

Analysis of Influencing Factors of Sleep Problems Based on Stata Regression

Ruimin Wang

Dalhousie University, Halifax, Canada

Keywords: Sleep, Stata, Ols regression, Reset, B-p and white's tests, F-test, Goodness-of-fit

Abstract: Based on the growing problem of sleep, this study focus on influencing factors of sleep problems and aim to improve the sleep situation in the society. OLS regression would illustrates the relationship between sleep time and other factors, such as age, earnings, health situation and so on. The related tests also prove the reliability and validy of result. This study will provide a guidance for public to understand the reasonable sleep time appropriately.

1. Introduction

In this project, I will discuss the factors which will influence the sleep time, including naps. I choose age, earnings, health situation, sex to be the independent variables while mins sleep, including naps per week to be dependent variable. These factors are part of essential parameters in influencing sleep duration, which would be meaningful for us to do the research. Through this project, we aim to improve current human sleep problems as much as possible under current conditions.

As the developing of the society, more and more people are concerned with the sleep problem. Will the time of working influence the sleep time? In our traditional views, there is positive effect between the total earnings and working time, so I try to use the total earnings to do the OLS regression. In the meantime, maybe the sleeping time is defferient from males and females, which is also an important factor for dependent variable. So study the factors that influence the sleeping time would be beneficial to our daily life and human health.

Maurice Ohayon and other scholars [1] thinks that millions of individuals are using commercially available sleep tracking devices. These devices purport to measure sleep quality and quantity. Therefore, there exists a need to define clearly both sleep quantity and quality using the best scientific evidence available.

About the data collection of sleep time, Chol Shin [2] believes that sleep quality affects health and the overall quality of life. As the factors that influence sleep quality and their relative importance vary among individuals, a self - report method is essential.

2. Methodology

I would use STATA to find the relationship between the independent variables and dependent variables. In this way, we could know the degree of influence of each independent variable on dependent variable by regression coefficient. I decide to use this modol by OLS regression to get the coefficient and other test results as follows:

$$\text{slnaps} = \alpha + \beta_1 \text{age} + \beta_2 \text{earn74} + \beta_3 \text{gdhlth} + \beta_4 \text{male} + u \quad (1)$$

According to traditional view, as the age increasing, people hope the sleep time of elder people is more than younger people. If the people spend more time to make money, their sleep time would be cut down, so the effect of earn74 on slnaps is negative. The influence of health situation on sleep time in uncertain, we would check it using the data later. As for sex, maybe it has small influence on sleep time, but to rule that out, I put it in the model, too. In my old opinion, the sleep time depends on the lifestyle basically. I use the data from Woodridge data set, which gives detailed data for us to do the regression. The reason I choose these variables is that they are typical and familiar to people. People can know the relationship between them easily and try to find the better method to improve ther sleep quality.

3. Description of the Data

The data source: Wooldridge, J. M. (2016). Introductory econometrics: A modern approach. Nelson Education.

The dependent variable are as follows:

slpnaps: mins sleep, including naps, per week.

The independent variable are as follows:

age: the age of the person who provide the data, for years

earns74: the total earnings of the person who provide the data, 1974

gdhlth: this is a dummy variable, the data=1 if in good or excellent health male: this is a dummy variable, the data=1 if the person is male

Table 1 the Description Of the Data

variable	obs	mean	std.dev.	min	max
slpnaps	706	3383.084	499.0469	1335	6110
age	706	38.81586	11.34264	23	65
earns74	706	9767.705	9323.588	0	42500
gdhlth	706	0.8909348	0.3119419	0	1
male	706	0.5665722	0.4958996	0	1

As we can see from table 1, there are 706 observations in total, the mean of slpnaps, age, earns74, gdhlth and male are 3383.08, 38.82, 9767.705, .8909, .5666. And the standard deviation are 499.05, 11.34, 9323.59, .3119, .4959. The minimum of slpnaps and age are 1335 and 23 while the minimum of other variables are zero. The maximum of slpnaps, age and earns74 are 6110, 65, 42500. Obviously, as the dummy variables, the maximum of gdhlth and male are 1. However, it makes little sense to discuss these character of dummy variables, so I will show the proportional distribution in the Fig.1 and Fig.2.



Fig. 1 The Proportional Distribution of Health

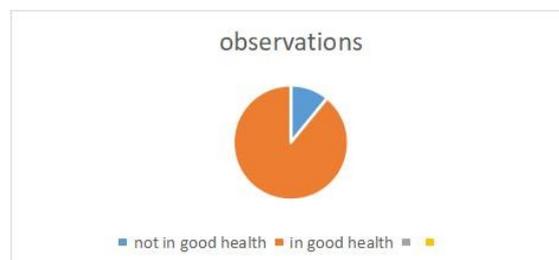


Fig. 2 The Proportional Distribution of Gender

4. Results

4.1 Misspecification Testing:

(1) Estimate the regression model in the original form:

Table 2 the Form Of the Regression Model

slpnaps	coef.	std.err.	t	p> t	95% Conf.Interval	
age	4.991668	1.646752	3.03	0.003	1.758512	8.224825
earns74	-0.0047319	0.0020281	-2.33	0.02	-0.0087138	-0.00075
gdhlth	-155.5288	60.23552	-2.58	0.01	-273.7925	-37.26516
male	-37.88975	37.76349	-1	0.316	-112.0328	36.25334
_cons	3395.581	90.43722	37.55	0	3218.021	3573.141
number of obs=706						
F(4,701)=6.47						
prob>F=0.0000						
R-squared=0.0356						
adj R-squared=0.0301						
Root MSE=491.47						

from the Table 2, We Could Know That the Regression Model is as Follows: $Slpnaps=4.99age-0.0047earns74-155.53gdhlth-37.89male+3395.58$ (2)

(2)RESET, B-P and White's tests:

Table 3 Reset Test

H0: model has no omitted variables
F(3,698)=2.06
prob>F=0.1045

Now we do the RESET test as table 3, as we can see, the F value is 2.06 while the p value is 0.1045 which is over 0.05. It means the we can not reject the H0 at 95% significance level. So we continue to do the B-P test that can be found in table 4, the results from B-P test shows that these residuals satisfy the homoscedasticity because the p value is 0.2207.

Table 4 B-P Test

H0: constant variance
variables: age earns74 gdhlth male
chi2(4)=5.72
prob>chi2=0.2207

For further verification, we also use White's test whose result is in table 5. Therefore, we could use OLS regression smoothly next step.

Table 5 Ols Regression

H0: homoskedasticity			
Ha: unrestricted heteroskedasticity			
chi2(12)=6.61			
prob>chi2=0.8821			
source	chi2	df	p
heteroskedasticity	6.61	12	0.8821
skewness	2.79	4	0.5934
kurtosis	4.97	1	0.0258
total	14.37	17	0.6407

4.2 Discussion of Regression Output:

(1)F-test

This result could be found in table 6.

Table 6 the Result Of f-Test

(1) age=0
(2) earns74=0
(3)gdhlth=0
(4) male=0
F(4, 701)=6.47

prob>F=0.0000

The H0: $u_i=0$, the result: $p=0.0000<0.0001$, so we reject the null hypothesis. The variables are significant which is efficient to dependent variable.

(2) Goodness-of-fit

Now we find the R-squared=0.0356 and adj R-squared=0.0301, which could be found in table 2. These figure is too small to have the explanatory power, so I decide to add more independent variables. However, after excluding the independent variables that are highly coincident with dependent variables (we could get a very big R-square but we can't because there is no meaning, it is shown in table 8) and doing OLS regression using the remaining variables, the R-square is also small which would be shown in table 7.

Table 7 the Data Result after Adding Independent Variables

slpnaps	coef.	Std. Err	t	P> t
earns74	-0.0026818	0.0020315	-1.32	0.187
age	-7.052442	7.773042	-0.91	0.365
gdhlth	-91.06414	58.83605	-1.55	0.122
male	102.3096	39.39329	2.6	0.01
union	-8.561602	43.75461	-0.2	0.845
marr	25.00875	57.22147	0.44	0.662
selfe	-17.99644	53.05412	-0.34	0.735
south	106.7398	45.35937	2.35	0.019
worknrm	-0.1963464	0.0202713	-9.69	0
workscnd	-0.0959944	0.1205122	-0.8	0.426
exper	11.6496	7.022461	1.66	0.098
yngkid	-39.53662	56.45349	-0.7	0.484
yrrsmarr	-2.726201	2,317,234	-1.18	0.24
_cons	3887.856	180.0029	21.6	0
Number of obs=706				
F(13,692)=9.96				
prob>F=0.0000		AdjR-squared=0.1419	Root MSE=462.3	
R-squared=0.1577				

But we can not deny that adding more variables(union, marr, selfe, south, worknrm, workscnd, exper, yngkid, yrsmarr) would increase the R-squared. So we could think that there is a lot of factors that will influence sleep time, and in this project, I just choose a part of them to explain the sleep time. Therefore, I don't change the model.

(3) Statistical interpretation

As we can see, the relationship between age and alpnaps are positive, which is the exact opposite of what I was thinking. This means the elder people have more sleeping time. Every year they get a year older, they get about 4.99 hours of sleep and the exact reason remains to be determined. What follows my expectation is the effect of earnings on sleep time including naps is negative. This strongly suggests that people who may work longer hours earning more one unit get less 0.0047 hours of sleep. The coefficient of health situation indicate that if the people is in excelent health, they need less sleep about 155.53 hours than the people who are sick or ill. And to my surprise, there is a strong connection between the sleep time and sex, in this model, the female have 37.89 hours of sleep more than male. Whether this is reliable or not remains to be checked.

(4) Economic Interpretation

In this original model, the health situation is an important factor. The normal person would have about 155.53 hours sleep time more than the person who is in good health.

In this model, the total earnings have a little effect on the sleep time, which means we can almost ignore it when we try to improve our sleep time. This does not mean that the data is without reference value because it also provide the relevance. Now we check the p-value:

The age: $p=0.003$ We reject H0 at 1% significance level. The earns74: $p=0.020$ We reject H0 at

5% significance level. The $gdhlth$: $p=0.010$ We reject H_0 at 10% significance level. The $male$: $p=0.316$ We do not reject null hypothesis.

Through the p value, we should drop the variable of $male$ because it is too large. But the other variable are significant in this model. So I regress the remaining variables, which is shown in table 8, it could also be a reference.

Table 8 the Data Of Regressing the Remaining Variables

number of obs=706
F(5,700)=1627.66
prob>F=0.0000
R-squared=0.9208
Adj R-squared=0.9202
Root MSE=140.95

5. Conclusion

So the model I estimate:

$$slpnaps = 4.99age - 0.0047earns74 - 155.53gdhlth - 37.89male + 3395.58$$

But when I do the OLS regression, I find the relationship between the earnings and sleep time is slight but effective. And the regression effect of sex is not significant, as I guessed before, there is many factors to influence the sleep time of a person, such as work efficiency, season, mind and so on. But what we can ensure is that there is strong correlation between the age and $slpnaps$, health situations and $slpnaps$.

In the future, if we would like to check our sleep time, we could start with our age and see how much sleep time we need. Try to improve the sleep time, we focus on the health situation of our body and also use other method. But remember, the earnings is not the main excuse for people to have less sleep. If they want to have more time to sleep while keeping the earnings unchanged, they could.

References

- [1] Ohayon M, Wickwire E M, Hirshkowitz M, et al. National sleep foundation's sleep quality recommendations: first report [J]. Sleep Health, 03 (1), pp.6-19, 2017.
- [2] Hyeryeon Y I, Shin K, Shin C. Development of the sleep quality scale [J]. Journal of Sleep Research, 15 (3), pp.89-93, 2006.