Matthew Ye

December 4, 2018

Effect of hours worked per week, gender, number of children, and class on income

For those working adults in the United States, there are a variety of factors that affect the amount of money one makes. It may be worthy to know which factors could increase or decrease annual income. This may reveal some trends in our society to be true when it comes to income, classism, family structure, and gender. How people choose to use this information is up to them.

My target population are people in the United States who may wish to see how different variables would affect their income. I hypothesize that factors such as Hours Worked per Week, Number of Children, Gender, and Number of Children will affect a person's annual income. This memo will show that these factors do in fact predict annual income better than the mean.

To investigate this question, I used the nationally representative data in the 2016 General Social Survey. Specifically, the target population are adults in the United States who have some sort of income. I will be looking at those who responded to a question about their yearly income, the number of hours worked they last week, how many children they have, their sex, and where they would place themselves in a social class. The respondent's yearly income will be used as the dependent variable. The number of hours worked per week, number of children, sex, and social class will be used as the independent variables. The income variable is the amount of money earned per year. The hours variable is the number of hours per week the respondent worked last week at all jobs. The number of children variable is the number of children the respondent has ever had. The sex variable is whether the respondent identifies as male or female. The class variable is which social class the respondent would place themselves in, going from Lower class, Working class, Middle class, and Upper class.

Table 1: Quantitative Summary

	Freq	Min	Max	Med.	Mean	SD
Hours worked	1646	1	89	40	40.91	14.41
Number of Children	2859	0	8	2	1.85	1.67
Income	1632	1000	170000	32500	45125	39225.93

For the hours variable, I changed the last level of "89+hr" to just 89. Therefore, the max number of hours worked is not just 89 hours, but potentially more than 89. For the children variable, I changed the last level of "EIGHT OR MORE" children to just "8." This means that max number of children could potentially exceed 8 children. Hours and children were both Interval-Ratio variables. The sex and class variables were dummy coded. This means that max number of children could potentially exceed 8

children. Hours and children were both Interval-Ratio variables. The sex and class variables were dummy coded. For sex, I created two dummies, one for each sex. The male dummy had 1 = male and 0 = female.

Table 2: Categorical Summary

	Freq.	Percent
Gender		
Female	1591	55.49%
Male	1276	44.51%
Total	2867	100%
Social Class		
Lower	286	10.06%
Working	1330	46.78%
Middle	1146	40.31%
Upper	81	2.85%
Total	2843	100%

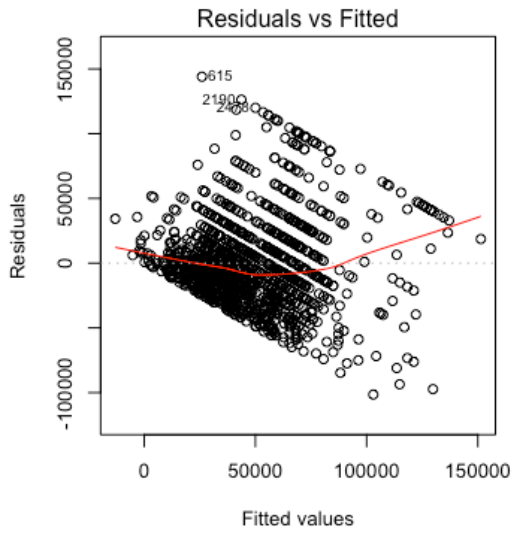
For the class variable, I created 4 dummy variables, one for each social class. For each, I set 1 equal to the class of interest, with all others being equal to 0. For all 26 original ranged levels of income (e.g., $10,000-$15,000), I recoded them so the levels are now coded to the midpoint. I also set the max value of 170,000+ to just 170,000. There was 1377 missing variables, which was less than 10% of the total sample, therefore not that significant.

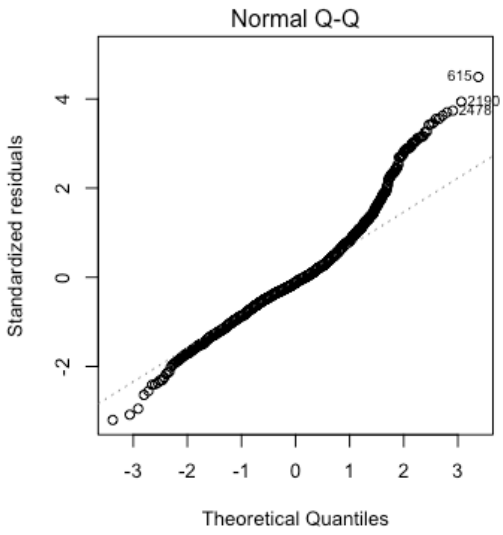
We performed a Multiple OLS Regression of income on hours worked last week, number of children, gender, and class at an alpha level of 0.01. $\widehat{Predicted\ Income} = -7130.5 + 765.6(Hours\ worked) + 1437.9(Number\ of\ children) - 9499.9(Female) + 13324.7(Working\ class) + 41842.6(Middle\ class) + 94824.4(Upper\ class)$. For the dummy variable Sex, Male was left out as the reference group. For the dummy variable Class, Lower Class was left out as the reference group. The entire model was statistically significant at the 0.01 alpha level with a F-statistic p-value < 0.01. We can reject the null hypothesis that the mean is sufficient to predict the dependent variable of income. Every independent variable, except number of children, was significant at the 0.01 alpha level. When running regression diagnostics, we found the following results. In the Residuals Vs Fitted plot test for linearity, the red line is somewhat flat, but has a dip in the middle and curved up towards the end. In the Normal QQ plot for normality, the data follows the line initially well, but deviates into an S-curve at the end. In the Scale=location plot, there are slight patterns of waves radiating from the origin, indicating some type of pattern, though it's unclear if this is significant. In the Residuals vs Leverage plot for influential outliers, there are no data points that cross Cook's distance, indicating not influential outliers. With the results considered, while there are no glaring problems, I would still be hesitant to generalize these findings to the entire population.

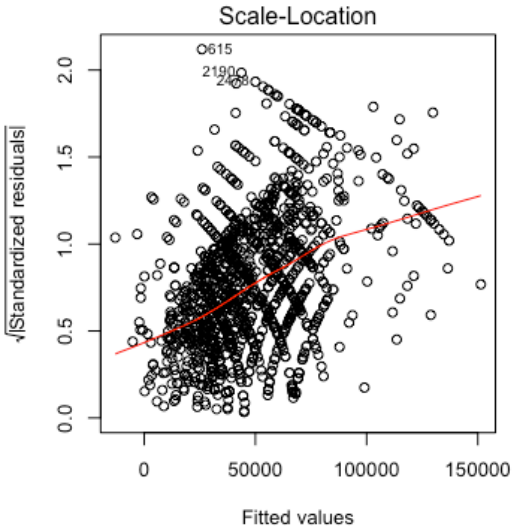
For every hour you work per week, you can expect an average increase of 765.6 dollars per year ($p<0.01$), holding all other variables constant. For every extra child you have, you can expect an average increase of 1437.9 dollars per year ($p = 0.015$), holding all other variables constant. If you are a female, you can expect to see an average decrease of 9499.9 dollars per year ($p<0.01$) compared to men, holding all other variables constant. If you consider yourself working class, you can expect to see an average increase of 13324.7 dollars per year ($p<0.01$), compared to lower class, holding all other variables constant. If you consider yourself middle class, you can expect to see an average increase of 41842.6

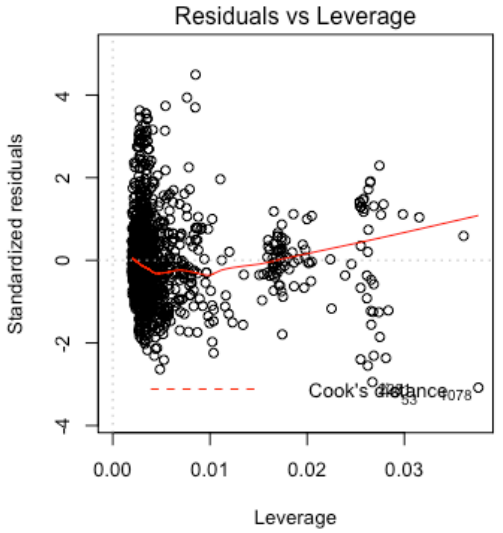
dollars per year ($p<0.01$), compared to lower class, holding all other variables constant. If you are upper class, you can expect to see an average increase of 94824.4 dollar increase per year ($p<0.01$), compared to lower class, holding all other variables constant.

I believe these coefficients to be meaningful. As each are significant at the 0.01 alpha, except for number of children (which is significant at the 0.05 level), they are good predictors of annual income. Our adjusted R^2 value of 0.34 means we can explain 34% of the variation in income by all the independent variables combined. Some weaknesses of this study include the ordinal nature of the Income variable that has been used as Interval-Ratio. This can lead to distinct levels of instead of a continuous range of numbers. Furthermore, the regression diagnostics indicate the data might not be perfectly normal or linear. While the errors aren't egregious, it makes our results less compelling to generalize to the population.


Residuals vs Fitted
Residuals
Fitted values
615
2190
2478
150000
50000
0
-100000
0
50000
100000
150000


Normal Q-Q
Standardized residuals
Theoretical Quantiles
615
2190
2478
4
2
0
-2
-3
-2
-1
0
1
2
3


Scale-Location
√|Standardized residuals|
Fitted values
615
2190
2478
2.0
1.5
1.0
0.5
0.0
0
50000
100000
150000


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
4078
4
2
0
-2
-4
0.00
0.01
0.02
0.03