A Regression Analysis of the Gender Pay Gap

Maria-Cristiana Gîrjău

Revised on 2019-09-20

Data wrangling

```
# Reading in the dataset
acs_sample_raw <- read_csv("data/acs230_3k.csv")
```

```
# Wrangling the data
acs_sample <- acs_sample_raw %>%

  mutate(ADJINC.x = ADJINC.x / 10^6, # adding decimal point to ADJINC
         HINCP = HINCP * ADJINC.x, # adjusting dollar amounts for inflation
         WAGP = WAGP * ADJINC.x) %>% # adjusting dollar amounts for inflation

  # selecting which variables to keep
  select(SEX, AGEP, CIT, RAC1P, MIL, DIS, # general demographics
         MAR, HUPARC, NRC, FER, # family and household
         SCHL, FOD1P, FOD2P, SCIENGP, # educational background
         ESR, COW, WKW, WKHP, NAICSP, # employment
         HINCP, WAGP, # income
         REGION.x, ST.x) %>% # location

  # renaming the variables
  rename(sex = SEX,
         age = AGEP,
         citizenship = CIT,
         race = RAC1P,
         military = MIL,
         disabled = DIS,
         married = MAR,
         children_age = HUPARC,
         children_no = NRC,
         gave_birth = FER,
         education = SCHL,
         degree_1 = FOD1P,
         degree_2 = FOD2P,
         stem_degree = SCIENGP,
         employment = ESR,
         worker_class = COW,
         weeks_worked = WKW,
         hours_worked = WKHP,
         industry = NAICSP,
         hh_income = HINCP,
         wage_income = WAGP,
         region = REGION.x,
         state = ST.x) %>%
```

```
  # converting inputs to an appropriate data type
  mutate(sex = as.factor(sex) %>%
           fct_recode(!!!sex_levels),
         citizenship = as.factor(citizenship) %>%
           fct_recode(!!!citizenship_levels),
         race = as.factor(race) %>%
           fct_recode(!!!race_levels),
         military = as.factor(military) %>%
           fct_recode(!!!military_levels),
         disabled = ifelse(disabled == 1, TRUE, FALSE),
         married = as.factor(married) %>%
           fct_recode(!!!married_levels),
         children_age = as.factor(children_age) %>%
           fct_recode(!!!children_age_levels),
         gave_birth = ifelse(gave_birth == 1, TRUE, FALSE),
         education = as.factor(education) %>%
           fct_recode(!!!education_levels),
         stem_degree = ifelse(stem_degree == 1, TRUE, FALSE),
         employment = as.factor(employment) %>%
           fct_recode(!!!employment_levels),
         region = as.factor(region) %>%
           fct_recode(!!!region_levels),
         state = as.factor(state) %>%
           fct_recode(!!!state_levels)) %>%

  # filtering the individuals to keep
  filter(!(is.na(wage_income)) & wage_income > 0, # salary income is positive
         employment %in% c("employed working", "employed not working",
                           "military working")) # employed and/or working

# - merge degree variables into one (after conversion)

# CHANGE DISABLED AND STEM DEGREE TO LOGICAL
```

```
# removing unused variables
rm(children_age_levels, citizenship_levels, degree_levels,
     disabled_levels, education_levels, employment_levels,
     gave_birth_levels, married_levels, military_levels,
     race_levels, region_levels, sex_levels, state_levels,
     stem_degree_levels, worker_class_levels)
```

Data exploration

```
# sex
acs_sample %>%
  group_by(sex) %>%
  tally() %>%
  kable(booktabs = TRUE) %>%
  kable_styling(latex_options = "striped")
```

sex	n
M	707
F	594

```
# race
acs_sample %>%
  group_by(race) %>%
  tally() %>%
  kable(booktabs = TRUE) %>%
  kable_styling(latex_options = "striped")
```

race	n
white	993
black	122
native	16
asian	84
pacific islander	4
other	52
multiple	30

```
# income
acs_sample %>%
  favstats(~wage_income, data = .) %>%
  kable(booktabs = TRUE) %>%
  kable_styling(latex_options = "striped")
```

min	Q1	median	Q3	max	mean	sd	n	missing
222.4616	20223.78	40447.56	70783.23	601657.5	54437.45	62163.78	1301	0

```
# region
acs_sample %>%
  group_by(region) %>%
  tally() %>%
  kable(booktabs = TRUE) %>%
  kable_styling(latex_options = "striped")
```

region	n
Northeast	246
Midwest	288
South	443
West	324

```
# acs_sample %>%
#   filter(wage_income != 0) %>%
#   ggplot(aes(x = wage_income, fill = sex)) +
```

```
#   geom_density(alpha = 0.5) +
#   scale_x_continuous(trans = "log2") +
#   geom_vline(xintercept = median(acs_sample %>% filter(wage_income!=0, sex == "male") %>% .$wage_inco
#   geom_vline(xintercept = median(acs_sample %>% filter(wage_income!=0, sex == "female") %>% .$wage_in
#
# # add in filtering on education, stem degrees, etc.
# acs_sample %>%
#   filter(hh_income != 0, wage_income != 0) %>%
#   group_by(sex) %>%
#   summarize(median_hh_income = median(hh_income, na.rm = TRUE),
#             median_wage_income = median(wage_income, na.rm = TRUE)) %>%
#   ggplot() +
#   geom_bar(aes(x = sex, y = median_hh_income), stat = "identity", fill = "blue") +
#   geom_bar(aes(x = sex, y = median_wage_income), stat = "identity", fill = "red")
```

Data analysis

Assessment

Current questions