Exploring the Links: Sleep, Age, Gender, Education, and Their Influence on Income

Joshua Besse

December 2021

Introduction

This research paper looks at the relationship between sleep, age, gender, education, and income. Specifically, my research question is: Is someone's income related to their sleep, age, gender, and education? The topic is important because almost every adult in the country earns some type of income. So, looking into what factors are associated with changes in one's income could prove to be very beneficial to people. Finding associations between income and the variables I will be addressing in my research could possibly help someone increase their chances of earning a higher income by changing habits as simple as sleep patterns.

Through my research, I found that there is not just one variable that is associated with an effect on someone's income. Instead, many variables share this association. After running a multiple linear regression, I found that a one-hour increase in sleep per week is associated with a $200.42 decrease in one's yearly income. I found that the subject being highly educated, 16 or more years of education (the amount of time it takes to earn a college degree), was associated with a yearly income higher than those who were not highly educated by $3763.10 on average. I found that male subjects were associated with an income that was $7086.70 higher, on average, than their female counterparts. Lastly, from my regression, I found that a one-year increase in a subject's age was associated with a $92.35 increase in their income. My research concluded that more sleep is associated with a loss in income on average, while higher age, more education, and being male are associated with higher levels of income on average.

These results are strictly correlational and not causal. This is because many other underlying variables have not been accounted for in this research. The observations used to make my conclusion and conduct my research were part of a dataset with only 706 observations. While this is a large sample and should provide representative results, it is not the entire US population, so I cannot know for certain the true effect of these variables on income. Due to the omitted variables and not knowing the population statistics, I can only find correlations between the variables based on the data, not causation.

Literature Review

In a 2016 article, by Jack Melling for the RAND Europe organization, the topic of the effect of sleep on a nation's economy is discussed. The article claims that sleep deprivation among workers in the US is costing the economy the equivalent of $411 billion and over 1.2 million days that could have been worked (RAND Europe 2016). These figures are derived from estimates of productivity loss in the workplace due to the combination of 'absenteeism and presenteeism,' which are the worker being absent from work and the worker working with low productivity, respectively. These results are produced when the worker receives less than six hours of sleep per night. Melling claims that a simple increase in the number of hours of sleep per night from below six hours to between six and seven hours can add $226 billion to the US economy (RAND Europe 2016). According to the article, these results are not just unique to the United States; "This was closely followed by Japan (up to $138 billion, which is 2.92 percent of its GDP, and around 600,000 working days lost) Germany (up to $60 billion, which is 1.56 percent of its GDP, and just over 200,000 working days lost) and the U.K (up to $50 billion, which is 1.86 percent of its GDP, and just over 200,000 working days lost) have similar losses" (RAND Europe 2016). In the context of my research, the regression that I performed indicates that there is an association between increasing the time you sleep and earning slightly less income. This is likely because my model does not take into account levels of productivity with regards to income, as the article mentioned does. The article is referring to potential income that could have been made if the worker had been more productive, while my regression is saying that sleeping more leaves less time in the day to work. While both are plausible, I believe that factoring in productivity is more applicable to the modern world. Unfortunately, this was not possible to account for in the dataset.

An additional study analyzing the relationship between sleep and income came from the Psychosomatic Medicine Journal, titled: *Socioeconomic Status Predicts Objective and Subjective Sleep Quality in Aging Women*. The results of the study suggest that there is a significant positive relationship between socioeconomic status (income and years of education) and sleep quality in the participants. Meaning that an increase in sleep quality is associated with an increase in earnings on average and that an increase in years of education was also associated with an increase in income. I found these results to be captivating and wanted to adjust the examined variables for several reasons. First, I wanted to look at the effects of sleep quantity as opposed to quality because it is much easier to record data on time spent sleeping rather than the quality of that sleep. This study also defined socioeconomic status as pre-tax household income and years of education, and I wanted to break this up into yearly income and the dummy variable highly_educated, which tells if someone is a college graduate or not, as the increase in earnings attributed to a college degree are more significant than just an additional year of education (simply put, a diploma matters).

Descriptive Analyses and Motivational Evidence

The data set that I used is sleep75.dta (the Stata command is *bcuse sleep75.dta*). To contextualize, the data is from 1975, so some of the numbers may look strange when thinking in a modern-day context. It includes 706 observations of data on age, race, job, education, earnings, health, gender, marital status, religion, employment, area of residence, and variations of sleep like minutes slept per night and in general (including naps). My dependent variable is income (incwage) and my explanatory variables are hours of sleep per week (sleep_hours), age (age), gender (male), and education (highly_educated).

Variable	Observations	Mean	Std. Deviation	Min	Max
incwage	532	9611.118	8889.848	0	102771.8
sleep_hours	706	56.38473	8.317449	22.25	101.8333
age	706	38.81586	11.34264	23	65
male	706	.5665722	.4958996	0	1
highly_educated	706	.2549575	.4361463	0	1

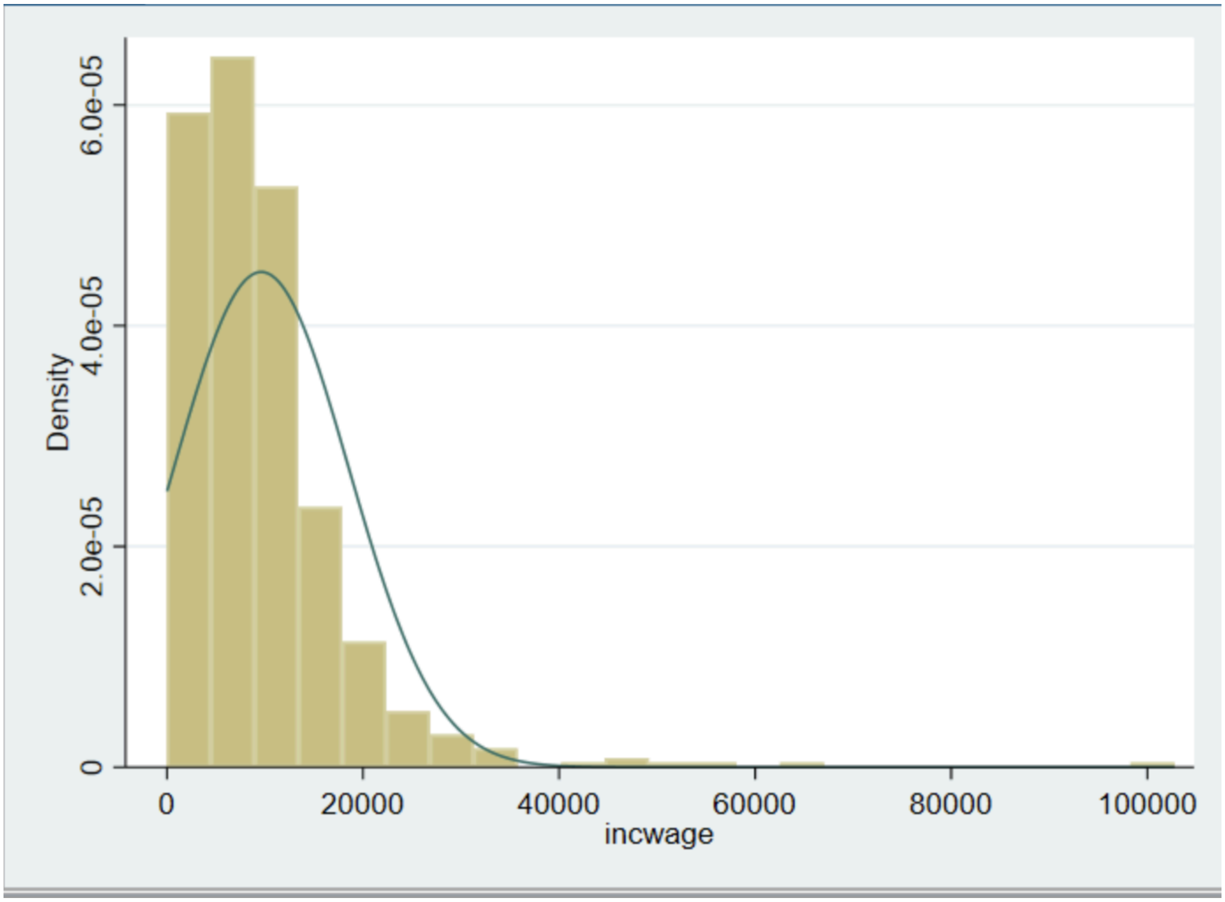
(Table 1: Mean, Standard Deviation, Min, and Max of Each Variable)

The average yearly income is $9,611.11. The average hours of sleep per week is 56.38, which is around 8 hours per night. The average age for the observations in the dataset is 39 years old. The male variable has a mean of .56, which means that 56% of the observations are male. The highly_educated variable, has a mean of .25, which means that 25% of the observations are highly educated. More specifically, 25% of the people have 16 or more years of schooling.

Variable	Median	Mean
incwage (yearly income)	7849.566	9611.118

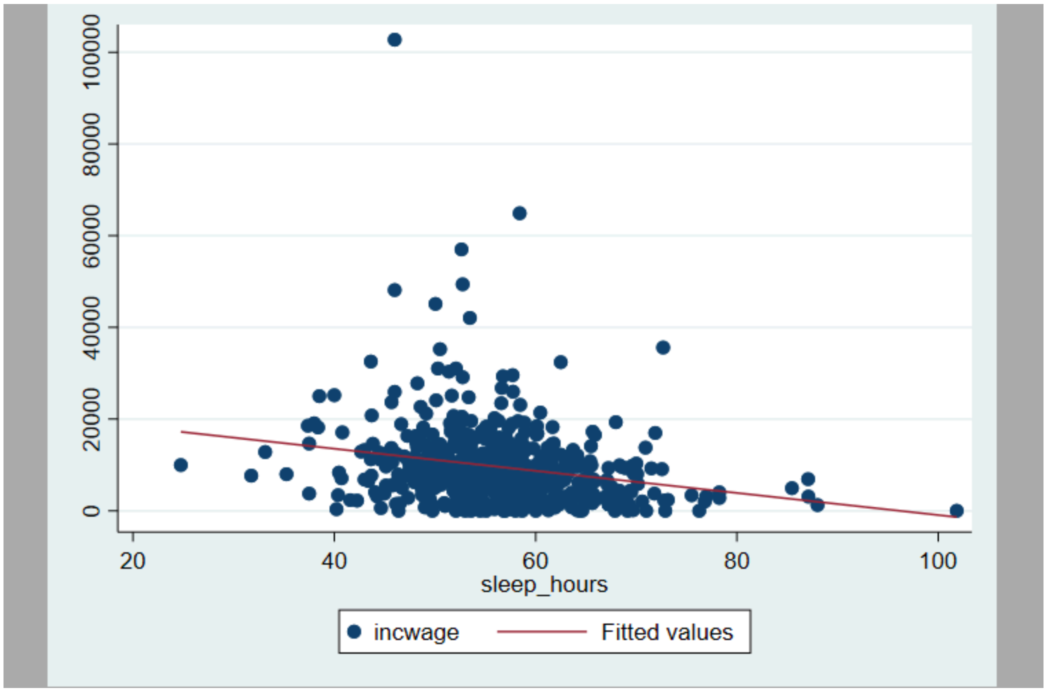
(Table 2: Mean and Median of Incwage)

The mean for yearly income is significantly higher than the median. This means that there are definitely outliers that I need to take into account. For example, the max incwage of $102,771.8 causes the mean to be higher than what is representative of the data. Therefore, the median is a better representation of the average income for observations in the dataset.



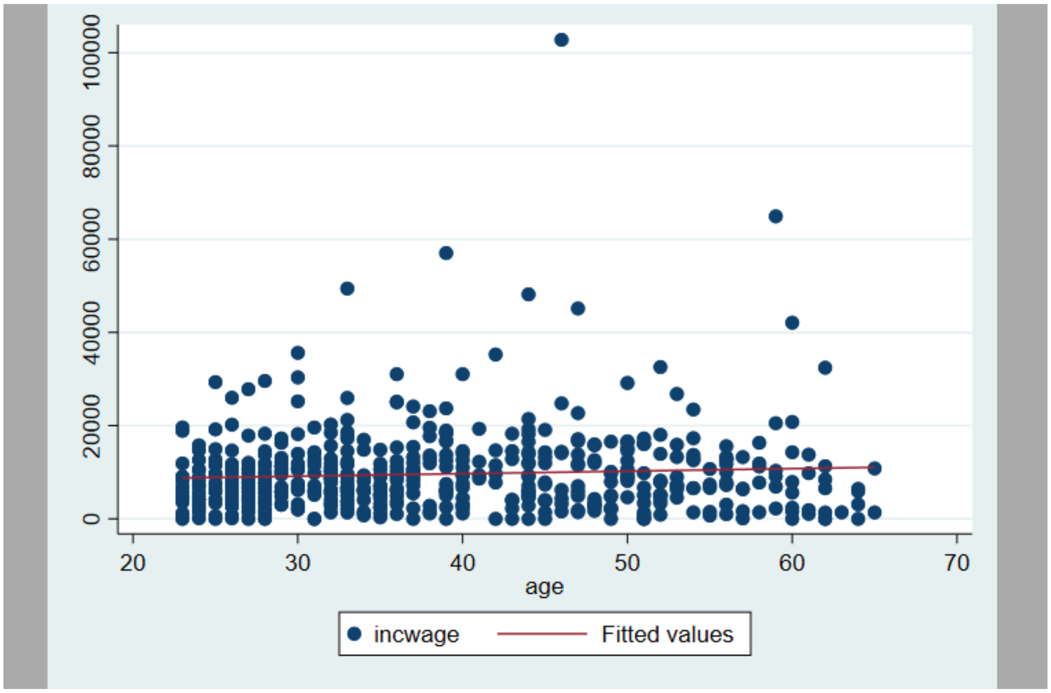
(Table 3: Histogram of Incwage)

Looking at Table 3, it is clear that the variable of yearly income is not normally distributed, but skewed to the right. This claim is supported by the fact that the mean of the incwage variable is $9,611.118 per year while the median of incwage is $7,849.566 per year. The mean is larger than the median which signifies that there are high outliers present in the data set, thus pulling the mean. We can also see some outliers in Table 3 (i.e. the observations around 60,000 and 100,000 are much higher than the rest of the data set).



(Table 4: Scatter Plot of Incwage and Sleep_Hours)

As we can see from Table 4, income decreases when sleep_hours increase. There is little variation in sleep_hours per week, which leads to the observations being mostly concentrated in one area. However, this makes sense considering variation in sleep is not usually very large.



(Table 5: Scatter Plot of Incwage and Age)

The main takeaway from Table 5 is that there is a very slight increase in income when age increases.

Correlations:

```
corr (incwage) (sleep_hours)
(obs=532)

             | incwage sleep_hours
-------------+------------------
     incwage |   1.0000
 sleep_hours |  -0.2241   1.0000
```

The correlation between the dependent variable (incwage) and the numerical explanatory variable of hours of sleep per week (sleep_hours) is r = -0.2241. This value indicates a negative linear relationship between yearly income and hours of sleep per week that is fairly low in strength for individuals in the data set, meaning that as income increases, hours of sleep per week is expected to decrease (variables move in opposite directions).

```
corr (incwage) (age)
(obs=532)

             | incwage      age
-------------+------------------
     incwage |   1.0000
         age |   0.0681   1.0000
```

The correlation between the incwage variable and the explanatory variable of age (age) is r = 0.0681. This value indicates a weak positive linear relationship between an individual's yearly income and their age, meaning that the values of the two variables generally move in the same direction (i.e. both increase or both decrease), but this relationship is not strong.

Table with mean of dependent variable conditional on the dummy variable:

```
 sum incwage if highly_educated == 1

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     incwage |        125    12799.89    12754.33          0   102771.8
```

```
sum incwage if highly_educated == 0

    Variable |        Obs        Mean    Std. Dev.       Min        Max
```

```
-------------+--------------------------------------------------------
    incwage |        407    8631.765    7041.706          0   64905.83
```

Hypothesis test:

H0: μ (incwage if highly_educated == 1) - μ (incwage if highly_educated == 0) = 0
H1: μ (incwage if highly_educated == 1) - μ (incwage if highly_educated == 0) > 0
$\alpha = 0.05$

```
 reg incwage if highly_educated == 1
------------------------------------------------------------------------------
    incwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      _cons |   12799.89   1140.782    11.22   0.000     10541.97    15057.82
------------------------------------------------------------------------------
```

```
reg incwage if highly_educated == 0

------------------------------------------------------------------------------
    incwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      _cons |   8631.765   349.0444    24.73   0.000     7945.605    9317.925
------------------------------------------------------------------------------
```

We can reject the null hypothesis that the mean yearly income of those who are highly educated is the same as the mean yearly income of those who are not highly educated. There is sufficient evidence to reject the null hypothesis because we can say with 95% confidence that the true average yearly income of the highly educated lies within the interval [10541.97, 15057.83], and that the true mean yearly income of those who are not highly educated lies in the interval [7945.605, 9317.925]. There is no overlap between the two intervals.

Empirical Strategy

In this paper, I'm hoping to determine whether or not there is a relationship between hours of sleep per week, gender, age, education, and income. Narrowing it down, I'm aiming to see if any of the previously mentioned variables hold any influence on a person's yearly income.

The multiple linear regression that I will be using to answer my research question will be "reg incwage sleep_hours highly_educated age male, robust." This regression makes sense in the context of my research because it will show how an additional unit of each numerical variable and presence of each dummy variable is associated with a change in income. Specifically, the regression will produce coefficients for each variable. These coefficients can be interpreted as the change in income associated with a one-unit change in the variable. The results of this regression will also help me determine whether these associations are statistically significant or not by taking into account the T value, p-value, and checking to see if the confidence interval includes zero. I check the confidence intervals to see if they include zero because the null hypothesis is that $\boldsymbol{\beta}_0 = 0$, meaning that the variables mentioned have no effect on one's annual income. So, if the confidence interval includes zero then there is a possibility that the specific variable actually has no effect on income, so I would not have sufficient evidence to reject the null hypothesis that each variable has an effect on income. I check the regression for a high T value, which would be a value with an absolute value larger than 1.96. This is because roughly 95% of the data occurs within two (1.96) standard deviations from the mean. If the b_1_hat I find has a T score larger than 1.96, it is very unlikely to have occurred by chance alone, so I am more confident that I can reject the null hypothesis that $\boldsymbol{\beta}_0 = 0$. I'm also looking for a small p-value, typically $p<0.05$ to give me more evidence that I can reject the null hypothesis. Similar to the T score, a p-value less than 0.05 would mean that these results have less than a 5% probability of occurring by chance alone. When I find that all three of these factors occur; high T, low p, and a confidence interval that does not include zero, I can reject the null hypothesis that $\boldsymbol{\beta}_0 = 0$.

I think that the regression coefficient for sleep_hours will be negative. The scatter plot of income and sleep_hours shows that when sleep_hours increase income decreases. Also, the correlation between income and sleep_hours is negative, which means that there is a negative linear relationship between the two variables. Both the scatter plot and the correlation indicate that the variables move in opposite directions. I predict that the regression coefficient for age will be positive. The scatter plot of income and age shows that there is a slight increase in income when age increases. The correlation indicates a weak positive relationship between income and age. Overall, the variable should move in the same direction. I anticipate that the regression coefficient for being highly educated will be positive. The correlation between income and being highly educated (see appendix: Table 7) is positive, which means the variables move in the same direction. Also, the bysort of being highly educated (see appendix: Figure 9) shows that the mean income for being highly educated (12799.89) is greater than not highly educated (8631.765). This leads to the prediction that the regression coefficient will be positive. Lastly, I think that the regression coefficient for being male will be positive because the correlation is positive (see appendix: table 6). Also, the bysort for being male (see appendix: Figure 8) displays that the mean income for men (12966.16) is greater than that of women (5498.034). Therefore, the regression coefficient will be positive. I expect all the variables to be statistically significant. When choosing what variables to test in relation to income, I specifically chose variables that I thought would be significant, compared to other variables in the data set that I thought would not be significant like if someone is Protestant or not.

I believe that the results will be correlational and not causal. This is because there could be other variables/factors that could affect both the dependent variable and the explanatory variables. Some possible omitted variables are race, quality of sleep, and experience. The race of the person could affect the income they earn. Depending on the race, the regression coefficients could be overestimated or underestimated. The quality of sleep could impact the sleep_hours variable. Quality of sleep could be just as important as how long someone sleeps when thinking about productivity and income. Regression does not take into account productivity. Experience is another omitted variable that could affect income. Experience could lead to the age and highly educated variables being overestimated or underestimated.

Results and Analysis

```
Linear regression                         Number of obs    =       532
                                            F(4, 527)       =     39.54
                                            Prob > F        =    0.0000
                                             R-squared       =    0.2514
                                           Root MSE         =    7720.7

------------------------------------------------------------------------------
                 |               Robust
         incwage |        Coef.     Std. Err.      t       P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
      sleep_hours |  -200.4241    39.6984    -5.05   0.000    -278.4107   -122.4376
highly_educated |   3763.076   1058.375     3.56   0.000     1683.924    5842.228
              age |   92.34729   33.51897     2.76   0.006     26.50009    158.1945
             male |   7086.697   613.2412    11.56   0.000         5882    8291.395
            _cons |   12570.04   2212.159     5.68   0.000     8224.311    16915.78
------------------------------------------------------------------------------
```

Estimated regression models:

income = 12570.04 - 200.4241 x sleep_hours + 92.34729 x age + 3763.076 x highly_educated + 7086.697 x male

highly_educated male:
income = 12570.04 - 200.4241 x sleep_hours + 92.34729 x age + 3763.076 x highly_educated + 7086.697 x male

highly_educated female:
income = 12570.04 - 200.4241 x sleep_hours + 92.34729 x age + 3763.076 x highly_educated

non highly_educated male:
income = 12570.04 - 200.4241 x sleep_hours + 92.34729 x age + 7086.697 x male

non highly_educated female:
income = 12570.04 - 200.4241 x sleep_hours + 92.34729 x age

The results of the regression analysis show that the regression coefficient for the sleep_hours variable is -200.4241. The negative sign indicates that income decreases when sleep_hours increase, and the value of 200.4241 means that each additional hour of sleep is associated with an average decrease in income of 200.4241 dollars, holding all other variables constant. The t-value for sleep_hours is -5.05. Since it is less than the critical value of -1.96, the variable is statistically significant. The regression coefficient for the highly educated variable is 3763.076. The positive sign conveys that income increases when someone is highly educated, and the value of 3763.076 conveys that being highly educated is associated with an increase in income of 3763.076 dollars on average, holding all other variables constant. The t-value for highly educated is 3.56, which is greater than the critical value of 1.96 and means that the variable is statistically significant. The regression coefficient for the age variable is 92.34729. The positive sign indicates that income increases when age increases, and the value of 92.34729 indicates that for each additional year of age is associated with an increase in income of 92.34729 on average, holding all other variables constant. The t-value for age is 2.76. Therefore, the variable is statistically significant since 2.76 is greater than the critical value of 1.96. Lastly, the regression coefficient for the male variable is 7086.697. The positive sign denotes that income increases when someone is male, and the value of 7086.697 denotes that being male is associated with an increase in income of 7086.697 on average, holding all other variables constant. The t-value for the male variable is 11.56, which is greater than the critical value of 1.96. This means that the variable is statistically significant.

The R-squared of the regression analysis is 0.2514. This value represents the variation around the regression line. In other words, the R-squared value indicates how much of the variation of the dependent variable can be explained by the independent variables. The R-squared is relatively low, which means that the independent variables do not explain a lot of the variation of the dependent variable. Generally, the regression model does not sufficiently fit the observations. However, this does not affect my interpretation of the significance of the variables. The regression coefficients estimate the trends of the variables, while the R-squared signifies the scatter around the regression line.

In conclusion, the highly_educated, age, and male variables are positively correlated with yearly income based on the positive sign of their regression coefficients. This means that as the value of those independent variables increase, the mean of the dependent variable (incwage) also tends to increase. The sleep_hours variable is negatively correlated with yearly income based on the negative sign of its regression coefficient. This means that as the value of the independent variable (sleep_hours) increases, the mean of the dependent variable (incwage) tends to decrease. I can also conclude the values that incwage changes given an additional unit of sleep_hours and age and how much incwage changes by being highly_educated and male. Yearly income decreases by $200.4241 with each additional hour of sleep per week. Yearly income increases by

$92.34729 for each additional year of age. Yearly income is higher by $3763.076 when someone is highly educated. Yearly income is higher by $7086.697 when someone is male. The regression coefficient values of my sample are estimates of the actual/true population parameters. I cannot say with certainty that sleep_hours, age, education, and gender have a causal relationship with yearly income, because there could be other factors that I have not included that affect both the independent and dependent variables. However, there is sufficient evidence that sleep_hours, age, education, and gender are connected with income.

Conclusion

Of the variables I analyzed, my findings suggest that the majority of them were positively correlated with annual income. The positively correlated variables with income include age, a college degree (highly_educated = 1), and being male (male = 1). To clarify, an increase in each of these variables is associated with a respective increase in annual income. The only variable that I found to be negatively correlated with income was the hours of sleep that someone got each week (sleep_hours). I found that an increase in the amount of sleep that someone got in a week is associated with a decrease in earnings. All of these correlations were significant at $\alpha = 0.05$.

My findings oppose what existing literature suggests in regards to the sleep variable. As described in the *Literature Review* section, existing studies shared a conclusion that an increase in sleep is associated with an increase in earnings, while my results showed an association to a decrease in income resulting from an increase in sleep. I believe that the contradicting results of my findings and existing research are due to the fact that the regression does not take into account the effects that sleep has on productivity, both positively and negatively. The regression eliminates the aspect of fluctuating human functionality as a result of factors such as sleep. My results did agree with those of existing studies in regards to my other variables. For example, I found that a person's level of education has a significant positive relationship with their respective earnings, and this association is supported by the literature review.

A policy implication from my findings is that, given the significant, positive relationship between being a male and earning more on average, we increase measures to get women in the workforce to assume higher-paying jobs. This could include an increased practice of prevention measures to stop applicant discrimination as a result of their gender. Also, given the positive relationship between the level of education and future earnings, we could work to raise awareness of the value and benefits of higher education and a college degree in society so that people can be fully informed before they decide to end their academic career.

In terms of future research, it would be beneficial to find a way to accurately and consistently both measure and quantify the quality of someone's sleep as a means to ensure more accurate and significant results. A further study of interest would be to look at the effects of one's health on income, and to examine this relationship alongside sleep.

References:

Friedman, Elliot M. PhD; Love, Gayle D. PhD; Rosenkranz, Melissa A.; Urry, Heather L. PhD; Davidson, Richard J. PhD; Singer, Burton H. PhD; Ryff, Carol D. PhD Socioeconomic Status Predicts Objective and Subjective Sleep Quality in Aging Women, Psychosomatic Medicine: September 2007 - Volume 69 - Issue 7 - p 682-691doi: 10.1097/PSY.0b013e31814ceada

Melling, J. (2016, November 30). *Lack of Sleep Costing US Economy Up to $411 billion per year*. EurekAlert!: https://www.eurekalert.org/news-releases/652979.

Appendix

Table #6: Correlation Income and Male

```
corr (incwage) (male)
(obs=532)

             |  incwage     male
-------------+------------------
     incwage |   1.0000
        male |   0.4183   1.0000
```

Table #7: Correlation Income and Highly Educated

```
corr (incwage) (highly_educated)
(obs=532)

             |  incwage highly_edcuated
-------------+------------------
     incwage |   1.0000
highly_edu~d |   0.1990   1.0000
```

Figure 8: Summary Statistics of income for males and females

```
bysort male: sum incwage

---------------------------------------------------------------------------------------------------------------
-------
-> male = 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     incwage |       239    5498.034     4807.32          0   32564.12

---------------------------------------------------------------------------------------------------------------
-------
-> male = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     incwage |       293    12966.16    9986.804          0   102771.8
```

Figure 9: Summary Statistics of income for highly educated and non-highly educated

```
 bysort highly_educated: sum incwage

------------------------------------------------------------------------------------------------------------------------
-------
-> highly_educated = 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     incwage |       407    8631.765    7041.706          0   64905.83

------------------------------------------------------------------------------------------------------------------------
-------
-> highly_educated = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
     incwage |       125    12799.89    12754.33          0   102771.8
```