The Analyzation of Worked Hours and Demographic In The Perspective of Education and Personal Income For United States In 2018

Haoxuan Cai, Matthew Ma, Zhihao Chen

20/03/2022

Abstract

A person's weekly working hours are closely related to his/her living standard, happiness degree and many other factors. In a typical developed country with a relatively different social status like the United States, it is of great representative significance for government, economy and social research to count the weekly working hours of the people. In this report, according to the General Social Survey(GSS) provided by the Us National Opinion Research Center, statistical information with both quality and quantity to systematically summarize, summarize and further study the average working hours of Native Americans in 2018. An analysis report of working hours and demographic distribution will be made from the perspective of education level and income. In addition, an appendix will be included in this report. A Social survey and fully reproducible method have helped other researchers understand the data in greater detail.

Introduction and Hypothesis

In 2018, with the increase of tariffs on the import of a variety of metal products, restrictions on investment from other countries, the Sino-US trade war and many other events related to the world economy and the local economy of the United States, the social structure and unemployment rate in the United States began. It was not only the most normal year people lived before the COVID-19 pandemic but also the most representative year for the-U.S. President Donald Trump's policies towards the U.S domestic economy.

Under the influence of the above situation, numerous jobs and employment rates in America have changed dramatically. Since the government's policy is proposed to allocate more jobs opportunity for the local United States of American citizens and guarantee more U.S citizens to get paid, many other nation people who live in the U.S are facing the problem of layoffs or higher working pressure. Even though many people got the jobs they needed, many others were unemployed. As a result, many people do not have enough money to support their living expenses due to reducing working hours and salaries, so they have to obtain more income through part-time jobs. On the other hand, as more jobs are available for local U.S citizens, the working hours will also change, and some people will face more work pressure that they did not have before, leading to a lower quality of life.

Therefore, it is imperative for the American people, government and society to analyze the total amount of people working in a certain period. The most intuitive measure is the number of hours worked in a week.

However, collecting data for this metric is difficult because working hours often fluctuate significantly based on different occupations, personal family circumstances, and working cycles, making it challenging to estimate. For example, for self-employed people, entrepreneurs, social service workers, and people with multiple occupations simultaneously, the most intuitive and effective statistical method is to conduct a questionnaire survey.

Attributed to the well-developed Social system in the United States, the General Social Survey (GSS), a sampling Survey method, emerged in 1972, which provides rich questionnaire data for the society. All sampling questions and data statistics were provided and integrated by the National Opinion Research Center. The vast majority of General Social Survey interview models are based on personal interviews, the respondents are all adults in the United States, and the questions studied are all high-quality scientific or Social science-related data with statistical significance. In the past few decades, the General Social Survey has made outstanding contributions to the Social sciences in the United States. It has effectively helped the government, statisticians, economists, sociologists, and a series of other related professions collect information about the adult population, including mental health assessment, happiness, income, race relations, Consumer habits, sexual recognition issues, etc. In addition, except for the first two years, the design of the General Social Survey is a full probability sample design. Therefore, compared with other similar Social questionnaires, General Social Survey has absolute advantages in output quality, output scale and output proportion.

Thus, General Social Survey made a survey on the weekly working hours of people in the United States in 2018 mentioned above and collected the results, which provided highly reliable, scientific and enormous data for the public and relevant researchers to study. By collecting and downloading data from General Social Survey official website, more than 2300 individuals were given feedback on weekly working hours.

Data

Data Summary

From 1972 to the current times, the GSS (General Social Survey) has been recording social transformations and making the research on the growth complexity of American society. The General Social Survey always provide a reliable source of data to help researchers, students, and journalists to conduct the analysis and explain trends in American behaviors, demographics, and opinions. In addition, GSS offers a convenient way to analyze, explore, extract and share.

Since GSS supplies the neat and direct dataset for all years, with tools that help you explore and save variables and analyze data online. In fact, we are pretty interested in working hours of people in America in 2018. Hence. We aim to explore the number of hours people, working at all jobs, last week on average. We select 4 variables, which are closely associated with the working hours and the relative topic.

Similar dataset

As a matter of fact, we can obtain numerous datasets in terms of our topic. But still we select to utilize the dataset created by GSS. Reasons are as followed. First of all, GSS is kind of official institution for the government, which is no doubt more reliable when compared with other datasets supported by some private local institutions. Meanwhile, the data resources are more stable and strict. statistically, GSS has higher rate of receiving the respondents than any other institutions. Besides, normally, any dataset provided by GSS has more than amount of 10,000 data, which is quite a lot in this area. Above all, we determine to select this stable, strict and reliable resources, given there are a lot more dataset regarding our topic.

In this study, we will clean and summarize the data obtained on the official website of GSS in 2018 and primarily analyze the relationship between the proportion of different working hours and the proportion of the population from the data distribution perspective. On top of this, we will use the auxiliary information provided by GSS to find out the proportion of different working hours and the reasons behind it from the perspective of education level and income and make some hypotheses to assist our experiment. Since the total number of respondents is more than 10,000, and each person has different working hours, we will divide the collected receipts into 10-hour units for subsequent analysis. In addition, we have developed a reproducible method for data extraction and analysis and summary. In addition, we have also introduced data obtaining methods in detail and developed a similar survey to further analyze the average weekly working hours of Americans in 2018.

Hypothsis

Before the experiment, we made an hypothesis. From the perspective of macro society, we believed that even if there were long hours in some particular industries, for the vast majority of population, the working hours of the whole society would appear in a relatively average normal distribution. We hypothesize a nearly perfect bell curve graph should be generated during the research. Nevertheless, we think that in the U.S, one of the most developed countries in the world, people do not need to work too hard for living. Thus, the we deduct the number of people who do less work is greater than people who works more, and there may have a right-skewed normal distribution.

Since the highly educated group is always a minority in society, we hypothesize that the less educated group will be the majority in any working time.

In addition, considering the national conditions of the United States, we believe that no matter in which time-length interval, the average income will be the mainstream of the region, and extremely low or high income will not appear in a particular interval.

Data cleaning

For sustaining our further analysis, we have to clean and process the data. To start with, we select four variables, describe the working hours of America and relative means of variable, to establish the new dataset, which are respectively `hrs1`, `earnrs`, `nateducy` and `finrela`. Then we utilize the *names* function to change their names with more understanding meaning.

Since the response variable working hours is numeric, which make us difficult to do explantory variable analysis. Hence, we first utilize the *tidyverse* function to transform the numeric hours to categorical. Meanwhile, we set up the criterion to rewrite the working hours. Specifically, we utilize "1" means 0-10 hours, "2" means 11-20 hours and so on, until "9" means 80-90 hours.

In addition, at the first beginning, we find that the dataset has numerous missing values, specifically as NA. In fact, we cannot make any constructable analysis, assume we cannot eliminate the existence of missing values. Hence, after we establish the new dataset, we simply apply the *na.omit* function to operate the new dataset. The final result is what we can utilie to conduct the further analysis.

We would exemplify the new dataset.

	hours	earn	education	income
4	4	2	2	4
7	4	2	1	2
10	4	1	2	4
11	4	1	3	2
12	2	1	1	1
14	4	1	1	2

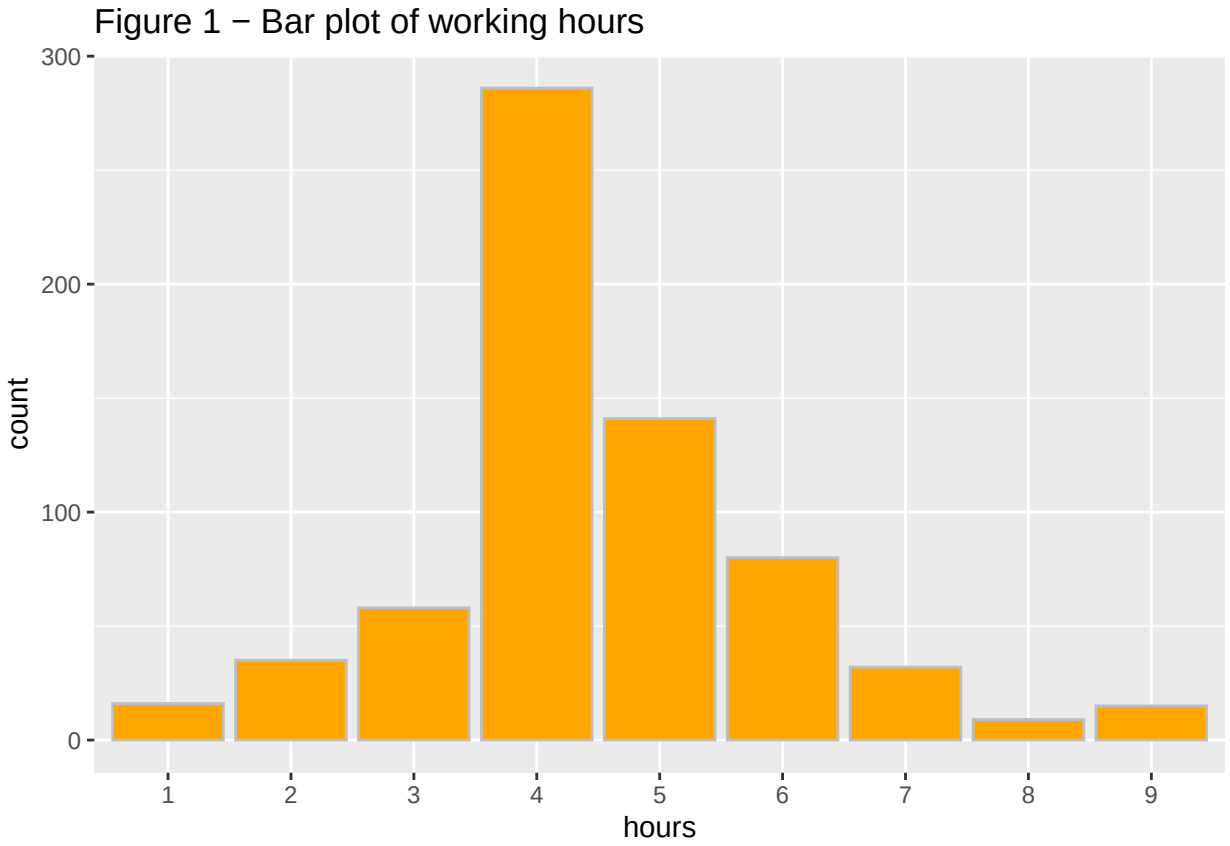
Specific variable

`hours`: This variable answers the question in the survey: How many hours did you work last week, at all jobs. It has 9 options, "1" means 0-10 hours and "9" means 80-90 hours in total.

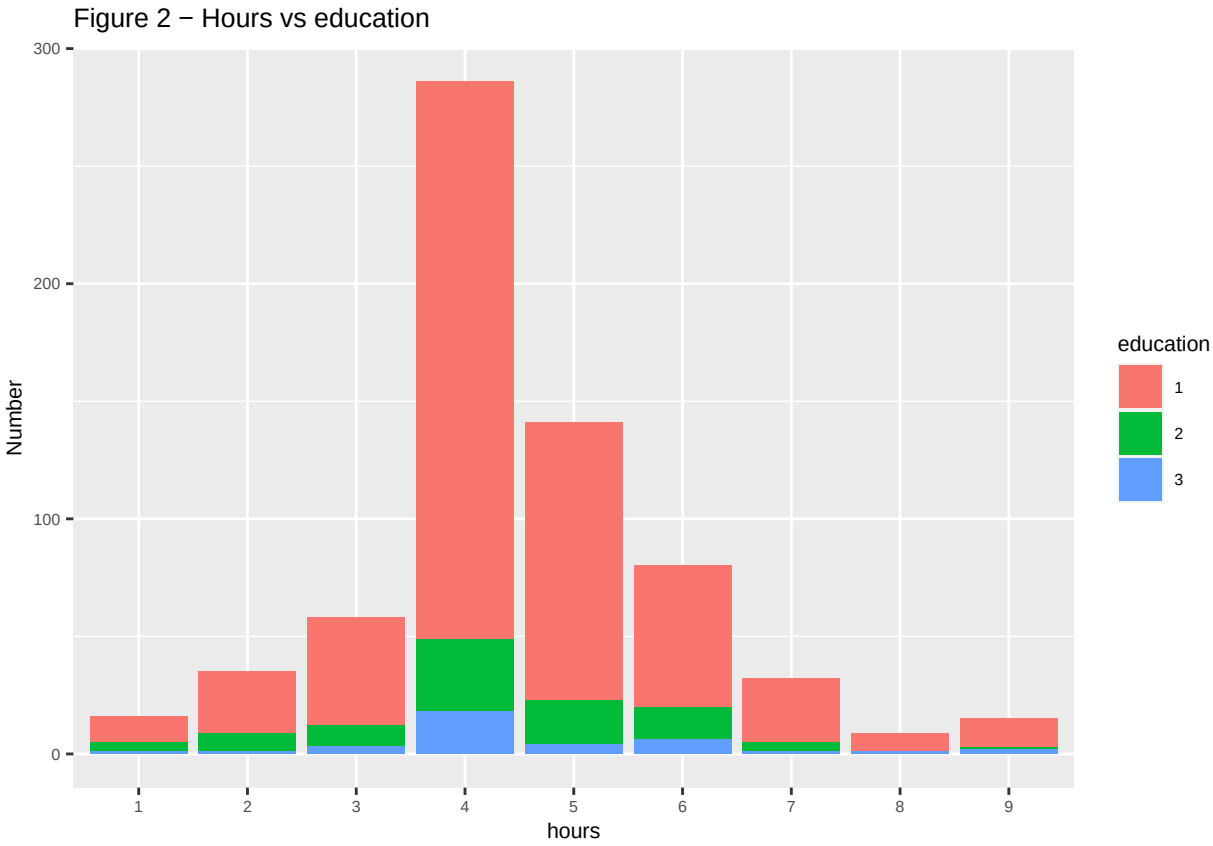
`earn`: This variable answers the question How many persons in the family (including yourself) earned any money last year (2017) from any job or employment. It is a numerical and consistent variable, which describes the number of people working in the family.

`education`: This variable answers the question about the satisfaction in terms of people's attitudes towards the domestic education supported by the government. It has three options. "1" represents too little, while "3" represents very good.

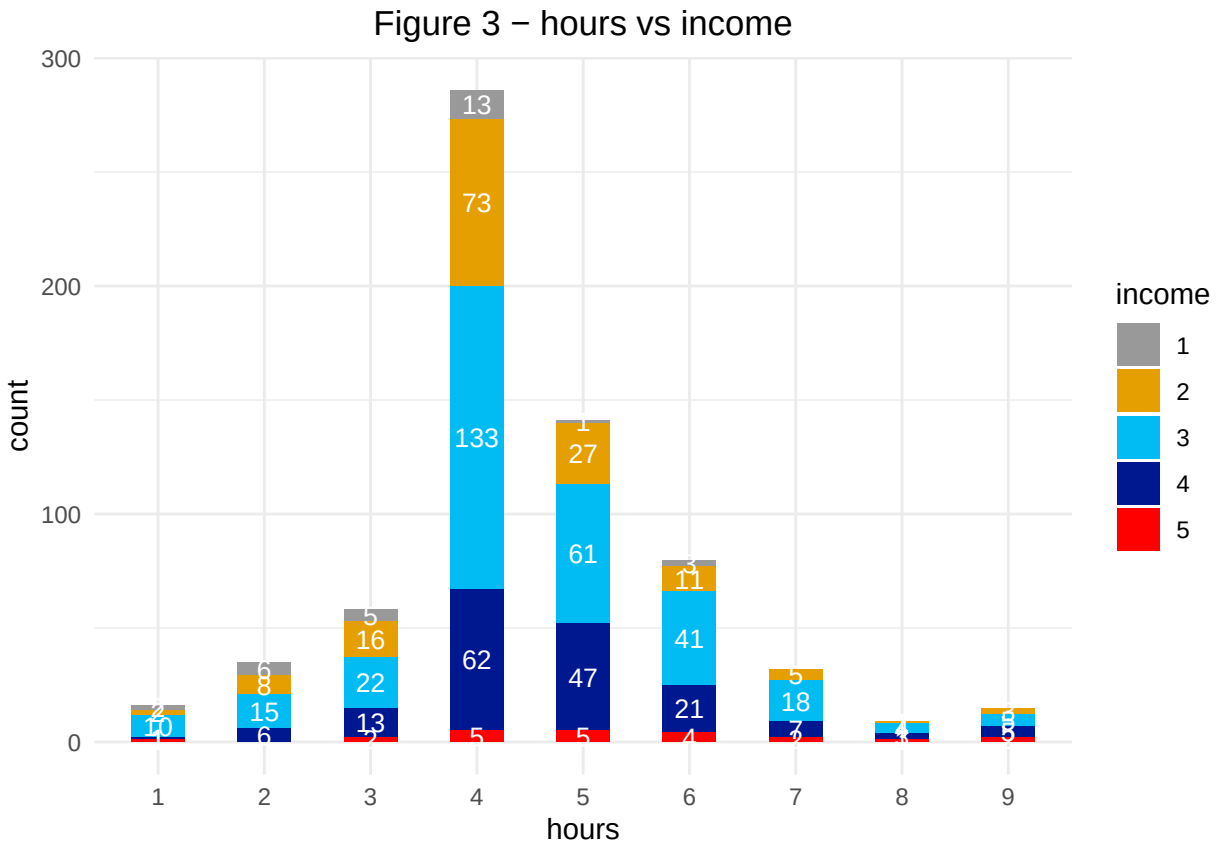
`income`: This variable answers the questions that compared with American families in general, would you say your family income is far below average, below average, average, above average, or far above average. The above is those five options.



According to figure1, we utilize the bar plot to describe the characteristics of `hours`. Apparently, it is symmetric and follows the normal distribution. We can judge that our response variable is quite good to be analysed. In addition, the majority of people just working approximately 4 hours per day on average and working at 8 hours has the least people.



As shown in the Figure2, It is the bar plot of working hours with the layer of education. The result is that whatever the people in the level of society, they all share the same point that the investment into the education is not enough. And the government needs to take care of this problem and take the measures to handle, given this common consensus.



As shown in the Figure3, it describes the relationship between the hours and income. The outcome is that majority of people is still with the average income, whatever how long they work. In fact, the working hours is closely related to the people's lifestyles. Analyzing the hours can help us get a more confound understanding of people common lives.

Result

Based on the information provided from the U.S. GSS on the number of hours worked per week in 2018 and the collated data, we found that the number of hours worked has a normal distribution with the population base.

According to figure1, we see a slightly right-skewed bell curve, and the left and right sides of the picture are basically symmetric. The x-axis presents the total working hours in one week from 0 to 10 hours to 90hours; there are nine sections in total, each of which is 10hours long. On the X-axis, the left-most and right-most sections are 0-10hours and 80-90hours, respectively. The Y-axis represents the corresponding numbers of people.

#The relationship between working hours and demographic

The people who work 30-40 hours are the most, accounting for about 40% of the total number. The second and third most are those who work 40-50 hours and 50-60 hours a week respectively, and the sum of the two indicators also accounts for nearly 30%. Based on these two points, we found that almost 70% of the population works from 30-to 60 hours a week, which is also the main reason for the normal distribution of the bell curve. On the other hand, those who work 70-80hours a week are the least. The second and third lowest hours are 0-10 hours and 80-90hours respectively, located on the left and right side of the image, respectively. The sum of the second and third highest values only reaches x% of the maximum value, that indicates it is a unimodal normal distribution. The other three data points are from the highest to the lowest for those working 20 to 30 hours a week, 10 to 20 hours, and 60 to 70 hours a week.

According to the above points, the three values corresponding to the total working hours and the population are respectively in the middle of the x-axis, and the highest values are the smallest and the leftmost of the three values, which is a characteristic of the right-skewed normal distribution. Based on this point, the three smallest values are located at the right and left sides of the image, respectively, which further verifies that this group of data appears in the above distribution pattern. In addition, the three intermediate values are in the middle of the highest and lowest values, respectively, and the highest values in these three groups are on the left side of the highest values, while the other two values are on the right side, which ultimately proves that this set of data followed by right-skewed distribution. Lastly, the extreme highest value is much greater than other high values, therefore, we conclude this data shows a unimodal right-skewed normal distribution, and it also verified that the hypothesis we made before the study was correct.

#Education level's proportion in every interval

Figure 2 is developed based on figure 1, introducing information about the education level in the different working length intervals. The education level is stratified into three leading groups, respectively represented in red, green, and blue, indicating education level from low to high. The majority of students with education level 1 take up the highest proportion Education Level 2 and 3 are very few. There is no exact interval in which the highly educated proportion of education is dominated by the highly educated. The overall level of education is almost the same regardless of the range. As can be seen intuitively from Figure 2, except for the 70-80 hours interval, the education level distribution of basically every working hour shows that education level1 is the largest, Education Level2 Next, education Level3 minimum. This fits very well with our hypothesis. In terms of proportion, education level2 accounts very few in the interval from 70-80hours and 80-90hours. The proportion of education Level3 is very less in the interval of 0 to 20 hours and 70 to 80 hours. The interval of working hours from 70-80 hours does not contain anyone from education level 2.

#The Income proportion

We classified income levels into five levels, respectively far below average, below average, average, above average, or far above average. The above five options correspond to the numbers 1 through 5 to observe with higher efficiency. No matter which interval works for a long time, it can be intuitively found that the middle class's average income is the majority. The average income accounts for 46.1% of the total, nearly equal to the sum of the other four. The above Average income and below Average income also account for 25.6% and 21.6%, respectively. These three together account for around 92% of the total. Far Above Average is the least in all the income, accounting for only 3%. The maximum portion of average income appears from the interval of 40-50 hours, but the highest proportion of that was found in 0-10 hours; in this range, average income accounted for as high as 62.5%. The interval contains the most proportion of far below average income is 10-20 hours at17%. Most of the above-average income appear from 30-60 hours, but in the interval of 80-90 hours, it reaches its highest proportion at 35.7%. The highest proportion of below average income is in 20-30 hours, at nearly 28%.

Besides that, we observed that the proportion of far below average income gradually decrease with the increase of working hours, and after the working hours exceed 60-70 hours a week, it disappears.

Discussion

Based on the information above, most of our hypothesis matches with the real situation after analyzing the result from this data set.

The result is roughly consistent with the hypothesis proposed before our experiment for the distribution type. Yet, although a normal distribution follows this data set, it still has a right-skewed trend. By comparing an ideal normal distribution type, that indicates the majority of people tend to work less. The increasing amplitude from 20-30 hours a week to 30-40 hours a week is exceptionally high, and then there exist a drastically decrease from 30-40 hours a week to 40-50 hours a week. We can infer from this that this is the overwhelming majority of Americans' favourite mode of work in 2018. Furthermore, in an ideal normal distribution situation, the number of people should gradually decrease when the working hours increasingly approaching to a high numbers. However, from 70-80 hours a week to 80-90 hours a week, the number has

gone up, which should not happen; this could be an error attributed to the sampling size being not big enough. The problem that caused the people who worked long hours to be too small to have statistical significance is also because of the sampling size. Although the sampling size exceeds ten thousand, it still cannot represent the entire population in the U.S in 2018.

The education level does not show too much influence on this research. The education level remains the same in any interval. Particularly in working hours form 70-80 hours a week, there has no information about it, that is caused by the data sampling size. However, the overall relationship between education and working hours/population matches our hypothesis.

The income is stratified at any interval. On the whole, the variation in each interval is slight. This result is consistent with our hypothesis to a great extent. The only difference is that there is no far below the average income for people who work 60-90 hours a week. We can see that the people who have a far below average income do not exist in any working time from 60-90 hours a week, although this is reasonable because people should not earn too less after doing many works, there still should have some outliers these intervals, which is due to the sampling size. Moreover, although the overall proportion in each interval does not change much, statistically speaking, except the two extreme values on the left and right are too few samples. The shorter hours worked, the more blow average income and far blow average income exist. Reversely, the more hours worked, the more above average income and far above average income appear. We did not expect that in our hypothesis.

Lastly, we can conclude that the entire distribution trend between population and working hours followed by a normal distribution. The education level remains the same in any working time length from 0 to 90 hours a week, and the total number of hours a person works per week does not significantly affect the overall income structure.

Appendix

Supplementary Survey

Survey's brief Introduction In order to make the investigation more effective, we generated a supplementary survey about the working hours as well.

The survey is designed on Google Form; users can easily use it to fill in their information without restriction. There are 11 questions in total, and we intend to use this survey to acquire basic information about one's working time length in a given period, education level, income situation, attitude toward his/her work and some other relevant information.

The URL Link below is the detail of the Survey; users can directly go to the Survey by clicking this URL.

Survey's URL :

https://forms.gle/uoGnHiNWHqYoGTqd6

Reference

University of Chicago, NORC at the. 2018. *Knitr: Individual Year Data Sets (Cross-Section Only)*. https://gss.norc.org/Documents/stata/2018_stata.zip.