Final Report

Carleena Ortega and Saelin Bjornson

15/03/2020

Contents

Adult Income

Introduction

The Dataset

Who: The data set was extracted by Barry Becker from the 1994 Census database and is donated by Silicon Graphics
What: This is a multivariate dataset with categorical and integer variables. It contains the predicted income of individuals from the census with attributes including age, marital status, work class, education, sex, and race.
When: The data is from a 1994 census.
Why: The data set is found in the University of California Irvine Machine Learning Repository, and was used for ML prediction of whether a person makes over or under 50K a year based on their attributes.
How: The census data was collected by survey.

The Variables

Variable	Type	Description
age	int	Age of individual
workclass	chr	e.g. private, self-emplowed, federal government, never worked, etc.
fnlwgt	int	Final weights: weighted sums of the socio-economic characteristics of the individual. People with similar demographics have similar weights.
education	chr	Highest education recieved
educationnum	factor	Numerical code for highest education recieved
marital_status	int	e.g. married, never married, divorced, etc.
occupation	chr	Occupation of individual
relationship	chr	Relation of individual in family. e.g. wife, child, husband, unmarried
race	chr	Asian-Pacific Islander, Native American, White, Black, other
sex	chr	Male or Female
capital_gain	int	Profit from capital assets such as investments, real estate, etc.
capital_loss	int	Loss from capital assets
hours_per_week		The number of hours that the individual works per week
country	chr	Country of origin
income	chr	Whether individual is predicted to make over or under 50K

The single group used in the following analysis includes divorced, widowed, and never married individuals while the married group includes individials currently married (whether separated, together, or etc.)

The Research Question and Method

Are the number of hours someone works per week correlated with their age, relationship, education level, race or sex?

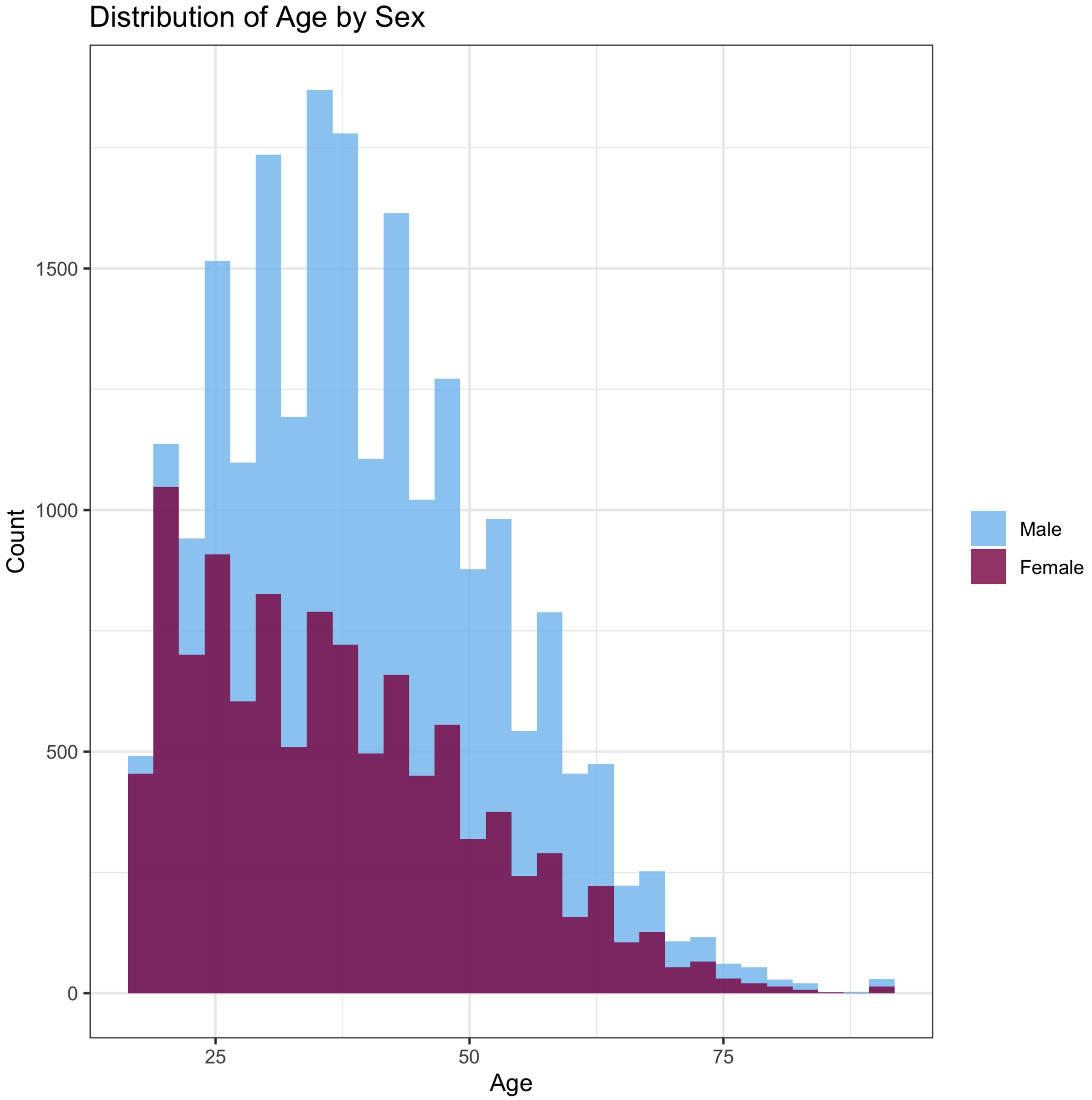
Plots showing the relationship between hours worked and each variable separately. For example, we will use the linear regression model to explore how hours at work is related to variables such as age, relationship, education level, and sex.

Exploratory Data Analysis

In this section, we will get to know our dataset better by exploring the relationship between certain factors.

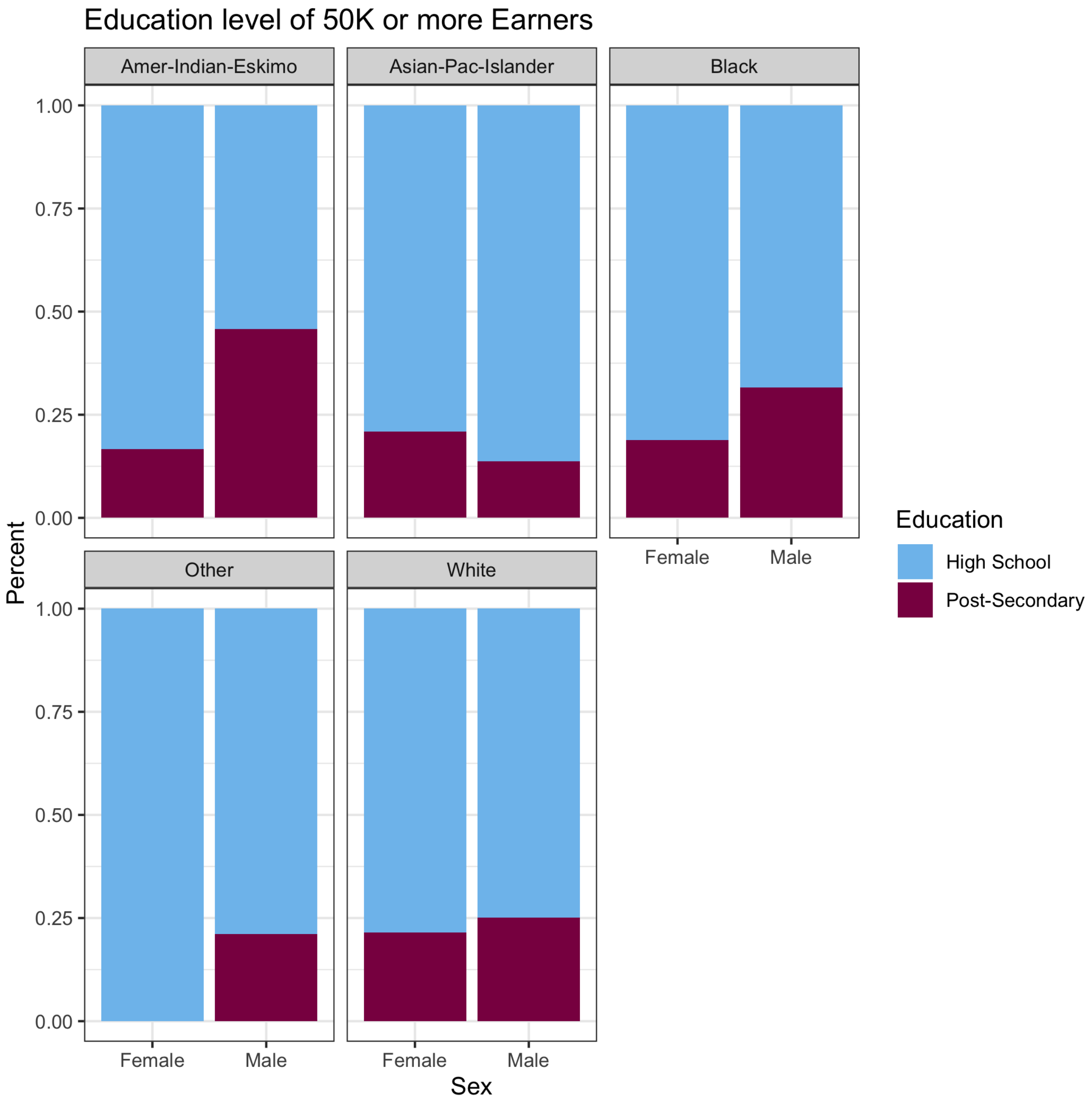
Age and Sex

The plot below shows that there are more male employees than female employees and that the majority of working males are older than working females since the male (blue curve) have a peak shifted to the right with respect to female (red peak).



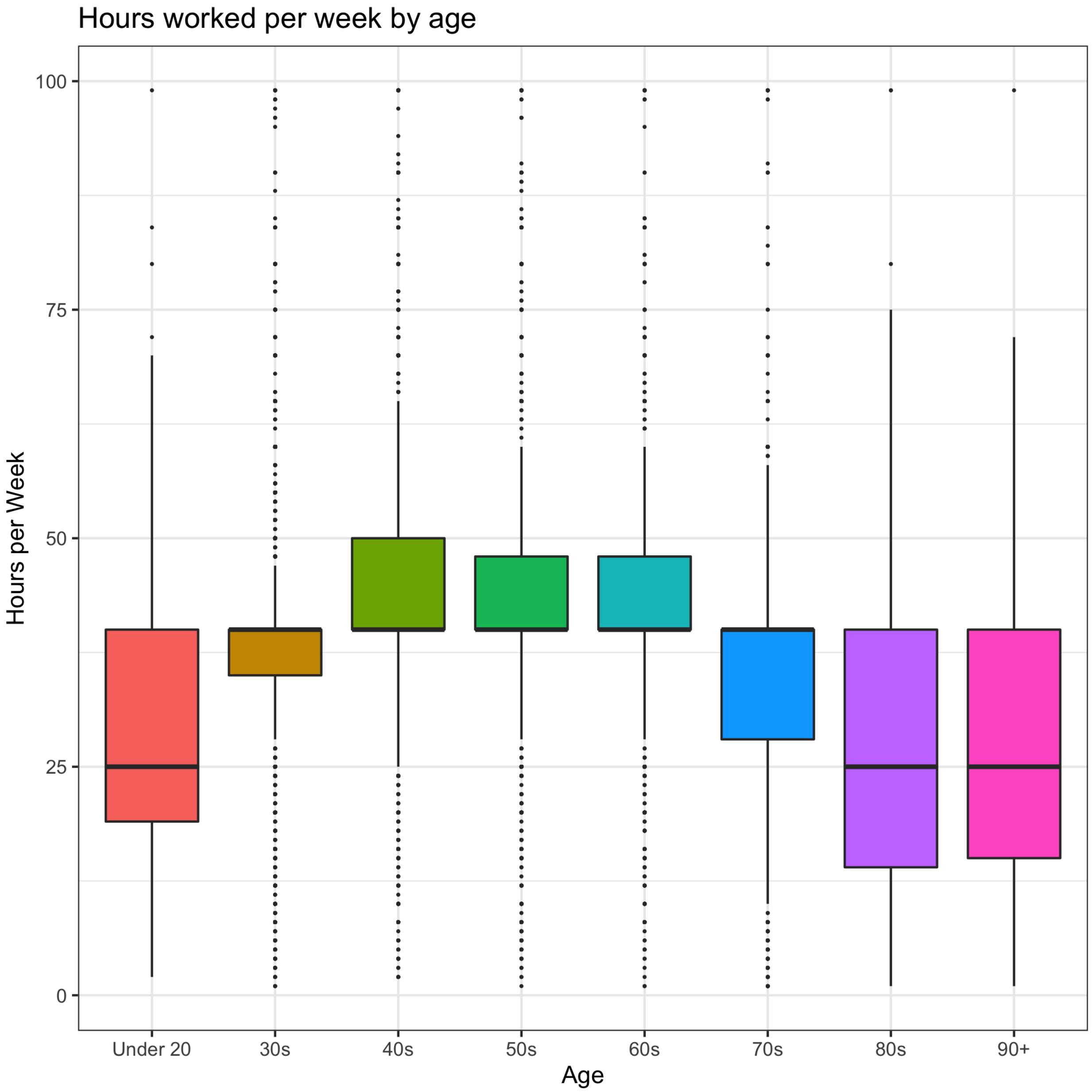
Educational Level and Income

We observe from the following graphs that a majority of individuals earning greater than $50,000 a year only accomplished high school irrespective of sex or ethnical background.



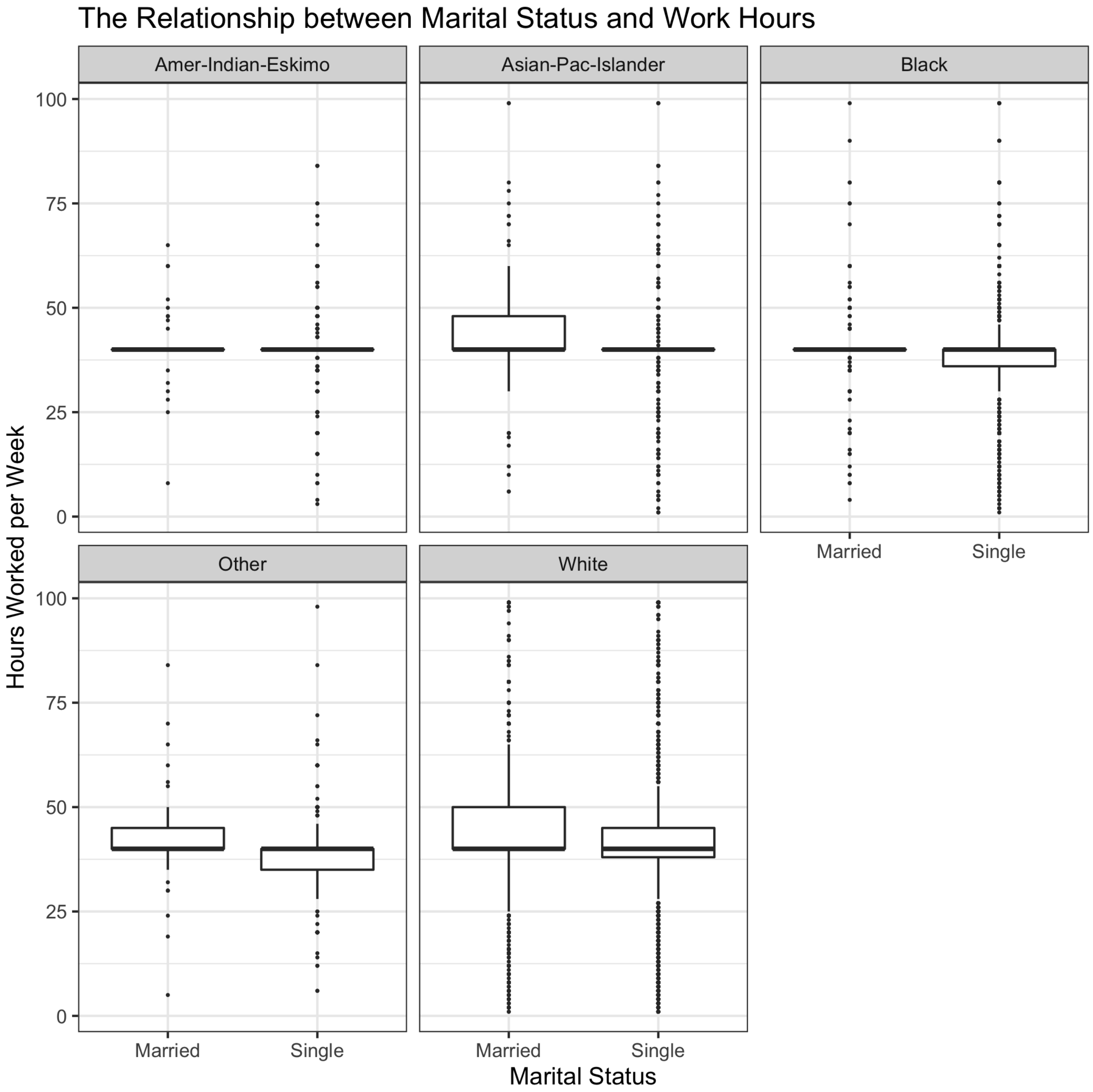
Number of Work Hours and Age

We can deduce from the graph below that individuals work the most hours between their 40's and 60's (probably full time at 40 hours or more a week) and that employees under 20 and over 80 years of age work the same number of hours (probably part time at 25 hours)



Marital Status and Number of Hours Worked

The plot below shows that the working hours between married individuals and single employees are similar.



Linear Regression

We have performed linear regression of hours per week vs. each variable separately, as well as linear regression using all these variables together.

This was done using the lm function of the purr package. For example: lm(hours_per_week~education,data)

For categorical variables sex, education and relationship, the intercept is the defaul reference group, where the "estimate" is the mean of that group, and the "estimates" of all other variables are the differences in means between that group and the reference. The statistic is the t-statistic comparing these means, with a given p-value reporting the significance of this difference.

Hypothesis and Results

Hours vs. Relationship

Hypotheses: Husbands work more than wives since wives are the primary care takers of the household. Unmarried and individuals not in a family work more than husbands or wives since they have more flexible schedules and more time.

Results:

```
## # A tibble: 22 x 5
##    term                 estimate std.error statistic  p.value
##    <chr>                   <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)            32.9       0.791    41.6   0.
##  2 sexMale                 5.69      0.141    40.4   0.
##  3 education11th          -2.73      0.518    -5.26  1.45e- 7
##  4 education12th          -1.04      0.687    -1.52  1.28e- 1
##  5 education1st-4th        0.756     0.991     0.763 4.46e- 1
##  6 education5th-6th        1.36      0.755     1.80  7.11e- 2
##  7 education7th-8th        1.66      0.607     2.73  6.30e- 3
##  8 education9th            0.734     0.649     1.13  2.58e- 1
##  9 educationAssoc-acdm     3.89      0.529     7.35  1.97e-13
## 10 educationAssoc-voc      4.77      0.500     9.54  1.53e-21
## # ... with 12 more rows
```

In this case, we are comparing the mean hours worked per week of husbands (intercept) to each other relationship category. It appears that husbands work more than any other age group, (as seen by the negative estimates), with significant p-values in each case.

```r
tidy(test_relationship)$p.value
```

```
## [1] NA  0
```

Comparing the full model to the model with no relationship info, we can see that relationship does have a significant effect on hours worked per week, with a p value of near 0

Hours vs. Sex

Hypothesis: Males work more hours than women given that women tend to be the care takers of the household.

Results:

```
## # A tibble: 26 x 5
##    term                 estimate std.error statistic  p.value
##    <chr>                   <dbl>     <dbl>     <dbl>    <dbl>
##  1 (Intercept)            45.7       0.793    57.6   0.
##  2 education11th          -2.40      0.508    -4.72  2.37e- 6
##  3 education12th          -0.449     0.674    -0.667 5.05e- 1
##  4 education1st-4th        0.362     0.972     0.372 7.10e- 1
##  5 education5th-6th        0.691     0.741     0.933 3.51e- 1
##  6 education7th-8th        1.33      0.595     2.24  2.49e- 2
##  7 education9th            0.561     0.636     0.883 3.77e- 1
##  8 educationAssoc-acdm     2.50      0.520     4.81  1.50e- 6
##  9 educationAssoc-voc     3.24      0.491     6.60  4.27e-11
## 10 educationBachelors     4.08      0.413     9.88  5.62e-23
## # ... with 16 more rows
```

From this output, we can see that the average hours worked per week for women is 46, and men work -2 more hours per week on average.

```
tidy(test_sex)$p.value
```

```
## [1]            NA 9.079086e-112
```

Comparing the full model to the model with no sex info, we can see that sex does have a significant effect on hours worked per week, with a p value of 9.07e-112

Hours vs. Age

Hypothesis: Individuals start to increase work hours in their 20s as they begin a career and keep increasing or plateauing depending on their marital status until it declines at retirement (around 60 years old).

Results:

```
## # A tibble: 26 x 5
##    term                estimate std.error statistic   p.value
##    <chr>                  <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)           37.3       0.779    47.9   0.
##  2 sexMale                4.22      0.179    23.6   7.57e-122
##  3 education11th         -2.01      0.506    -3.97  7.14e-  5
##  4 education12th         -0.254     0.671    -0.378 7.05e-  1
##  5 education1st-4th      -0.218     0.967    -0.225 8.22e-  1
##  6 education5th-6th       0.337     0.737     0.456 6.48e-  1
##  7 education7th-8th       0.672     0.591     1.14  2.55e-  1
##  8 education9th           0.263     0.633     0.416 6.78e-  1
##  9 educationAssoc-acdm    2.94      0.517     5.68  1.35e-  8
## 10 educationAssoc-voc     3.66      0.489     7.47  8.00e- 14
## # ... with 16 more rows
```

```
tidy(test_age)$p.value
```

```
## [1]           NA 2.507908e-51
```

Comparing the full model to the model with no age info, we can see that age does have a significant effect on hours worked per week, with a p value of 2.507e-51

Hours vs. Education

Hypothesis: Higher educational attainment provides job security and stability leading to more work hours for higher educated individuals.

Results:

```
## # A tibble: 12 x 5
##    term                       estimate std.error statistic   p.value
##    <chr>                         <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)                 43.3      0.732     59.2    0.       
##  2 sexMale                      3.96     0.180     22.0    2.74e-106
##  3 relationshipNot-in-family   -2.04     0.186    -10.9    9.97e- 28
##  4 relationshipOther-relative  -6.04     0.399    -15.1    1.27e- 51
##  5 relationshipOwn-child      -10.4      0.234    -44.6    0.       
##  6 relationshipUnmarried       -2.08     0.266     -7.82   5.38e- 15
##  7 relationshipWife            -3.55     0.360     -9.86   6.68e- 23
##  8 age                         -0.0733   0.00537  -13.6    2.76e- 42
##  9 raceAsian-Pac-Islander      -0.0222   0.751     -0.0295 9.76e-  1
## 10 raceBlack                   -0.738    0.691     -1.07   2.85e-  1
## 11 raceOther                   -0.682    0.965     -0.706  4.80e-  1
## 12 raceWhite                    0.0972   0.663      0.147  8.83e-  1
```

In this analysis, the default reference group (intercept) is a 10th grade education. It appears that those with an 11th grade education work 4 hours less (significant p-value), whereas all other with no more than a high school education work the same amount (no significant p-values).

Every other education level higher than highschool worked significantly more hours, as seen by positive estimates of each group and low p-values.

```r
tidy(test_education)$p.value
```

```
## [1]            NA 1.039446e-148
```

Comparing the full model to the model with no education info, we can see that education does have a significant effect on hours worked per week, with a p value of 1.039e-148

Hours vs. Race

Hypothesis: *Insert Hypothesis*

Results:

```
## # A tibble: 23 x 5
##    term                estimate std.error statistic   p.value
##    <chr>                  <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)          40.6       0.496     81.9   0.       
##  2 sexMale               4.04      0.179     22.6   1.23e-112
##  3 education11th        -2.26      0.504     -4.49  7.31e-  6
##  4 education12th        -0.536     0.668     -0.801 4.23e-  1
##  5 education1st-4th      0.305     0.964      0.317 7.51e-  1
##  6 education5th-6th      0.575     0.734      0.783 4.34e-  1
##  7 education7th-8th      1.34      0.590      2.27  2.30e-  2
##  8 education9th          0.430     0.631      0.681 4.96e-  1
##  9 educationAssoc-acdm   2.78      0.516      5.39  7.20e-  8
## 10 educationAssoc-voc    3.55      0.488      7.28  3.34e- 13
## # ... with 13 more rows
```

```r
tidy(test_race)$p.value
```

```
## [1]         NA 0.06580715
```

Comparing the full model to the model with no race info, race does *not* have a significant effect on hours worked per week, with a p value of 0.065

Discussion

Our data is quite biased in that there are much more men (21790) than women (10771) in data set. This is especially apparent at older ages, for instance there are 742 women over 60 and 1590 men over 60.

Conclusion

Education, age, sex and relationship all have a very significant effect on the number of hours worked per week, while race does not.