ECON - 540

Kushal Chaudhari

Project Report

How high school students use time: a visual essay

Instructor's Name

Dr. Tristan Potter

INTRODUCTION

Students in high school have many things to keep up with in today's world. They have to run on a very tight schedule. From waking up early and being at their classes on time, to scrambling pockets of it to keep up with their health, leisure and social activities. These demands are also affected by factors like the age and sex of an individual. Data from the American Time Use Survey (ATUS) show how much time, on average, high school students devote to various activities per day in the United States. These activities include leisure activities, sleeping, various household activities, time spent on work, and playing sports. ATUS data also reveal differences in the way students' use of time on weekdays compared to weekends along with what months were they mostly in schools or which days were holidays in a given year.

High school students are an interesting demographic because many consumer companies are looking to market their products directly to them. And one of the critical factors that businesses consider is how these students allocate time. That would tell them what they like doing, how do they usually do it. In what directions are the trends going that can be capitalised on and so on.

This project aims to replicate a few results from a paper titled "How high school students spend time: a visual essay" by Mary Dorinda Allard, an economist at the US Bureau of Labor Statistics. Three graphs are replicated from the paper. After which there is an extension that answers a few important questions that interested me personally for this demographics.

DATA SOURCES AND VARIABLES OF INTEREST

The survey is administered to individuals age 15 and older. The respondents are asked about the activities they performed "yesterday". The survey includes information about the respondent's primary activities that they carried out during the previous day. For this essay, the data is restricted to the years 2003 to 2007 obtained for the months from September through May which is when most high school students attend school. All the data in this essay refers to students between the ages 15 and 19 who were enrolled full-time in high school. The data set is downloaded from the US Bureau of Labor Statistics website. It is referred to as a multi-year dataset on the website available from 2003-2020.

The ATUS respondent file and the ATUS activity file is then merged. To do that the data is first sorted by a variable named tucaseid which is assigned to each respondent interview and then merged with another file. This is done using the command - 'use merge 1:1 tucaseid using file_2.dta '. The dataset created is then saved and used as a default for the replication exercises and the extension.

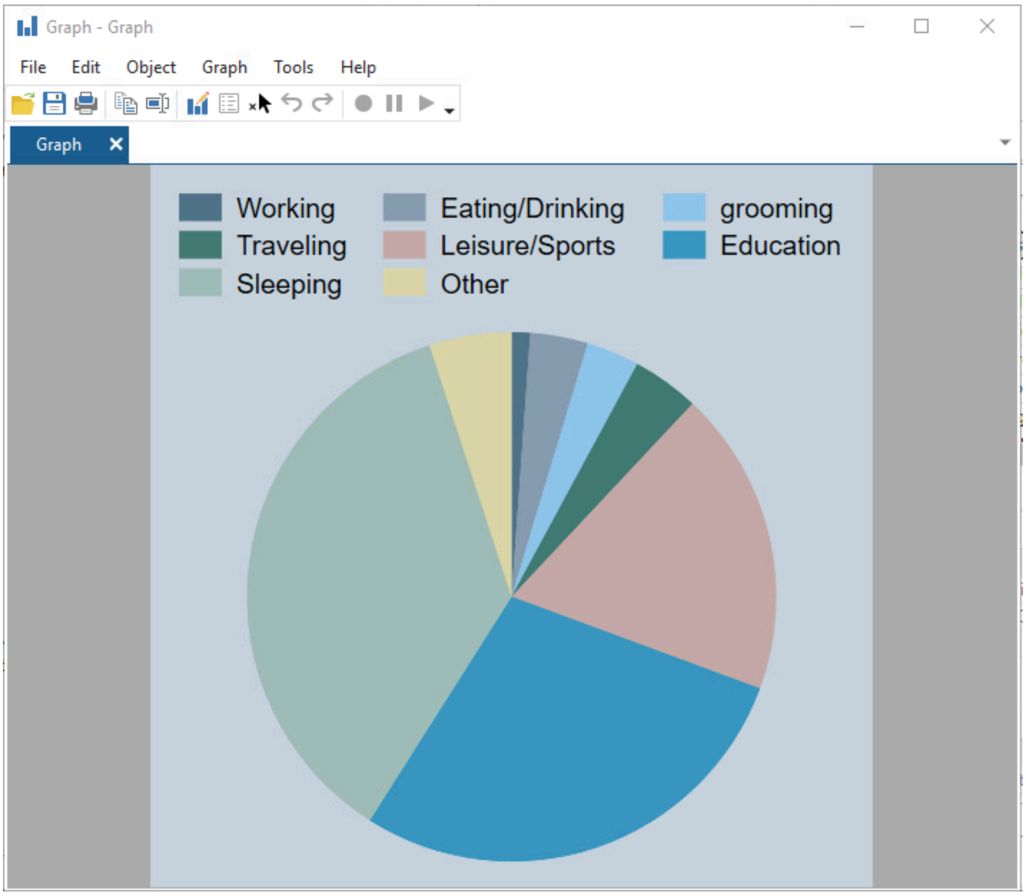
Since the variables of interest are not mentioned in the paper, one has to ask questions like - What is the demographics one is interested in? What does the data look like? Which variables are needed to look at high school students working on a particular day? The questions by searching through the variables in the ATUS dictionaries for multiyear datasets provided on the US Bureau of Labor Statistics Website which answer those questions. Along with it, the ATUS lexicon website can be used in conjunction with categories of variables storing time spent on different activities. After finding the variables, one has to take into account the coding of the variables as given in the data dictionaries. Using this information, we begin to clean our data by eliminating the redundant information and answer the question. Using the process mentioned, below is the summary of some of the important variables that are used in this paper.

Coding of that Variable in the ATUS data set	Value of Interest	Summary of the variable/ Question answered.
teschnr	1	Respondent's enrolment in high school, college or university.
teschlvl	1	Is it high school or college?
teage	15 - 19	Age of respondent
tumonth	September - May	Month of the day respondent was interview
tuyear	2003 - 2007	Year of the day respondent was interviewed
peeduca	34 - 38	What is the highest level of schooling for respondent or degree received?
trholiday	0	Flag to check if the diary day is a holiday
tudiaryday	2 - 6	Day of the week
teschft	1	Flag to check full time status of the student
tesex	Male or Female	Sex of the respondent
t01*	Time in minutes	Category for time spent on personal care activities from ATUS time use multi-year dataset's activity coding lexicon.
t01*	Time in minutes	Time spent on work or work related activities.

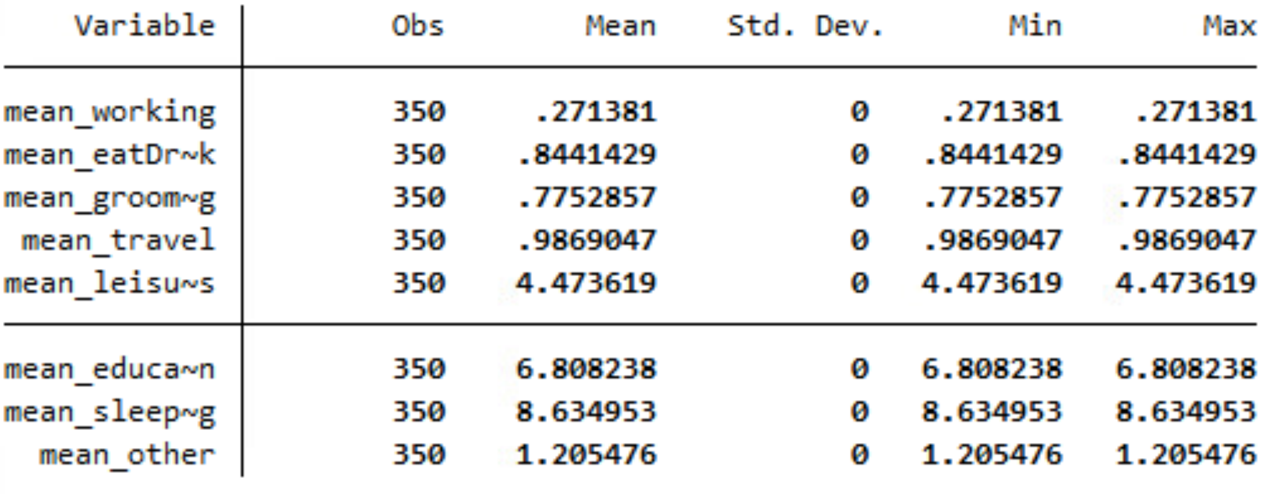
REPLICATION 1

HOW HIGH SCHOOL STUDENTS ALLOCATE TIME FOR DIFFERENT ACTIVITIES

The replication is a pie chart representing the amount of time that high school students allocate time to various activities on an average school day. The estimates are for months of September through May, from the year 2003 to 2007. Schooldays are non-holiday weekdays. And the demographic is high school students of age group 15 to 19 who are enrolled full-time.



The data is cleaned by keeping and dropping the variables to get the required demographics. The variables teschnr, teschlvl, teschft are set to 1 to filter students who attend high school and are enrolled full-time. We keep the data if the respondent is between 15 years or 19 years of age using the variable teage. We drop all the data for when students have holidays by using variables tumonth, tudiaryday and trholiday. And since we are interested in years 2003-2007, we drop the remaining data. Then we combine it into a single variable for that activity. Then we find the average time spent on that activity by taking a mean. The time in minutes is converted into hours by dividing it by 60. We then graph the results.



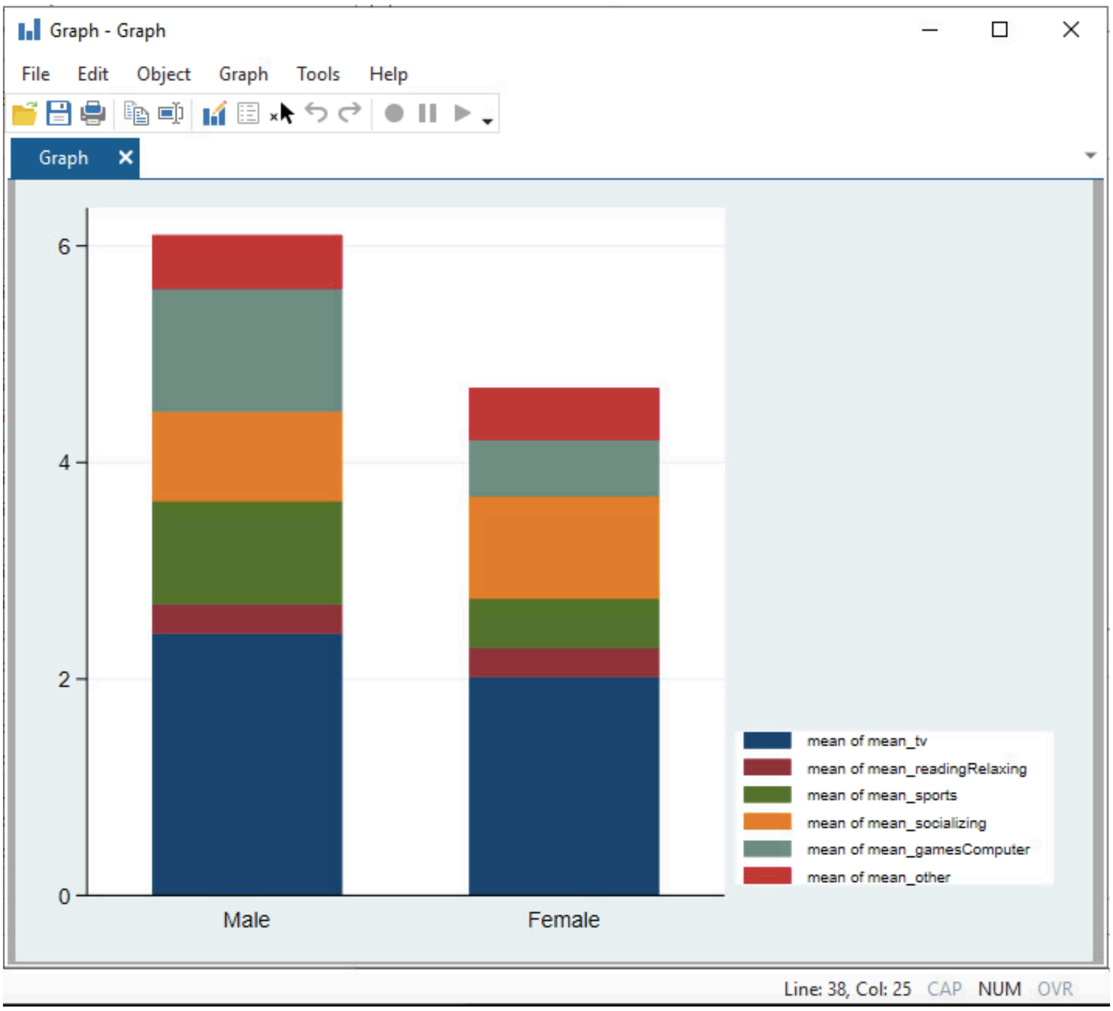
```
    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
mean_working |        350     .271381           0    .271381    .271381
mean_eatDr~k |        350    .8441429           0   .8441429   .8441429
mean_groom~g |        350    .7752857           0   .7752857   .7752857
 mean_travel |        350    .9869047           0   .9869047   .9869047
mean_leisu~s |        350    4.473619           0   4.473619   4.473619
-------------+---------------------------------------------------------
mean_educa~n |        350    6.808238           0   6.808238   6.808238
mean_sleep~g |        350    8.634953           0   8.634953   8.634953
  mean_other |        350    1.205476           0   1.205476   1.205476
```

We can see from the graph as well as the statistics above, on average, students slept for 8.6 hours and performed educational activities, such as attending class and doing homework, for about 7 hours. Remaining time was used among a range of activities such as travelling (almost an hour), leisure and sports (4.5 hours), grooming (0.77 hours), eating and drinking(0.84 hours); working (0.27 hours); and other activities, such as volunteering, shopping, and doing household activities for about 1.2 hours. The values are a bit off from the paper because the variables are not mentioned in the paper. One could never truly guess all the variables that the author might have looked at while replicating to get the results obtained. This is why the variables in sense are guesses as to what variables the author might have looked at to answer the question we seek an answer to.

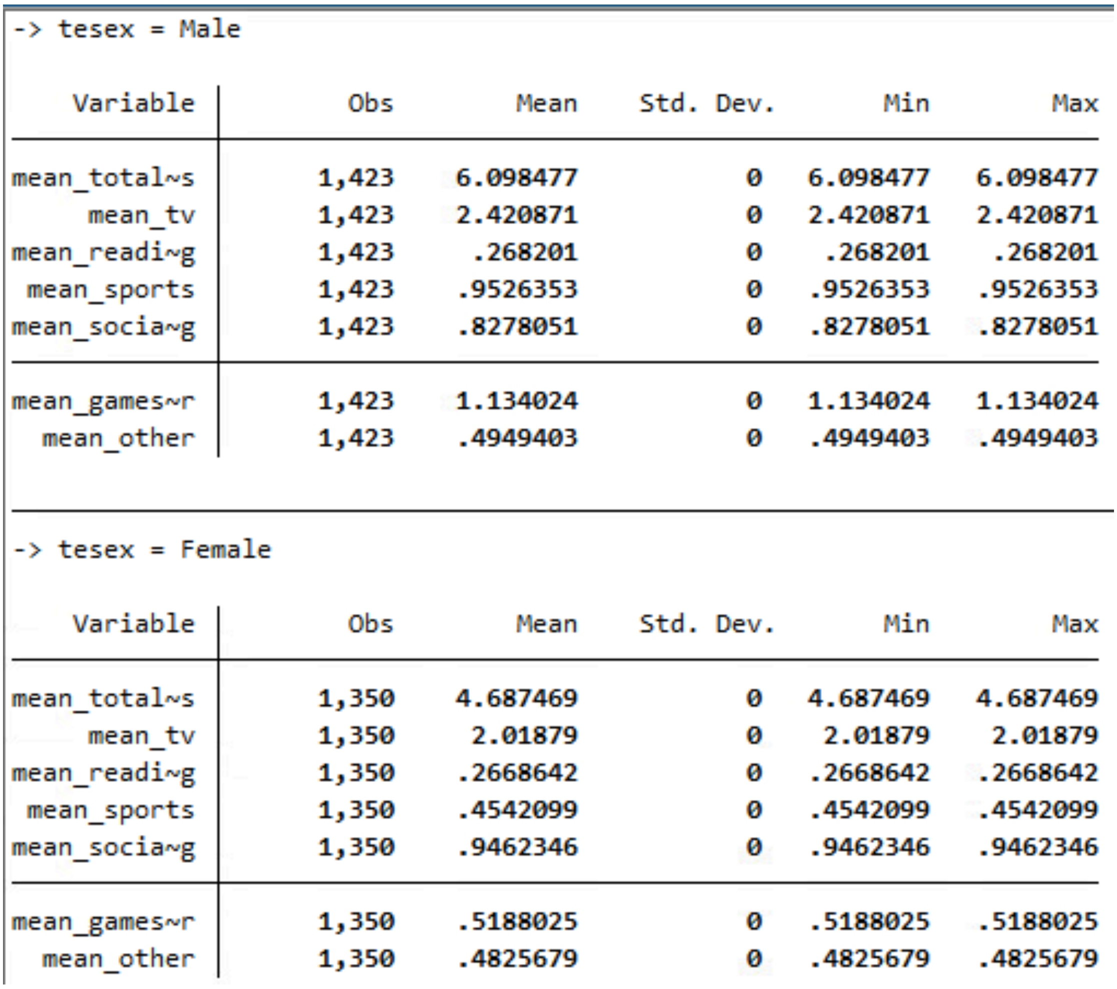
REPLICATION 2

HOW MALE AND FEMALE HIGH SCHOOL STUDENTS SPENT THEIR TIME DIFFERENTLY ON VARIOUS ACTIVITIES

The replication is a bar graph. The question we seek to answer is how differently do male and female high school students allocate their time to different leisure activities. The estimates are for months of September through May, from year 2003 to 2007. Schooldays are non-holiday weekdays. And the demographic is high school students of age group 15 to 19 who are enrolled full-time.



The data is cleaned by keeping and dropping the variables to get the required demographics. The variables teschnr, teschlvl, teschft are set to 1 to get students who attend high school and are enrolled full-time. We keep data if the respondent is between 15 years of age or 19 years using variable teage. And since we are interested in years 2003-2007, we drop the remaining data. Then we combine various categories under a single variable for that activity. Then we sort the data by sex and then find the average time spent on that activity by taking a mean and graphing it. The time in minutes is converted into hours by dividing it by 60.



```
-> tesex = Male

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
mean_total~s |      1,423    6.098477           0   6.098477   6.098477
     mean_tv |      1,423    2.420871           0   2.420871   2.420871
mean_readi~g |      1,423     .268201           0    .268201    .268201
  mean_sports |      1,423    .9526353           0   .9526353   .9526353
mean_socia~g |      1,423    .8278051           0   .8278051   .8278051
-------------+---------------------------------------------------------
mean_games~r |      1,423    1.134024           0   1.134024   1.134024
  mean_other |      1,423    .4949403           0   .4949403   .4949403

-----------------------------------------------------------------------
-> tesex = Female

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
mean_total~s |      1,350    4.687469           0   4.687469   4.687469
     mean_tv |      1,350     2.01879           0    2.01879    2.01879
mean_readi~g |      1,350    .2668642           0   .2668642   .2668642
  mean_sports |      1,350    .4542099           0   .4542099   .4542099
mean_socia~g |      1,350    .9462346           0   .9462346   .9462346
-------------+---------------------------------------------------------
mean_games~r |      1,350    .5188025           0   .5188025   .5188025
  mean_other |      1,350    .4825679           0   .4825679   .4825679
```

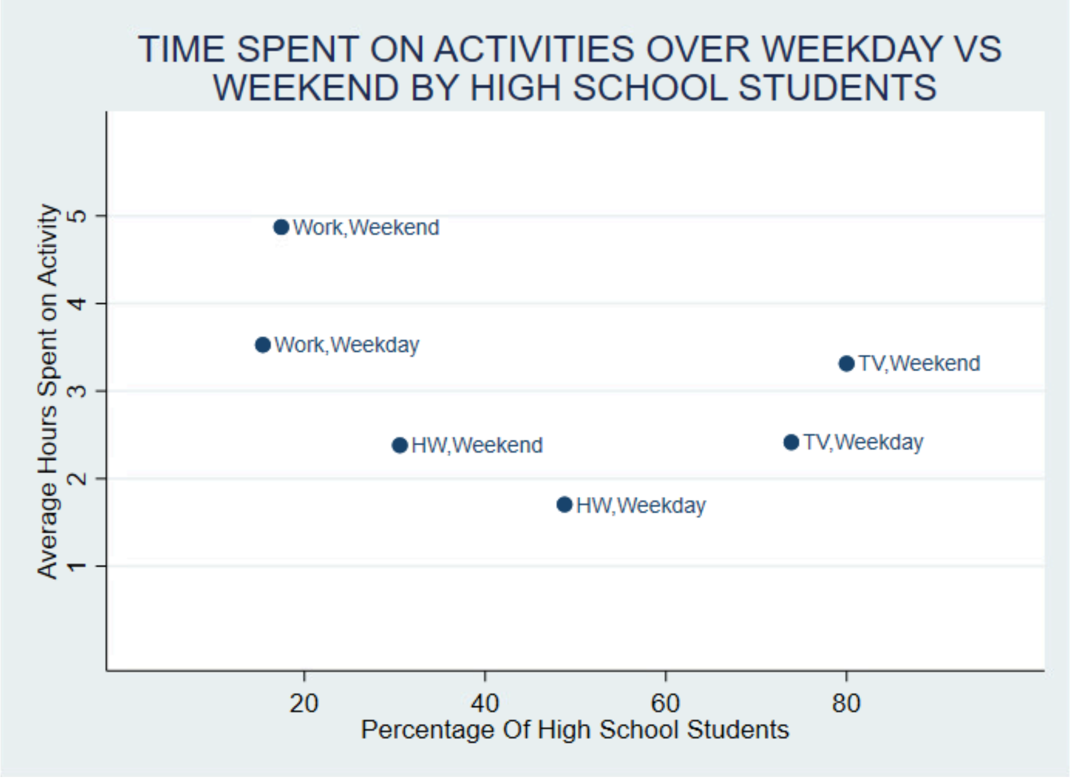
We can see from the statistics above that on average, male high school students spent 1.3 hours more doing leisure activities on an average day than their female counterparts (6.0 hours, compared to 4.7 hours).

Also, male high school students spent more time than female high school students watching TV (2.42 hours compared to 2 hours), playing games and/or using a computer (1.13 hours compared to 0.51 hour), and doing sports activities (0.95 hour compared to 0.45 hour). Female high school students spent slightly more time socializing (1 hour) than their male counterparts (0.82 hour).

REPLICATION 3

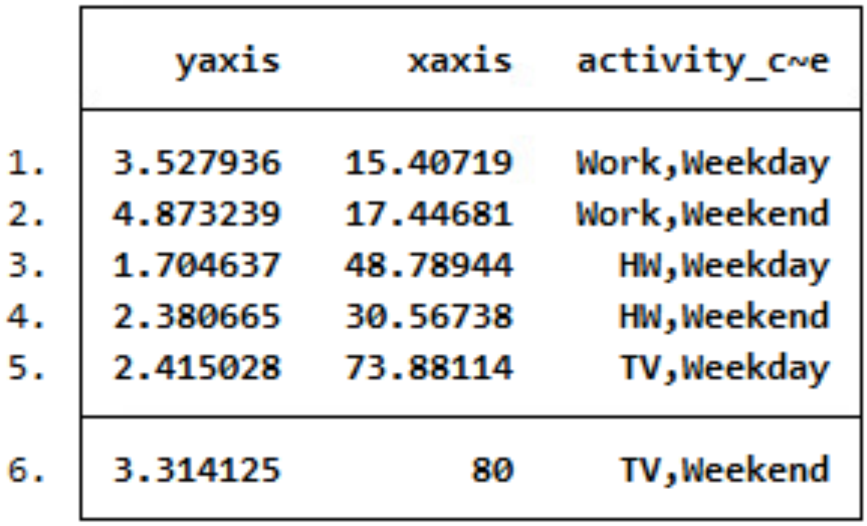
HOW TIME SPENT ON DIFFERENT ACTIVITIES VARIED FOR WEEKDAYS COMPARED TO WEEKENDS FOR HIGH SCHOOL STUDENTS

The replication is a bar graph. The question we seek to answer is how differently do high school students allocate their time to different activities for weekends compared to the weekdays. The estimates are for months of September through May, from year 2003 to 2007. Schooldays are non-holiday weekdays. And the demographic is high school students of age group 15 to 19 who are enrolled full-time.



The data is cleaned by keeping and dropping the variables to get the required demographics. The variables teschnr, teschlvl, teschft are set to 1 to get students who attend high school and are enrolled full-time. We keep data if the respondent is between 15 years of age or 19 years using variable teage. And since we are interested in years 2003-2007, we drop the remaining data. Then we create a variable weekday that is essentially a flag if the day of a week is a weekday. After summarising minutes spent by activity, we find mean of activities over weekends and weekdays.

This was by far the most difficult replication I had to do. The reason for it being hard was my inability to categorise and use the newer variables that I created. I eventually was able to collapse the data to focus on the variables I needed to work on. On top of that, for a lot of time, I was unable to group them so that I could make a scatterplot. I was able to overcome that when I used the stack command to rearrange data into a data structure suitable to build a scatterplot along with labelled individual categories as follows.



```
     +------------------------------------+
     |    yaxis      xaxis   activity_c~e |
     |------------------------------------|
  1. | 3.527936   15.40719   Work,Weekday |
  2. | 4.873239   17.44681   Work,Weekend |
  3. | 1.704637   48.78944     HW,Weekday |
  4. | 2.380665   30.56738     HW,Weekend |
  5. | 2.415028   73.88114     TV,Weekday |
     |------------------------------------|
  6. | 3.314125         80     TV,Weekend |
     +------------------------------------+
```

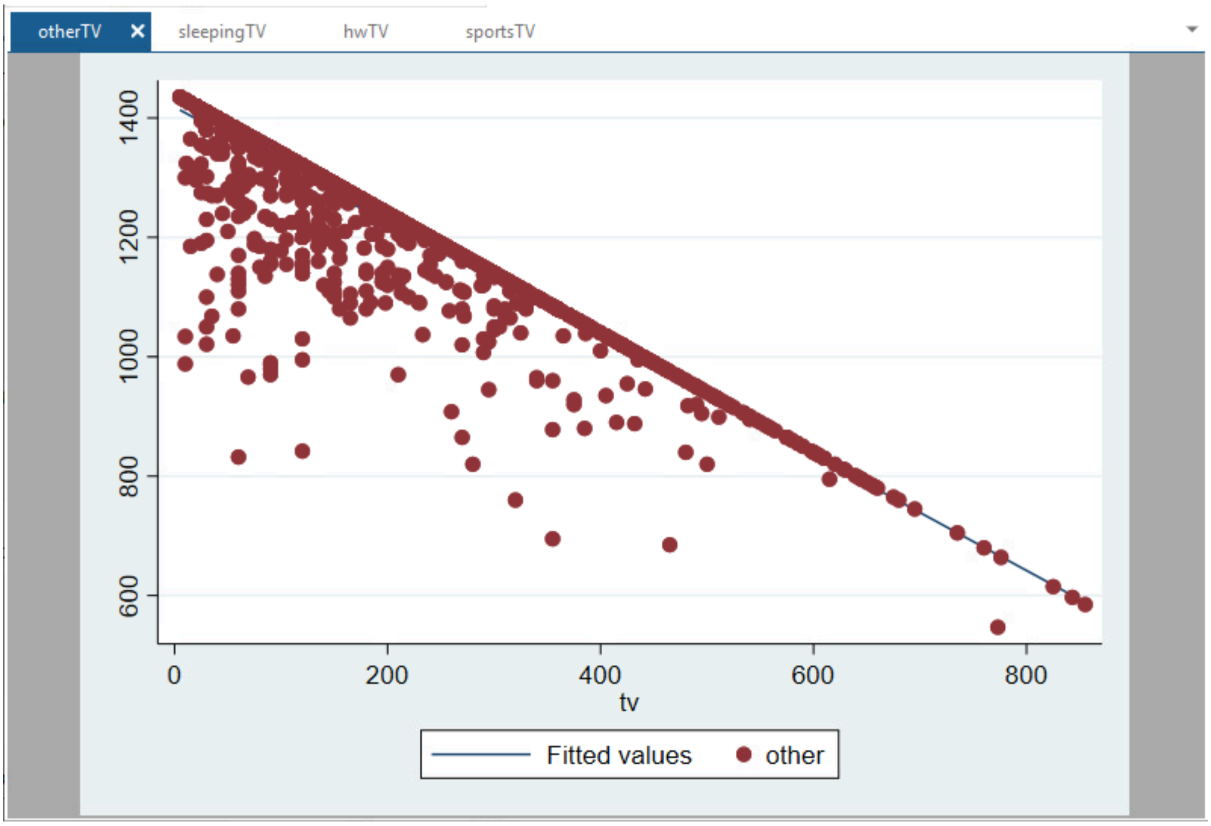
We can see from the statistics and the graph above that on average, about 16% of high school students worked on an average weekday and an average weekend. They spent more time working on weekends than on weekdays (4.8 hours compared to 3.5 hours).

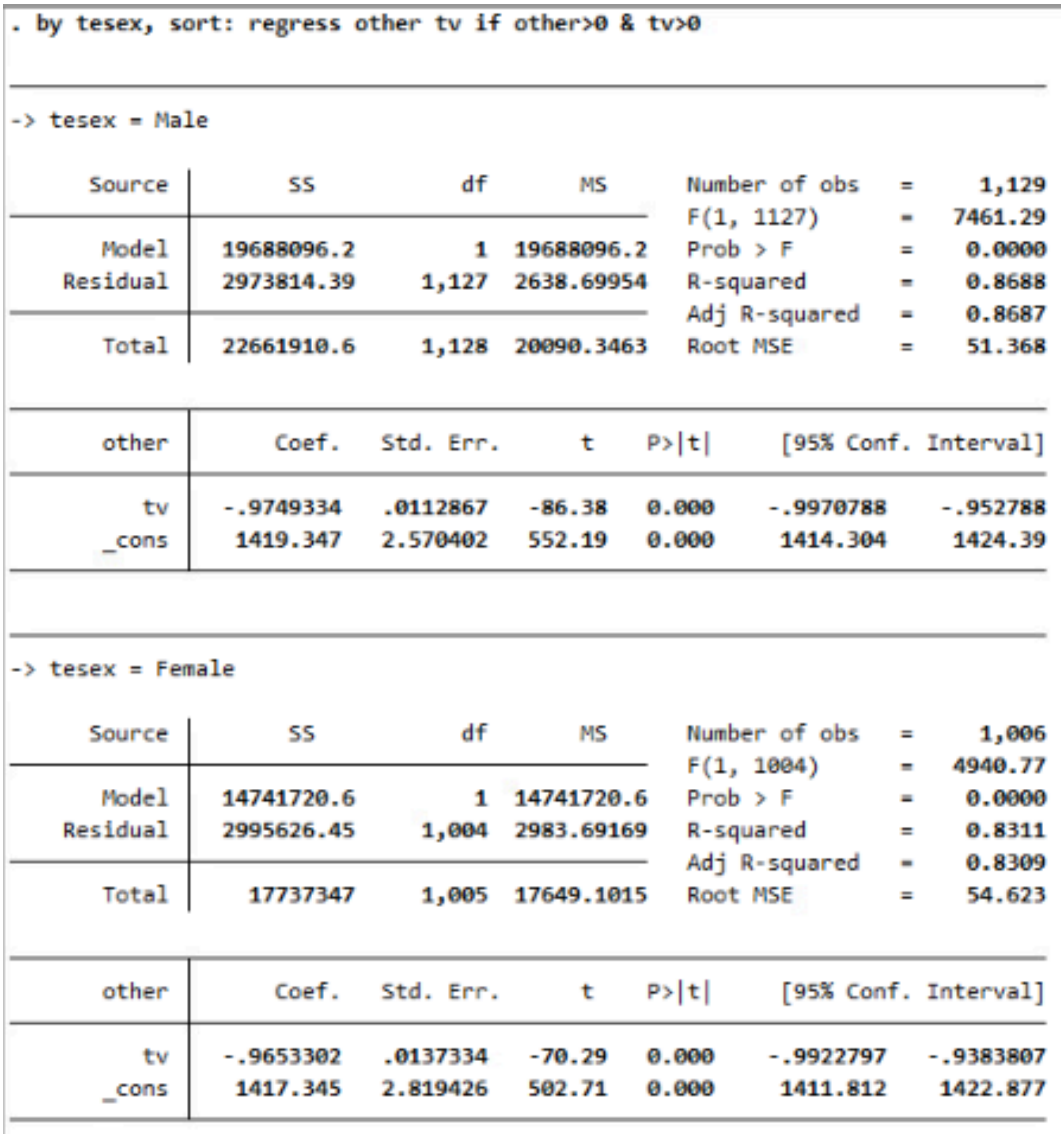
48.79% of high school students did homework on an average weekday, compared to 30% on an average weekend. On days that students did their homework, they studied for 2.38 hours on weekends and 1.7 hours on average on weekdays. Also, about 74% of high school students watched TV on an average weekday, compared to 80 % on an average weekend. High school students who did watch TV spent almost an hour more on weekends (3.3 hours) than they did on weekdays (2.4 hours).

EXTENSION

The extension I have chosen to extend the authors work is asking an economic question. The question is - What is the marginal impact on activities namely Sleeping, Doing homework, and Playing Sports for every incremental hour that high school students spend on watching TV. And how does this differ across demographics of high school students categorised by their sex.

I start the analysis by observing the answer to an obvious question. An hour spent more on watching TV should reduce the hour spent on other activities combined(combined in variable other using rowtotal functionality of egen command) by a factor of 1. This is achieved by regressing other(dependent variable) with tv(independent variable) if time was spent on that activity for male and female high school students. The result was as follows:

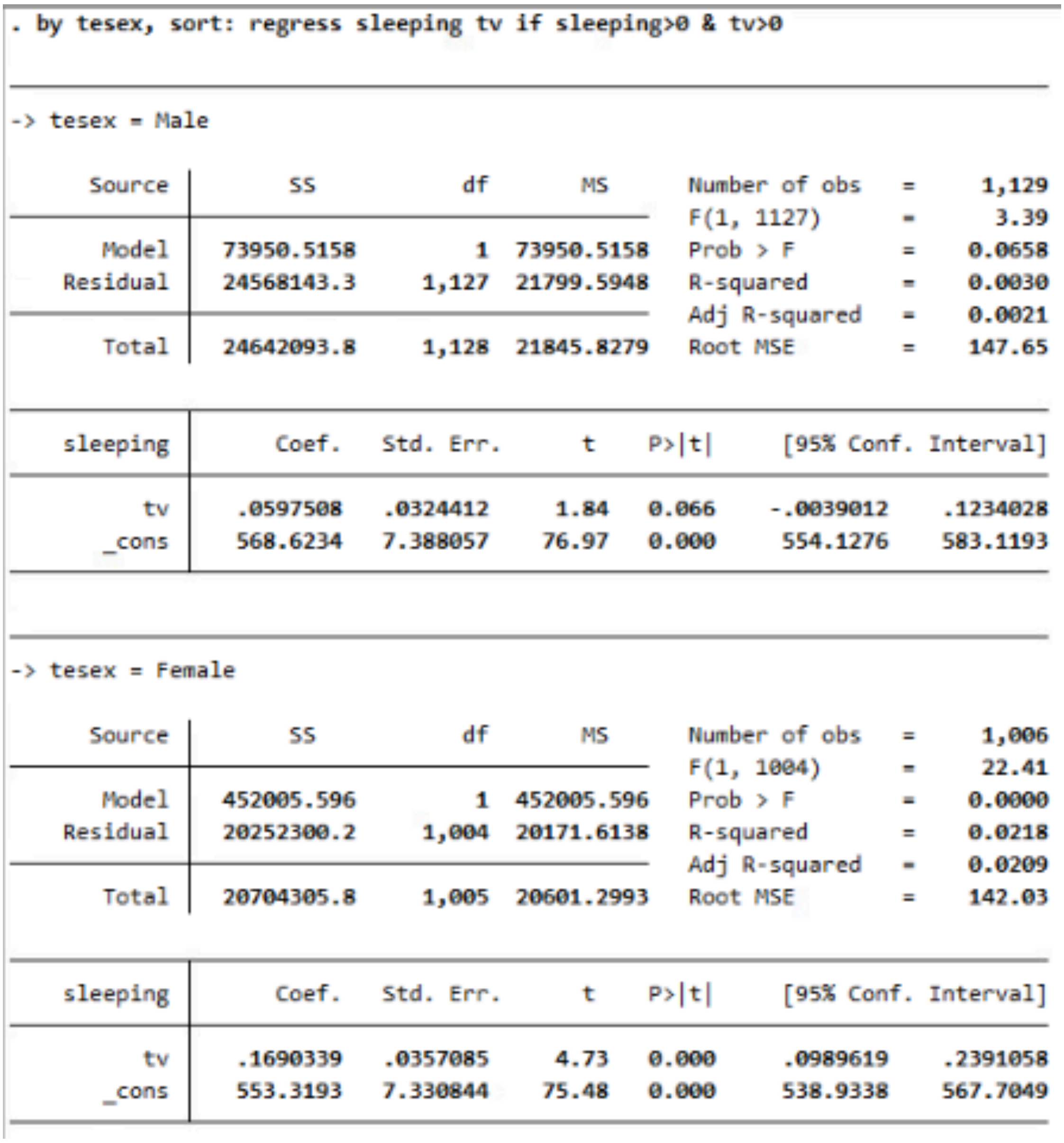




```
. by tesex, sort: regress other tv if other>0 & tv>0

-------------------------------------------------------------------------------
-> tesex = Male

      Source |       SS           df       MS      Number of obs   =     1,129
-------------+----------------------------------   F(1, 1127)      =   7461.29
       Model |  19688096.2         1  19688096.2   Prob > F        =    0.0000
    Residual |  2973814.39     1,127  2638.69954   R-squared       =    0.8688
-------------+----------------------------------   Adj R-squared   =    0.8687
       Total |  22661910.6     1,128  20090.3463   Root MSE        =    51.368

------------------------------------------------------------------------------
       other |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tv |  -.9749334   .0112867   -86.38   0.000    -.9970788    -.952788
       _cons |   1419.347   2.570402   552.19   0.000     1414.304     1424.39
------------------------------------------------------------------------------

-------------------------------------------------------------------------------
-> tesex = Female

      Source |       SS           df       MS      Number of obs   =     1,006
-------------+----------------------------------   F(1, 1004)      =   4940.77
       Model |  14741720.6         1  14741720.6   Prob > F        =    0.0000
    Residual |  2995626.45     1,004  2983.69169   R-squared       =    0.8311
-------------+----------------------------------   Adj R-squared   =    0.8309
       Total |    17737347     1,005  17649.1015   Root MSE        =    54.623

------------------------------------------------------------------------------
       other |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tv |  -.9653302   .0137334   -70.29   0.000    -.9922797   -.9383807
       _cons |   1417.345   2.819426   502.71   0.000     1411.812    1422.877
------------------------------------------------------------------------------
```

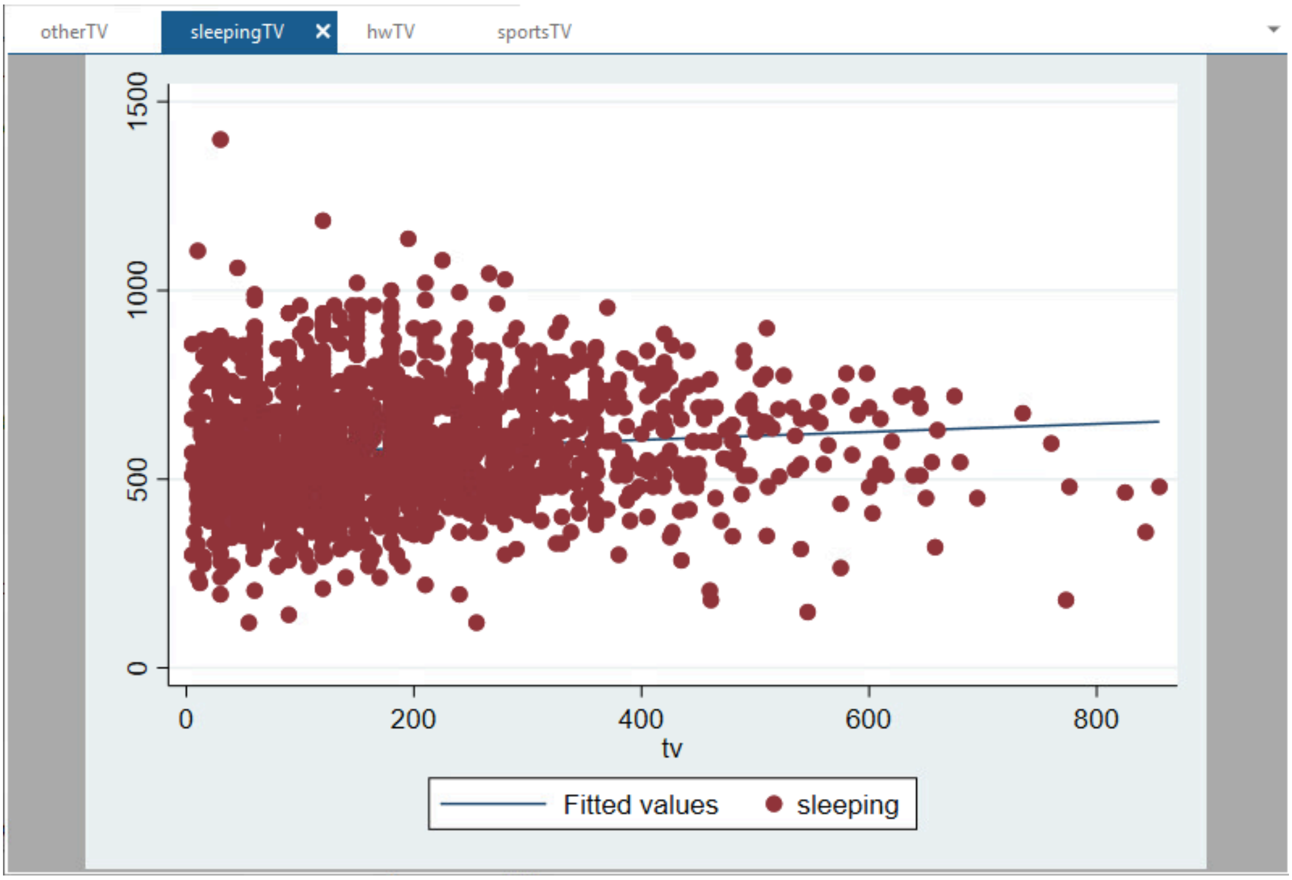
From the result we can see that the pvalue of F is very low i.e its <0.001 in both for male and female high school students. Hence we can reject the null hypothesis and accept the alternative hypothesis that model is statistically significant. Also R-squared = 0.8688(Male) and 0.8311(Female). We know when it is closer to 1 better the model(fit). In this case, about 86% for male and 83% for female variation in dependent variable is explained by independent variables. P-value of each t-test is very low i.e 0.001. Hence we reject the null(that states that the coefficient of independent variable is 0) and accept the alternative that the independent variables have a significant effect on dependent variable. Also, the coefficient for tv is negative, it indicates a strong negative correlation. In other words, one unit hour increase in time spent on watching TV will cause 0.97 hour decrease for male high school students and 0.96 hour decrease for female high school students doing other activities(dependent variable).

After verifying the obvious, I move on to find out the marginal effect of an additional hour of watching TV on time spent sleeping. This is achieved by regressing sleeping(dependent variable) with tv(independent variable) if time was spent on that activity for male and female high school students. The results were as follows:

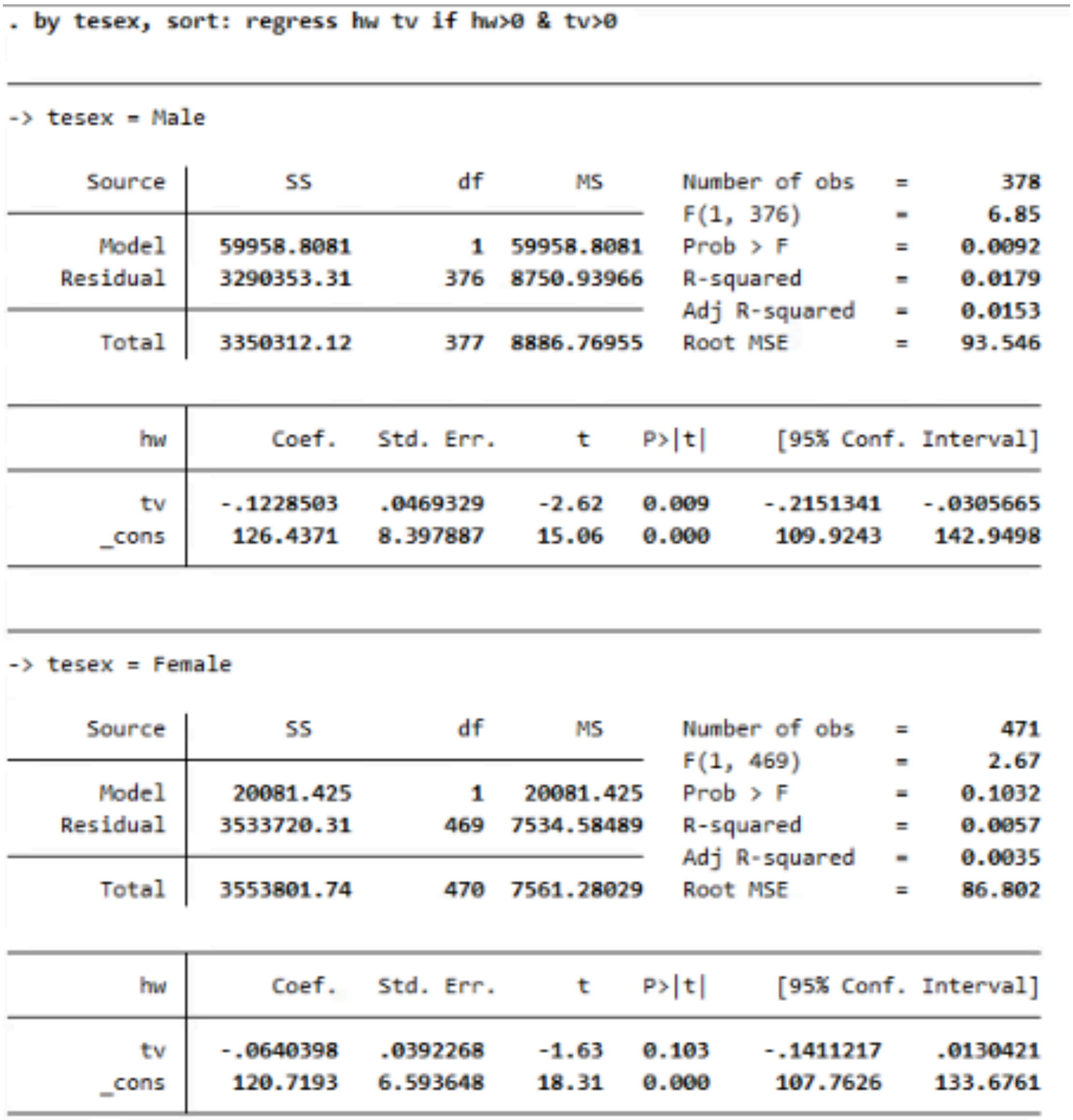


```
. by tesex, sort: regress sleeping tv if sleeping>0 & tv>0

-------------------------------------------------------------------------------
-> tesex = Male

      Source |       SS           df       MS      Number of obs   =     1,129
-------------+----------------------------------   F(1, 1127)      =      3.39
       Model |  73950.5158         1  73950.5158   Prob > F        =    0.0658
    Residual |  24568143.3     1,127  21799.5948   R-squared       =    0.0030
-------------+----------------------------------   Adj R-squared   =    0.0021
       Total |  24642093.8     1,128  21845.8279   Root MSE        =    147.65

------------------------------------------------------------------------------
    sleeping |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tv |   .0597508   .0324412     1.84   0.066    -.0039012    .1234028
       _cons |   568.6234   7.388057    76.97   0.000     554.1276    583.1193
------------------------------------------------------------------------------

-------------------------------------------------------------------------------
-> tesex = Female

      Source |       SS           df       MS      Number of obs   =     1,006
-------------+----------------------------------   F(1, 1004)      =     22.41
       Model |  452005.596         1  452005.596   Prob > F        =    0.0000
    Residual |  20252300.2     1,004  20171.6138   R-squared       =    0.0218
-------------+----------------------------------   Adj R-squared   =    0.0209
       Total |  20704305.8     1,005  20601.2993   Root MSE        =    142.03

------------------------------------------------------------------------------
    sleeping |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tv |   .1690339   .0357085     4.73   0.000     .0989619    .2391058
       _cons |   553.3193   7.330844    75.48   0.000     538.9338    567.7049
------------------------------------------------------------------------------
```

From the result we can see that the pvalue of F is very low in both cases - for male and female high school students. Hence we can reject the null hypothesis and accept the alternative hypothesis that model is statistically significant. Also R-squared = 0.003(Male) and 0.022(Female). We know when it is closer to 1 better the model(fit). But this case, only 0.03-0.2% of variation for male and female in dependent variable is explained by independent variables. So there might be other variables we may look at too that would better explain the variation. P-value of each t-test is very low i.e 0.001. Hence we reject the null(that states that the coefficient of independent variable is 0) and accept the alternative that the independent variables have a significant effect on dependent variable. Also, the coefficient for tv is positive, it indicates a weakly positive correlation. In other words, one hour increase in time spent on watching TV will cause 0.06(for male) 0.17(for female) hour increase on sleeping(dependent variable) for male high school students and female high school students respectively.

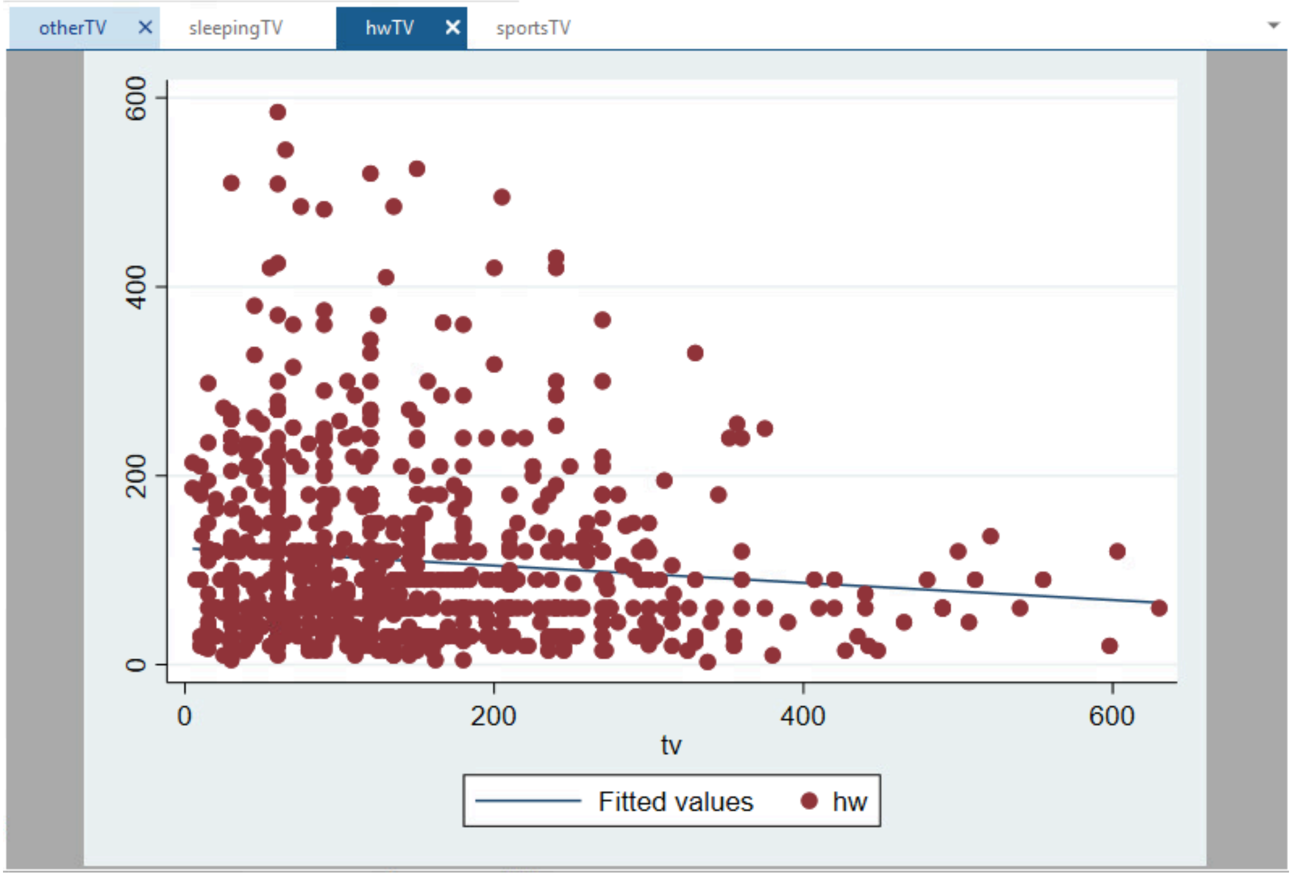


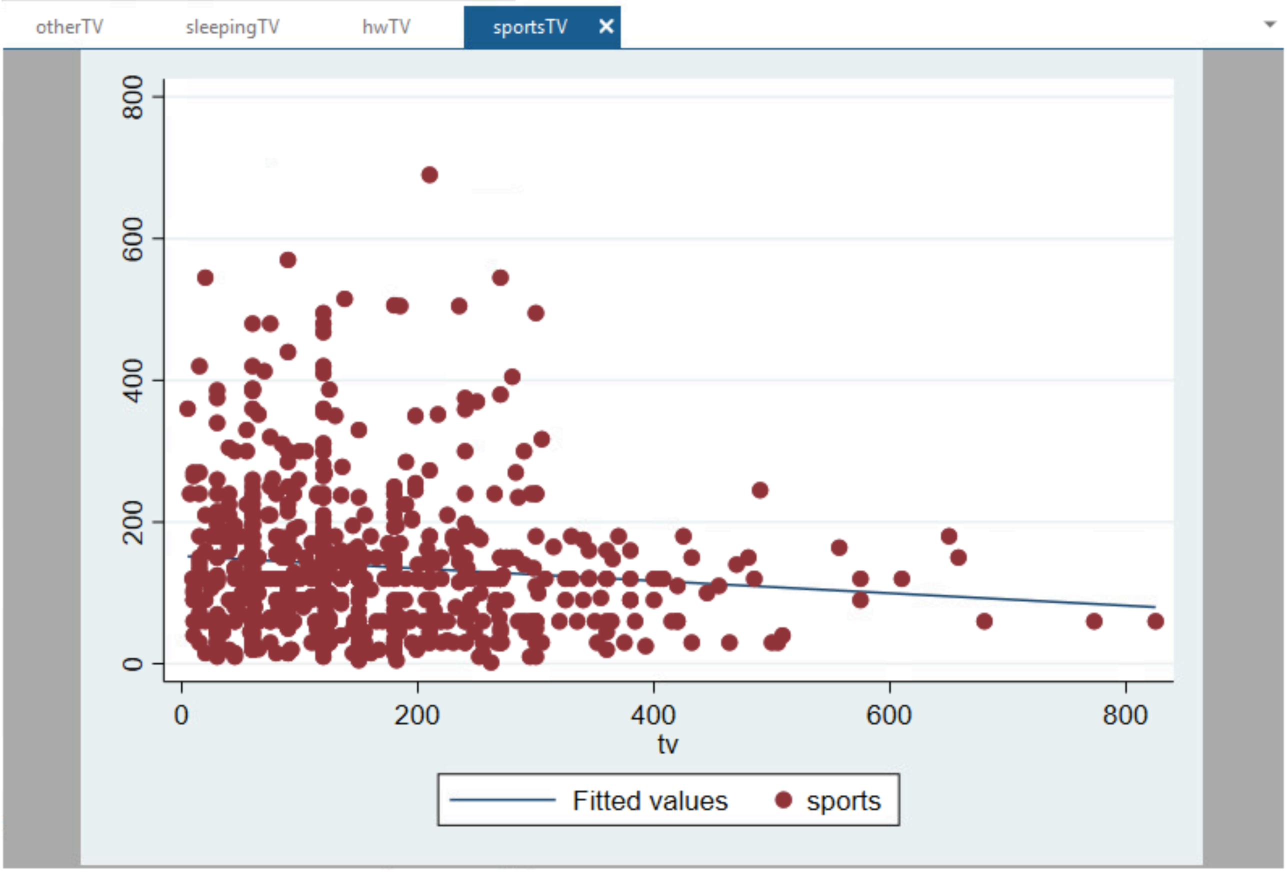
Similarly, we find out the marginal effect of an additional hour of watching TV on time spent doing homework.



```
. by tesex, sort: regress hw tv if hw>0 & tv>0

-------------------------------------------------------------------------------
-> tesex = Male

      Source |       SS           df       MS      Number of obs   =       378
-------------+----------------------------------   F(1, 376)       =      6.85
       Model |  59958.8081         1  59958.8081   Prob > F        =    0.0092
    Residual |  3290353.31       376  8750.93966   R-squared       =    0.0179
-------------+----------------------------------   Adj R-squared   =    0.0153
       Total |  3350312.12       377  8886.76955   Root MSE        =    93.546

------------------------------------------------------------------------------
          hw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tv |  -.1228503   .0469329    -2.62   0.009    -.2151341   -.0305665
       _cons |   126.4371   8.397887    15.06   0.000     109.9243    142.9498
------------------------------------------------------------------------------

-------------------------------------------------------------------------------
-> tesex = Female

      Source |       SS           df       MS      Number of obs   =       471
-------------+----------------------------------   F(1, 469)       =      2.67
       Model |   20081.425         1   20081.425   Prob > F        =    0.1032
    Residual |  3533720.31       469  7534.58489   R-squared       =    0.0057
-------------+----------------------------------   Adj R-squared   =    0.0035
       Total |  3553801.74       470  7561.28029   Root MSE        =    86.802

------------------------------------------------------------------------------
          hw |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tv |  -.0640398   .0392268    -1.63   0.103    -.1411217    .0130421
       _cons |   120.7193   6.593648    18.31   0.000     107.7626    133.6761
------------------------------------------------------------------------------
```

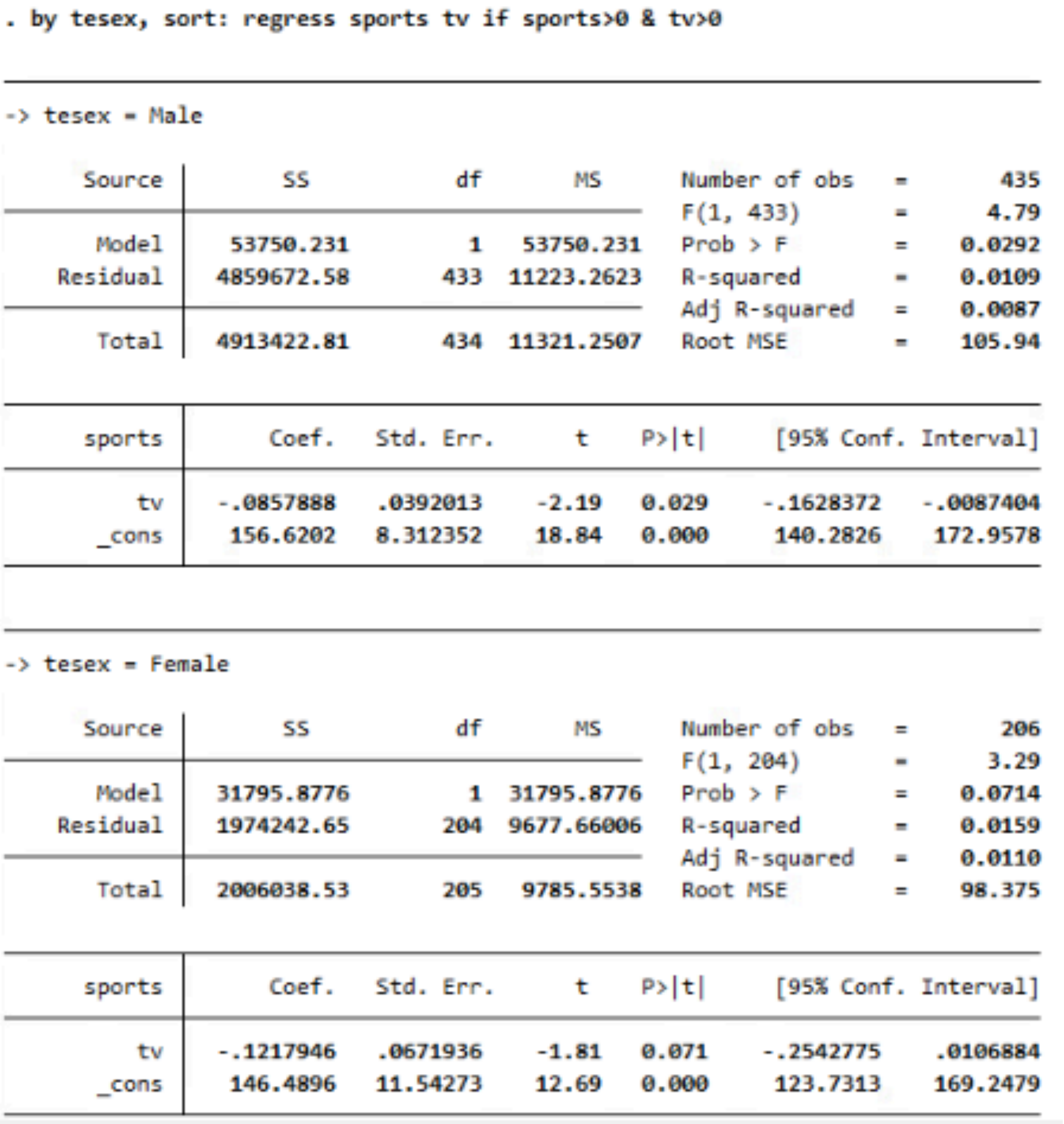
From the result we can see that the pvalue of F is very low in both cases - for male and female high school students. Hence we can reject the null hypothesis and accept the alternative hypothesis that model is statistically significant. Also R-squared = 0.018(Male) and 0.0057(Female). We know when it is closer to 1 better the model(fit). But this case, only 0.018-0.0057% of variation for male and female in dependent variable is explained by independent variables. So there might be other variables we may look at too that would better explain the variation. P-value of each t-test is very low i.e 0.001. Hence we reject the null(that states that the coefficient of independent variable is 0) and accept the alternative that the independent variables have a significant effect on dependent variable. Also, the coefficient for tv is negative, it indicates a weakly negative correlation. In other words, one hour increase in time spent on watching TV will cause 0.123(for male) & 0.64(for female) hour decrease on doing homework(dependent variable) for male high school students and female high school students.





Finally, in a similar way, we analyse the marginal effect of an additional hour of watching TV on time spent playing sports.

From the result below, we can see that the pvalue of F is very low in both cases - for male and female high school students. Hence we can reject the null hypothesis and accept the alternative hypothesis that model is statistically significant. Also R-squared = 0.0109(Male) and 0.00159(Female). We know when it is closer to 1 better the model(fit). P-value of each t-test is very low. Hence we reject the null(that states that the coefficient of independent variable is 0) and accept the alternative that the independent variables have a significant effect on dependent variable. Also, the coefficient for tv is negative, it indicates a weakly negative correlation. In other words, one hour increase in time spent on watching TV will cause 0.123(for male) & 0.64(for female) hour decrease on playing sports(dependent variable) for male high school students and female high school students.



```
. by tesex, sort: regress sports tv if sports>0 & tv>0

-------------------------------------------------------------------------------
-> tesex = Male

      Source |       SS           df       MS      Number of obs   =       435
-------------+----------------------------------   F(1, 433)       =      4.79
       Model |   53750.231         1   53750.231   Prob > F        =    0.0292
    Residual |  4859672.58       433  11223.2623   R-squared       =    0.0109
-------------+----------------------------------   Adj R-squared   =    0.0087
       Total |  4913422.81       434  11321.2507   Root MSE        =    105.94

------------------------------------------------------------------------------
      sports |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tv |  -.0857888   .0392013    -2.19   0.029    -.1628372   -.0087404
       _cons |   156.6202   8.312352    18.84   0.000     140.2826    172.9578
------------------------------------------------------------------------------

-------------------------------------------------------------------------------
-> tesex = Female

      Source |       SS           df       MS      Number of obs   =       206
-------------+----------------------------------   F(1, 204)       =      3.29
       Model |  31795.8776         1  31795.8776   Prob > F        =    0.0714
    Residual |  1974242.65       204  9677.66006   R-squared       =    0.0159
-------------+----------------------------------   Adj R-squared   =    0.0110
       Total |  2006038.53       205   9785.5538   Root MSE        =    98.375

------------------------------------------------------------------------------
      sports |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          tv |  -.1217946   .0671936    -1.81   0.071    -.2542775    .0106884
       _cons |   146.4896   11.54273    12.69   0.000     123.7313    169.2479
------------------------------------------------------------------------------
```

LEARNINGS

I was also surprised to learn of a weakly positive marginal impact on sleep for every additional hour spent watching TV. I always thought that an hour spent more on watching TV would reduce the hour spent sleeping given the finite amount of minutes in a day. But I was surprised that it wasn't the case. On the contrary, an hour spent more on watching TV meant the same person(male or female) would spend more time sleeping. In a way a person who spends more time on unproductive activity like watching TV is more likely to spend some more time sleeping. There is of course a need for further analysis of this result with different variable that is beyond the scope of this data set that needs to be carried out before we can reach at any judgement about the particular individual.

One of the key observations and learnings for this project for me was that cleaning, decluttering and prepping data for analysis takes a lot more time than actually analysing the data. Another learning was the fact that - if you don't know the variables to look for, it take a huge amount of time to come up with variable that author might be looking at to come up with his/her statistics. To minimise it, the framework is therefore to ask questions like what is the goal, what information will be essential for the questions one seeks answer to and then figure out the variables needed for the actual analysis. This highlights the need for technical documentation for the statistical analysis along with proper variable dictionaries to refer to that can reduce the toll on future replicators.