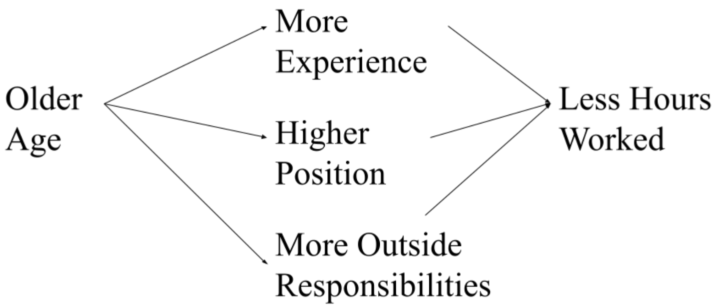
The Relationship Between Age and Hours Worked In the Last Week

Robert James

Paper Assignment #3

Data 202 - Fall 2023, Section 1

The theory explored is that as the age of workers increases the number of hours worked in a week will decrease.This paper will examine the relationship between the age of American workers and how many hours they work in a week. Younger people may work longer hours because they have more time to work and less experience to be able to afford to work less than older people.



This exploration is based on data from the 2018 GSS survey data. The variables were age with a range from 18-89 and hours worked in the last week with a range from 1-89

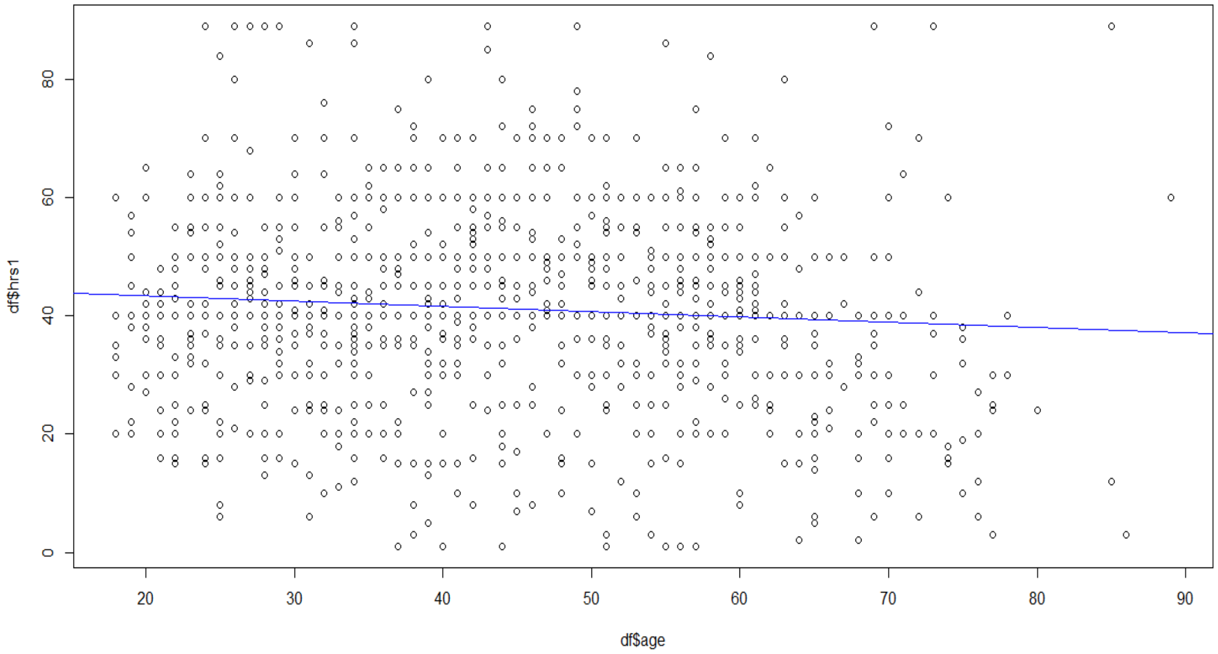
```
age              hrs1
 Min.   :18.00   Min.   : 1.0
 1st Qu.:32.00   1st Qu.:35.0
 Median :42.00   Median :40.0
 Mean   :43.52   Mean   :41.3
 3rd Qu.:55.00   3rd Qu.:50.0
 Max.   :89.00   Max.   :89.0
```

The study hypothesis was that the age of workers is a factor in the number of hours worked in a week. The null hypothesis is that age has no effect on the number of hours worked in a week while the alternative hypothesis is that age has a significant effect on the number of hours worked in a week.

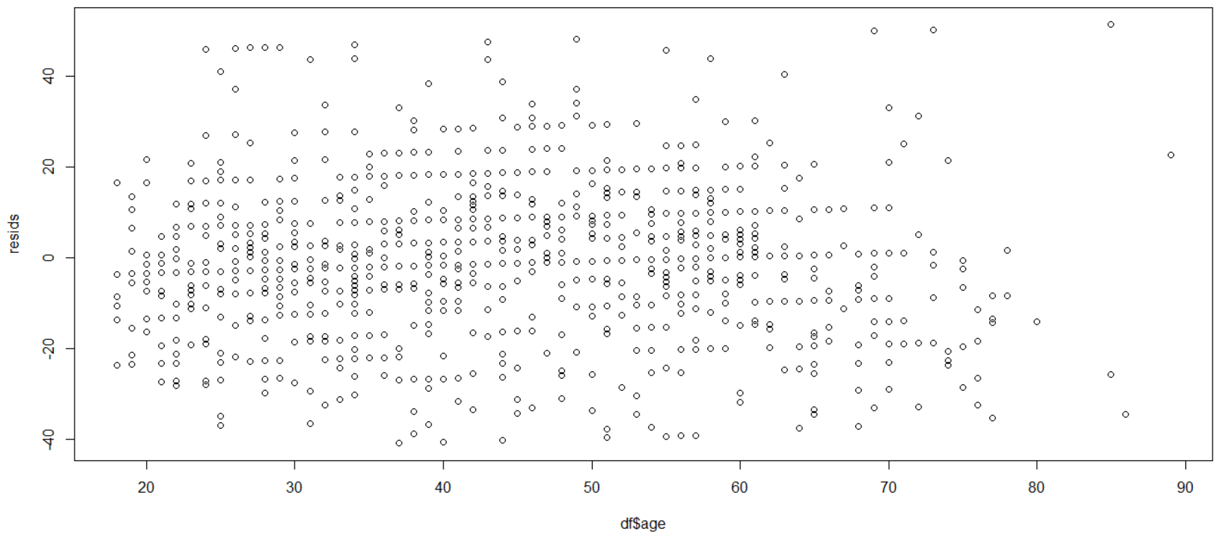
A linear regression model was used to explore how the change in age predicts the hours worked in a week . The summary of the model is below:

```
> model <- lm(df$hrs1 ~ df$age)
> summary(model)

Call:
lm(formula = df$hrs1 ~ df$age)

Residuals:
    Min      1Q  Median      3Q     Max 
-40.874  -6.136  -0.902   7.612  51.368 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 45.14312    1.26126  35.792  < 2e-16 ***
df$age      -0.08837    0.02756  -3.206  0.00138 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.44 on 1373 degrees of freedom
Multiple R-squared:  0.00743,	Adjusted R-squared:  0.006707 
F-statistic: 10.28 on 1 and 1373 DF,  p-value: 0.001378
```

The analysis shows there is a weak but statistically significant relationship between age and hours worked in a week. For every year that age increases the number of hours goes down by .08837 hours. The multiple R-squared value of .00743 is low suggesting that age is not the only factor in the number of hours worked and possibly has other factors that contribute more. Education, industry, and gender could play a bigger part than age but were not included in this analysis.



The above scatterplot visualizes the data modeled with each point representing an individual and their age and number of hours worked. The is a wide distribution of ages and hours worked with the main clusters being people under the age of 60 and working between 30 and 50 hours a week. The model is within this cluster but there are many data points that fall outside of this general trend further suggesting that age is not a main factor in the number of hours worked.



The regression plot above does not have a clear pattern suggesting that the linear model was appropriate to predict a relationship between age and hours worked.

```
> anova(model)
Analysis of Variance Table


Response: df$hrs1
            Df Sum Sq Mean Sq F value   Pr(>F)
df$age       1   2144 2144.08  10.277 0.001378 **
Residuals 1373 286445  208.63
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The ANOVA test conducted shows that age does have an effect on the number of hours worked but further exploration with more variables is necessary to have a full scope in what influences the number of hours worked in a week.

The null hypothesis can be rejected as there is a slight negative relationship between age and hours worked.

Code

```
## Name: Robert James
## Assignment: Paper #3
## Date: 11/12/2023
## Purpose: Explore the Relationship between age and hours worked in the last week
# install packages
# load libraries
library(gssr)
library(dplyr)
library(tidyr)

# load the master documentation files
data(gss_all) #  large file of all GSS data
data(gss_doc) # documentation for the GSS data
# use the dictionary to get information in a different format
data(gss_dict)
gss_dict

df_2018 <- gss_all %>% #filter for only the year 2018
  filter(year == 2018)
df_2018

gss_doc %>% filter(id == "age") %>% # get information of age variable
```

```
  select(id, description, text)

gss_doc %>% filter(id == "hrs1") %>% # get inforrmation of hrs1 variable
  select(id, description, text)

df <- df_2018 %>% #income at 16 and years in school, and wtssall
  select(age, hrs1, wtssall,) %>%
  drop_na() #remove missing values


sapply(df, function(x) sum(is.na(x))) #count missing values
df

# run linear regression model
model <- lm(df$hrs1 ~ df$age)
summary(model)

plot(df$age, df$hrs1) #plots age and hours worked on scatterplot
abline(model, col="blue") # Plot the regression line
cor(df$age, df$hrs1) # correlation between age and hours worked
df

# Check residuals
```

```
resids <- residuals(model)

plot(df$age, resids)


anova(model) # ANOVA test
```