A Regression Analysis of the Gender Pay Gap

Maria-Cristiana Gîrjău

Revised on 2019-09-17

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
         SEMP = SEMP * ADJINC.x, # adjusting dollar amounts for inflation
         hours_worked = WKHP * WKW) %>% # total number of hours worked

  # 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, hours_worked, NAICSP, # employment
         HINCP, WAGP, # income
         REGION.x, ST.x) %>% # location

  # renaming the variables
  rename(sex = SEX,
         age = AGEP,
         citizenship = CIT,
         race = RAC1P,
         military = MIL,
         disability = 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,
         industry = NAICSP,
         hh_income = HINCP,
         wage_income = WAGP,
         region = REGION.x,
         state = ST.x)
```

```
# Remaining work to be done:
# - convert inputs to appropriate data types (e.g. factors)
# - merge degree variables into one (after conversion)
```

Data exploration

Data analysis

Assessment

Current questions