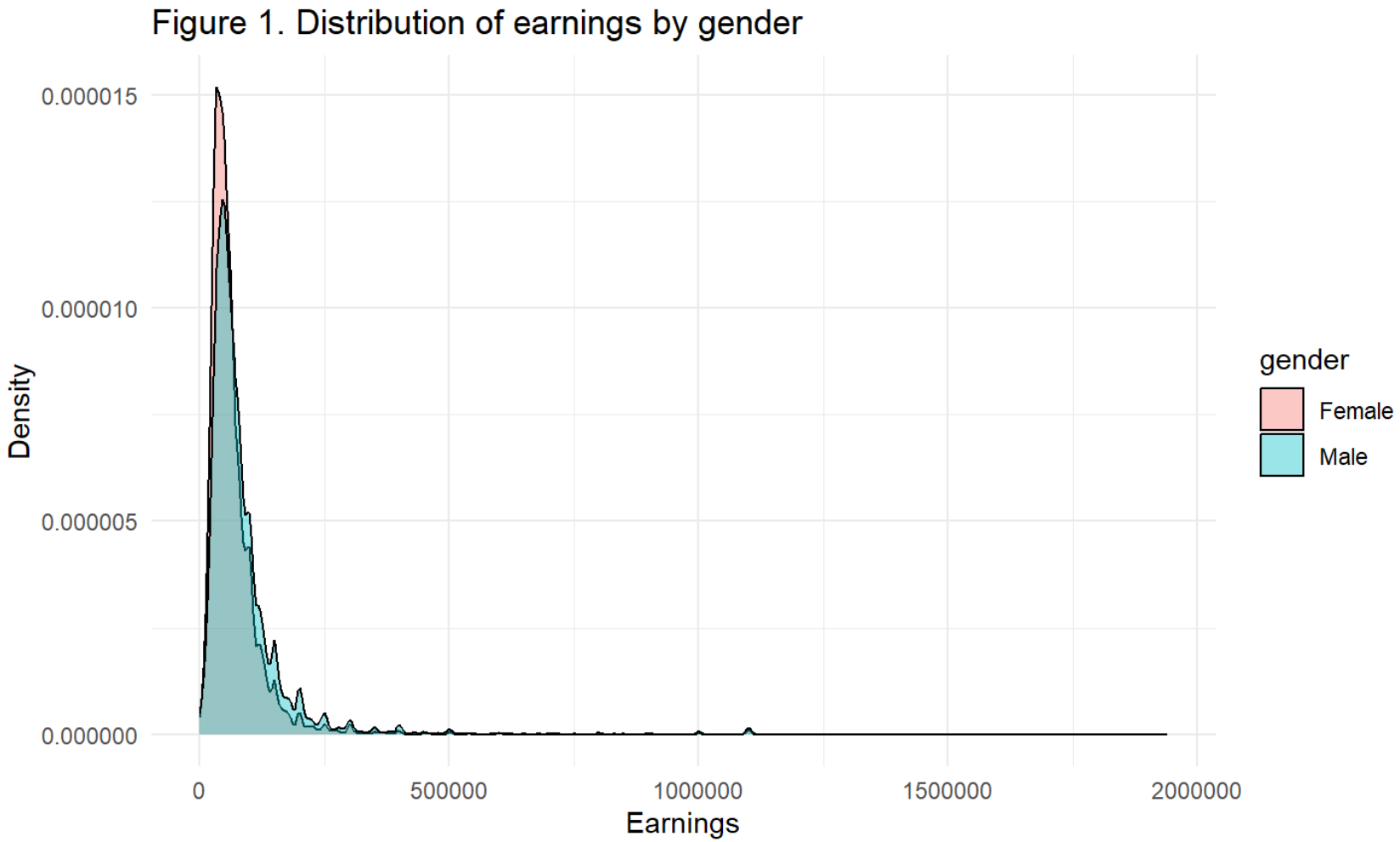
Plotting earnings distributions

```
figure1 <- ggplot(cpsmar_a, aes(x = earnings, group = gender, fill = gender)) +
  geom_density(alpha = 0.4) +
  labs(
    title="Figure 1. Distribution of earnings by gender",
    x="Earnings",
    y="Density"
    )+
  theme_minimal()
earnings_fvm <- cpsmar_a %>%
  group_by(gender) %>%
  summarize(avg_earnings = round(mean(earnings, na.rm = TRUE),0))

avg_earnings_f <- earnings_fvm %>%
  filter(gender == "Female") %>%
  pull(avg_earnings) # `pull` extracts the "avg_earnings" value for "Female" from earnings_fvm, a single value since the data
only record two genders.

avg_earnings_m <- earnings_fvm %>%
  filter(gender == "Male") %>%
  pull(avg_earnings) # `pull` extracts the "avg_earnings" value for "Male" from earnings_fvm, a single value since the data only
record two genders.
```

Distribution of earnings by gender



Baseline comparisons

The most important fact communicated by Figure 1 is that average male earnings are higher than average female earnings. Female average earnings are $67646, while male average earnings are $86881. The dollar gap is $19235, which translates to an approximate 28% pay gap. The figure also shows that male earnings have greater variability than female earnings.

The career gender gap

Wages and hours differences

Table 1. Wages and hours by gender

	Female			Male		
	N	Mean	SD	N	Mean	SD
wage	20030	30.65	30.92	26164	37.91	41.07
hours	20030	42.26	6.10	26164	44.03	7.77

Documenting the differences

Table 1 shows that there are more men than women in the sample. It also shows that men, on average, have higher hourly wages ($37.91 v. $30.65), work more hours (44.03 v. 42.26), and that their hours and wages are more variable than those for women.

Plotting career log wage profiles

```
cef_fvm_w <- cpsmar_a %>%
  group_by(age_centered, gender) %>%
  summarize(avg_lwage = mean(lwage, na.rm = TRUE))

figure2 <- ggplot(cef_fvm_w, aes(x = age_centered, y = avg_lwage, color = gender, linetype = gender, linewidth = gender)) +
  geom_point() +
  geom_line() +
  scale_linetype_manual(values = c("Female" = "longdash", "Male" = "solid")) +
  scale_linewidth_manual(values = c("Female" = 0.7, "Male" = 0.5)) +
  guides(linewidth = "none") +
  labs(
    title="Figure 2. Career log-wage profiles for women and men",
    x="Year",
    y="Average log wage"
    )+
  theme_minimal()
```

Career log wage profiles

Estimating wage differences over a career

```
males <- cef_fvm_w %>%
  filter(gender == "Male") %>%
  rename(avg_lwage_male = avg_lwage) %>%
  select(-gender)
females <- cef_fvm_w %>%
  filter(gender == "Female") %>%
  rename(avg_lwage_female = avg_lwage) %>%
  select(-gender)

diff_fvm <- inner_join(males, females, by = "age_centered") %>%
  filter(age_centered <= 30) %>%
  mutate(
    diff = avg_lwage_male - avg_lwage_female,
    age_group = cut(
      age_centered,
      breaks = c(-1, 10, 20, 30),
      labels = c("1-10", "11-20", "21-30"))
    ) %>%
  group_by(age_group) %>%
  summarize(mean_diff = mean(diff)*100)

table2 <- kable(
  diff_fvm,
  digits = 2,
  col.names = c("Year Range", "Avg Pct Difference"),
  align = "cc",
  caption = "Table 2. Percent wage differences, first 30 years",
  ) %>%
  kable_styling(position = "center")
```

Evolution of the gender wage gap

Table 2. Percent wage differences, first 30 years

Year Range	Avg Pct Difference
1-10	8.64
11-20	16.42
21-30	18.00

Discussing the gender wage gap evolution

Figure 2 and Table 2 show that the male-female wage gap in average log wages increases over the course of a career. Table 2 shows the gap in ten year increments whereas Figure 2 displays the quadratic nature of the gap increase more incrementally.

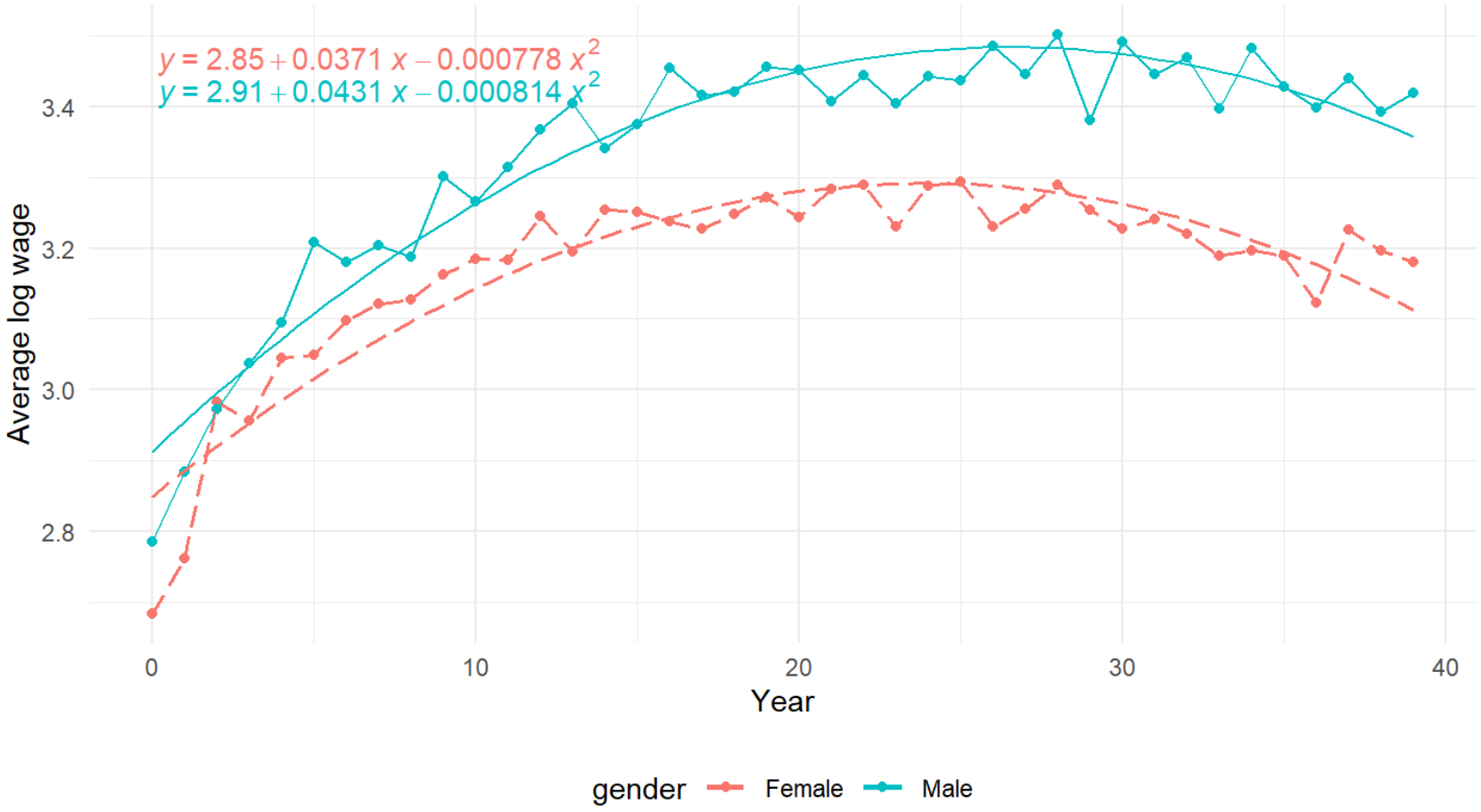
Explaining the gender wage gap

Fitting the log wage profiles

```
formula <- y ~ x + I(x^2)
figure3 <- figure2 +
  geom_smooth(
    method = "lm",
    formula = formula,
    aes(group = gender),
    se = FALSE
    ) +
  stat_poly_eq(
    aes(label =  after_stat(eq.label)),
    formula = formula,
    parse = TRUE
    ) +
  labs(
    title="Figure 3. Career log-wage profiles with quadratic fits",
    x="Year",
    y="Average log wage"
    )+
  theme_minimal() +
  theme(legend.position = "bottom")
```

Log wage profiles with quadratic fits

Figure 3. Career log-wage profiles with quadratic fits



Gender differences in education

Table 3. Educational attainment by gender

	Female			Male		
	N	Mean	SD	N	Mean	SD
HSGrad	20030	0.20	0.40	26164	0.28	0.45
SomeColl	20030	0.25	0.44	26164	0.24	0.43
CollDeg	20030	0.51	0.50	26164	0.41	0.49

Gender differences in demographics

Table 4. Demographic characteristics by gender

	Female			Male		
	N	Mean	SD	N	Mean	SD
Black	20030	0.13	0.34	26164	0.09	0.29
hisp	20030	0.17	0.38	26164	0.21	0.40
south	20030	0.38	0.49	26164	0.37	0.48
city	20030	0.68	0.47	26164	0.68	0.47
married	20030	0.58	0.49	26164	0.66	0.48
child_u6	20030	0.19	0.39	26164	0.23	0.42

Documenting differences in characteristics

Table 3 shows that females in the sample are less likely to have a high school degree than males, about equally likely to have some college, but more likely to have a college degree. Table 4 shows that males in the sample are more likely to be Hispanic while females are more likely to be Black. Table 4 also shows that males are more likely to be married and have at least one child under six in the household. Finally, residence in the South and in urban areas shows little to no gender difference.

Controlling for education and demographic characteristics

```
singles <- cpsmar_a %>%
  filter(
    married==0,
    child_u6==0
    )
models <- list(
  "Baseline"      = lm(lwage ~ female +
                       age_centered + I(age_centered^2),
                       data = cpsmar_a),
  "Add Education" = lm(lwage ~ female +
                       age_centered + I(age_centered^2) + HSGrad + SomeColl + CollDeg,
                       data = cpsmar_a),
  "Add Person"    = lm(lwage ~ female +
                       age_centered + I(age_centered^2) + HSGrad + SomeColl + CollDeg +
                       Black + hisp + south + city,
                       data = cpsmar_a),
  "Add Household" = lm(lwage ~ female +
                       age_centered + I(age_centered^2) + HSGrad + SomeColl + CollDeg +
                       Black + hisp + south + city +
                       married + child_u6,
                       data = cpsmar_a),
  "Only Singles"  = lm(lwage ~ female +
                       age_centered + I(age_centered^2) + HSGrad + SomeColl + CollDeg +
                       Black + hisp + south + city,
                       data = singles)
)
```

Reporting the results

```
cm <- c(
  'female'            = 'Female',
  'age_centered'      = 'Age',
  'I(age_centered^2)' = 'Age$^2$',
  '(Intercept)'       = 'Constant'
)
gm <-  tibble::tribble(
  ~raw, ~clean, ~fmt,
  "nobs", "$N$", 0,
  "r.squared", "$R^2$", 2
)
rows <- tribble(~term, ~Baseline, ~Add_Education, ~Add_Person, ~Add_Household, ~Only_Singles,
  'Education controls',   ' ',         'X',           'X',          'X',           'X',
  'Demographic controls', ' ',         ' ',           'X',          'X',           'X',
  'Household controls',   ' ',         ' ',           ' ',          'X',           'X'
)
attr(rows, 'position') <- c(9, 10, 11) # Positions where you want these rows to appear

table5 <- modelsummary(
  models,
  add_rows = rows,
  coef_map = cm,
  gof_map = gm,
  vcov = c("robust","robust","robust","robust","robust"),
  title = "Table 5. OLS estimates of the gender wage gap",
  notes = "Robust standard errors in parentheses.",
  escape = FALSE
  )
```

Explaining the gender wage gap

Table 5. OLS estimates of the gender wage gap

	Baseline	Add Education	Add Person	Add Household	Only Singles
Female	-0.162	-0.236	-0.227	-0.214	-0.123
	(0.007)	(0.006)	(0.006)	(0.006)	(0.011)
Age	0.039	0.032	0.031	0.025	0.025
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Age^2	-0.001	-0.001	-0.001	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Constant	2.977	2.479	2.520	2.472	2.495
	(0.010)	(0.017)	(0.018)	(0.019)	(0.031)
Education controls		X	X	X	X
Demographic controls			X	X	X
Household controls				X	X
N	46194	46194	46194	46194	15378
R^2	0.04	0.21	0.22	0.23	0.17

Robust standard errors in parentheses.

Documenting the findings

The baseline regression model shows a gender gap of 16.2%. This estimate increases to 23.6% when educational controls are added, indicating that the baseline model likely suffers from omitted variable bias. Estimates decrease slightly to 22.7% as personal demographic controls are added and yet more to 21.4% when household controls are used. Restricting the analysis to singles shows a much lower gap of 12.3%. All estimates of the female coefficient are statistically significant at the 99% confidence level.

Conclusion

Summary

This project examines gender pay differences using a March 2022 CPS sample of workers age 23 to 62 who made positive earnings. Statistically significant regression estimates show that men earn more than women. These estimates range from 12.3% for singles to 23.6% for the entire sample when only educational controls are used. The gap estimate lessens to 21.4% when all controls are used. It is also shown that the gap increases at a decreasing rate over the course of a career, almost doubling between the first and second 10 years, and continuing to grow throughout.

Appendix

Data documentation

```
# Define the variables and their descriptions
variables <- data.frame(
  Variable = c(
    "age",
    "earnings",
    "hours",
    "race",
    "marital",
    "HSGrad",
    "SomeColl",
    "CollDeg",
    "region",
    "female",
    "hisp",
    "fulltime"
    ),
  Definition = c(
    "years; capped at 85",
    "earnings; greater than 0",
    "hours worked per week",
    "respondent’s race (1 = White only, 2 = Black only, 3 = AI only, 4 = Asian only, 5 = Hawaiian/Pacific Islander only (HP), 6 = White-Black, 7 = White-AI, 8 = White-Asian, 9 = White-HP, 11 = Black-Asian, 12 = Black-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 = White-Black-AI-Asian-HP, 25 = Other 3 race comb., 26 = Other 4 or 5 race comb.)",
    "marital status (1 = Married civilian, 2 = Married AF, 3 = Married absent, 4 = Widowed, 5 = Divorced, 6 = Separated, 7 = Never married)",
    "= 1, if high school graduate",
    "= 1, if some college",
    "= 1, if college degree",
    "household region (1 = Northeast, 2 = Midwest, 3 = South, 4 = West)",
    "= 1, if female",
    "= 1, if Hispanic, Spanish, or Latino",
    "= 1, if works full time"
  )
)
```

List of main variables with definitions

This is a list of the main variables used in this project with their definitions.

Variable	Definition
age	years; capped at 85
earnings	earnings; greater than 0
hours	hours worked per week
race	respondent's race (1 = White only, 2 = Black only, 3 = AI only, 4 = Asian only, 5 = Hawaiian/Pacific Islander only (HP), 6 = White-Black, 7 = White-AI, 8 = White-Asian, 9 = White-HP, 11 = Black-Asian, 12 = Black-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 = White-Black-AI-Asian-HP, 25 = Other 3 race comb., 26 = Other 4 or 5 race comb.)
marital	marital status (1 = Married civilian, 2 = Married AF, 3 = Married absent, 4 = Widowed, 5 = Divorced, 6 = Separated, 7 = Never married)
HSGrad	= 1, if high school graduate
SomeColl	= 1, if some college
CollDeg	= 1, if college degree
region	household region (1 = Northeast, 2 = Midwest, 3 = South, 4 = West)
female	= 1, if female
hisp	= 1, if Hispanic, Spanish, or Latino
fulltime	= 1, if works full time