Matthew Ye
November 17, 2018

Investigating the connection between number of hours worked per week and yearly income

As working adults, we may want to think about the amount of time spent at work and the cost-benefit of that time. Spending more time at work means there's less time to spend on leisure activities like family or friends. This can have a toll on mental health if working adults are not incorporating time for self-care. It might not be worth it if working more hours is not paying off monetarily. The 40hr work week is the standard, but is there a relationship between the number of hours you work per week, and your total income?

My target population are people in the United States, and may wish to see how different variables could affect their income. I hypothesize that there is a positive linear relationship between the number of hours worked and yearly income and that the hours worked per week can explain the variation in income. This memo will show that there is indeed a positive relationship between the number of hours worked per week and yearly income. When we regressed income on hours worked, we found that the more hours worked per week, the higher the yearly income.

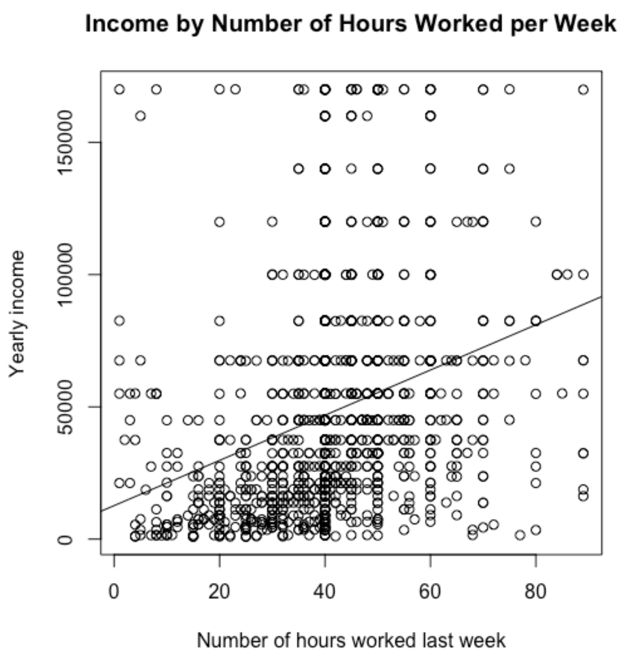
To investigate this question, I used the nationally representative data in the 2016 General Social Survey. Specifically, the target population are working class adults in the U.S. I will be looking at those who responded to a question about their yearly income and how many hours they worked last week. The respondent's yearly income will be used as the Dependent Variable, and the number of hours worked per week will be used as the Independent Variable. The income variable is the amount of money earned per year. The hours variable is the number of hours per week they worked last week at all jobs.

**Table 1: Summary Statistics**

| | Freq | Min | Max | Med. | Mean | SD |
|---|---|---|---|---|---|---|
| Average hours worked per week | 1646 | 1 | 89 | 40 | 40.91 | 14.41 |
| Yearly income | 1632 | 1000 | 170000 | 32500 | 45125 | 39225.93 |

I recoded the hours variable into numeric and changed the level of "89+ hr" to just 89. Therefore, the max number of hours worked is not just 89 hours, but potentially more than 89. I also recoded the income variable so that the levels are now coded to the midpoint. I also set the max value of 170,000+ to just 170,000. There were a substantial number of missing values, about half, but the sample size was large enough that this not that influential, and the amount missing for both variables was very small.

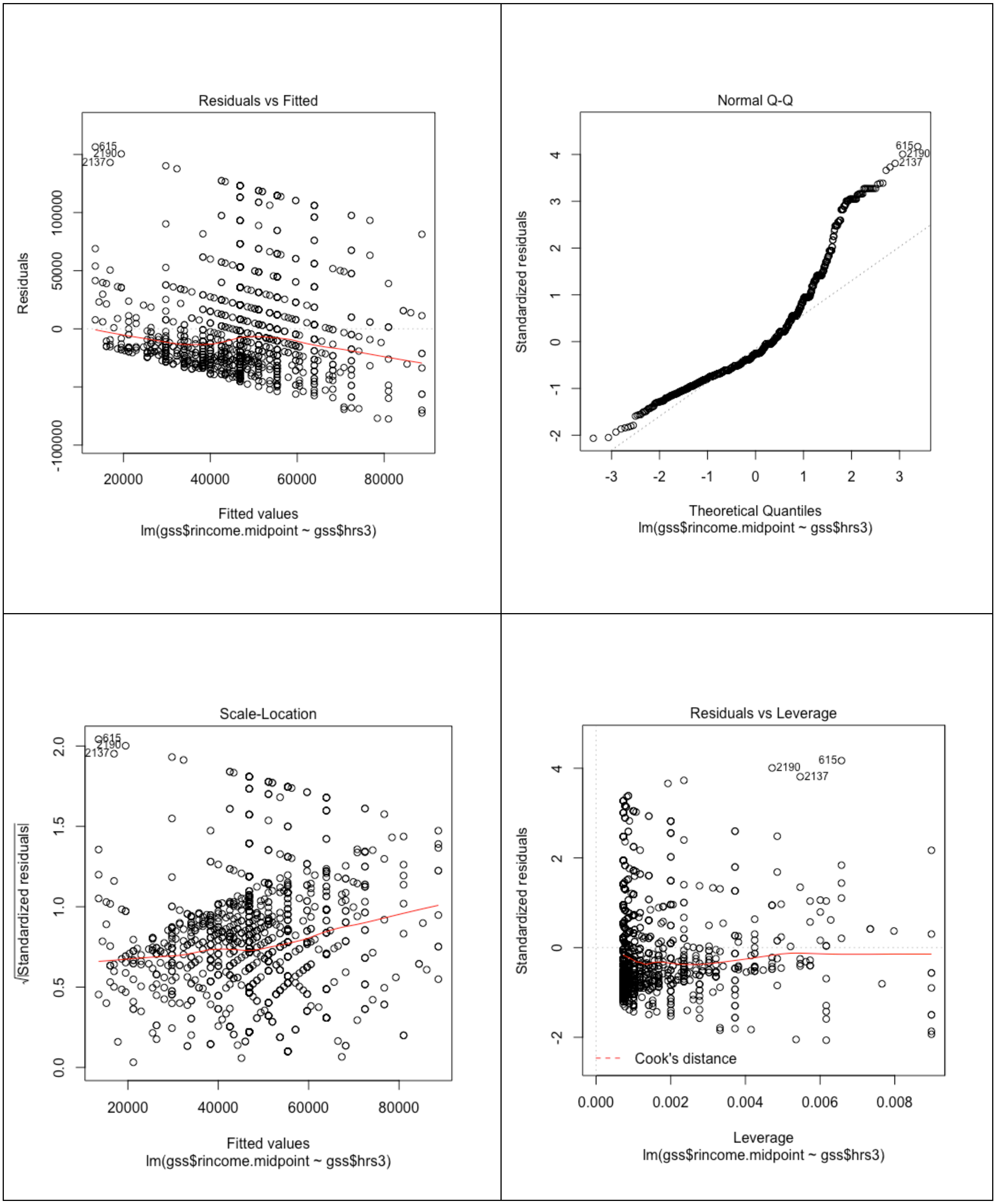
We performed a Simple OLS Regression of Income on Hours worked Last week with an alpha level of 0.10. Our regression model has an equation: $\widehat{Predicted\ Income} = 12669.3 + 854.2(Hours\ worked\ per\ week)$ The entire model is statistically significant at the 0.01 alpha level, with a F-Statistic p-value <0.01. The independent variable is also statistically significant.




We can reject the null hypothesis that the mean is sufficient to predict the dependent variable of yearly income. At the 0.01 significance level, we can generalize our findings to the target population of working class adults in the U.S. For every extra hour you work per week, you can expect an average increase of 854.2 dollars (p<0.01). We believe that the coefficients are meaningful to a degree. The y-intercept coefficient is not meaningful, because the x-value (hours worked last week) never reaches zero, so the y-intercept is extrapolated. There is a statistically significant Rsquared value of 0.094. This means we can explain 9.4% of the variation in income by number of hours worked. $R^2 = 0.093$, $F(1, 1377) = 142.3$, $p < 0.01$.

When running Regression Diagnostics, we found the following results. In the Residuals vs Fitted plot testing for linearity, the red line is relatively flat with some dips. There is a lot of data clustered underneath the red line, which means the data could do better for linearity. In the Normal QQ plot for normality, there is a slight S-curve. The data is somewhat linear near the middle, but gravitates towards the higher ends. In the Scale-location plot, there are no obvious visible patterns, meaning there is decent homoscedasticity of the data. In the Residuals vs Leverage plot test for influential outliers, there are no data points that cross Cook's distance, meaning there are no influential outliers to influence our data.

Some weaknesses in our study include the fact that a lot of the data points for average number of hours worked per week clusters around 40 hours, since that is the average work week. Furthermore, those who work 40 hours per week usually have a full-time job with wildly ranging salaries. Thus, our results may not be representative. We can remedy this by changing some of the variables slightly. For example, we can look at incomes of people with part time jobs and the average number of hours worked for that population. This should show a stronger regression as part-time jobs' incomes usually depend directly on the number of hours worked. We can also use a truly I-R Dependent Variable, one that is not limited to ranges of incomes, so we avoid the flat lines of the "same" income in our data.



Residuals vs Fitted
Residuals
Fitted values
lm(gss$rincome.midpoint ~ gss$hrs3)
615
2190
2137
Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(gss$rincome.midpoint ~ gss$hrs3)
615
2190
2137
Scale-Location
√|Standardized residuals|
Fitted values
lm(gss$rincome.midpoint ~ gss$hrs3)
615
2190
2137
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
lm(gss$rincome.midpoint ~ gss$hrs3)
2190
615
2137