Predictors of Personal Income

Rana Gahwagy, Irene Foster, Anne Lepow, Wayne Ndlovu

5/20/2021

```
## Use of data from IPUMS USA is subject to conditions including that users should
## cite the data appropriately. Use command 'ipums_conditions()' for more details.
```

Abstract

The goal for our project is to explore the relationship between the number of hours worked per week and total personal income between married and single people. We also wanted to explore the relationship between time travelling to work and total personal income between married and single people. To test our hypotheses, we created a linear regression model using data from the American Community Survey. The variables in our model were total personal income, hours worked per week, transit time, and marital status as well as other variables we decided to adjust for such as sex, age, and race. From our model, we were able to conclude that hours worked per week is statistically significant in predicting total personal income after controlling for age, binary sex, marital status and race. However, we found that transit time was not statistically significant in predicting total personal income after controlling for age, binary sex, marital status and race.

Introduction

The male marriage premium, or the fact that married men tend to make more than their single counterparts is a well documented phenomenon in the field of Marriage and Family studies (McDonald, 2020). Research has shown that the reason for this wage disparity primarily depends upon the type of male individuals who get married. Indeed, studies suggest that men may "self-select" into marriage based not only on current wage level, but also on the potential for wage growth. That is to say men with higher incomes and greater predicted rates of income increase are overrepresented among married men (Hopcroft, 21). Simultaneously, research has been done showing declines in the probability of ever marrying (Bloome & Ang, 20), with especially steep declines predicted in low-income groups and with even steeper declines seen among black people compared to white people.

Existing research has focused on the decline in marriage formation as it relates to racial and socioeconomic differences between groups, assuming that the relationship between marriage and income has a similar relationship across all groups. The results of these studies indicate the potential for marriage to become concentrated among white people. The question then becomes if marriage has a significant relationship with income, especially with regard to race. Specifically, we want to look at the relationship between total personal income and marital status with regard to race, sex, age, and hours worked per week. We expect that there will be a significant relationship between income and marital status based on hours worked, and that relationship will vary with regard to demographic factors, especially with the consideration of race and sex. The implications of this are great: if true, the institution of marriage has the potential to reproduce and worsen existing income inequalities, especially along race lines.

We will also consider a secondary model including transit time as a predictor of income with respect to marital status. Specifically, we want to look at the relationship between total personal income and marital status with regard to race, sex, age, hours worked per week, and weekly transit time to and from work. We expect that the relationship between income and marital status will vary based on demographic factors and based on transit time to work, as previous studies have shown commute time to be an effective indicator of socioeconomic status (Titheridge et. al., 14).

Any significant differences in the relationship between income and marital status with regard to personal income and marital status with regard to demographic factors, specifically race and sex, will be of great relevance to current conversations about the demographics of marriage. Most importantly, any significant relationship between income and marital status that varies by the race, sex, or transit time (considering transit time as a rough estimator of socioeconomic status: lower income populations tend to have increased transit times to and from work) can be thought of in terms of current research which suggests that marriage will become concentrated among white persons. In concert, these findings might suggest that marriage not only continues income inequalities between different groups, but that marriage actually worsens and contributes to increased inequality across different groups.

Methods

For our model we used data from the 2019 American Community Survey, which is housed on by IPUMS USA (https://usa.ipums.org/usa/). The American Community Survey is a survey that collects data about the United State's population every year. The population of the ACS is the United State's population, but in our data we filtered for adults (aged 18+). In 2019 the sample size was 3,544,301 housing units with a nationwide response rate of 86% as well as a sample size of 167,187 people in group quarters with a response rate of 90.9%. The ACS collects data from housing units and group quarters. To collect data from housing units, the ACB begins with a mailed request to respond to the survey online, and then moves onto a mailing survey, a telephone call, or a personal visit if there is no response. To collect data from GQ (group quarters), U.S. Census Bureau Field Representatives interview the GQ facility contact person or the Administrator and then interview a sample of individuals from the facility.

Since we wanted to model the relationship between income and marital status, we chose to have total personal income (INCEARN) as the response variable. This is the individual's total pre-tax personal income

in nominal US dollars during the twelve months immediately preceding the survey. In our original model we decided have a linear model with hours worked per week (UHRSWORK), travel time to work (TRANTIME), marital status (MARST), sex of the participant (SEX), age (AGE), and race of the participant (RACE) as predictor variables. Marital status, sex, and race are all indicator variables. Hours worked per week is the usual number of hours the participant worked last year as a numerical variable. Travel time to work is the approximate number of minutes it took the respondent to travel from home to work in the previous week and it is also a numerical variable. Age is how old the participant is in years. Since our model is focused on married versus single participants, we filtered our data and only included entries where marital status is 1 (married) or 6 (single). Then, to make our model easier to understand, we changed single to 0.

Once we created our model, we noticed that it failed some of the conditions for linear regression so we log transformed our response variable. This helped, but our data still is not perfectly linear and does not have a normal distribution. Our data set was also very large, so we took a random sample of 1,000 people. There was no missing data in our dataset because IPUMS houses complete data. We then performed individual t-tests on UHRSWORK, TRANTIME, MARST, and SEX to see if they were significant in the model. We also performed a Nested F-Test on RACE to see if any of the race indicator variables were significant.

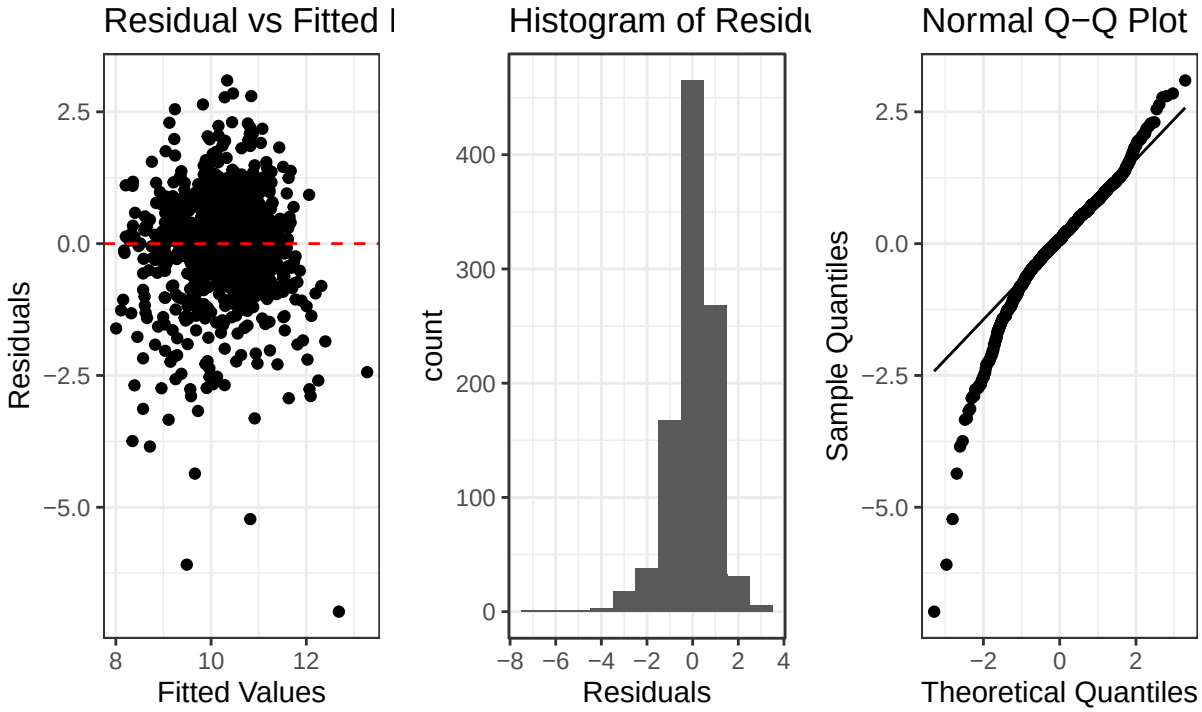

Fig 1. Testing the linearity of the model

Results

The income data shows a discrepancy between female and male-identifying persons as well as married and single people as seen in Fig 2.a. Both married men and women have a higher income than their single counterparts, However, there seems to be little difference in distribution between single men and married women. On the other hand, married men have the highest median and greatest outliers. When comparing income based on the race of the person (Fig 2.b), the largest median is found in people identifying with three or more major races, however, the highest outliers are found among white people.

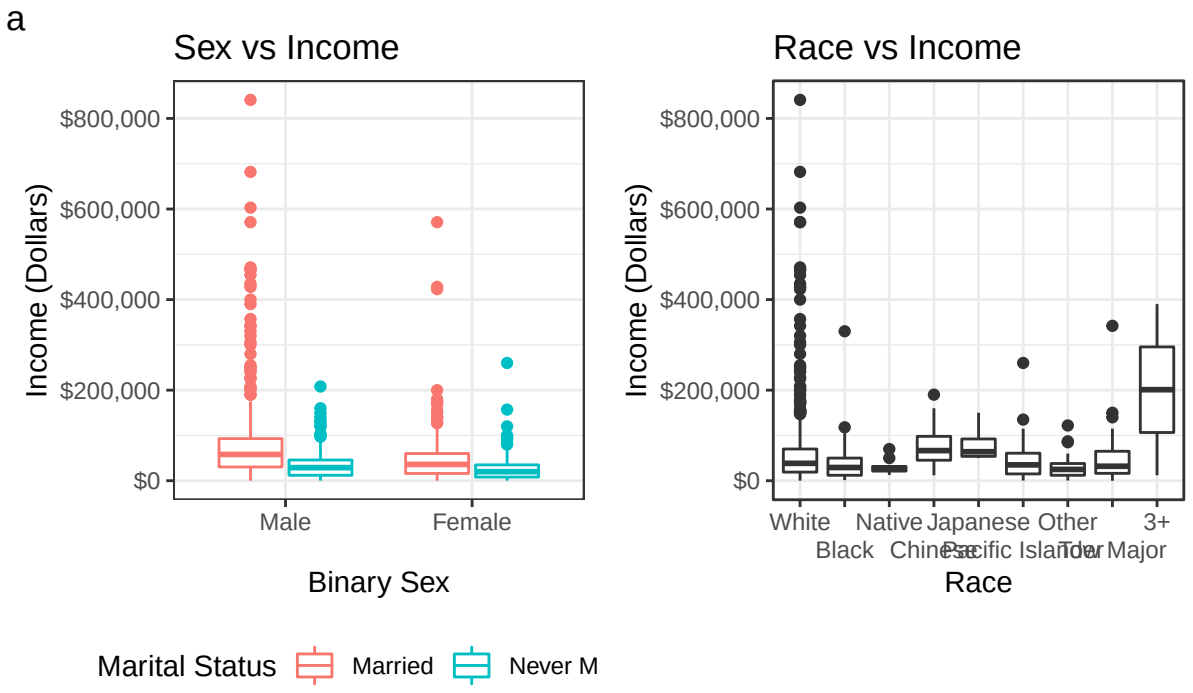

Fig 2. Income disturbution of sample data by (a) sex with regard to marital statu

In Fig 3.a, a positive association between log of income and age, and a simple linear regression reveals that on average men have a higher income across all ages. There is a stronger association between hours worked and income where the average income for married individuals is higher than people who never married (Fig 3.b).

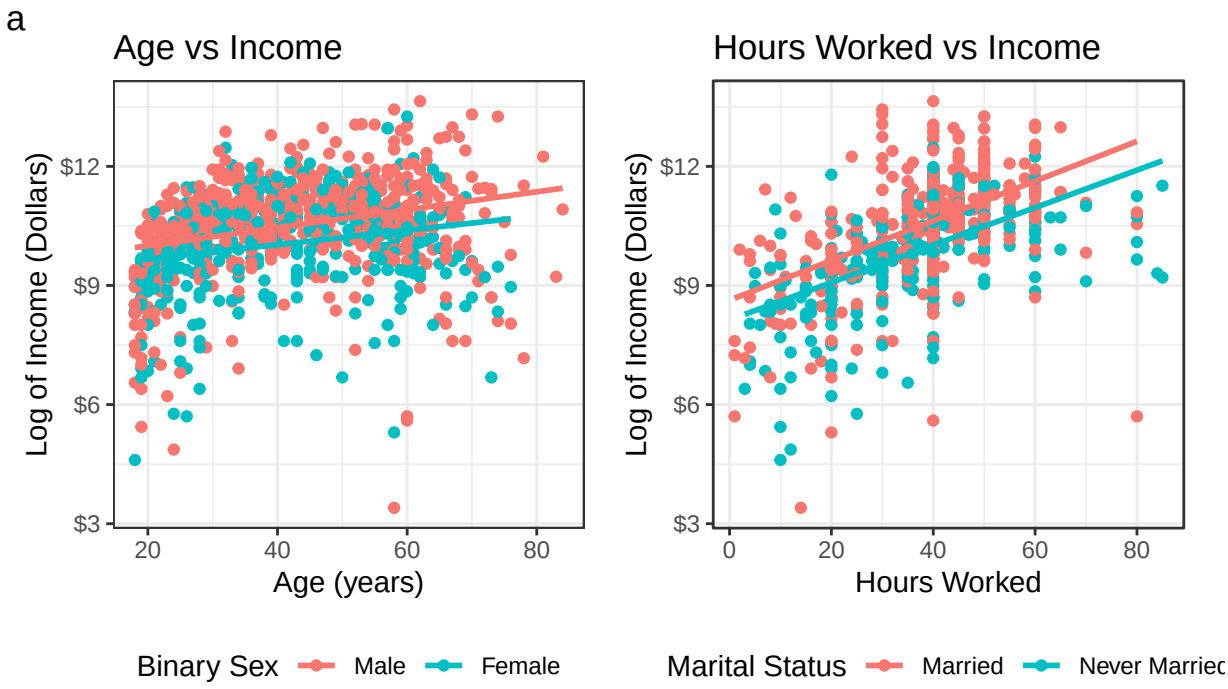

Fig 3. Log income relationship with (a) age by sex and (b) hours worked by ma

In order to understand the various factors that contribute to personal income, a multiple regression model was created which is (INSERT fancy model). The log transformation was incorporated in order to meet the liberty assumptions. An ANOVA test was conducted to choose which variable to include in the model. We found no evidence that transportation time is associated with log income so it does not contribute significantly to the model (p-value = 0.08), therefore, it was not included in the model. We conducted an ANOVA test for the race indicator variables on the untransformed data as well as the transformed data. Although race was found to be not significant after the log transformation, it was included as a controlling factor since it was found to be significant before the log transformation.

```
## Analysis of Variance Table
##
## Model 1: log(INCEARN) ~ AGE + as.factor(SEX) + UHRSWORK + as.factor(MARST)
## Model 2: log(INCEARN) ~ AGE + as.factor(RACE) + as.factor(SEX) + UHRSWORK +
##     as.factor(MARST)
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    995 1017.1
## 2    987 1005.4  8    11.661 1.4309 0.1792
```

```
## Analysis of Variance Table
##
## Model 1: INCEARN ~ AGE + as.factor(MARST) + as.factor(SEX) + UHRSWORK +
##     TRANTIME
## Model 2: INCEARN ~ AGE + as.factor(RACE) + as.factor(MARST) + as.factor(SEX) +
##     UHRSWORK + TRANTIME
##   Res.Df        RSS Df  Sum of Sq      F  Pr(>F)
## 1    994 4.8751e+12
```

```
## 2    986 4.7990e+12  8 7.6099e+10 1.9544 0.04913 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The summary of the result is presented in Table 1 in detail. Based on the t-test output, a p-value smaller than 0.05 is observed for age and hours worked which means that each of those variables are significantly different from zero. Sex, material status, and at least one of the races are too less than 0.05, that implies that difference between male and female, married and single, and white and at least one of the races (Chinese) is significantly different than zero. We noticed that with each increase of ages by one year the income log increases by 0.01 holding all variables constant. Also for each additional hour a person works while all else is constant, the income log increases by 0.05. One the other hand, both single material status and female persons decrease the income. By holding other variables constant, women's log income decreases by 0.21 while for single people it decreases by 0.44. There were also 9 self-identified of which Black/African American, American Indian or Alaska Native, Chinese, Japanese, or three or more major races rise the income log of the person while people identifying as other Asian or Pacific Islander, two major races, or other race experience a decrease in their income log. The overall fit of the model using an F- statistic has a p-value of less than 2.2e-16 which signals that at least one of the variables is helpful to predict the total personal income.

Table 1. Model out put

```
##
## Call:
## lm(formula = log(INCEARN) ~ AGE + as.factor(RACE) + as.factor(SEX) +
##     UHRSWORK + as.factor(MARST), data = final_sample)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9802 -0.4336  0.0799  0.5908  3.0949
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         8.350563   0.171153  48.790  < 2e-16 ***
## AGE                 0.010206   0.002513   4.061 5.26e-05 ***
## as.factor(RACE)2    0.053102   0.122953   0.432  0.66592
## as.factor(RACE)3    0.025191   0.323448   0.078  0.93794
## as.factor(RACE)4    0.644958   0.255957   2.520  0.01190 *
## as.factor(RACE)5    0.608856   0.507361   1.200  0.23041
## as.factor(RACE)6   -0.058877   0.156901  -0.375  0.70756
## as.factor(RACE)7   -0.120077   0.171502  -0.700  0.48400
## as.factor(RACE)8   -0.103460   0.177332  -0.583  0.55974
## as.factor(RACE)9    1.090361   0.715723   1.523  0.12797
## as.factor(SEX)2    -0.212391   0.066246  -3.206  0.00139 **
## UHRSWORK            0.046513   0.002532  18.371  < 2e-16 ***
## as.factor(MARST)1  -0.437732   0.081391  -5.378 9.40e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.009 on 987 degrees of freedom
## Multiple R-squared:  0.3649, Adjusted R-squared:  0.3571
## F-statistic: 47.25 on 12 and 987 DF,  p-value: < 2.2e-16
```

Discussion:

We hypothesised that the usual hours worked per week and the transit time (approximate number of minutes it takes a person to travel from home to work) are good predictors for total personal income of married and single people. We predicted that single people earn less as compared to married people. Our results only support one of our hypotheses - there was a statistically significant difference in total personal income based on the usual hours a person worked per week after controlling for age, binary sex, marital status and race. On the other hand, there was statistically no significant difference in total personal income based on transit time after controlling for age, binary sex, marital status and race. Since there was no statistical significance of transit time, we failed to reject the null hypothesis that transit time does not influence total personal income. As a result, we removed the transit time variable from our model. Overall, our analysis shows that single people tend to earn less than married people.

Although the final model shows that only one race (three or more major races) is statistically significant for predicting total personal income, the anova test with the untransformed data shows that race as a group is significant. Our model accounts for race and shows that people who self identified as being 3 or more major races, Chinese or Japanese tend to earn more. Although the gender variable is binary, our model also shows evidence of the gender pay gap as it shows that people identifying as female tend to earn less than those identifying as male. However, the model has some limitations; it violated the linearity and normality assumptions, and this lowers our confidence of its ability to predict the total personal income. Additionally, only 36% of the variability in income is explained by the model based on age, binary sex, hours worked, marital status and race ($R2 = 0.3649$). We believe future models would benefit from including other variables such as education level, years of experience, type of job and the minimum wage of the state someone lives in.

References

Bloome, D.; Ang, S. (2020). Marriage and Union Formation in the United States: Recent Trends Across Racial Groups and Economic Backgrounds. Demography, 57(5), 1753–1786. https://doi.org/10.1007/s13524-020-00910-7 Hopcroft, R. L. (2021). High income men have high value as long-term mates in the U.S.: Personal income and the probability of marriage, divorce, and childbearing in the U.S. Evolution and Human Behavior, S1090513821000222. https://doi.org/10.1016/j.evolhumbehav.2021.03.004 McDonald, P. (2020), The Male Marriage Premium: Selection, Productivity, or Employer Preferences?. J. Marriage Fam, 82: 1553-1570. https://doi.org/10.1111/jomf.12683 Ruggles, S; Flood, S; Foster, S; Goeken, R; Pacas, J; Schouweiler, M; Sobek, M;. IPUMS USA: Version 11.0 [dataset]. Minneapolis, MN: IPUMS, 2021. https://doi.org/10.18128/D010.V11.0 Titheridge, H; Mackett, RL; Christie, N; Oviedo Hernández, D; Ye, R;. (2014). Transport and Poverty: A review of the evidence. https://doi.org/10.13140/RG.2.1.1166.8645 United States Census Bureau. (Last Accessed: 2021, May 13). American Community Survey (ACS). The United States Census Bureau. https://www.census.gov/programs-surveys/acs