# GSS Data: Socioeconomic Status and Labor 1971-2022

Final Project STAT 365

Gavin Martinez

6/10/2022

## Introduction

The state of the economy is a concern for many people throughout the country. Our generation has seen a fluctuation in the price of homes, gasoline, and various items we deem essential to our day-to-day lives. With a changing economy like the one we have experienced, many lifestyles have changed. GSS, also known as the General Social Survey has been collecting data on socioeconomic issues –such as the shift in our economy– for decades. GSS has been one of the leading factors in policy-making for politicians at all levels, as it reflects the unbiased viewpoints of the American people. We used their data in an attempt to effectively use the GSS data to answer the following three questions about economic and labor trends in America:

**1. Is the average number of household members who earn money significantly larger in 2010 than in 2000? If so, by how much?** Inflation from 2000 to 2010 is reported to be up to 26.23% according to CPI Inflation Calculator I was quite curious to examine whether or not this has forced more family members to work in order to combat the negative effects of inflation. A proper analysis will reveal a more appropriate conclusion, however GSS data reveals shows the average number of household members employed for respondents in 2010 to be about 1.41 and the average number of household earners for respondents in 2000 to be 1.36. Initially, this does not seem to be a large difference in the number of household earners between 2020 and 2000. Also, it is worth noting that this time period has no specific relevance except for the fact that I was alive for a majority of the early 2000s and they are personally relevant to me.

**2. Are respondents working more hours per week?** This question will follow the trend of weekly hours reported by respondents. Following the theme of the American economy and labor trends, I aim to see whether those surveyed are putting in the same amount of work hours consistently throughout the years. As we know, the standard full-time work week is typically 40 hours/week. We can expect to see the average number of weekly hours close to 40, but a proper analysis can suggest otherwise.

**3. Is there an association between income group and highest level of education achieved?** Changing viewpoints on subjects such as education and career interests are seen as a direct reflection of the shifting economy by many. The reason being many jobs that previously provided a respectable wage have failed to combat the increase in inflation from year to year, or have been slow in the process of doing so at the hands of suffering corporations and stagnant workers unions. And so the new generation of students have been encouraged to go to college more than ever, in an attempt to earn higher paying jobs and better job security. This question is aimed at testing whether this theory of higher levels of education yielding higher paying jobs holds true.

# Methods

Participation in the GSS data study is strictly voluntary. However, the General Social Survey uses an “area probability design” which randomly selects respondents from various areas of the country. A majority of the interviews are done via an in-person, 90 minute interview, but there are instances in which a phone interview will be conducted. The survey has been conducted since 1972, and every even year since 1994. The questions for the corresponding variables were phrased as follows.

- **year** - *year of survey*
  - 1972-2014
- **degree** - *“If finished 9th-12th grade: Did you ever get a high school diploma or a GED certificate?”*
  - 0 - less than high school
  - 1 - high school
  - 2 - associate/junior college
  - 3 - bachelor’s
  - 4 - graduate
- **age** - *age of respondent*
  - 18-99
- **hrs2** - *“If with a job, but not at work: How many hours a week do you usually work, at all jobs?”*
  - 0-89
- **rincome** - *“Did you earn any income from [OCCUPATION DESCRIBED IN Q2] last year? a. If yes: In which of these groups did your earnings from [OCCUPATION IN Q2] for last year fall? That is, before taxes or other deductions.”*
  - 1 - under $1,000
  - 2 - $1,000-$2,999
  - 3 - $3,000-$3,999
  - 4 -$4,000-$4,999
  - 5 - $5,000-$5,999
  - 6 - $6,000-$6,999
  - 7 - $7,000-$7,999
  - 8 - $8,000-$9,999
  - 9 - $10,000-$14,999
  - 10 - $15,000-$19,999
  - 11 - $20,000-$24,999
  - 12 - $25,000 or more
  - 13 - refused
- **earnrs** - *“Just thinking about your family now – those people in the household who are related to you – how many persons in the family, including yourself, earned any money last year from any job or employment?”*
  - 0-8

# Analysis

***Question 1: Is the average number of household members who earn money significantly larger in 2010 than in 2000?***

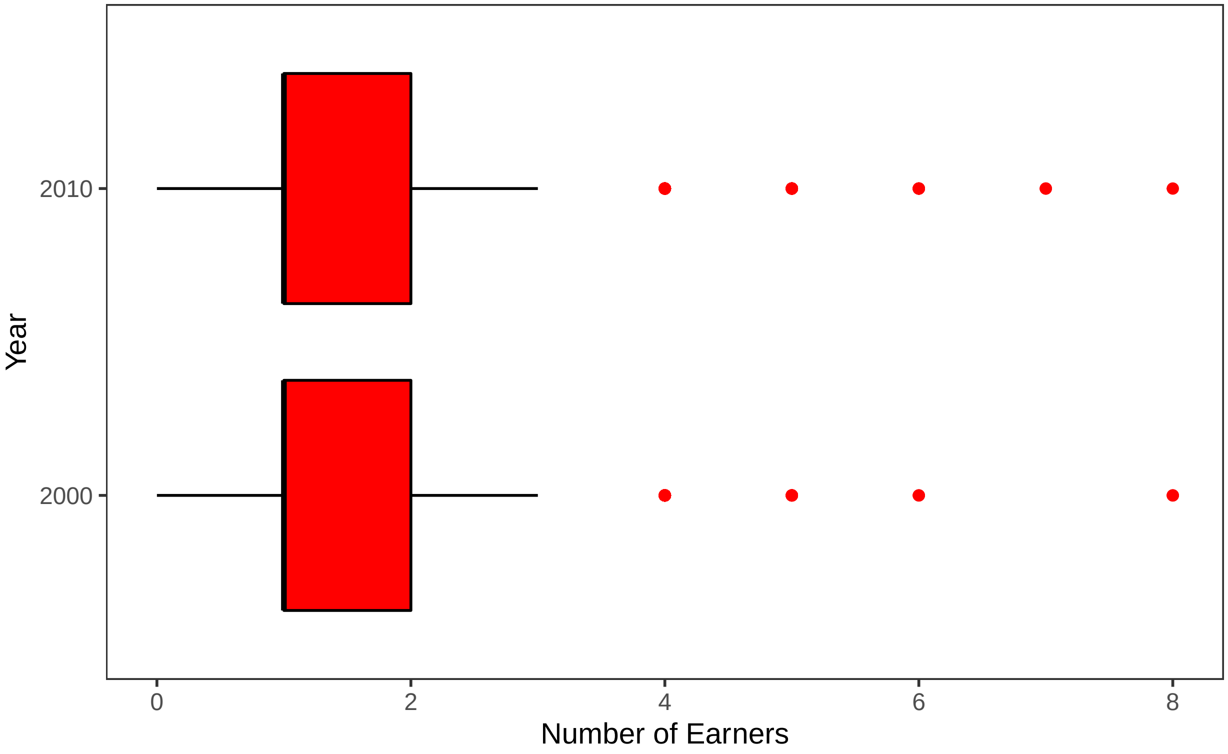
### Summary Statistics

**Table 1. Descriptive Statistics for Number of Household Earners. 2010 reported an average number of earners of 1.41 and 2000 reported a slightly larger average number of earners with 1.36.**

| year | n | average | sd |
|---|---|---|---|
| 2000 | 2810 | 1.362278 | 0.9310898 |
| 2010 | 2030 | 1.406404 | 1.0345716 |

Figure 1. Distrbutions of Household Earners

For the years 2000 and 2010



### Conditions and Assumptions

```
## econ2$year: 2000
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.86569, p-value < 2.2e-16
##
## ------------------------------------------------------------
## econ2$year: 2010
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.86285, p-value < 2.2e-16
```

```
##
##  F test to compare two variances
##
## data:  earnrs by year
## F = 0.80996, num df = 2809, denom df = 2029, p-value = 2.724e-07
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.7469050 0.8778638
## sample estimates:
## ratio of variances
##          0.8099571
```

Both Shapiro-Wilk Tests for normality for the number of household earners reported a p-value of less than .05. The p-values were 2.2e-16 for both 2000 and 2010. A failure of the Shapiro-Wilk tests were expected, however, as we can see clear skew in the boxplots shown in Figure 1, with a handful of outliers pulling the average to greater values. We find significant evidence that the number of respondents' household earners are not normally distributed. Thus, the normality condition is not met. Similarly, the F test for equal variances reveals a ratio of variances for 2000 and 2010 of .81, and F value of .809 and a p-value of 2.724-e7, providing very strong evidence that the two samples do not have equal variances. Thus, the equal variance condition also cannot be met. We must move along with the analysis with caution.

### Hypothesis Testing

```
##
##  Welch Two Sample t-test
##
## data:  econ2$earnrs by econ2$year
## t = -1.5263, df = 4087.3, p-value = 0.0635
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf 0.003436785
## sample estimates:
## mean in group 2000 mean in group 2010
##           1.362278           1.406404
```

We find evidence that the average number of household earners in 2010 is higher than that of 2000. However, based on our .05 significance level, we fail to find significant evidence that the average in 2010 is greater than 2000 (p-value = .0635). We are 95% confident that the average number of household earners in 2010 is anywhere from infinitely greater than to .0034 less than the average number of household earners in 2000. The presence of 0 in this interval supports our decision to fail to reject the null hypothesis.

***Question 2: Are respondents working more hours per week?***

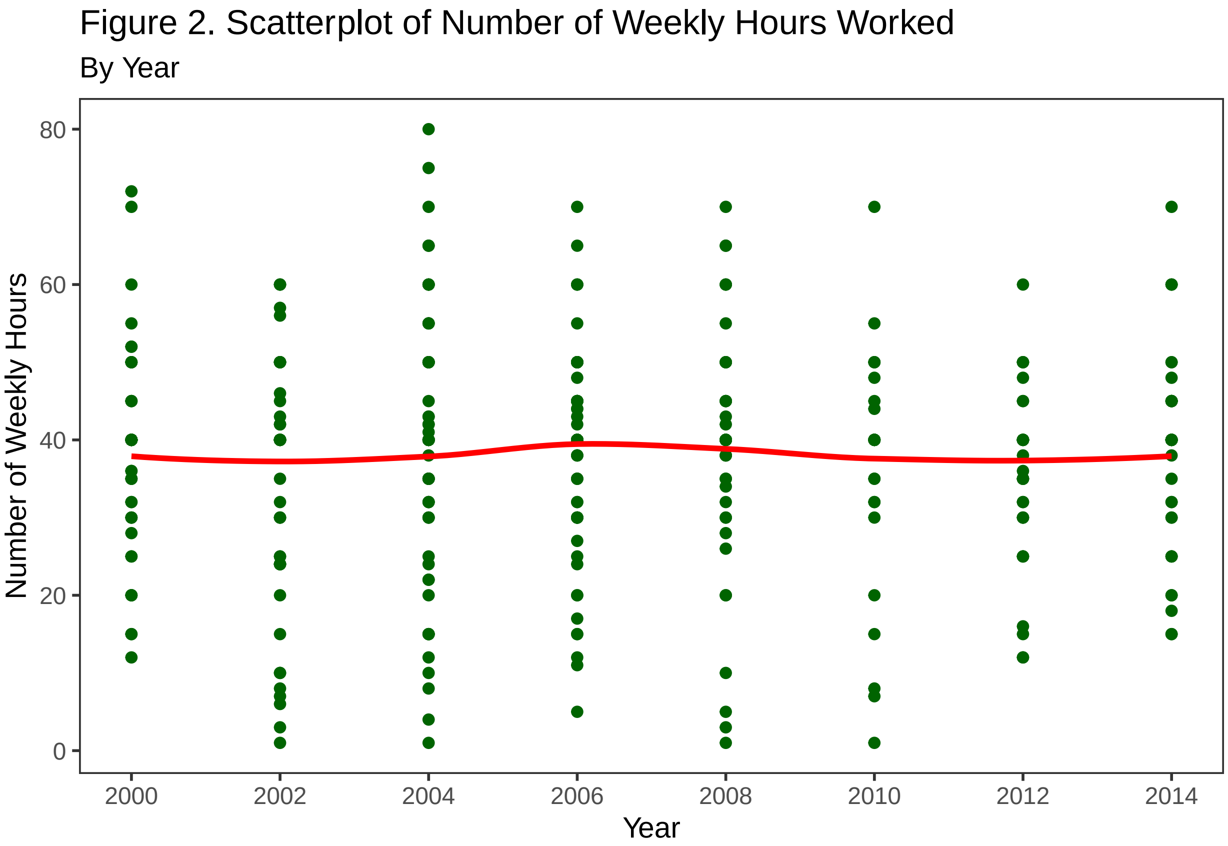


Figure 2. Scatterplot of Number of Weekly Hours Worked

By Year

In Figure 2 we fail to see any apparent trend in the number of weekly hours worked by respondents as the years go by. The plotted line also shows no obvious slope. I would assume an r correlation coefficient fairly close to 0, however, I did not believe a simple linear regression model would reveal very much to us, considering the state of the scatter plot. Another model can be seen below, which attempts to see if there is a trend with number of weekly hours worked and year of survey after adjusting for age of the respondent.

### Hypothesis Testing

### GSS Linear Model

```
##
## Call:
## lm(formula = hrs2 ~ year + age, data = econ)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -38.810  -3.611   0.505   5.061  50.025
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 194.955279  67.037880   2.908  0.00371 **
## year         -0.078105   0.033665  -2.320  0.02051 *
## age          -0.002154   0.030764  -0.070  0.94420
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.59 on 1127 degrees of freedom
##   (58469 observations deleted due to missingness)
## Multiple R-squared:  0.004786,   Adjusted R-squared:  0.00302
## F-statistic:  2.71 on 2 and 1127 DF,  p-value: 0.06699
```

Our F-test for multiple regression has an overall a p-value of .07 initially leading us to believe that neither age nor year are effective predictors for weekly hours worked, indicating no significant trend over time. However, the individual t-test for the slope coefficient for year is shown to be statistically significant at the .05 significance level after adjusting for age of respondent. Based on our model, for every increase of 1 year, we can expect to see an *decrease* of about .078 in the average number of reported weekly hours worked from respondents. That is, after adjusting for age. It is not expected to find a significant effect from year, considering our overall F-test did not reveal any significant effects in the model. A closer look at our conditions and assumptions below may reveal more.

### Conditions and Assumptions

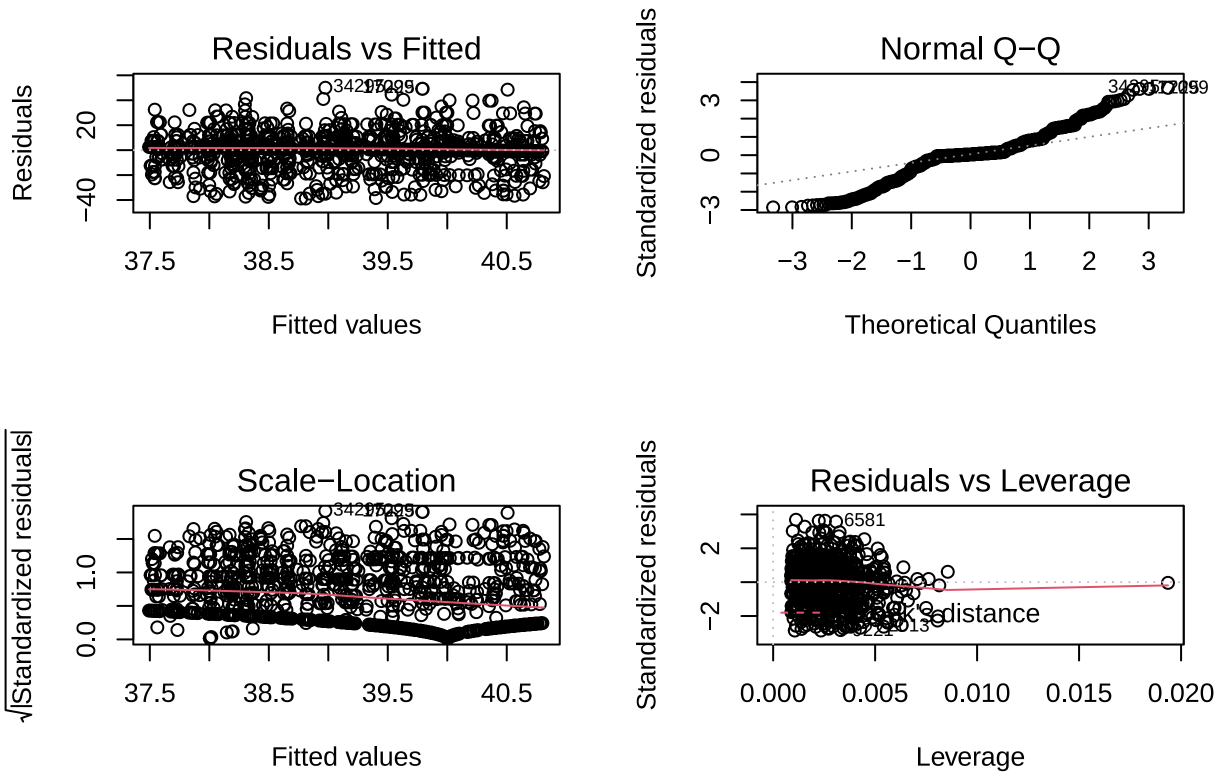


**Figure 3. Residual and QQ-Plots for Conditions and Assumptions of work-week hours model.**

1. **Linearity**: Based on the upper left plot in Figure 3, we can see no clear curvature or pattern in the residual vs. fitted values plot. Thus, we can assume that our residuals follow a linear pattern.

2. **Independence**: Since the respondents were selected randomly, we can assume that all residuals are independent of each other.

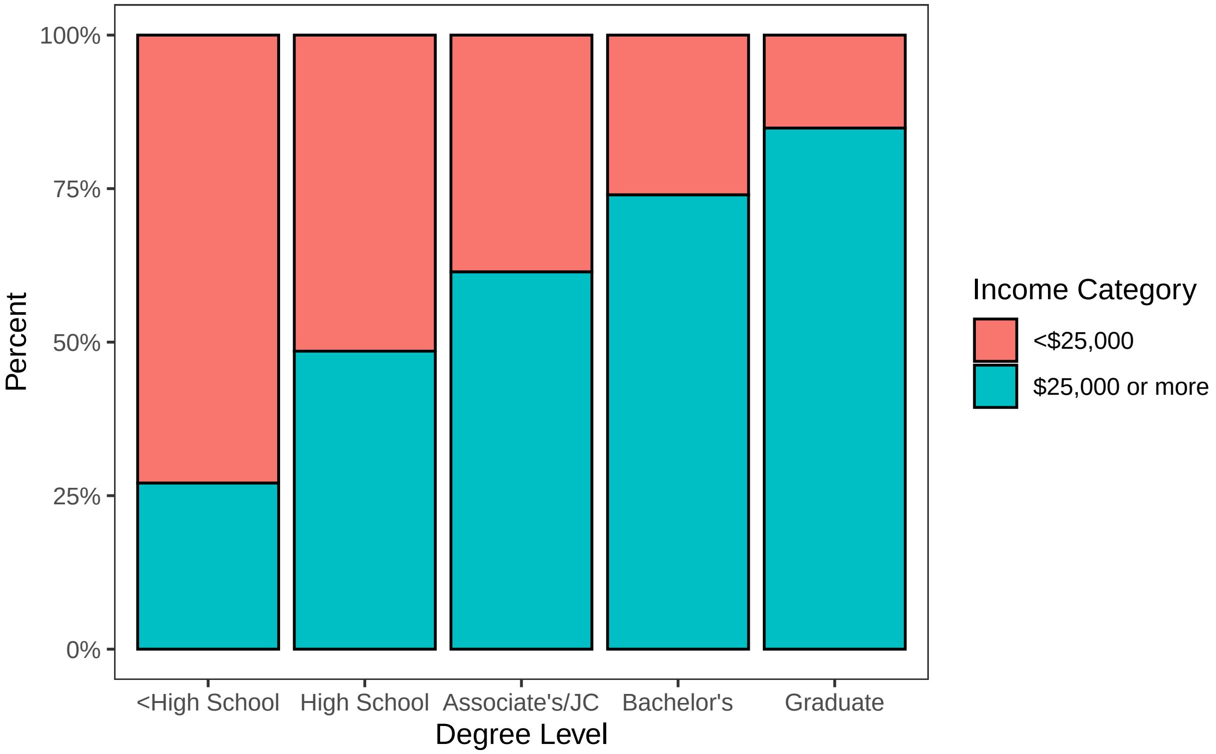
3. **Normality**: The upper right plot in Figure 3, the QQ-plot, shows that the residuals are not distributed normally, due to the fact that they do not follow the linear trend of the normal quantile line.

4. **Equal Variance** : We can assume that the equal variance assumption is met, due to the lack of fanning/ unequal band width in the residuals vs. fitted values plot

***Question 3: Is there an association between income group and highest level of education achieved?***

Figure 4. Proportions of Earners Greater than $25,000

By Highest Level Degree Obtained



**Tables 2 and 3. Count and Proportions of Income Category by Degree Obtained**

| Income Category | <High School | High School | Associate's/JC | Bachelor's | Graduate |
|---|---:|---:|---:|---:|---:|
| <$25,000 | 931 | 3403 | 443 | 666 | 221 |
| $25,000 or more | 349 | 3211 | 706 | 1890 | 1227 |

| Income Category | <High School | High School | Associate's/JC | Bachelor's | Graduate |
|---|---:|---:|---:|---:|---:|
| <$25,000 | 0.73 | 0.50 | 0.36 | 0.26 | 0.15 |
| $25,000 or more | 0.27 | 0.47 | 0.57 | 0.74 | 0.85 |

**Conditions and Assumptions for Chi-Squared Test for Independence**

- Assumption 1: Both variables are categorical.
- Assumption 2: All observations are independent.
- Assumption 3: Cells in the contingency table are mutually exclusive.
- Assumption 4: Expected value of cells should be 5 or greater in at least 80% of cells.

### Hypothesis Testing

```
##
##  Pearson’s Chi-squared test
##
## data:  table(econ3$income2cat, econ3$degreeCat)
## X-squared = 1419.6, df = 4, p-value < 2.2e-16
```

We find significant evidence that the type of degree obtained is associated with whether or not one makes more or less than $25,000 per year (p-value = 2.2e-16).

# Conclusion

**Findings** Results of our tests failed to show a significant increase in the number of household earners between 2000 and 2010. It is plausible to assume that within the first decade of the 2000s the average number of household earners was consistently under 1.5 in the United States. In context, this is time period experienced the recession of 2008, which many would expect to influence the number of household earners necessary from home to home. However, our analysis reveals otherwise. Similarly, our conjecture regarding an increase of weekly hours worked as year increased was halted by what we saw in Figure 2, and the lack of any strong positive association between year and weekly hours reported by respondents. A slightly more complex model factoring in age of respondent was made, and the individual t-test for year did show a significant effect of year on weekly hours after adjusting for age. However, much like our first test we did not see what we expected, as weekly reported hours saw about a .078 hour decrease with every increase of 1 year. This shows a negative trend in weekly hours worked from 2000-2014. Finally, a chi-squared test of independence yielded very strong evidence that there was an association between degree obtained and whether or not an individual made more than $25,000 for their annual income.

**Discussion** Although our tests did not reveal the trends that we expected for the first two questions, and we did see what we expected for the last question, it is important to recognize some of the limitations of these tests. First, the t-test results comparing the number of average household earners in 2000 and 2010 failed both the Shapiro-Wilk test for normality (for both 2000 and 2010) and the F-test for equal variances. An effect size of .045 was reported as well. Because of these failed conditions, we cannot trust the results of our t-test's analysis. Similarly, the conditions of our linear model appeared valid, except for the normality of the residuals, showed by a non-linear QQ-plot in Figure 3. We cannot trust the results of our overall F-test or the individual t-tests for the slope coefficients of our multiple regression model. This may explain the F-test showing no evidence of any significant effects in the model, despite a significant p-value for the year variable. The assumptions for the Chi-Squared test, however, appear to be valid, and our analysis can be trusted.