

DUBLIN INSTITUTE OF TECHNOLOGY

Title: Factors Affecting satisfaction of sleep

**Probability and statistics**

(PSICA part-2)

Faculty: Dierdre Lawless

NAME: ASHIS SAHU

STUDENT ID: **D17129721**

COURSE: **DT-228A ( DATA ANALYTICS )**

# Introduction

The initial exploration was done in continuous assessment Part I assignment, which was focused on identifying factors influencing satisfaction of sleep. Understanding these factors will help an individual to have changed or schedule particular events in their particular habits of sleeping. This paper will continue and build upon that analysis identifying additional factors, which contribute towards satisfaction of sleep and then combining all investigated factors to build ordinal logistic regression model capable of predicting satisfaction of sleep of individual with various factors related to sleep.

There can be various factors affecting ones satisfaction of sleep, it can be the quality of sleep, drinking caffeine or alcohol drinks, fit rate, health rate of an individual or certain habits such as drinking alcohol, smoking, etc. Though there can be various factors affecting sleep, various variables have been taken for the research from sleep dataset which has 271 records which were collected from a survey using the sleep questionnaire. The main constraint variable or outcome variable in this research is the satisfaction in sleep which is a scale variable ranging from 1 to 10. Normality test has been done with continuous variables and variables with normalized scores have been undergone hypothesis testing and reporting with respect to the satisfaction of sleep. The variables under consideration are quality of sleep, alcohol, fit rate, health rate, caffeine, weeknight hours, weekend hours, and quality sleep record. The correlation hypothesis test is used for ordinal variables for satsleep and qualslp, as both were ordinal variables. Pearson correlation is done with alcohol and fitrate as both were continuous variables. The statistical tests have been produced with respect to each variable type and reported after each test. This paper will continue and build upon that analysis identifying factors affecting satisfaction of sleep, and then combining all investigated factors to build ordinal logistic regression model capable of predicting a individuals satisfaction of sleep.

The initial analysis found that factors such as quality of sleep, sleep hours in the weekend, sleep hours in weekdays all had a small but significant contribution towards satisfaction of sleep of individual. Moreover, an additional nominal variable quality sleep record (qualsleeprec) has been taken in the additional model, in order to record the statistical difference between the two model due to additional of nominal variable and change in accuracy due to the changes.

The sleep dataset is a real data file condensed from a study conducted to explore the prevalence and impact of sleep problems on various aspects of people’s lives. Staff from a university in Melbourne, Australia were invited to complete a questionnaire containing questions about their sleep behavior (e.g. hours slept per night), sleep problems (e.g. difficulty getting to sleep) and the impact that these problems have on aspects of their lives (work, driving, relationships). The sample consisted of 271 respondents (55% female, 45% male) ranging in age from 18 to 84 years (mean=44yrs).

The statistically significant variables to as a factor of satisfaction of sleep have been considered for the paper. An additional factor, quality sleep record has been taken as a nominal variable in order for comparison of the previous test in order to form proceeding ordinal logistic regression model.

## 1.1 Hypotheses correlation test

1. There is a no correlation between Quality of sleep and satisfaction of sleep of individual.

2. There is no correlation between satisfaction of sleep and fitrate for respondents

3 There is no correlation between in satisfaction of sleep and healthrate for respondents.

4. There is no correlation between satisfaction of sleep and caffeine.

5. There is no correlation between satisfaction of sleep and hours of sleep during week end.

6. There is no correlation between satisfaction of sleep and hours of sleep during week night.

## 1.2 Ordinal Logistic regression Model

Model1: Prediction of satisfaction of sleep with hourwend, hourwnit.

Model2: Prediction of satisfaction of sleep with hourwend, hourwnit, qualslp.

Model3: Prediction of satisfaction of sleep with hourwend, hourwnit, qualslp and qualsleeprec.

# Methodology

The paper is the research of the factors affecting satisfaction of sleep which is a variable taken from sleep data set which was collected by questionnaire .The sleep dataset is a real data file condensed from a study conducted to explore the prevalence and impact of sleep problems on various aspects of people’s lives. Staff from a university in Melbourne, Australia were invited to complete a questionnaire containing questions about their sleep behavior (e.g. hours slept per night), sleep problems (e.g. difficulty getting to sleep) and the impact that these problems have on aspects of their lives (work, driving, relationships).

## 2.1 Materials

### 2.1.1 Quality of sleep

Quality of sleep (m=`3.69`, sd=`1.14`, n=`268`) is represented by an ordinal variable in the sleep dataset which was obtained from the sleep questionnaire. It consists of quality of sleep experienced by respondent on a range of scores from 1 t0 6. The maximum quality of sleep a respondent can register is 6 and minimum of 1 score. Quality of sleep is the predictor variable in the data set. Histogram of quality of sleep has been constructed and descriptive statistics have been presented. The main motive of the paper is to find the factors affecting satisfaction of sleep, the quality of sleep comes in as a factor of satisfaction of sleep.

### 2.1.2 Satisfaction in sleep

Satisfaction of sleep (m=`5.54`, sd=`2.5`, n=`268`) is represented by an ordinal variable in the sleep dataset which was obtained from the sleep questionnaire. It consists of satisfaction of sleep experienced by the respondent on a range of scores from 1 to 10. The maximum satisfaction of sleep a respondent can register is 10 and minimum of 1 score. Satisfaction of sleep is the outcome variable in the dataset. Histogram of satisfaction of sleep has been constructed and descriptive statistics have been presented. The main motive of the paper is to find the factors affecting satisfaction of sleep.

### 2.1.3 Caffeine

Caffeine (m=`2.94`, sd=`1.92`, n=`270`) is represented by a continuous variable in the sleep dataset which was obtained from the sleep questionnaire. It consists of the number of caffeine drink the respondent consumes each day. A number of caffeine drinks in the dataset ranges from 0 to 10. Caffeine is the predictor variable in the dataset. Histogram of caffeine has been presented and descriptive statistics have been presented. Normality analysis has been also presented with finding the skew and kurtosis. The main motive of the paper is to find the factors affecting satisfaction of sleep, a number of caffeine drinks come in as a factor of satisfaction of sleep.

Caffeine is represented by a continuous variable in the dataset which was calculated from the sleep questionnaire. Inspection of the histogram and normality plot shows that the distribution does not conform for a normal distribution. Inspection of standardized normal scores for skewness and kurtosis indicated that the skewness fall with an acceptable range of +/-2 obatined from (Field, A., Miles, J., & Field, Z. (2012)), (skewness of `0.67` SE=`1.51`, the kurtosis of `0.56` SE=`0.30`. Further inspection of the variable, showed that skew and kurtosis scores were inside the acceptable range. Caffeine will, therefore, be treated as a normal within this analysis.

### 2.1.4 Health rate

Health rate (m=`7.79`, sd=`1.57`, n=`267`) is represented by an ordinal variable in the sleep dataset which was obtained from the sleep questionnaire. It consists of health rate of the respondent on a range of scores from 1 to 10. The maximum health rate of a respondent can register is 10 and minimum of 1 score. Health rate is the predictor variable in the paper. Histogram of health rate has been presented and descriptive statistics have been presented. The main motive of the paper is to find the factors affecting satisfaction of sleep, the health rate of the individual comes in as a factor of satisfaction of sleep.

### 2.1.5 Alcohol

Alcohol(m=`5.54`, sd=`2.5`, n=`271`) is represented as a continuous variable in the dataset which was obtained from the sleep questionnaire. Inspection of standardized normal scores for skewness and kurtosis indicated that the skewness falls out of the acceptable range of +/-2 obatined from (Field, A., Miles, J., & Field, Z. (2012)), (skewness of `3.63` SE=`0.15`, the kurtosis of `22.4` SE=`0.30`. Further inspection of the variable using standardized scores showed that 0% of standardized scores were outside the acceptable range of +/3.29. Therefore, Alcohol will not be undergone normality test and further, it will not undergo hypothesis testing or included in the model as a factor affecting the satisfaction of sleep.

### 2.1.6 Hourwend

Hourwend (m=`8`, sd=`1.33`, n=`269`) is represented by a continuous variable in the sleep dataset which was obtained from the sleep questionnaire. It consists of a number of hours slept by the respondent each weekend. A number of hour slept by the respondent each weekend in the dataset ranges from 4 to 14. Hour went is the predictor variable in the dataset. Histogram of an hour went has been presented and descriptive statistics have been presented. Normality analysis has been also presented with finding the skew and kurtosis. The main motive of the paper is to find the factors affecting satisfaction of sleep, hour slept by the respondent each weekend comes as a factor of satisfaction of sleep.

Inspection of the histogram and normality plot shows that the distribution does not conform for a normal distribution. Inspection of standardized normal scores for skewness and kurtosis indicated that the skewness and kurtosis fall with an acceptable range of +/-3, (skewness of `0.09` SE=`0.14`, the kurtosis of `2.42` SE=`2.98`. Further inspection of the variable, showed that the variable tends to attain normality. Therefore, it treated as a normal within this analysis.

### 2.1.7 Hourwnit

Hourwnit (m=`6.97`, sd=`1.067`, n=`270`) is represented by a continuous variable in the sleep dataset which was obtained from the sleep questionnaire. It consists of a number of hours slept by the respondent each weekend. A number of hours slept by the respondent each weekend in the dataset ranges from 3 to 10. Hourwend is the predictor variables in the data set. Histogram of hourwend has been presented and descriptive statistics have been presented obatined from (Field, A., Miles, J., & Field, Z. (2012)). Normality analysis has been also presented with finding the skew and kurtosis. The main motive of the paper is to find the factors affecting satisfaction of sleep, hour slept by the respondent each weekend comes as a factor of satisfaction of sleep.

Inspection of the histogram and normality plot shows that the distribution does not conform for a normal distribution. Inspection of standardized normal scores for skewness and kurtosis indicated that the skewness and kurtosis fall with an acceptable range of +/-3 obatined from (Field, A., Miles, J., & Field, Z. (2012)), (skewness of `0.40` SE=`0.14`, the kurtosis of `0.68` SE=`0.29`. Further inspection of the variable, showed that the variables tend to attain normality. Therefore, it treated as a normal within this analysis.

### 2.1.8 Qualsleeprec

Qualsleeprec (m=`2.70`, sd=`0.99`, n=`268`) is represented by a nominal variable in the sleep dataset which was obtained from the sleep questionnaire. Qualsleeprec is the record of quality of sleep by an individual respondent. It consists of 6 options (1=very poor, 2=poor, 3=fair, 4=good,5=very good, 6=excellent). Qualslp is a nominal variable, so, it is a predictor variable in the paper. Histogram of qualsleeprec has been presented and descriptive statistics have been presented. The main motive of the paper is to find the factors affecting satisfaction of sleep, therefore qualsleeprec has been introduced in the logistic regression model.

## 2.2Hypothesis correlation testing results

### 2.2.1. There is a correlation between Quality of sleep and satisfaction of sleep of individual.

The relationship between Satisfaction in sleep and Quality of sleep (derived from the sleep questionnaire) was investigated using a Pearson correlation. A strong positive correlation was found (r =.576, p<.001). The correlation test provided evidence to reject the null hypothesis, that there is no relationship between the Satisfaction of sleep and Quality of sleep.P-value < 2.2e-16, shows there is a statistical significance and correlation between satisfaction of sleep and quality of sleep. The p-value tells us that the probability of this correlation being due to random chance is very low (close to zero in fact). Hence, this relationship is genuine. Therefore, qualslp is suitable for the logistic regression model predicting the satisfaction of sleep.

### 2.2.2 There is no correlation between satisfaction of sleep and fitrate for respondents

The relationship between Satisfaction in sleep and fitrate (derived from the sleep questionnaire) was investigated using a Kendal correlation. Statistics from this hypothesis testing generated are (r =0.33, p = 0.47). After conducting the test, we got z = '0.71', n='261'. The probability is higher than the conventional 5%, the correlation coefficient will be called statistically insignificant. There is a evidence to retain the null hypothesis. The P-value > 0.05 obatined from (Field, A., Miles, J., & Field, Z. (2012)), shows there is a no statistically significant correlation between the respondents with satisfaction in sleep and fitrate. Therefore, fitrate is not suitable for the logistic regression model predicting satisfaction of sleep.

### 2.2.3 There is no correlation between in satisfaction of sleep and health rate for respondents.

The statistical relationship between Satisfaction in sleep and health rate (derived from the sleep questionnaire) was investigated using a Kendall correlation, as both are ordinal variables. After conducting the test, following scores were observed z = '1.79', n='261', p= 0.07. The probability is higher than the conventional 5%, the correlation coefficient will be called statistically insignificant. Therefore, there is no evidence to accept the alternative hypothesis and retain the null hypothesis. The P-value > 0.05 obatined from (Field, A., Miles, J., & Field, Z. (2012)), shows there is a no statistically significant correlation between the both. Therefore, healthrate is not suitable for the logistic regression model predicting the satisfaction of sleep.

### 2.2.4. There is no correlation between satisfaction of sleep and caffeine.

The statistical relationship between Satisfaction in sleep and caffeine (derived from the sleep questionnaire) was investigated using a Pearson correlation. After conducting the test, following scores were observed r = '-0.081', n='257', p= 0.18. The probability is higher than the conventional 5%, the correlation coefficient will be called statistically insignificant obatined from (Field, A., Miles, J., & Field, Z. (2012)). Therefore, there is no evidence to accept the alternative hypothesis and therefore the null hypothesis is retained. The P-value > 0.05, shows there is a no statistically significant correlation between the both. Therefore, caffeine is not suitable for the logistic regression model predicting the satisfaction of sleep.

### 2.2.5. There is no correlation between the satisfaction of sleep and hours of sleep during the weekend.

The relationship between Satisfaction in sleep and hours of sleep during the weekend (derived from the sleep questionnaire) was investigated using a Pearson correlation. A positive correlation was found (r =.26, p = 1.645e-05). The correlation test provided evidence to reject the null hypothesis, that there is no relationship between the Satisfaction of sleep and Quality of sleep and accept the alternate hypothesis. The P-value = 1.645e-05 shows there is a statistical significance and correlation between satisfaction of sleep and hourwnit. The p-value tells us that the probability of this correlation being due to random chance is very low (close to zero in fact) obatined from (Field, A., Miles, J., & Field, Z. (2012)). Hence, this relationship is genuine. Therefore, hourwend is suitable for the logistic regression model predicting the satisfaction of sleep.

### 2.2.6. There is no correlation between the satisfaction of sleep and hours of sleep during a weeknight.

The relationship between Satisfaction in sleep and hours of sleep during weeknight (derived from the sleep questionnaire) was investigated using a Pearson correlation. A strong positive correlation was found (r =.376, p = 2.011e-10). The correlation test provided evidence to reject the null hypothesis, that there is no relationship between the Satisfaction of sleep and Quality of sleep and accept the alternate hypothesis. P-value = 2.011e-10 shows there is a statistical significance and correlation between satisfaction of sleep and hourwnit. The p-value tells us that the probability of this correlation being due to random chance is very low (close to zero in fact) obatined from (Field, A., Miles, J., & Field, Z. (2012)). Hence, this relationship is genuine. Therefore, hourwnit is suitable for the logistic regression model predicting the satisfaction of sleep.

# 3.Methodology

## 3.1 Data cleaning

The dataset had many NA or missing variables, which are properly handled by omitting the rows with the NA or missing variables. As the variable under consideration were ordinal data obtained from Gaston Sanchez (2018).

## 3.2 Data Slicing

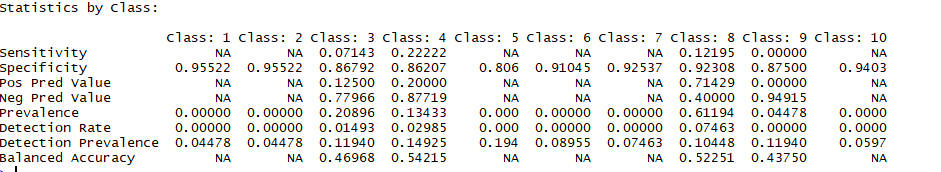
The data has been sliced in 75% train data (199 records) and 25% test data(67 records) of the sleep dataset for proceeding in the regression model. The satsleep has been taken as a factor as the model, not to consider it as a numeric variable obtained from (Andy Field , Jeremy Miles , Zoe Field.(2018)). The seed has been set to 123 as a random variable as a parameter of data slicing. Codes obtained from (Garrett Grolemund , Hadley Wickham.(2017)).

# 4. Results

## 4.1 Ordinal Logistic Regression model 1

A ordinal linear regression was calculated to predict a satisfaction of sleep based on the ordinal variables hour sleep weeknight, hour sleep weekend. A significant regression equation was found with Residual Deviance: 843.17 and AIC: 865.17. Satisfaction of sleep is equal to 0.152(hourwnit) + 0.99(hourwend).By observation of p-values, hourwend(p<0.01) and hourwnit (p< 0.05) were significant predictors of satisfaction of sleep obtained from (Andy Field , Jeremy Miles , Zoe Field.(2018)).. The ordinal linear regression model has Pr(chisq)<2.2e-16 and chisq = 105.59 when undergone a likelihood ratio test. The coefficients for the model are given by hourwnit = 1.830,and hourwend = 1.103.

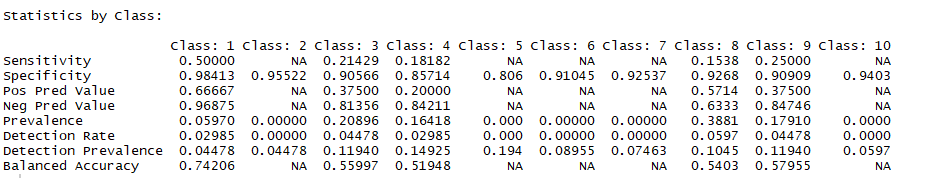
The predicted values have been computed on the test data set and confusion matrix has been formed with an accuracy of 11.94%. As the outcome variable satsleep is the ordinal variable, the statistics by class has been presented in the paper.



## 4.2 Ordinal Logistic Regression model 2

A ordinal linear regression was calculated to predict a satisfaction of sleep based on the ordinal variables quality of sleep, hour sleep weeknight, hour sleep weekend. A significant regression equation was found with Residual Deviance: 768.9973 and AIC: 792.9973. Satisfaction of sleep is equal to 0.146(qualslp) + 0.147(hourwnit) + 0.118(hourwend).By observation of p-values, Qualslp (p< 0.1), hourwend(p<0.01) and hourwnit (p< 0.05) were significant predictors of satisfaction of sleep obtained from (Andy Field , Jeremy Miles , Zoe Field.(2018)).. The ordinal linear regression model has Pr(chisq)<2.2e-16 and chisq = 105.59 when undergone a likelihood ratio test. The coefficients for the model are given by qualsp = 3.21, hourwnit = 1.7682498, and hourwend = 0.886.

The predicted values have been computed on the test data set and confusion matrix has been formed with an accuracy of 20.9%. As the outcome variable satsleep is ordinal variable, the statistics by class has been presented in the paper.



## 4.3 Ordinal Logistic Regression model 3

A ordinal linear regression was calculated to predict a satisfaction of sleep based on the ordinal variables quality of sleep, hour sleep weeknight, hour sleep weekend and nominal variable qualsleeprec. A significant regression equation was found with Residual Deviance: 760.043 and AIC: 786.043.By interpretation of model, Satisfaction of sleep is equal to 0.601(qualslp) + 0.156(hourwnit) + 0.133(hourwend) + 0.66(qualsleeprec) .By observation of p-values, Qualslp (p< 0.1), hourwend(p<0.01) and hourwnit (p< 0.05) were significant predictors of satisfaction of sleep obtained from (Andy Field , Jeremy Miles , Zoe Field.(2018)). The ordinal linear regression model has Pr(chisq)<2.2e-16 and chisq = 115.59 when undergone a likelihood ratio test. The coefficients for the model are given by qualsp = 2.36, hourwnit = 1.73,and hourwend = 0.94, qualsleeprec = 1.64.

The predicted values have been computed on the test data set by the model and confusion matrix has been formed with an accuracy of 20.9%. As the outcome variable satsleep is the ordinal variable, the model statistics by class has been presented in the paper

.

## 4.4 Model comparison

The likelihood ratio test of ordinal regression models with satsleep as response varaible conveys that the first ordinal logistic model has residual deviance = '843.1702', df = '188'. Second ordinal logistic model, has residual deviance = '760.5940', df = '187'. Third ordinal logistic model has residual deviance = '760.0430', df = '186'.The likelyhood ratio of first and second model is given by 82.5761 and second and third model is given by 0.5510.

# 

# 5. Discussion and conclusion

The satisfaction of sleep is the major attribute in this paper, there are three ordinal logistic regression models predicting the satisfaction of sleep. The accuracy of the first model which had two continuous variables hourwend and hourwnit is 11.49% and degree of freedom is 188. In the second model which had two continuous variables, hourwend and hourwnit and one ordinal variable satsleep are 20.9% and degree of freedom is 187. There is a significant difference in accuracy has been observed in the second model when an ordinal variable added to the model. In the third model, which had two continuous variables hourwend and hourwnit and one ordinal variable satsleep and one nominal variable qualsleeprec, has an accuracy of 20.9% and degree of freedom is 186. There is no difference in accuracy observed in the second model and the third model. But, the residual deviance and degree of freedom of the third model are smaller than the second model. Considering, all we can say the third model would give an accurate and efficient result as compared to all three models considering that, there is no difference in the accuracy of the second or third model. There is no difference in adding the nominal variable in the ordinal regression model.

The continuous variable has undergone normality analysis and the variables which didn’t have normal distribution has been used in the ordinal logistic regression model. All the variables have undergone hypothesis testing leaving the nominal variable qualsleeprec. The variable caffeine, fitrate, and healthrate of sleep

Dataset has no statistical significance with the satisfaction of sleep, therefore these variable not been reproduced in the model.

The paper concludes that the quality of sleep, sleep hour’s weekend and weeknight is the main factor affecting the satisfaction of sleep. Quality of sleep signifies major the portion for an individual to acquire satisfaction of sleep.

# 6. REFERENCES

Field, A., Miles, J., & Field, Z. (2012). Discovering statistics using R. Sage publications.

George, D. (2011). SPSS for windows step by step: A simple study guide and reference, 17.0 update, 10/e. Pearson Education India.

Andy Field , Jeremy Miles , Zoe Field.(2018). Discovering Statistics Using R. Sage Publications Ltd.

Garrett Grolemund , Hadley Wickham.(2017). R for Data Science. OReilly Media Inc USA

Gaston Sanchez (2018). Handling Missing Values, Department of Statistics, UC–Berkeley