Does hourly wage influence sleep?

ECN 4330 Final Project - Group 1

Austin Halvorsen, Chase Dean Harward, Suneet Dash

12/8/2020

# Introduction

How does someone’s hourly wage impact the amount of sleep that they get on a given night?

* Determining what role hourly wage plays in sleep is important because it can help us to understand an aspect of sleep that isn’t explored as often as opposed to just looking at hours of work itself.
* Through this question we can even see how different income classes sleep, such as how many hours does someone in the top income percentile sleep as opposed to someone in the middle class or someone making minimum wage.

## Initial Assumptions

* On one side of the question we assumed that people making minimum wages would typically sleep more, as a lot of these people are just entering the labor market and would be younger (high school/college). Also they wouldn’t have to deal with as many high-stress situations as someone who is making a significantly higher amount of money.
* On the other side we considered that someone making minimum wages could be sleeping less, as they might be working a second job or something like that to make their quality of life better. And that someone making a higher wage would be in a comfortable position in life where financial issues weren’t impacting their sleep.

## Ethical Importance

This question poses some interesting ethical discussions.

* We typically look towards financial positions, family status, and where people live as basic indications of a quality of life. Something we don’t often consider is what type of quality of life information we can get from sleep data.
* Another example would be in terms of government programs allocating resources towards low-income workers, can we make a conclusion on the quality of life they have based on the relationship between the hourly wage and sleep?

# Data

* We will be using the dataset “sleep75” found in the Wooldridge package and originally used for the paper “Sleep and the Allocation of Time” by Jeff Biddle and Daniel Hamermesh, published in the Journal of Political Economy in 1998.
* The goal of the original paper was to show a relationship between an increase in time in the labor market and how much sleep someone gets.
* Previous questions involving the “sleep75” data set typically included some analysis on a tradeoff between sleep and time worked, so we aimed to ask a question that hasn’t really been explored with this data.
* This data set includes 706 observations on 34 variables, notably hourly wage, sleep, years of experience, marital status, spouse wage, age, children or no children, type of employment, and quality of health.

## Variables

As we began our research for this project, we noticed that so many of these variables had fairly low levels of correlation, and the variation in any given variable didn’t explain much of the variation in the others save for only a few exceptions in variables like totwork that we have previously covered in homework assignments. What we decided to then do was ensure that we included enough variables to try and paint as compelling of a picture as we could in order to try and make some useful conclusions about our research question. They are outlined below.

### Explanatory Variables

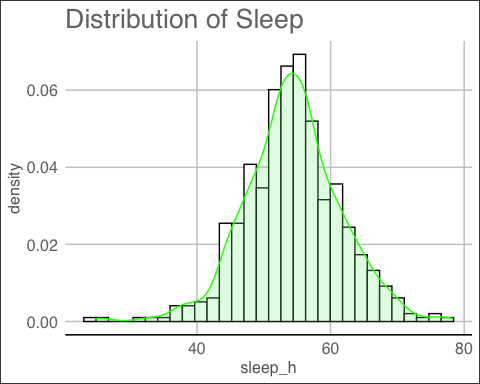
* age = years old of the individual  
  Our story for age being included is that this probably does effect the amount of sleep someone gets every night. As age increase, we would probably expect to see that someone gets more sleep.
* educ = years of education We felt like someone’s level of education did have some impact on hours of sleep per week, and so we included this.
* hrwage = the hourly rate of the individual This is our most important variable and the main explanatory variable being considered here. We want to know if this has some effect on someone’s hours of sleep per week.
* marr = Is the person married or not? We absolutely felt that this needed to be included, as we have all seen numerous studies done that offer some information on sleeping patterns between singles and couples. This is a dummy variable.
* yrsmarr = time in years the person was married (if married)  
  This goes along with our marr variable to help see if the years that someone is married has anything to do with sleep as well.
* spsepay = amount of pay for spouse if they are paid If we are looking at someone’s hourly pay rate, it would also make sense that their spouse’s hourly pay rate would also have some effect on that person’s sleep.
* male = is the person male or not  
  While we didn’t have any specific belief about what the sign of the coefficient for this variable could be, we felt that it is important to account for someone’s gender here. This is a dummy variable.
* gdhlth = is the individual in good health or not This was a very interesting variable that we saw included in the dataset, and so we felt like it would probably be a good idea to account for this one as well in the regression. This is a dummy variable.
* totwrk = amount of time spent working per week Of course, the variable that we’ve looked at before and already know has a statistically significant effect on sleep.
* yngkid = 1 if children less than 3 years old is present We felt curious to explore the statistical significance of this specific variable, as many parents would be inclined to say that it *does* have an effect on someone’s sleep.

### Dependent Variable

Our Dependent variable in our study is sleep.

* As part of our manipulation for our data, we changed sleep from “minutes” to “hours” per week to have a simpler interpretation of our regression. Not only does this help to make our data look a little bit cleaner, but it also helps to solve any possible heteroskedasticity issues we could run into in the data. Without this important transformation, our model is heteroskedastic. This transformation is the solution. Typically, we would transform our model further and try to change our “hours per week” into “hours per day”, but as it is later shown in the regression results, this would not end up being a useful choice in trying to visualize the data.
* What is important to also note about our transformation from minutes into hours for sleep is that this doesn’t actually affect the at all. The explained part of the variation will not change at all.
* Additionally since we were only analyzing individuals who are working, we filtered our data to only include those who are in the labor force and have an hourly wage.

## Distribution



We will explore the implications of this distribution later as we discuss the various assumptions that we need to make for our regression analysis, but it is fairly straightforward in understanding what this distribution is. There is a clear normal distribution, and this is a very good thing for us. As we’ll touch on later, this isn’t absolutely necessary, but this is good to see.

## Summary Stats

sleep\_h age educ hrwage   
 Min. :24.75 Min. :23.00 Min. : 1.00 Min. : 0.350   
 1st Qu.:50.06 1st Qu.:29.00 1st Qu.:12.00 1st Qu.: 2.890   
 Median :54.19 Median :36.00 Median :12.00 Median : 4.380   
 Mean :54.32 Mean :38.32 Mean :12.73 Mean : 5.083   
 3rd Qu.:58.58 3rd Qu.:47.00 3rd Qu.:15.00 3rd Qu.: 6.210   
 Max. :78.25 Max. :65.00 Max. :17.00 Max. :35.510   
 yrsmarr spsepay male gdhlth   
 Min. : 0.00 Min. : 0 Min. :0.0000 Min. :0.0000   
 1st Qu.: 0.00 1st Qu.: 0 1st Qu.:0.0000 1st Qu.:1.0000   
 Median : 8.00 Median : 1000 Median :1.0000 Median :1.0000   
 Mean :11.18 Mean : 5270 Mean :0.5508 Mean :0.8853   
 3rd Qu.:18.25 3rd Qu.: 9395 3rd Qu.:1.0000 3rd Qu.:1.0000   
 Max. :43.00 Max. :50000 Max. :1.0000 Max. :1.0000   
 totwrk marr yngkid   
 Min. : 0 Min. :0.0000 Min. :0.000   
 1st Qu.:1611 1st Qu.:1.0000 1st Qu.:0.000   
 Median :2300 Median :1.0000 Median :0.000   
 Mean :2161 Mean :0.8177 Mean :0.141   
 3rd Qu.:2700 3rd Qu.:1.0000 3rd Qu.:0.000   
 Max. :6415 Max. :1.0000 Max. :1.000

What we can take away from our summary statistics being shown here is that it could be useful to look at the mean for sleep. With the mean and median being so close together, it wouldn’t seem that there are any crazy outliers there. We can also generally see that the average amount of sleep reported was about 24 hours per week. For age, we see an average of around 38 years. Education has an average around 12 years. Hourly wage is quite interesting to look at as well in that its mean is at about $5 per hour.

# Emperical Framework

**True Model**

**Estimated Model**

**Regular Standard Errors:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| (Intercept) | hrwage | educ | age | yrsmarr |
| 2.540 | 0.09044 | 0.1237 | 0.03636 | 0.03978 |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| spsepay | male | gdhlth | totwrk | marr | yngkid |
| 4.951e-05 | 0.739 | 0.9642 | 3.499e-04 | 1.052 | 0.9211 |

* Our “engine” we used was **OLS**
  1. We have normally distributed data.
  2. Our data is homoskedastic
  3. OLS is BLUE for our data.
* We did not include any Robust Standards Errors since we are working with homoskedastic data.

## Classical Linear Model (CLM) Assumptions

*While it is true that we really only need to prove for the Gauss-Markov Assumptions, our normal distribution allows us to go a step farther. As will be discussed in more detail, the asymptotic properties of the data tells us that we don’t need normality.*

**1. MLR.1 Linear in Parameters**

**2. MLR.2 Random Sampling**

**3. MLR.3 No Perfect Collinearity**

**4. MLR.4 Zero Conditional Mean**

**5. MLR.5 Homoskedasticity**

**6. MLR.6 Normality**

### MLR.1 Linear in Parameters

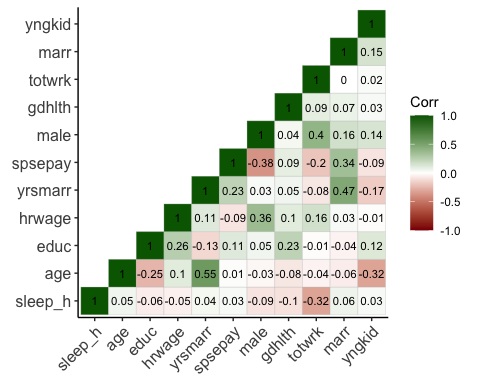
**True Model**

* All of our parameters are linear

### MLR.2 Random Sampling

* Data is taken from Wooldridge dataset sleep75, which was randomly sampled with 706 observations

### MLR.3 No Perfect Collinearity



* No evidence of perfect collinearity
* Nothing is perfectly correlated with our dependent variable “sleep”
* No independent variable is constant.

hrwage educ age yrsmarr spsepay male gdhlth totwrk   
1.293238 1.278630 1.921456 2.299962 1.535915 1.710383 1.089305 1.228595   
 marr yngkid   
1.904797 1.185977

* When running our VIF test in R, the results give us an identical story to what the correlation matrix showed us. There is no perfect collinearity, and there’s even the case that almost none of our variables have a fairly low correlation with each other.

### MLR.4 Zero Conditional Mean

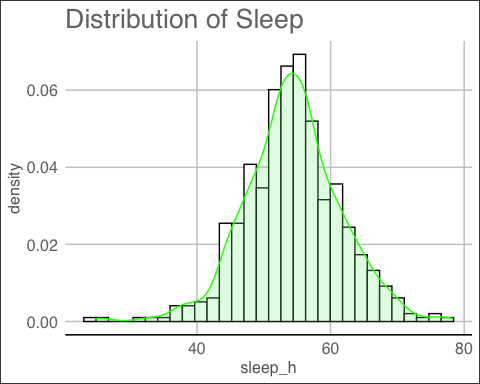
* We included multiple important variables in order to avoid violating the Zero Conditional Mean Assumption.

### MLR.5 Homoskedasticity

studentized Breusch-Pagan test  
  
data: model  
BP = 12.537, df = 10, p-value = 0.2507

* We can prove that the error has the same variance across all values of the explanatory variables
* According to the results of our Breusch-Pagan test, we have a p value of .2507, and so we fail to reject the null hypothesis

### MLR.6 Normality



* Normal distribution of our data is shown here, but we don’t like this assumption. With a sample size this large, we can get around this assumption and ignore it.
* It is important to clarify that this assumption is not necessary. In fact, it is quite a strong assunption to make, and we typically want to make as weak of assumptions as we can in validating these analyses. What is comforting is that it isn’t a necessary assumption to make because of the fact that we have such a large sample size of 700+. This introduces asymptotic properties into the model to help validate our regression results.

# Results

Call:  
lm(formula = sleep\_h ~ hrwage + educ + age + yrsmarr + spsepay +   
 male + gdhlth + totwrk + marr + yngkid, data = df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-29.6065 -4.0830 -0.1541 4.0329 21.3023   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 5.934e+01 2.540e+00 23.359 < 2e-16 \*\*\*  
hrwage 1.248e-02 9.044e-02 0.138 0.8903   
educ -1.157e-01 1.237e-01 -0.935 0.3501   
age 5.096e-02 3.636e-02 1.402 0.1616   
yrsmarr -4.272e-02 3.978e-02 -1.074 0.2833   
spsepay -3.551e-05 4.951e-05 -0.717 0.4736   
male 3.155e-01 7.739e-01 0.408 0.6836   
gdhlth -1.193e+00 9.642e-01 -1.237 0.2165   
totwrk -2.614e-03 3.499e-04 -7.472 3.37e-13 \*\*\*  
marr 1.866e+00 1.052e+00 1.774 0.0767 .   
yngkid 8.092e-01 9.211e-01 0.878 0.3801   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 6.789 on 521 degrees of freedom  
Multiple R-squared: 0.1235, Adjusted R-squared: 0.1067   
F-statistic: 7.344 on 10 and 521 DF, p-value: 6.266e-11

* Initial view of results: only two of our variables are statistically significant!
* is low at 12.35%, while adjusted is also low at 10.67%
* This tells us that the variation in our explanatory variables explain very little of the variation in our explained variable (sleep)
* We shouldn’t focus too much on the of the regression, but this could be a sign that other factors we haven’t analyzed are affecting sleep.

## Coefficients

Call:  
lm(formula = sleep\_h ~ hrwage + educ + age + yrsmarr + spsepay +   
 male + gdhlth + totwrk + marr + yngkid, data = df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-29.6065 -4.0830 -0.1541 4.0329 21.3023   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 5.934e+01 2.540e+00 23.359 < 2e-16 \*\*\*  
hrwage 1.248e-02 9.044e-02 0.138 0.8903   
educ -1.157e-01 1.237e-01 -0.935 0.3501   
age 5.096e-02 3.636e-02 1.402 0.1616   
yrsmarr -4.272e-02 3.978e-02 -1.074 0.2833   
spsepay -3.551e-05 4.951e-05 -0.717 0.4736   
male 3.155e-01 7.739e-01 0.408 0.6836   
gdhlth -1.193e+00 9.642e-01 -1.237 0.2165   
totwrk -2.614e-03 3.499e-04 -7.472 3.37e-13 \*\*\*  
marr 1.866e+00 1.052e+00 1.774 0.0767 .   
yngkid 8.092e-01 9.211e-01 0.878 0.3801   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 6.789 on 521 degrees of freedom  
Multiple R-squared: 0.1235, Adjusted R-squared: 0.1067   
F-statistic: 7.344 on 10 and 521 DF, p-value: 6.266e-11

* On average, in 1975, a one dollar increase in someone’s hourly wage is associated with an increase of 0.0125 hours of sleep per week
* This would equate to, on average, only 45 seconds extra sleep per week for a one dollar increase in hourly wage
* We could break down that even further to say that on average, this is only 6.43 seconds extra sleep per DAY. That’s not at all economically statistic, and we will explore in more detail whether or not this is truly statistically significant when we do hypothesis testing.
* This helps paint a clearer picture for our motivation behind not making a transformation into hours per day. The effect is so small that it is difficult to visualize at first glance.
* Rest of our coefficients show similar results in that they appear to, on average, have a small impact on sleep
* Only exception to this is totwrk and also marr, which we can see are statistically significant at varying significance levels
* Strangely enough, our variable for yngkid was *not* statistically significant.

## F and t Statistics

Call:  
lm(formula = sleep\_h ~ hrwage + educ + age + yrsmarr + spsepay +   
 male + gdhlth + totwrk + marr + yngkid, data = df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-29.6065 -4.0830 -0.1541 4.0329 21.3023   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 5.934e+01 2.540e+00 23.359 < 2e-16 \*\*\*  
hrwage 1.248e-02 9.044e-02 0.138 0.8903   
educ -1.157e-01 1.237e-01 -0.935 0.3501   
age 5.096e-02 3.636e-02 1.402 0.1616   
yrsmarr -4.272e-02 3.978e-02 -1.074 0.2833   
spsepay -3.551e-05 4.951e-05 -0.717 0.4736   
male 3.155e-01 7.739e-01 0.408 0.6836   
gdhlth -1.193e+00 9.642e-01 -1.237 0.2165   
totwrk -2.614e-03 3.499e-04 -7.472 3.37e-13 \*\*\*  
marr 1.866e+00 1.052e+00 1.774 0.0767 .   
yngkid 8.092e-01 9.211e-01 0.878 0.3801   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 6.789 on 521 degrees of freedom  
Multiple R-squared: 0.1235, Adjusted R-squared: 0.1067   
F-statistic: 7.344 on 10 and 521 DF, p-value: 6.266e-11

* The F statistics show us unsurprising results: it is statistically significant
* We ran iterations of our model that weren’t even jointly significant, so we made sure to include enough variables to fix this
* The T statistics are almost all very small with the exception of totwrk and marr

## Hypothesis Testing

Call:  
lm(formula = sleep\_h ~ hrwage + educ + age + yrsmarr + spsepay +   
 male + gdhlth + totwrk + marr + yngkid, data = df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-29.6065 -4.0830 -0.1541 4.0329 21.3023   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 5.934e+01 2.540e+00 23.359 < 2e-16 \*\*\*  
hrwage 1.248e-02 9.044e-02 0.138 0.8903   
educ -1.157e-01 1.237e-01 -0.935 0.3501   
age 5.096e-02 3.636e-02 1.402 0.1616   
yrsmarr -4.272e-02 3.978e-02 -1.074 0.2833   
spsepay -3.551e-05 4.951e-05 -0.717 0.4736   
male 3.155e-01 7.739e-01 0.408 0.6836   
gdhlth -1.193e+00 9.642e-01 -1.237 0.2165   
totwrk -2.614e-03 3.499e-04 -7.472 3.37e-13 \*\*\*  
marr 1.866e+00 1.052e+00 1.774 0.0767 .   
yngkid 8.092e-01 9.211e-01 0.878 0.3801   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 6.789 on 521 degrees of freedom  
Multiple R-squared: 0.1235, Adjusted R-squared: 0.1067   
F-statistic: 7.344 on 10 and 521 DF, p-value: 6.266e-11

The Null Hypothesis is that hourly wage has no effect on the number of hours of sleep someone gets in a week

The Alternative Hypothesis is that hourly wage does have an effect on the amount of sleep that someone gets in a week

* With a p-value of .8903, hrwage is only statistically significant as low as the 89.03% significance level
* We fail to reject the Null Hypothesis
* Take note: totwrk and marr are statistically significant

## Confidence Interval

2.5 % 97.5 %  
(Intercept) 54.3455321034 6.432614e+01  
hrwage -0.1651915849 1.901593e-01  
educ -0.3586376739 1.272984e-01  
age -0.0204640791 1.223776e-01  
yrsmarr -0.1208613794 3.541735e-02  
spsepay -0.0001327792 6.176626e-05  
male -1.2047347629 1.835809e+00  
gdhlth -3.0871052962 7.011281e-01  
totwrk -0.0033012088 -1.926634e-03  
marr -0.2009297351 3.932696e+00  
yngkid -1.0003294763 2.618655e+00

* There is a 95% probability that the true population mean lies within our confidence interval here
* Unfortunately (but unsurprising, 0 lies within our 95% confidence interval), further providing evidence to reject the null hypothesis

# Conclusion

Does one’s hourly wage have some relationship with the amount of sleep that they get in a given week?

## The Downside

Even after exploring multiple potential transformations of our original model, adding more explanatory variables, and close analysis, our conclusion is that someone’s hourly wage has little to no effect on the amount of sleep that someone gets in a given week.

Moreover, the only variables that we could conclude had **ANY** statistically significant effect on sleep were how much someone works in a week and their marital status. We were originally discouraged by these results, as it made us feel like our original research question was a boring one.

## The Upside

These results, however, raise more interesting questions that call for further investigation. While knowing someone’s hourly wage might not help us that much in predicting how much sleep that they get in a week, perhaps the reason for this is that our preferences are more individualistic in a way that cannot be captured simply by how much money that someone makes.

Two different CEOs may have wildly different preferences on sleeping, and we could see a similarly complex situation at lower hourly wages too. So if we can make any conclusion from these results, it’s that we need to know **MORE.**

## Further Investigation

* Some other aspects of our data that would like to investigate would be researching individual preferences. Specifically, how does someone perceive their wage. Is it fair to them or not? Does this perception have a larger influence on sleep than wage?
* Are there other interactions that we could investigate that go into sleeping more than just wage?
* There is, however, one other piece of this analysis that we can explore right now…

### New Dependent Variable?

* Let’s consider utilizing a different dependent variable to see if we can increase the statistical significance of our individual slope coefficients. This variable is slpnaps, which is defined as the minutes of sleep when incorporating naps. We will run a quick regression with all of the same explanatory variables as well as the same data mutation to turn minutes into hours for sleep.

### Results

**True Model 2.0**

Call:  
lm(formula = slpnaps\_h ~ hrwage + educ + age + yrsmarr + spsepay +   
 male + gdhlth + totwrk + marr + yngkid, data = df)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-31.567 -4.527 -0.089 4.345 38.564   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 6.391e+01 2.891e+00 22.104 < 2e-16 \*\*\*  
hrwage -9.138e-02 1.029e-01 -0.888 0.3751   
educ -1.770e-01 1.408e-01 -1.257 0.2092   
age 7.277e-02 4.138e-02 1.759 0.0792 .   
yrsmarr -4.062e-02 4.527e-02 -0.897 0.3700   
spsepay -3.384e-05 5.636e-05 -0.600 0.5484   
male 1.030e+00 8.808e-01 1.169 0.2427   
gdhlth -1.624e+00 1.097e+00 -1.480 0.1396   
totwrk -3.194e-03 3.982e-04 -8.021 6.96e-15 \*\*\*  
marr 8.674e-01 1.197e+00 0.724 0.4692   
yngkid 1.448e-01 1.048e+00 0.138 0.8902   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 7.727 on 521 degrees of freedom  
Multiple R-squared: 0.141, Adjusted R-squared: 0.1245   
F-statistic: 8.554 on 10 and 521 DF, p-value: 5.485e-13

* There is certainly some useful information from this new model, but it does not give us the statistical significance that we’d like.
* While our hrwage slope coefficient in the original model had a p-value of nearly 0.90, this new hrwage slope coefficient has shot down to 0.3751. This is good news.
* Our has increased slightly
* Other variables like age have actually decreased greatly in their p-value! While before it was at about 0.16, it is now at 0.0792. This nearly gives it statistical significance, and so this new dependent variable is a clear sign that we are headed at least in the right direction.