A Regression Analysis of the Gender Pay Gap (Technical Report)

Maria-Cristiana Gîrjău

Revised on 29 October 2019

Contents

1 Data Wrangling

1.1 Data Import

Importing both the provided sample and the full original PUMS dataset (merged, pre-wrangled, and filtered appropriately in an .R script). The provided sample will be used for analysis, while the full dataset might come in handy for certain visualizations such as maps, which require more data to be meaningful.

```
# Sample of 3000 to be wrangled in RMarkdown
ACSSmallRaw <- readr::read_csv("data/ACS3k.csv")

# Full dataset (already wrangled, see .R script for details)
ACSBig <- readRDS("data/ACSBig.Rds")
```

1.2 Preparation

Preparing the levels for the factor variables we intend to recode.

```
# Setting factor levels
sex_levels <- c(M = "1",
                "F" = "2")

# will only be used for tallying
race_levels <- c(White = "1",
                 Black = "2",
                 Native = "3",
                 Native = "4",
                 Native = "5",
                 Asian = "6",
                 Native = "7",
                 Other = "8",
                 Multiple = "9")

# will only be used for tallying
marital_status_levels <- c(Married = "1",
                           Widowed = "2",
                           Divorced = "3",
                           Separated = "4",
                           Single = "5")

# collapsing degree
degree_levels <- c("Agriculture" = "11",
                   "Environmental" = "13",
                   "Architecture" = "14",
                   "Ethnic Studies" = "15",
                   "Media and Journalism" = "19",
                   "Communication" = "20",
                   "Computer Science and IT" = "21",
                   "Cosmetology and Gastronomy" = "22",
                   "Education" = "23",
```

```
                   "Engineering" = "24",
                   "Engineering" = "25",
                   "Languages" = "26",
                   "Consumer Sciences" = "29",
                   "Law and Policy" = "32",
                   "English Literature" = "33",
                   "Liberal Arts" = "34",
                   "Library Science" = "35",
                   "Biological Sciences" = "36",
                   "Mathematics and Statistics" = "37",
                   "Military" = "38",
                   "Interdisciplinary" = "40",
                   "Fitness" = "41",
                   "Philosophy" = "48",
                   "Theology" = "49",
                   "Physical Sciences" = "50",
                   "Physical Sciences" = "51",
                   "Psychology" = "52",
                   "Law and Policy" = "53",
                   "Law and Policy" = "54",
                   "Social Sciences" = "55",
                   "Construction" = "56",
                   "Construction" = "57",
                   "Transportation" = "59",
                   "Arts" = "60",
                   "Medicine" = "61",
                   "Business and Finance" = "62",
                   "History" = "64")

# will be used for filtering
employment_levels <- c("Employed working" = "1",
                       "Employed not working" = "2",
                       Unemployed = "3",
                       "Military working" = "4",
                       "Military not working" = "5",
                       "Not in labor force" = "6")

region_levels <- c(Northeast = "1",
                   Midwest = "2",
                   South = "3",
                   West = "4",
                   "Puerto Rico" = "9")

state_levels <- c(AL = "1",
                  AK = "2",
                  AZ = "4",
                  AR = "5",
                  CA = "6",
                  CO = "8",
                  CT = "9",
                  DE = "10",
                  DC = "11",
                  FL = "12",
```

```r
                   GA = "13",
                   HI = "15",
                   ID = "16",
                   IL = "17",
                   IN = "18",
                   IA = "19",
                   KS = "20",
                   KY = "21",
                   LA = "22",
                   ME = "23",
                   MD = "24",
                   MA = "25",
                   MI = "26",
                   MN = "27",
                   MS = "28",
                   MO = "29",
                   MT = "30",
                   NE = "31",
                   NV = "32",
                   NH = "33",
                   NJ = "34",
                   NM = "35",
                   NY = "36",
                   NC = "37",
                   ND = "38",
                   OH = "39",
                   OK = "40",
                   OR = "41",
                   PA = "42",
                   RI = "44",
                   SC = "45",
                   SD = "46",
                   TN = "47",
                   TX = "48",
                   UT = "49",
                   VT = "50",
                   VA = "51",
                   WA = "53",
                   WV = "54",
                   WI = "55",
                   WY = "56",
                   "Puerto Rico" = "72")

# collapsing industry
industry_levels <- c("Agriculture, Forestry, Fishing and Hunting" = "11",
                     "Mining, Quarrying, and Oil and Gas Extraction" = "21",
                     "Utilities" = "22",
                     "Construction" = "23",
                     "Manufacturing" = "31",
                     "Manufacturing" = "32",
                     "Manufacturing" = "33",
                     "Manufacturing" = "3M",
                     "Wholesale Trade" = "42",
```

```
                     "Retail Trade" = "44",
                     "Retail Trade" = "45",
                     "Transportation and Warehousing" = "48",
                     "Transportation and Warehousing" = "49",
                     "Information" = "51",
                     "Finance and Insurance" = "52",
                     "Real Estate and Rental and Leasing" = "53",
                     "Professional, Scientific, and Technical Services" = "54",
                     "Management of Companies and Enterprises" = "55",
                     "Administrative and Support and Waste Management
                     and Remediation Services" = "56",
                     "Educational Services" = "61",
                     "Health Care and Social Assistance" = "62",
                     "Arts, Entertainment, and Recreation" = "71",
                     "Accommodation and Food Services" = "72",
                     "Other Services (except Public Administration)" = "81",
                     "Public Administration" = "92")
```

1.3 Variable Selection

Selecting the variables of interest (see the Data Dictionary in Appendix A for details), renaming and mutating them appropriately, and filtering based on specific criteria.

```
# Wrangling the data
ACSSmall <- ACSSmallRaw %>%

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

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

  # Renaming the variables
  rename(sex = SEX,
         age = AGEP,
         citizenship = CIT,
         race = RAC1P,
         military = MIL,
         disabled = DIS,
         people_in_household = NP,
         marital_status = MAR,
         children = NRC,
         education = SCHL,
         degree_1 = FOD1P,
         degree_2 = FOD2P,
```

```
        stem_degree = SCIENGP,
        employment = ESR,
        hours_per_week = WKHP,
        industry = NAICSP,
        wage_income = WAGP,
        region = REGION.x,
        state = ST.x) %>%

  # Converting entries to the appropriate data types
  mutate(sex = as.factor(sex) %>%
           forcats::fct_recode(!!!sex_levels),

         employment = as.factor(employment) %>%
           forcats::fct_recode(!!!employment_levels),

         race = as.factor(race) %>%
           forcats::fct_recode(!!!race_levels),

         marital_status = as.factor(marital_status) %>%
           forcats::fct_recode(!!!marital_status_levels),

         region = as.factor(region) %>%
           forcats::fct_recode(!!!region_levels),

         state = as.factor(state) %>%
           forcats::fct_recode(!!!state_levels),

         # for citizenship, 5 stands for not a citizen
         citizenship = ifelse(citizenship == 5, "No", "Yes") %>%
           as.factor(),

         privilege = ifelse(race %in% c("White", "Asian"),
                            "Yes", "No") %>%
           as.factor(),

         # for military, 4 stands for never enlisted
         military = ifelse(military == 4, "No", "Yes"),
         military = ifelse(is.na(military), "No", military) %>%
           as.factor(),

         ever_married = ifelse(marital_status == "Single", "No", "Yes") %>%
           as.factor(),

         # for education, 22, 23, and 24 stand for graduate degrees
         grad_degree = ifelse(education %in% c(22, 23, 24), "Yes", "No"),
         grad_degree = ifelse(is.na(grad_degree), "No", grad_degree) %>%
           as.factor(),

         # for STEMdegree, 1 stands for a STEM degree
         stem_degree = ifelse(stem_degree == 1, "Yes", "No"),
         stem_degree = ifelse(is.na(stem_degree), "No", stem_degree) %>%
           as.factor(),
```

```r
        disabled = ifelse(disabled == 1, "Yes", "No") %>%
          as.factor(),

        # Filling empty industry fields with NAs
        industry = ifelse(industry == "", NA, industry),
        # Collapsing industry codes into broad NAICS sectors (only taking the first
        # two digits of the NAICS code, which represent the broader industry sectors)
        industry = substr(industry, start = 1, stop = 2) %>%
          as.factor() %>%
          fct_recode(!!!industry_levels),

        # Collapsing education codes into broader fields (only taking the first
        # two digits of each code, which represent the broader field)
        degree_1 = substr(degree_1, start = 1, stop = 2) %>%
          as.factor() %>%
          fct_recode(!!!degree_levels),
        degree_2 = substr(degree_2, start = 1, stop = 2) %>%
          as.factor() %>%
          fct_recode(!!!degree_levels),
        # Merging the two degree variables, taking care of NAs and repeated values
        degree = ifelse(is.na(degree_1) | is.na(degree_2),
                        as.character(degree_1),
                        ifelse(as.character(degree_1) == as.character(degree_2),
                               as.character(degree_1),
                               paste(degree_1, degree_2, sep = " and ")))) %>%

# Filtering the individuals of interest
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
       age >= 18) %>% # only people over 18

# Removing merged variables and those used for filtering or joining
select(-employment, -education, -degree_1, -degree_2)  %>%

# Dropping unused factor levels
mutate_if(is.factor, fct_drop)
```

2 Data Exploration

In this section, we will be exploring distributions and associations graphically and numerically, as a preparation for our model fitting and in order to better understand our data.

Let's take a look at the variables in our dataset.

```
names(ACSSmall) %>% cat(sep = "\n")
```

```
sex
age
citizenship
race
military
disabled
people_in_household
marital_status
children
stem_degree
hours_per_week
industry
wage_income
region
state
privilege
ever_married
grad_degree
degree
```

```
setequal(names(ACSBig), names(ACSSmall))
```

```
[1] TRUE
```

Variables to be used for graphical exploration only (because they have too many levels): state, industry, degree, region.

Variables to be used for our regression: wage_income (NUMERIC RESPONSE), sex, age (NUMERIC), citizenship, privilege, military, disabled, ever_married, children (NUMERIC), stem_degree, grad_degree, hours_per_week (NUMERIC), people_in_household (NUMERIC)

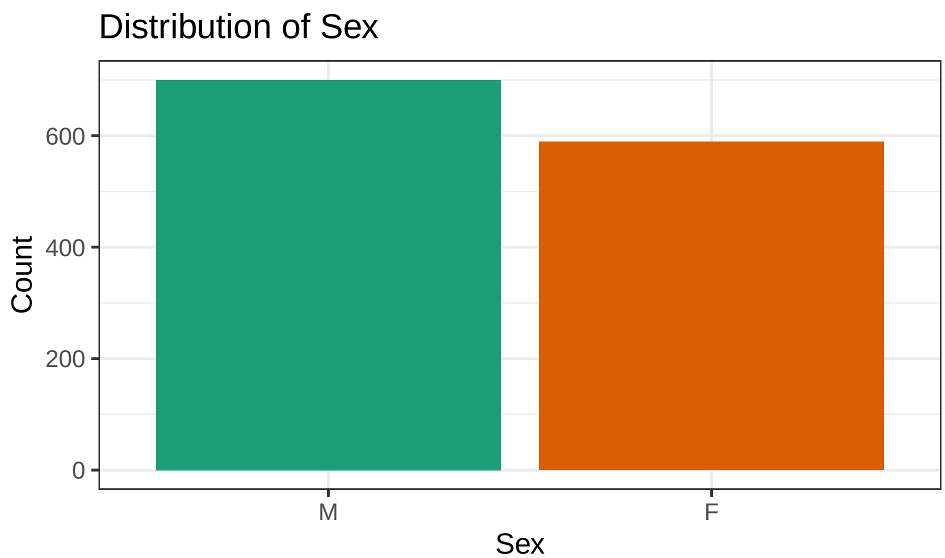
2.1 Univariate Data Exploration

In this section we explore univariate distributions of some of the most important demographic variables, especially those that we aim to include in the regression, to see whether we would benefit from any log transformations. The code is trivial and takes up unnecessary space, so it has been deferred to the appendix (see appendix B - Univariate Data Exploration).

2.1.1 Sex

Table 1: Distribution of Sex

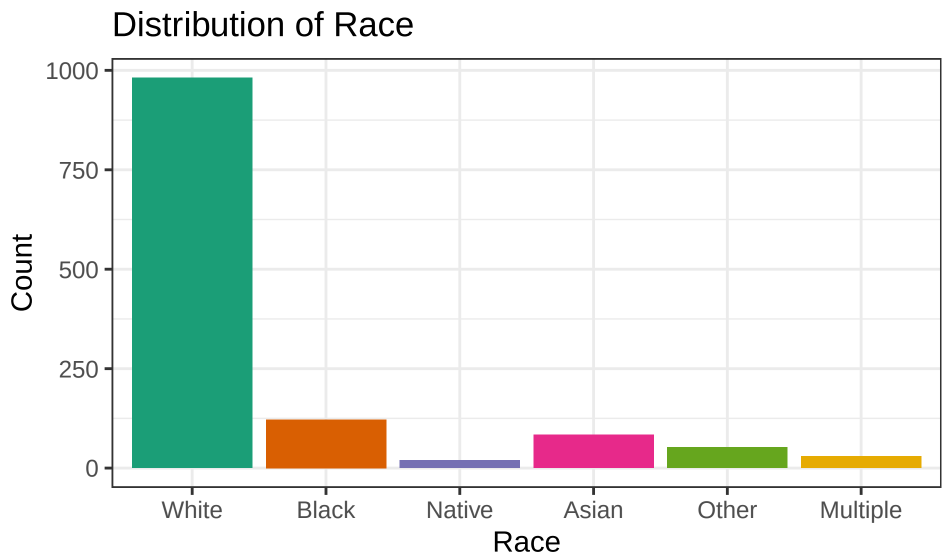
Sex	Tally
M	700
F	589



2.1.2 Race

Table 2: Distribution of Race

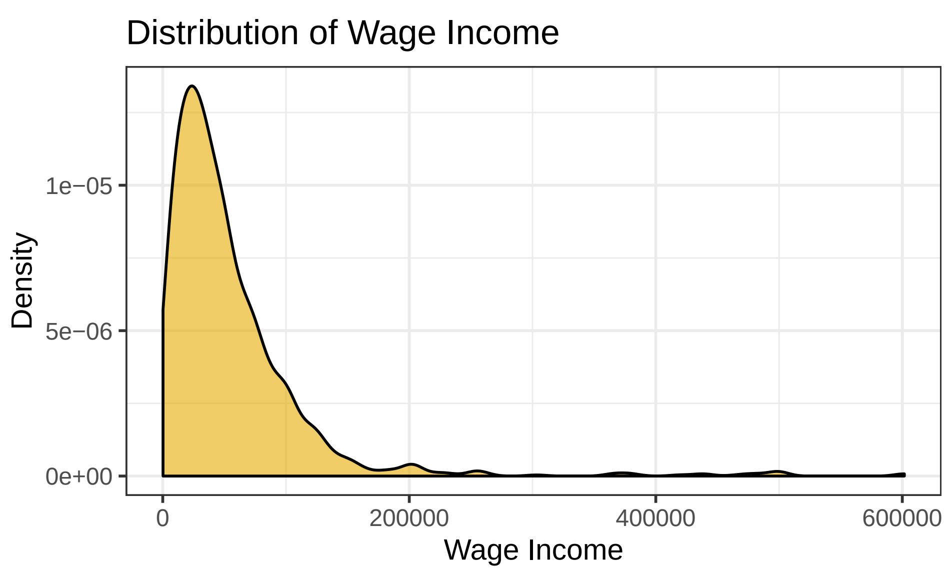
Race	Tally
White	981
Black	122
Native	20
Asian	84
Other	52
Multiple	30



2.1.3 Income

Table 3: Distribution of Wage Income

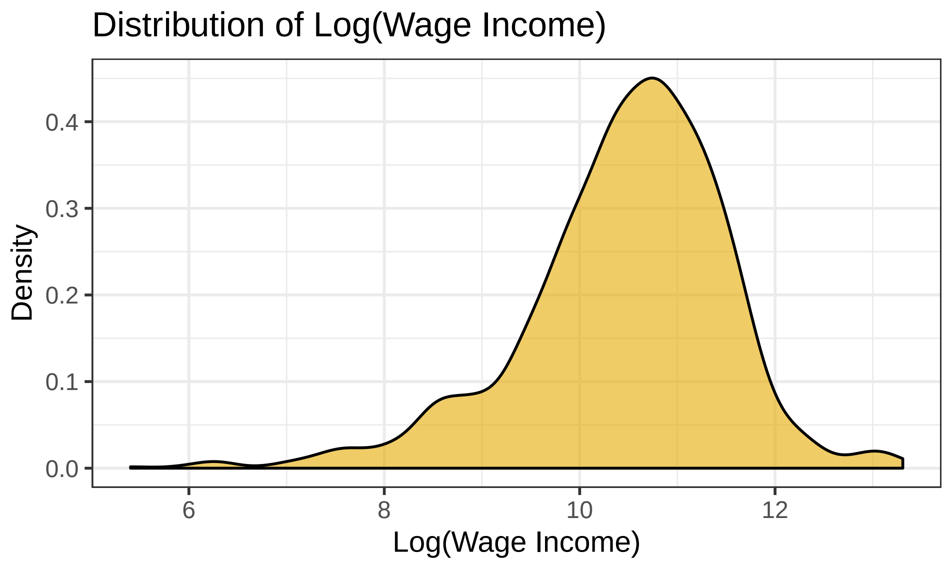
Min	Q1	Median	Q3	Max	Mean	SD	N	Missing
222	20224	40448	70783	601657	54908.96	62257.94	1289	0



We notice that the distribution of wage income is strongly skewed right, so it would be wise to consider a $\log_{10}$ transformation.

Table 4: Distribution of Log(Wage Income)

Min	Q1	Median	Q3	Max	Mean	SD	N	Missing
5.402677	9.914625	10.60777	11.16737	13.30744	10.46028	1.042913	1289	0

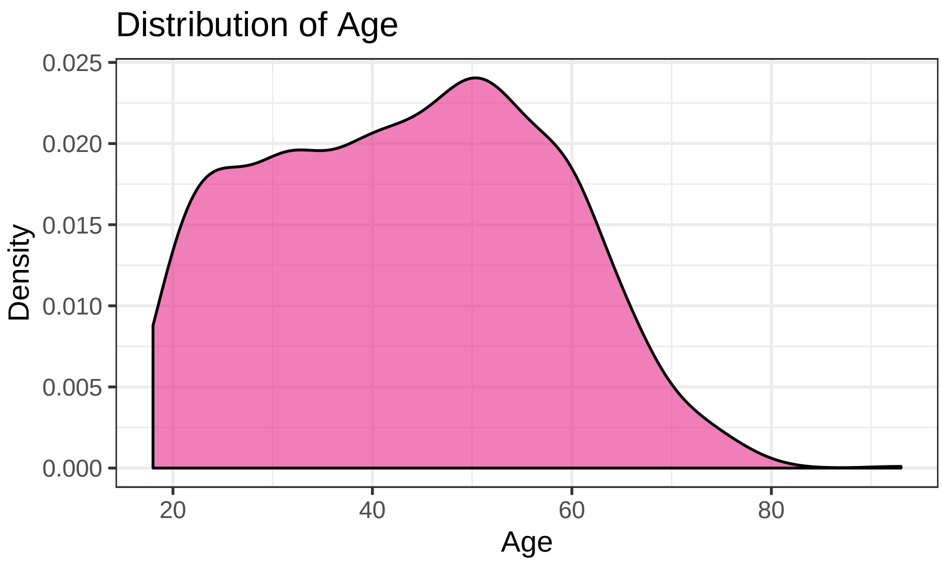


This looks much better, despite a slight left skew. We shall use `log_wage_income` as our response variable, partly because of skewness, but also because income is better considered on a multiplicative rather than additive scale. In other words, \$1,000 is worth a lot more to a poor person than a millionaire because \$1,000 is a much greater fraction of the poor person's wealth.

2.1.4 Age

Table 5: Distribution of Age

Min	Q1	Median	Q3	Max	Mean	SD	N	Missing
18	31	44	54	93	43.45617	14.28057	1289	0

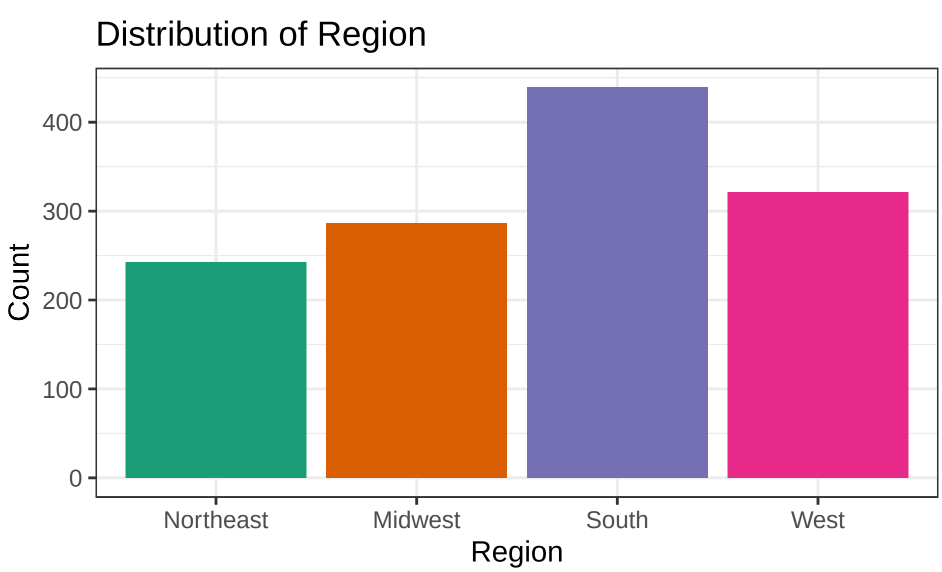


Understandably, the number of individuals in our sample starts to fall dramatically past the age of 60, since that is usually when the sweet release of death is upon us.

2.1.5 Region

Table 6: Distribution of Region

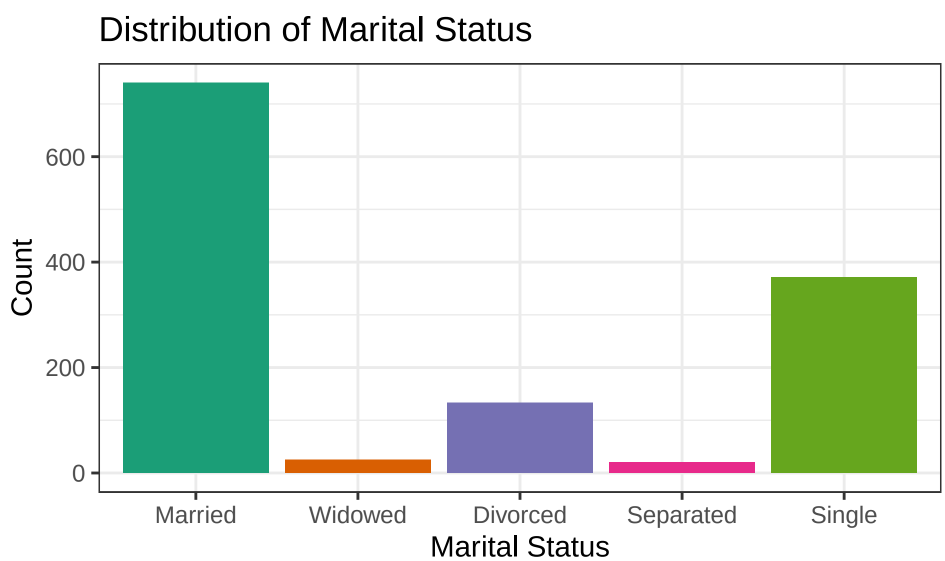
Region	Tally
Northeast	243
Midwest	286
South	439
West	321



2.1.6 Marital Status

Table 7: Distribution of Marital Status

Marital Status	Tally
Married	740
Widowed	25
Divorced	133
Separated	20
Single	371



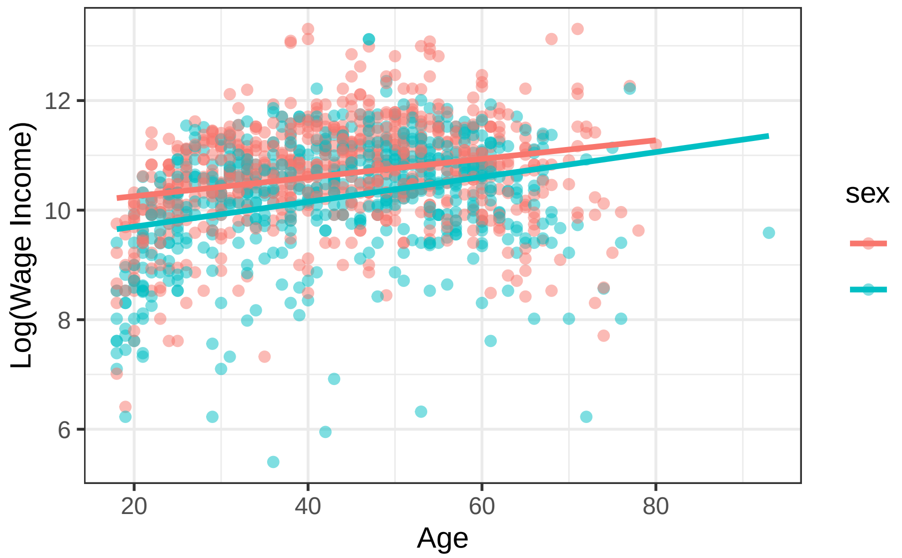
2.2 Multivariate Data Exploration

In this section we take a multivariate approach to exploring our data.

```
ACSSmall %>%
  group_by(privilege, sex) %>%
  summarize(average_wage = mean(wage_income))
```

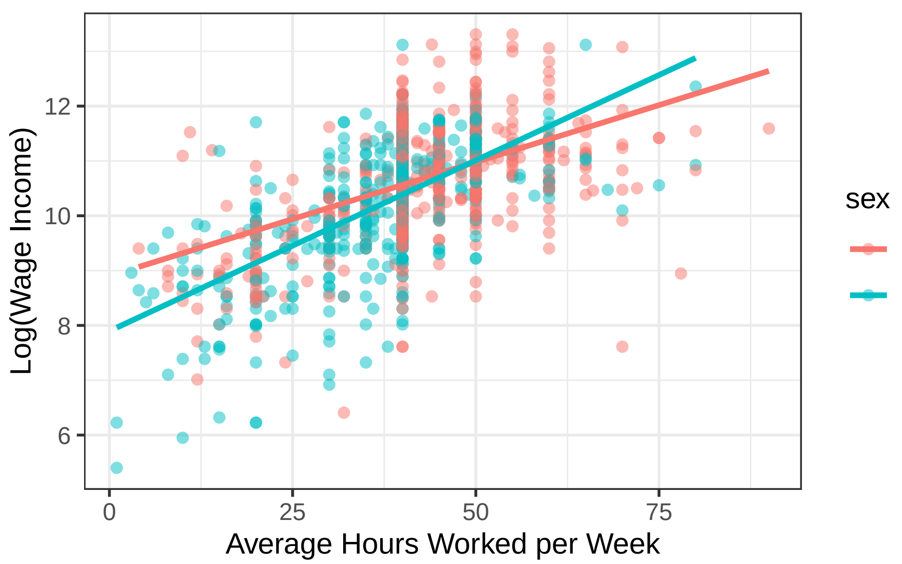
```
# A tibble: 4 x 3
# Groups:   privilege [2]
  privilege sex   average_wage
  <fct>     <fct>        <dbl>
1 No        M           44661.
2 No        F           35232.
3 Yes       M           68847.
4 Yes       F           44966.
```

```r
ACSSmall %>%
  ggplot(aes(x = age, y = log_wage_income, color = sex)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Age", y = "Log(Wage Income)")
```



It seems like men earn on average more than women of the same age, but the rate of change of log_wage_income as age increases is relatively the same for the two genders. We will later do a nested F test to see whether the rates of change are actually different, and whether we need two regression lines at all.

```r
ACSSmall %>%
  ggplot(aes(x = hours_per_week, y = log_wage_income, color = sex)) +
  geom_point(alpha = 0.5) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Average Hours Worked per Week", y = "Log(Wage Income)")
```



It seems like the rate of change of log_wage_income as age increases is higher for women than for men. We will later do a nested F test to see whether the rates of change are indeed different, and whether we need two regression lines at all.

```
ACSBig %>%
  group_by(industry, sex) %>%
  summarize(average_income = mean(wage_income)) %>%
  spread(sex, average_income) %>%
  rename(male_income = M, female_income = "F") %>%
  mutate(difference = male_income - female_income) %>%
  ggplot(aes(x = female_income, y = male_income, color = industry, size = difference)) +
    geom_point() +
    geom_abline(slope = 1, intercept = 0, alpha = 0.5) +
    geom_text_repel(aes(label = industry), size = 4,
                    segment.color = "transparent") +
    expand_limits(x = 0, y = 0) +
    labs(x = "Average Female Income", y = "Average Male Income") +
    ggtitle("Average Income Difference between Men and Women, by Industry") +
    theme(legend.position = "none")
```

3 Modeling

Having completed our variable exploration, we suspect that there is a significant association between wage income and gender. Now let us proceed by fitting a regression model, in order to find out which variables help us account for the variability in income, and more specifically whether gender is indeed a significant predictor thereof.

3.1 Choosing Predictors

It is time to select predictors for our model. We shall use the best subsets procedure in order to find a model that accounts for the most variability in income, while also keeping matters simple.

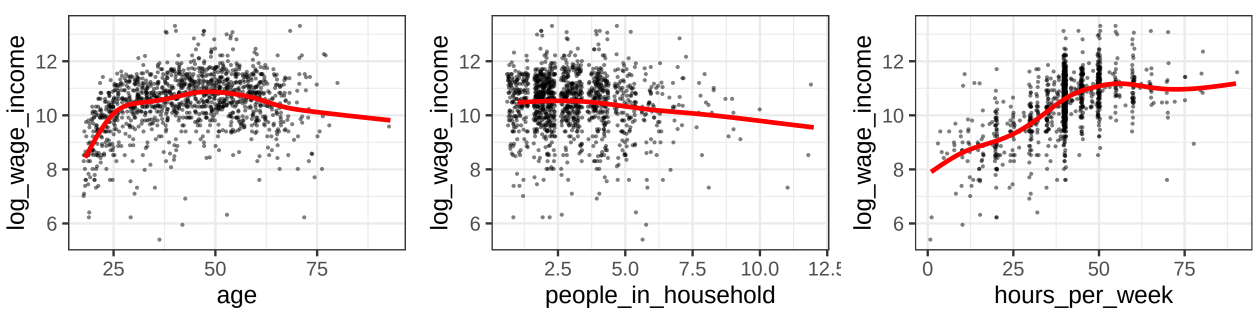
```
# Creating a dataset to be used for our predictor selection procedure
ACSReg1 <- ACSSmall %>%
  # selecting the potential predictors...
  select(sex, age, citizenship, privilege, military, disabled, ever_married, children,
         stem_degree, people_in_household, hours_per_week, grad_degree,
         # ... and the response
         log_wage_income)
```

```
# Using best subsets method
output <- leaps::regsubsets(log_wage_income ~ ., nbest = 1, data = ACSReg1)

with(summary(output), data.frame(adjr2, cp, bic, outmat)) %>%
  arrange(desc(adjr2)) %>%
  rename(R2_adj = adjr2,
         Cp = cp,
         BIC = bic) %>%
  mutate(R2_adj = round(R2_adj * 100, 2),
         Cp = round(Cp, 2),
         BIC = round(BIC, 2)) %>%
  t() %>%
  as.data.frame() %>%
  kable(booktabs = TRUE, col.names = c("Model 1", "Model 2", "Model 3",
                                       "Model 4", "Model 5", "Model 6",
                                       "Model 7", "Model 8")) %>%
  kable_styling(latex_options = c("striped", "HOLD_position"), font_size = 9)
```

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
R2_adj	42.67	42.45	42.19	41.50	40.75	39.46	35.68	30.71
Cp	16.64	20.59	25.39	39.43	55.04	82.57	165.02	274.00
BIC	-645.40	-646.60	-646.98	-638.33	-628.34	-607.29	-537.11	-449.36
sexF	*	*	*	*				
age	*	*	*	*	*	*	*	
citizenshipYes								
privilegeYes	*	*	*					
militaryYes								
disabledYes	*	*						
ever_marriedYes	*							
children								
stem_degreeYes	*	*	*	*	*			
people_in_household								
hours_per_week	*	*	*	*	*	*	*	*
grad_degreeYes	*	*	*	*	*	*		

Our adjusted R squared is not bad, at 42.6%. Mallows' Cp is a bit concerning, at 16.6 lowest. Let's see if any of our quantitative variables might benefit from the usage of a polynomial model.

```
p1 <- ggplot(ACSReg1, aes(x = age, y = log_wage_income)) +
  geom_jitter(size = 0.2, alpha = 0.5) +
  geom_smooth(se = FALSE, color = "red")

p2 <- ggplot(ACSReg1, aes(x = people_in_household, y = log_wage_income)) +
  geom_jitter(size = 0.2, alpha = 0.5) +
  geom_smooth(se = FALSE, color = "red")

p3 <- ggplot(ACSReg1, aes(x = hours_per_week, y = log_wage_income)) +
  geom_jitter(size = 0.2, alpha = 0.5) +
  geom_smooth(se = FALSE, color = "red")

cowplot::plot_grid(p1, p2, p3, nrow = 1)
```



The relationship between `log(income)` and `age` is visibly curved, so we might want to consider a quadratic model.

```
cor(ACSReg1 %>% select(log_wage_income, children, people_in_household, age,
                       hours_per_week), use = "complete.obs") %>%
  kable(digits = 3, booktabs = TRUE, caption = "Correlation Matrix for Numeric Variables") %>%
  kable_styling(latex_options = c("striped", "HOLD_position"), font_size = 9)
```

Table 8: Correlation Matrix for Numeric Variables

	log_wage_income	children	people_in_household	age	hours_per_week
log_wage_income	1.000	-0.040	-0.144	0.238	0.555
children	-0.040	1.000	0.744	-0.198	-0.035
people_in_household	-0.144	0.744	1.000	-0.237	-0.107
age	0.238	-0.198	-0.237	1.000	0.025
hours_per_week	0.555	-0.035	-0.107	0.025	1.000

The correlation matrix for our quantitative predictors also shows that `people_in_household` and `children` are highly correlated (0.744), so we might as well remove one of them to avoid multicollinearity in our final model. We remove `children` since it is less correlated with our response, `log_wage_income` (-0.040), than `people_in_household` (-0.144).

```
ACSReg2 <- ACSReg1 %>%
  # Adding in squared age as a predictor
  mutate(ageSq = age^2) %>%
  # Removing highly correlated predictor
  select(-children)
```

Now let's run the best subsets procedure again:

```
output <- leaps::regsubsets(log_wage_income ~ ., nbest = 1, data = ACSReg2)

with(summary(output), data.frame(adjr2, cp, bic, outmat)) %>%
  arrange(desc(adjr2)) %>%
  rename(R2_adj = adjr2,
         Cp = cp,
         BIC = bic) %>%
  mutate(R2_adj = round(R2_adj * 100, 2),
         Cp = round(Cp, 2),
         BIC = round(BIC, 2)) %>%
  t() %>%
  as.data.frame() %>%
  kable(booktabs = TRUE, col.names = c("Model 1", "Model 2", "Model 3",
                                       "Model 4", "Model 5", "Model 6",
                                       "Model 7", "Model 8")) %>%
  kable_styling(latex_options = c("striped", "HOLD_position"), font_size = 9)
```

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
R2_adj	49.04	48.76	47.96	47.04	45.90	42.51	37.81	32.31
Cp	10.62	16.53	35.79	58.08	85.85	170.40	288.22	426.61
BIC	-812.46	-811.69	-797.78	-781.27	-760.01	-687.91	-592.83	-489.68
sexF	*	*	*					
age	*	*	*	*	*	*	*	
citizenshipYes	*							
privilegeYes	*	*						
militaryYes								
disabledYes								
ever_marriedYes								
stem_degreeYes	*	*	*	*				
people_in_household								
hours_per_week	*	*	*	*	*	*	*	*
grad_degreeYes	*	*	*	*	*			
ageSq	*	*	*	*	*	*		

The adjusted R squared for the best model has increased from 42.7% to 49%, and Mallows' Cp has decreased to 10.6. The BIC has also decreased significantly from -645 to -812. This looks promising! Let's fit the corresponding model now.

```
initialModel <- lm(log_wage_income ~ sex + age + ageSq + citizenship + privilege +
              stem_degree + hours_per_week + grad_degree, data = ACSReg2)
msummary(initialModel)
```

```
                  Estimate  Std. Error t value Pr(>|t|)    
(Intercept)     5.83127098  0.19963808  29.209  < 2e-16 ***
sexF           -0.21633367  0.04273615  -5.062 4.75e-07 ***
age             0.11171849  0.00888981  12.567  < 2e-16 ***
ageSq          -0.00109013  0.00009944 -10.962  < 2e-16 ***
citizenshipYes  0.23469629  0.08348647   2.811  0.00501 ** 
privilegeYes    0.23062495  0.05552735   4.153 3.49e-05 ***
stem_degreeYes  0.34677111  0.06754743   5.134 3.28e-07 ***
hours_per_week  0.04091504  0.00197064  20.762  < 2e-16 ***
grad_degreeYes  0.41218860  0.06428703   6.412 2.02e-10 ***

Residual standard error: 0.7445 on 1280 degrees of freedom
Multiple R-squared:  0.4935,    Adjusted R-squared:  0.4904 
F-statistic: 155.9 on 8 and 1280 DF,  p-value: < 2.2e-16
```

This looks good! All of our predictors are significant individually (see p-values for the t-statistics), as is our model as a whole (see p-value for the F-statistic). Let's now play around now using manual forward selection and see if any interaction terms might be of use. We will add the interaction between `sex` and `hours_per_week` since it adds the biggest increase in adjusted R squared.

```
model <- lm(log_wage_income ~ sex * hours_per_week + age + ageSq + citizenship +
              privilege + stem_degree + grad_degree, data = ACSReg2)
msummary(model)
```

```
                          Estimate  Std. Error t value Pr(>|t|)
```

```
(Intercept)          6.13847740  0.21082322  29.117  < 2e-16 ***
sexF                -0.85632199  0.15507447  -5.522 4.05e-08 ***
hours_per_week       0.03369606  0.00258105  13.055  < 2e-16 ***
age                  0.11175199  0.00882996  12.656  < 2e-16 ***
ageSq               -0.00109449  0.00009878 -11.080  < 2e-16 ***
citizenshipYes       0.24002195  0.08293367   2.894  0.00387 **
privilegeYes         0.23423660  0.05515993   4.246 2.33e-05 ***
stem_degreeYes       0.33655526  0.06713488   5.013 6.11e-07 ***
grad_degreeYes       0.40182656  0.06389985   6.288 4.39e-10 ***
sexF:hours_per_week  0.01616058  0.00376628   4.291 1.91e-05 ***

Residual standard error: 0.7395 on 1279 degrees of freedom
Multiple R-squared:  0.5007,	Adjusted R-squared:  0.4972
F-statistic: 142.5 on 9 and 1279 DF,  p-value: < 2.2e-16
```

This model looks good! We have 7 predictors. We could cut some for the interest of simplicity without too much of an impact on the adjusted R squared, but 7 predictors is not excessive so we shall leave all of them in.

3.2 Fitting Model

We now fit our final model.

```r
bestModel <- lm(log_wage_income ~ sex * hours_per_week + age + ageSq + citizenship +
                  privilege + stem_degree + grad_degree, data = ACSReg2)
msummary(bestModel)
```


```
                       Estimate  Std. Error t value Pr(>|t|)
(Intercept)          6.13847740  0.21082322  29.117  < 2e-16 ***
sexF                -0.85632199  0.15507447  -5.522 4.05e-08 ***
hours_per_week       0.03369606  0.00258105  13.055  < 2e-16 ***
age                  0.11175199  0.00882996  12.656  < 2e-16 ***
ageSq               -0.00109449  0.00009878 -11.080  < 2e-16 ***
citizenshipYes       0.24002195  0.08293367   2.894  0.00387 **
privilegeYes         0.23423660  0.05515993   4.246 2.33e-05 ***
stem_degreeYes       0.33655526  0.06713488   5.013 6.11e-07 ***
grad_degreeYes       0.40182656  0.06389985   6.288 4.39e-10 ***
sexF:hours_per_week  0.01616058  0.00376628   4.291 1.91e-05 ***

Residual standard error: 0.7395 on 1279 degrees of freedom
Multiple R-squared:  0.5007,	Adjusted R-squared:  0.4972
F-statistic: 142.5 on 9 and 1279 DF,  p-value: < 2.2e-16
```


3.3 Confidence intervals

```r
(100 * (confint(bestModel) %>% exp() - 1))
```

```
                             2.5 %          97.5 %
```

```
(Intercept)         30539.6520381 69969.64127181
sexF                  -68.6686270   -42.42568958
hours_per_week          2.9046354     3.95205787
age                     9.9031331    13.77752167
ageSq                  -0.1287443    -0.09002982
citizenshipYes          8.0388802    49.58923620
privilegeYes           13.4307947    40.83945980
stem_degreeYes         22.7339330    59.72155732
grad_degreeYes         31.8462807    69.41592614
sexF:hours_per_week     0.8810390     2.38288218
```

(exp(0.24002195) - 1) * 100 (exp(0.23423660) - 1) * 100 etc.

4 Analysis

4.1 Interpretation

Let's interpret our model. Each of the coefficients are significant, so we interpret.

Since our response is log-transformed, we will give our interpretation in terms of percentage change.

Sex: coefficient of -0.85632199 so if you're a woman (adjusting for the other predictors) that is associated with a $(e^{-0.85632199} - 1) \cdot 100 = -57.52787\%$ lower wages than for men.

log(income for man) = sth log(income for woman) = sth - 0.856 log(income for woman) - log(income for man) = -0.856 log(income for woman / income for man) = -0.856 income for woman / income for man = e^(-0.856) so income for woman = income for man * e^(-0.856) = income for man * 0.4247213

For a man, every additional 5 hours worked per week are associated with $(e^{5 \cdot 0.03369606} - 1) \cdot 100 = 18.35049\%$ higher wages. But for women, every additional 5 hours worked per week are only associated with $(e^{5 \cdot (0.03369606 - 0.01616058)} - 1) \cdot 100 = 28.31054\%$ higher wages! Interesting.

4.2 Nested F-test

Now let's come back to our question of interest - does sex matter in predicting wage?

```r
bestModel <- lm(log_wage_income ~ sex * hours_per_week + age + ageSq + citizenship +
                  privilege + stem_degree + grad_degree, data = ACSReg2)
reducedModel <- lm(log_wage_income ~ age + ageSq + hours_per_week + citizenship +
                     privilege + stem_degree + grad_degree, data = ACSReg2)
anova(reducedModel, bestModel)
```

```
Analysis of Variance Table

Model 1: log_wage_income ~ age + ageSq + hours_per_week + citizenship +
    privilege + stem_degree + grad_degree
Model 2: log_wage_income ~ sex * hours_per_week + age + ageSq + citizenship +
    privilege + stem_degree + grad_degree
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1   1281 723.71
2   1279 699.44  2    24.272 22.192 3.353e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
reducedModel <- lm(log_wage_income ~ age + ageSq + sex + hours_per_week + citizenship +
                     privilege + stem_degree + grad_degree, data = ACSReg2)
anova(reducedModel, bestModel)
```

```
Analysis of Variance Table

Model 1: log_wage_income ~ age + ageSq + sex + hours_per_week + citizenship +
    privilege + stem_degree + grad_degree
Model 2: log_wage_income ~ sex * hours_per_week + age + ageSq + citizenship +
```

```
    privilege + stem_degree + grad_degree
  Res.Df    RSS Df Sum of Sq      F     Pr(>F)
1   1280 709.51
2   1279 699.44  1    10.069 18.411 0.00001914 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
fullModel <- lm(log_wage_income ~ sex * hours_per_week + sex * age + age + ageSq +
                  citizenship + privilege + stem_degree + grad_degree,
                data = ACSReg2)
reducedModel <- lm(log_wage_income ~ sex * hours_per_week + age + ageSq + citizenship +
                  privilege + stem_degree + grad_degree, data = ACSReg2)
anova(reducedModel, fullModel)
```

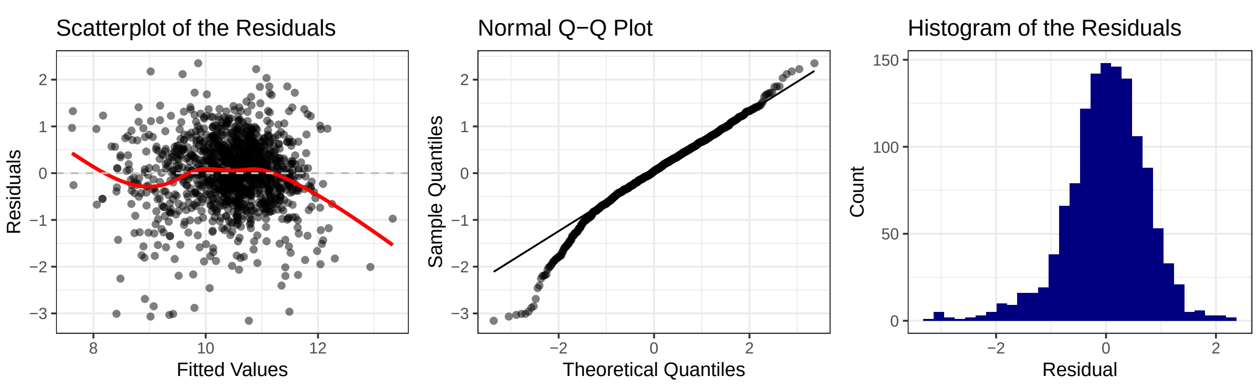
```
Analysis of Variance Table

Model 1: log_wage_income ~ sex * hours_per_week + age + ageSq + citizenship +
    privilege + stem_degree + grad_degree
Model 2: log_wage_income ~ sex * hours_per_week + sex * age + age + ageSq +
    citizenship + privilege + stem_degree + grad_degree
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1   1279 699.44
2   1278 699.42  1  0.017468 0.0319 0.8582
```

5 Assessment

5.1 Conditions

Diagnostic plots to check conditions:

```
diagnostic <- function(model, binwidth) {

  # Scatterplot of the residuals
  r1 <- ggplot(data = model, aes(x = .fitted, y = .resid)) +
    geom_point(alpha = 0.5) +
    geom_smooth(color = "red", se = FALSE) +
    geom_hline(yintercept = 0, linetype = 2, color = "grey") +
    xlab("Fitted Values") +
    ylab("Residuals") +
    ggtitle("Scatterplot of the Residuals")

  # Normal QQ-plot
  r2 <- ggplot(data = model, aes(sample = .resid)) +
    geom_qq(alpha = 0.5) +
    geom_qq_line() +
    xlab("Theoretical Quantiles") +
    ylab("Sample Quantiles") +
    ggtitle("Normal Q-Q Plot")

  # Histogram of the residuals
  r3 <- ggplot(data = model, aes(x = .resid, y = ..count..)) +
    geom_histogram(fill = "navy", binwidth = binwidth) +
    xlab("Residual") +
    ylab("Count") +
    ggtitle("Histogram of the Residuals")

  # Plotting all three side by side
  cowplot::plot_grid(plotlist = list(r1, r2, r3), nrow = 1)
}

diagnostic(bestModel, binwidth = NULL)
```



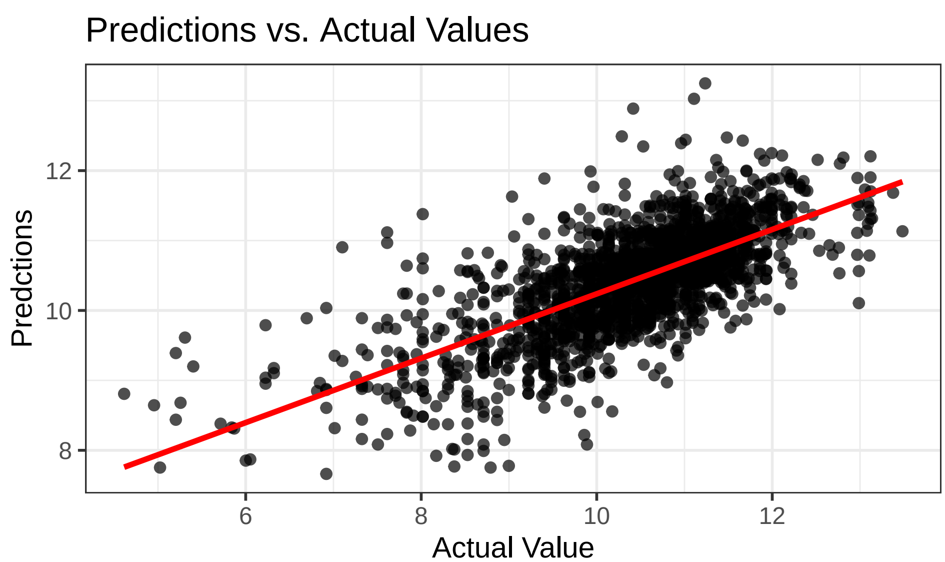
Nice model - we like !!

5.2 Cross-validation Correlation

```
set.seed(1)

# Creating test dataset
ACSTest <- sample(ACSBig %>% mutate(ageSq = age^2), size = 2000)

# Computing predictions
ACSTestPred <- ACSTest %>%
  mutate(prediction = predict(bestModel, newdata = .)) %>%
  select(log_wage_income, prediction)

# Finding cross validation correlation
cross_val_cor <- cor(ACSTestPred$log_wage_income, ACSTestPred$prediction)
cross_val_cor
```

```
[1] 0.7002889
```

```
# Plotting actual values against predictions
ggplot(ACSTestPred, aes(x = log_wage_income, y = prediction)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "lm", se = FALSE, color = "red") +
  labs(x = "Actual Value", y = "Predctions") +
  ggtitle("Predictions vs. Actual Values")
```




```
r_squared <- 0.5007
shrinkage <- r_squared - cross_val_cor^2
shrinkage
```

```
[1] 0.01029542
```

The cross validation correlation is not bad - over 70%. Our model is accurate less than half the time This means that our predictive model is expected to perform decently in practice (i.e. when applied to data that was not used in our estimations).

5.3 Logistic Regression

```
ACSLogit <- ACSSmall %>%
  # selecting the potential predictors...
  select(sex, age, citizenship, privilege, military, disabled, ever_married, children,
         stem_degree, people_in_household, hours_per_week, grad_degree,
         # ... and the response
         wage_income) %>%
  # Adding in squared age as a predictor
  mutate(ageSq = age^2) %>%
  # Dichotomizing outcome
  mutate(wealthy = ifelse(wage_income >= 120000, 1, 0) %>%
           as.factor()) %>%
  select(-wage_income)

tally(ACSLogit$wealthy ~ ACSLogit$sex)
```

```
                ACSLogit$sex
ACSLogit$wealthy   M   F
               0 626 563
               1  74  26
```

```
model1 <- glm(wealthy ~ ., data = ACSLogit, family = binomial(logit),
              na.action = na.exclude)
msummary(model1)
```

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -17.051128   2.621667  -6.504 7.83e-11 ***
sexF                 -0.855970   0.267794  -3.196 0.001392 **
age                   0.420235   0.102372   4.105 4.04e-05 ***
citizenshipYes       -0.553136   0.432161  -1.280 0.200570
privilegeYes          0.927774   0.424866   2.184 0.028985 *
militaryYes           0.176460   0.384410   0.459 0.646204
disabledYes          -0.916610   0.752712  -1.218 0.223321
ever_marriedYes       0.598767   0.424717   1.410 0.158599
children              0.228686   0.177027   1.292 0.196421
stem_degreeYes        1.291614   0.276190   4.677 2.92e-06 ***
people_in_household   0.004045   0.132633   0.031 0.975667
hours_per_week        0.055978   0.011864   4.718 2.38e-06 ***
grad_degreeYes        1.181822   0.267817   4.413 1.02e-05 ***
ageSq                -0.003838   0.001026  -3.742 0.000183 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 698.75  on 1260  degrees of freedom
Residual deviance: 511.50  on 1247  degrees of freedom
  (28 observations deleted due to missingness)
AIC: 539.5

Number of Fisher Scoring iterations: 7
```

Let's remove what seems most insignificant - citizenship and ever_married. People in household and children are highly correlated so we'll remove one.

```
ACSLogit2 <- ACSLogit %>%
  select(-ever_married, - citizenship)

model2 <- glm(wealthy ~ ., data = ACSLogit2, family = binomial(logit),
              na.action = na.exclude)
msummary(model2)
```

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)         -17.459310   2.565984  -6.804 1.02e-11 ***
sexF                 -0.871464   0.266793  -3.266  0.00109 **
age                   0.433987   0.099857   4.346 1.39e-05 ***
privilegeYes          0.846854   0.412009   2.055  0.03984 *
militaryYes           0.141978   0.382287   0.371  0.71035
disabledYes          -0.926411   0.751036  -1.234  0.21738
children              0.204237   0.172798   1.182  0.23723
stem_degreeYes        1.305093   0.275463   4.738 2.16e-06 ***
people_in_household   0.052264   0.125506   0.416  0.67710
hours_per_week        0.055152   0.011763   4.689 2.75e-06 ***
grad_degreeYes        1.203622   0.267058   4.507 6.58e-06 ***
ageSq                -0.003944   0.001007  -3.916 9.02e-05 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 698.75  on 1260  degrees of freedom
Residual deviance: 515.14  on 1249  degrees of freedom
  (28 observations deleted due to missingness)
AIC: 539.14

Number of Fisher Scoring iterations: 7
```

```
model2a <- glm(wealthy ~ .-children, data = ACSLogit2, family = binomial(logit),
              na.action = na.exclude)

model2b <- glm(wealthy ~ .-people_in_household, data = ACSLogit2, family = binomial(logit),
              na.action = na.exclude)
msummary(model2a)
```

```
Coefficients:
                       Estimate  Std. Error z value Pr(>|z|)
(Intercept)         -17.7811173   2.5281560  -7.033 2.02e-12 ***
sexF                 -0.8700505   0.2668513  -3.260  0.00111 **
age                   0.4405309   0.0987737   4.460 8.20e-06 ***
privilegeYes          0.8603876   0.4109600   2.094  0.03630 *
militaryYes           0.1331174   0.3787428   0.351  0.72523
disabledYes          -0.9481486   0.7505415  -1.263  0.20649
stem_degreeYes        1.3192083   0.2753338   4.791 1.66e-06 ***
people_in_household   0.1673562   0.0742255   2.255  0.02415 *
hours_per_week        0.0559230   0.0117809   4.747 2.07e-06 ***
```

```
grad_degreeYes        1.2442603   0.2649010   4.697 2.64e-06 ***
ageSq                -0.0040456   0.0009971  -4.057 4.96e-05 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 698.75  on 1260  degrees of freedom
Residual deviance: 516.59  on 1250  degrees of freedom
  (28 observations deleted due to missingness)
AIC: 538.59

Number of Fisher Scoring iterations: 7
```

```
msummary(model2b)
```

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -17.330659   2.551685  -6.792 1.11e-11 ***
sexF            -0.875860   0.266524  -3.286  0.00102 **
age              0.434683   0.100038   4.345 1.39e-05 ***
privilegeYes     0.842412   0.412113   2.044  0.04094 *
militaryYes      0.144956   0.382224   0.379  0.70451
disabledYes     -0.931671   0.751385  -1.240  0.21500
children         0.261894   0.103863   2.522  0.01168 *
stem_degreeYes   1.296392   0.274425   4.724 2.31e-06 ***
hours_per_week   0.054564   0.011656   4.681 2.85e-06 ***
grad_degreeYes   1.192694   0.265523   4.492 7.06e-06 ***
ageSq           -0.003949   0.001009  -3.914 9.07e-05 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 698.75  on 1260  degrees of freedom
Residual deviance: 515.31  on 1250  degrees of freedom
  (28 observations deleted due to missingness)
AIC: 537.31

Number of Fisher Scoring iterations: 7
```

We remove people in household, as well as disabled and military.

```
ACSLogit3 <- ACSLogit2 %>%
  select(-people_in_household, -disabled, -military)

model3 <- glm(wealthy ~ ., data = ACSLogit3, family = binomial(logit),
              na.action = na.exclude)
msummary(model3)
```

```
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -17.448186   2.563235  -6.807 9.96e-12 ***
sexF            -0.892400   0.262193  -3.404 0.000665 ***
age              0.437004   0.100443   4.351 1.36e-05 ***
privilegeYes     0.863990   0.412296   2.096 0.036122 *
```

```
children          0.268985   0.103254   2.605 0.009185 **
stem_degreeYes    1.280143   0.273442   4.682 2.85e-06 ***
hours_per_week    0.055034   0.011450   4.807 1.54e-06 ***
grad_degreeYes    1.207763   0.265588   4.548 5.43e-06 ***
ageSq            -0.003972   0.001012  -3.924 8.70e-05 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 698.75  on 1260  degrees of freedom
Residual deviance: 517.36  on 1252  degrees of freedom
  (28 observations deleted due to missingness)
AIC: 535.36

Number of Fisher Scoring iterations: 7
```

Try to remove some more?

```
ACSLogit4 <- ACSLogit3 %>%
  select(-privilege, -children)

model4 <- glm(wealthy ~ ., data = ACSLogit4, family = binomial(logit),
              na.action = na.exclude)
msummary(model4)
```

```
Coefficients:
                 Estimate  Std. Error z value Pr(>|z|)
(Intercept)    -16.5299044   2.4397345  -6.775 1.24e-11 ***
sexF            -0.9293942   0.2607856  -3.564 0.000365 ***
age              0.4542464   0.0975530   4.656 3.22e-06 ***
stem_degreeYes   1.2269853   0.2715363   4.519 6.22e-06 ***
hours_per_week   0.0531789   0.0112440   4.730 2.25e-06 ***
grad_degreeYes   1.2959211   0.2641273   4.906 9.28e-07 ***
ageSq           -0.0042548   0.0009853  -4.318 1.57e-05 ***

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 703.32  on 1288  degrees of freedom
Residual deviance: 530.09  on 1282  degrees of freedom
AIC: 544.09

Number of Fisher Scoring iterations: 7
```

No no, AIC grows. model3 looked good. Interactions? Trying some, no.

```
model5 <- glm(wealthy ~ . + hours_per_week * stem_degree, data = ACSLogit3, family = binomial(logit),
              na.action = na.exclude)
msummary(model5)
```

```
Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -17.117756   2.586012  -6.619 3.61e-11 ***
```

```
sexF                             -0.925743   0.265403  -3.488 0.000487 ***
age                               0.438046   0.100895   4.342 1.41e-05 ***
privilegeYes                      0.890442   0.415740   2.142 0.032208 *
children                          0.266621   0.103200   2.584 0.009779 **
stem_degreeYes                   -0.100023   1.253531  -0.080 0.936402
hours_per_week                    0.046595   0.013732   3.393 0.000691 ***
grad_degreeYes                    1.225565   0.266845   4.593 4.37e-06 ***
ageSq                            -0.003976   0.001016  -3.913 9.11e-05 ***
stem_degreeYes:hours_per_week     0.030072   0.026604   1.130 0.258320

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 698.75  on 1260  degrees of freedom
Residual deviance: 516.03  on 1251  degrees of freedom
  (28 observations deleted due to missingness)
AIC: 536.03

Number of Fisher Scoring iterations: 7
```

```r
# Amy Wagaman's function to calculate coconcordance
getConcordance <- function(model){
  Con_Dis_Data <- cbind(model$y, model$fitted.values)
  ones <- Con_Dis_Data[Con_Dis_Data[,1] == 1,]
  zeros <- Con_Dis_Data[Con_Dis_Data[,1] == 0,]
  conc <- matrix(0, dim(zeros)[1], dim(ones)[1])
  disc <- matrix(0, dim(zeros)[1], dim(ones)[1])
  ties <- matrix(0, dim(zeros)[1], dim(ones)[1])
  for(j in 1:dim(zeros)[1]){
    for(i in 1:dim(ones)[1]){
      if(ones[i,2]>zeros[j,2]){
        conc[j,i]=1
      }else if(ones[i,2]<zeros[j,2]){
          disc[j,i]=1
      }else if(ones[i,2]==zeros[j,2]){
        ties[j,i]=1
      }
    }
  }
  Pairs <- dim(zeros)[1]*dim(ones)[1]
  PercentConcordance <- (sum(conc)/Pairs)*100
  PercentDiscordance <- (sum(disc)/Pairs)*100
  PercentTied <- (sum(ties)/Pairs)*100
  return(list("Percent Concordance" = PercentConcordance,
              "Percent Discordance" = PercentDiscordance,
              "Percent Tied" = PercentTied,
              "Pairs" = Pairs))
}
```

```r
getConcordance(model3)
```

```
$`Percent Concordance`
[1] 86.10508
```

```
$`Percent Discordance`
[1] 13.86994

$`Percent Tied`
[1] 0.02497847

$Pairs
[1] 116100
```

```
ACSLogit3 <- mutate(ACSLogit3, predProb = predict(model, type = "response"),
                    predComplete = ifelse(predProb > 0.9, 1, 0))
# cross-tally predicted successes with actual success
tally(data = ACSLogit3, wealthy ~ predComplete, margins = TRUE)
```

```
       predComplete
wealthy    1
  0     1189
  1      100
  Total 1289
```

ADD EMPIRICAL LOGIT PLOTS

Testing

```
library(lmtest)
lrtest(model3)
```

```
Likelihood ratio test

Model 1: wealthy ~ sex + age + privilege + children + stem_degree + hours_per_week +
    grad_degree + ageSq
Model 2: wealthy ~ 1
  #Df  LogLik Df  Chisq Pr(>Chisq)
1   9 -258.68
2   1 -349.37 -8 181.39  < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model3_nosex <- glm(wealthy ~ age + privilege + children + stem_degree +
                      hours_per_week + grad_degree + ageSq,
                    data = ACSLogit3, family = binomial(logit),
                    na.action = na.exclude)
lrtest(model3_nosex, model3)
```

```
Likelihood ratio test

Model 1: wealthy ~ age + privilege + children + stem_degree + hours_per_week +
    grad_degree + ageSq
Model 2: wealthy ~ sex + age + privilege + children + stem_degree + hours_per_week +
    grad_degree + ageSq
  #Df  LogLik Df  Chisq Pr(>Chisq)
```

```
1   8 -264.94
2   9 -258.68  1 12.525  0.0004016 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6 Appendices

6.1 Appendix A - Data Dictionary

A number of variables have been identified as potentially relevant to the issue of the gender pay gap. Optimally, a careful consideration of each of them might provide us with a more precise understanding of the relationship between gender and income. However, only some of the variables below will make it into our final regression models - some will be used for filtering (e.g. **employment**), others are intended for creating meaningful visualizations (e.g. **state**, **industry**), and others might prove to have an insignificant effect on the relationship between gender and income.

In the list below, variables have been grouped under general topics, and we have included their names as they appear in the original data set, their new names as assigned for our analysis, as well as their respective descriptions from the Data Dictionary. For each individual, we will look at:

1. General demographics:
   (a) **SEX** (renamed to **sex**) - *"Sex"*
   (FACTOR WITH TWO LEVELS)
   (b) **AGEP** (renamed to **age**) - *"Age"*
   (c) **CIT** (renamed to **citizenship**) - *"Citizenship Status"*
   (d) **RAC1P** (renamed to **race**) - *"Recoded detailed race code"*
   (e) **MIL** (renamed to **military**) - *"Military service"*
   (f) **DIS** (renamed to **disabled**) - *"Disability recode"*
2. Family and household:
   (a) **MAR** (renamed to **married**)- *"Marital status"*
   (b) **NRC** (renamed to **children_no**) - *"Number of related children in household (unweighted)"*
3. Educational background:
   (a) **SCHL** (renamed to **education**) - *"Educational attainment"*
   (b) **FOD1P** (will be merged with **FOD2P** to create **degree**) - *"Recoded field of degree - first entry"*
   (c) **FOD2P** (will be merged with **FOD1P** to create **degree**) - *"Recoded field of degree - second entry"* [1]
   (d) **SCIENGP** - (renamed to **stem_degree**) *"Field of Degree Science and Engineering Flag - NSF Definition"*
4. Employment:
   (a) **ESR** (renamed to **employment**) - *"Employment status recode"*
   (b) **WKHP** (renamed to **hours_per_week**) - *"Usual hours worked per week past 12 months'*

[1] e.g. for double majors or dual degrees

(c) **NAICSP** (renamed to **industry**) - *"NAICS Industry recode for 2013 and later based on 2012 NAICS codes"*

5. Income:

   (a) **WAGP** (renamed to **wage_income**) - *"Wages or salary income past 12 months (use ADJINC to adjust WAGP to constant dollars)"*

   (b) **ADJINC** (not renamed, will be used during data wrangling to adjust dollar amounts, then discarded) - *"Adjustment factor for income and earnings dollar amounts"*

6. Location:

   (a) **REGION** (renamed to **region**) - *"Region code based on 2010 Census definitions"*

   (b) **ST** (renamed to **state**) - *"State Code based on 2010 Census definitions"*

6.2 Appendix B - Code for Univariate Data Exploration

The code from the Univariate Data Exploration section appears here.

6.2.1 Sex

```
ACSSmall %>%
  group_by(sex) %>%
  tally() %>%
  kable(booktabs = TRUE, caption = "Distribution of Sex",
        col.names = c("Sex", "Tally")) %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))

ggplot(ACSSmall, aes(x = sex, fill= sex)) +
  geom_bar() +
  labs(x = "Sex", y = "Count") +
  ggtitle("Distribution of Sex") +
  theme(legend.position = "none") +
  scale_fill_brewer(palette = "Dark2")
```

6.2.2 Race

```
ACSSmall %>%
  group_by(race) %>%
  tally() %>%
  kable(booktabs = TRUE, caption = "Distribution of Race",
        col.names = c("Race", "Tally")) %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))

ggplot(ACSSmall, aes(x = race, fill= race)) +
  geom_bar() +
  labs(x = "Race", y = "Count") +
```

```
  ggtitle("Distribution of Race") +
  theme(legend.position = "none") +
  scale_fill_brewer(palette = "Dark2")
```

6.2.3 Income

```
ACSSmall %>%
  favstats(~wage_income, data = .) %>%
  kable(booktabs = TRUE, caption = "Distribution of Wage Income",
        col.names = c("Min", "Q1", "Median", "Q3", "Max",
                      "Mean", "SD", "N", "Missing")) %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))

ggplot(ACSSmall, aes(x = wage_income)) +
  geom_density(fill = "#E6AB02", alpha = 0.6) +
  labs(x = "Wage Income", y = "Density") +
  ggtitle("Distribution of Wage Income")
```

```
ACSSmall <- ACSSmall %>%
  mutate(log_wage_income = log(wage_income))

ACSBig <- ACSBig %>%
  mutate(log_wage_income = log(wage_income))

ACSSmall %>%
  favstats(~log_wage_income, data = .) %>%
  kable(booktabs = TRUE, caption = "Distribution of Log(Wage Income)",
        col.names = c("Min", "Q1", "Median", "Q3", "Max",
                      "Mean", "SD", "N", "Missing")) %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))

ggplot(ACSSmall, aes(x = log_wage_income)) +
  geom_density(fill = "#E6AB02", alpha = 0.6) +
  labs(x = "Log(Wage Income)", y = "Density") +
  ggtitle("Distribution of Log(Wage Income)")
```

6.2.4 Age

```
ACSSmall %>%
  favstats(~age, data = .) %>%
  kable(booktabs = TRUE, caption = "Distribution of Age",
        col.names = c("Min", "Q1", "Median", "Q3", "Max",
                      "Mean", "SD", "N", "Missing")) %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))

ggplot(ACSSmall, aes(x = age)) +
  geom_density(fill = "#E7298A", alpha = 0.6) +
  labs(x = "Age", y = "Density") +
  ggtitle("Distribution of Age")
```

6.2.5 Region

```r
ACSSmall %>%
  group_by(region) %>%
  tally() %>%
  kable(booktabs = TRUE, caption = "Distribution of Region",
        col.names = c("Region", "Tally")) %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))

ggplot(ACSSmall, aes(x = region, fill= region)) +
  geom_bar() +
  labs(x = "Region", y = "Count") +
  ggtitle("Distribution of Region") +
  theme(legend.position = "none") +
  scale_fill_brewer(palette = "Dark2")
```

6.2.6 Marital Status

```r
ACSSmall %>%
  group_by(marital_status) %>%
  tally() %>%
  kable(booktabs = TRUE, caption = "Distribution of Marital Status",
        col.names = c("Marital Status", "Tally")) %>%
  kable_styling(latex_options = c("striped", "HOLD_position"))

ggplot(ACSSmall, aes(x = marital_status, fill= marital_status)) +
  geom_bar() +
  labs(x = "Marital Status", y = "Count") +
  ggtitle("Distribution of Marital Status") +
  theme(legend.position = "none") +
  scale_fill_brewer(palette = "Dark2")
```