


# 19088325 Portfolio

I have difficulty getting sufficient sleep. I consistently and autonomously wake up early in the morning regardless of what time I fell asleep, so my length of sleep is typically short. I want to investigate how I could improve my length of sleep. I researched possible methods, and chose to measure my water intake, based on a news article from the Independent, and my room temperature and stress levels, based on the commonly used Pittsburgh Sleep Quality Index questionnaire, to determine if they could impact my sleep quality.

 INDEPENDENT


Subscribe

LOGIN

☰

NEWSINDEPENDENT TVCLIMATEFOOTBALLVOICESCULTUREPREMIUMINDY/LIFEINDYBESTINDY100VOUCHERSCOMPARE





🔍🌐




**NOT GETTING ENOUGH SLEEP?**  
**DRINKING MORE WATER COULD HELP**  
**YOU**

A study involving more than 20,000 people claims that hydration is the key to a good night's sleep

Olivia Petter | @oliviapetter1 | Thursday 08 November 2018 15:13 | comments





Here, I will analyse a dataset which contains daily data about my length of sleep (in hours), water intake (in mL), stress levels (on a scale from 1 to 5, 1 being lowest), and room temperature upon waking up (in celsius).

```
library(readxl)
library(modelr)
```

```
## Warning: package 'modelr' was built under R version 4.0.5
```

```
mydata <- read_excel("C:/Users/jasmn/Documents/uni/STAT601 Statistical Methods/SLEEP DATA FINAL.xlsx")
```

```
mydata
```

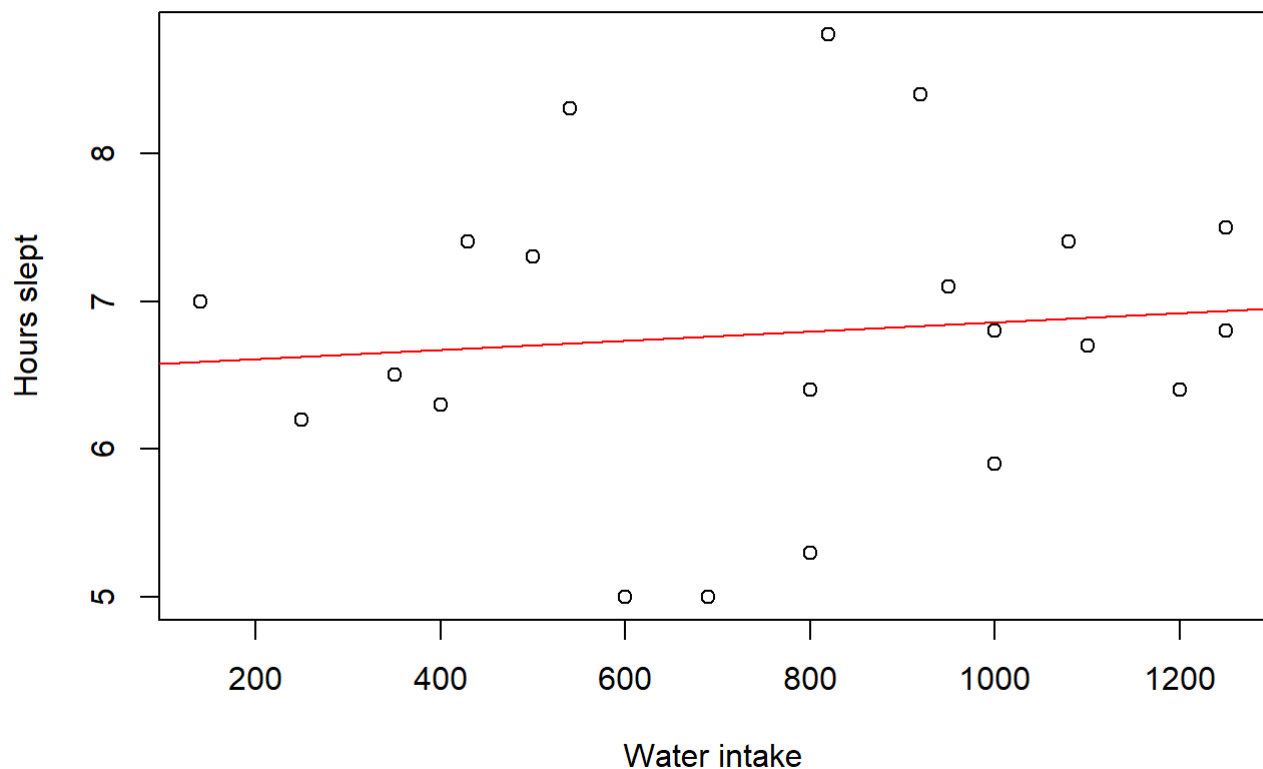
```
## # A tibble: 21 x 5
##   Day `Hours slept` `Water consumption` `Stress level` `Room temperature`
##   <dbl>         <dbl>         <dbl>         <dbl>         <dbl>
## 1     1           5.3           800           0           19.5
## 2     2           7.3           500           1           22
## 3     3           7.4           430           2           20.3
## 4     4           6.3           400           4           19.4
## 5     5           5           600           3           18.5
## 6     6           8.3           540           2           18.9
## 7     7           7.4          1080           1           20
## 8     8           7           140           0           20.2
## 9     9           5           690           2           18
## 10    10          6.4          1200           4           19
## # ... with 11 more rows
```

## Linear regression

I am first going to plot linear models of the variables against length of sleep, to see if any have a linear relationship worth investigating.

```
plot(mydata$`Hours slept` ~ mydata$`Water consumption`, xlab="Water intake", ylab="Hours slept",
     main="Water intake against sleep") +
  abline(lm(mydata$`Hours slept` ~ mydata$`Water consumption`), col="red")
```

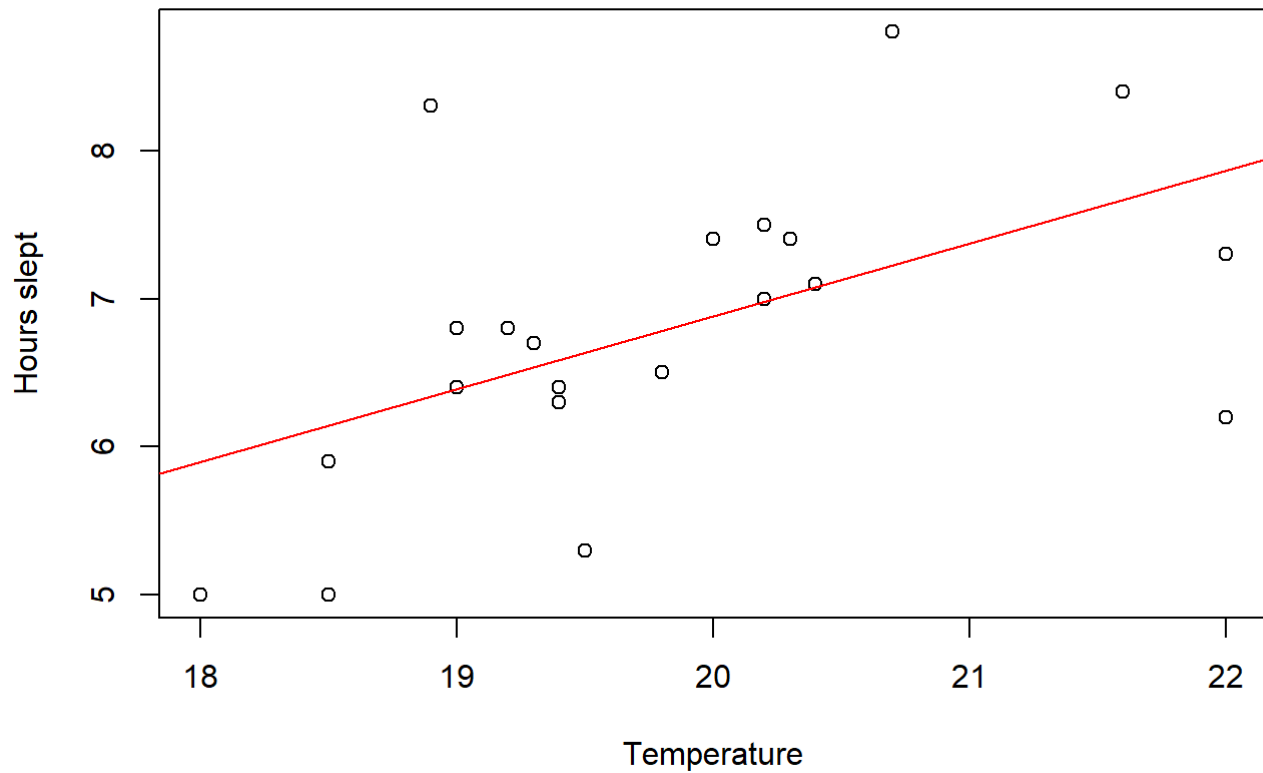
**Water intake against sleep**



```
## integer(0)
```

```
plot(mydata$`Hours slept` ~ mydata$`Room temperature`, xlab="Temperature", ylab="Hours slept"
, main="Room temperature against sleep") +
  abline(lm(mydata$`Hours slept` ~ mydata$`Room temperature`), col="red")
```

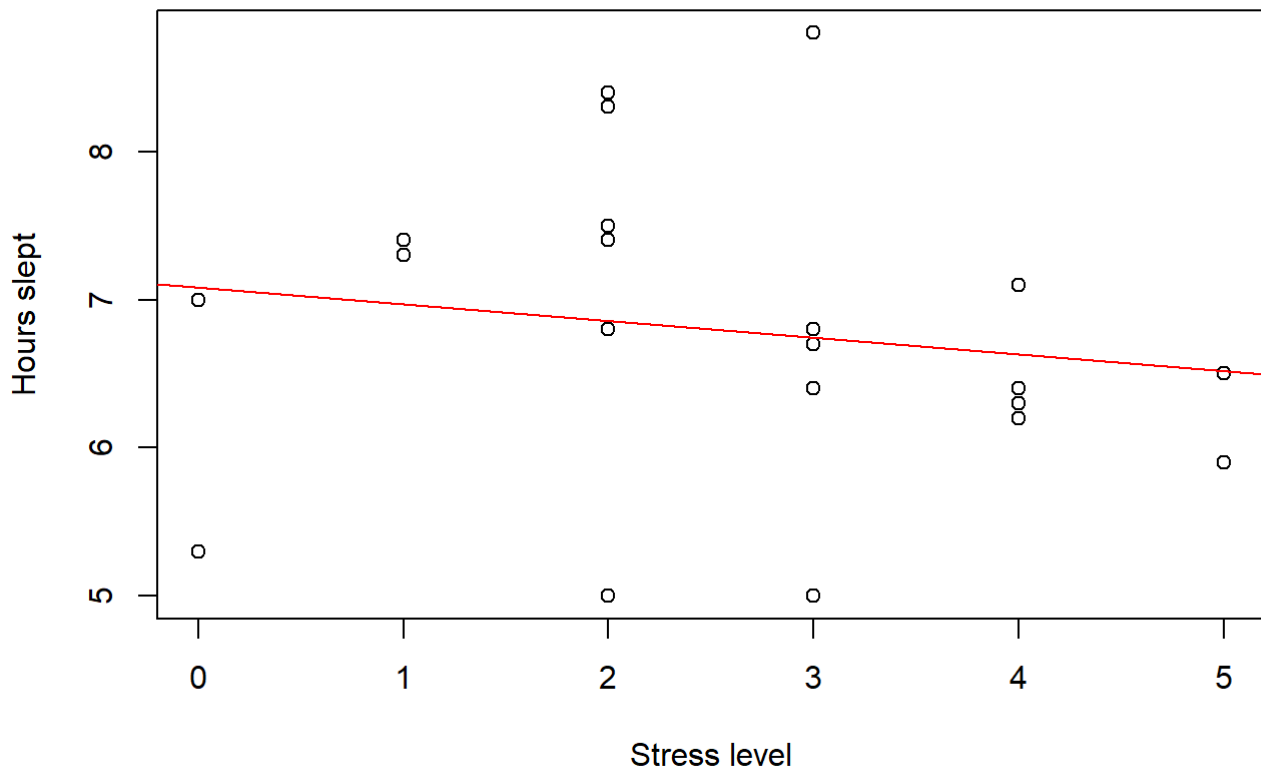
### Room temperature against sleep



```
## integer(0)
```

```
plot(mydata$`Hours slept` ~ mydata$`Stress level`, xlab="Stress level", ylab="Hours slept", m
ain="Stress level against sleep") +
  abline(lm(mydata$`Hours slept` ~ mydata$`Stress level`), col="red")
```

## Stress level against sleep



```
## integer(0)
```

Out of all three variables, a linear trend appears appropriate only for room temperature - room temperature appears most likely to have a non-independent relationship with length of sleep.

```
summary(lm(mydata$`Hours slept` ~ mydata$`Room temperature`))
```

```
##
## Call:
## lm(formula = mydata$`Hours slept` ~ mydata$`Room temperature`)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.66416 -0.28687  0.02012  0.40964  1.95877
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2.9437     3.5738  -0.824   0.4203
## mydata$`Room temperature`  0.4913     0.1802   2.726   0.0134 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8906 on 19 degrees of freedom
## Multiple R-squared:  0.2812, Adjusted R-squared:  0.2434
## F-statistic: 7.434 on 1 and 19 DF, p-value: 0.0134
```

The null hypothesis is that there is no linear trend/relationship (variables are independent) -  $H_0: \beta = 0$ .

The p-value is 0.01. The p-value is the probability, given that the null is true, of obtaining the observation or one more extreme. Here, more extreme means a  $\beta$ -value different from 0.

Since the p-value is less than 0.05, I have evidence to reject the null. This suggests that there is no independence between length of sleep and room temperature.

I will perform a Fisher's exact test, to further investigate the independence of their relationship.

## Fisher's exact test

I need to determine what a significantly low and high temperature will be, using a confidence interval of 95%.

```
temp_mean <- mean(mydata$`Room temperature`)  
sleep_mean <- mean(mydata$`Hours slept`)
```

```
t.test(mydata$`Room temperature`, alternative="two.sided")
```

```
##  
## One Sample t-test  
##  
## data: mydata$`Room temperature`  
## t = 82.118, df = 20, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 19.30168 20.30784  
## sample estimates:  
## mean of x  
## 19.80476
```

### Room temperature

- Lower 5% = 19.30
- Upper 5% = 20.31

```
t.test(mydata$`Hours slept`, alternative="two.sided")
```

```
##  
## One Sample t-test  
##  
## data: mydata$`Hours slept`  
## t = 30.371, df = 20, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 6.319660 7.251769  
## sample estimates:  
## mean of x  
## 6.785714
```

### Length of sleep

- Lower 5% = 6.32
- Upper 5% = 7.25

```
temp_l5 <- 19.30
temp_u5 <- 20.31
```

```
sleep_l5 <- 6.32
sleep_u5 <- 7.25
```

```
#High room temp and high length of sleep
table(mydata$`Room temperature` > temp_u5 & mydata$`Hours slept` > sleep_u5)
```

```
##
## FALSE TRUE
##    18    3
```

```
#High room temp and low length of sleep
table(mydata$`Room temperature` > temp_u5 & mydata$`Hours slept` < sleep_l5)
```

```
##
## FALSE TRUE
##    20    1
```

```
#Low room temp and high length of sleep
table(mydata$`Room temperature` < temp_l5 & mydata$`Hours slept` > sleep_u5)
```

```
##
## FALSE TRUE
##    20    1
```

```
#Low room temp and low length of sleep
table(mydata$`Room temperature` < temp_l5 & mydata$`Hours slept` < sleep_l5)
```

```
##
## FALSE TRUE
##    18    3
```

```
#Creating two-way table
temp_table <- matrix(c(3,1,1,3), 2, 2)

colnames(temp_table) <- c("High temp", "Low temp")
rownames(temp_table) <- c("High sleep", "Low sleep")

temp_table <- as.table(temp_table)

temp_table
```

```
##           High temp Low temp
## High sleep         3         1
## Low sleep          1         3
```

```
fisher.test(temp_table, alternative="two.sided")
```

```
##
## Fisher's Exact Test for Count Data
##
## data: temp_table
## p-value = 0.4857
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##    0.2117329 621.9337505
## sample estimates:
## odds ratio
##    6.408309
```

For this test, the null hypothesis is that the two variables - length of sleep and room temperature - are independent.

The p-value is greater than 0.05. Hence, I cannot reject the null.

This gives two contradicting results. The linear regression p-value suggests that there is no independence, but Fisher's exact test suggests that there is independence.

I want to test to see if I have made any type 1 or type 2 errors.

For the linear regression, I rejected the null. It is possible for me to have committed a type 1 error.

For the Fisher's exact test, I accepted the null. It is possible for me to have committed a type 2 error.

## Calculating Type 1 Error

```
set.seed(0)

#Get p-value of linear regression (lm)
lmp <- function(modelobject) {
  if (class(modelobject) != "lm") stop("Not an object of class 'lm' ")
  f <- summary(modelobject)$fstatistic
  p <- pf(f[1],f[2],f[3],lower.tail=F)
  attributes(p) <- NULL
  return(p)
}

#Resample function
resample_a <- function(a, b){
  A <- sample(a, size=length(a), replace=TRUE)
  B <- sample(b, size=length(b), replace=TRUE)

  p <- lmp(lm(A~B))

  return(p)
}

#1000 resamplings
p_val <- replicate(1000, resample_a(mydata$`Hours slept`, mydata$`Room temperature`))

#Probability of committing a type 1 error
(table(p_val < 0.05)/1000)['TRUE']
```

```
## TRUE
## 0.045
```

The probability of committing a type 1 error for my linear regression is 0.045 -  $\alpha = 0.045$ .  $\alpha$  is small,  $\beta$  is large. Since  $\alpha$  is small and  $\beta$  is large, it is unlikely that my rejection of the null hypothesis ( $\beta = 0$ , independent) is an error. It is more likely that the null is false - more likely that variables are not independent.

```
#Resample function #####
resample_b <- function(sleep, temperature){

  S <- sample(sleep, size=length(sleep), replace=TRUE)
  T <- sample(temperature, size=length(temperature), replace=TRUE)

  bind <- cbind(S, T)
  colnames(bind) <- c("Sleep", "Temp")

  resample <- as.data.frame(bind)

  return(resample)
}
```

```
#Get 95% CI values #####
t.test_p <- function(D){

  #Lower limit
  lower <- t.test(D)$conf.int[1]

  #Upper limit
  upper <- t.test(D)$conf.int[2]

  limits <- c(lower, upper)

  return(limits)
}
```



```
#Get values to create 2-way table #####
```

```
table_function <- function(D){

  s_lim <- t.test_p(D$Sleep)
  t_lim <- t.test_p(D$Temp)

  #High sleep, high temp
  a <- (table(D$Sleep >= s_lim[2] & D$Temp >= t_lim[2])['TRUE'])

  if(is.na(a)) {
    a <- 0
  }
  else {
    a <- a
  }

  #Low sleep, high temp
  b <- (table(D$Sleep <= s_lim[1] & D$Temp >= t_lim[2])['TRUE'])

  if(is.na(b)) {
    b <- 0
  }
  else {
    b <- b
  }

  #High sleep, Low temp
  c <- (table(D$Sleep >= s_lim[2] & D$Temp <= t_lim[1])['TRUE'])

  if(is.na(c)) {
    c <- 0
  }
  else {
    c <- c
  }

  #Low sleep, Low temp
  d <- (table(D$Sleep <= s_lim[1] & D$Temp <= t_lim[1])['TRUE'])

  if(is.na(d)) {
    d <- 0
  }
  else {
    d <- d
  }

  f_matrix <- matrix(c(a, b, c, d), 2, 2)

  colnames(f_matrix) <- c("High temp", "Low temp")
  rownames(f_matrix) <- c("High sleep", "Low sleep")

  return(f_matrix)
}
```

```
#Get p-value of Fisher's exact test #####
fisher_p <- function(D){
  fisher.test(as.table(D), alternative="two.sided")$p.value
}
```

```
#COMPLETE FUNCTION
final <- function(sleep_col, temp_col){

  sample <- resample_b(sleep_col, temp_col)

  sample_table <- table_function(sample)

  sample_fisher_p <- fisher_p(sample_table)

  return(sample_fisher_p)
}
```

## Calculating Type 2 Error

```
set.seed(1)

#Get p-values from Fisher's test over 1000 samples
fishers <- replicate(1000, final(mydata$`Hours slept`, mydata$`Room temperature`))

#Probability of committing a Type 2 error
(table(fishers > 0.05)/1000)['TRUE']
```

```
## TRUE
## 0.986
```

The probability of committing a type 2 error for my Fisher's exact test is 0.986 -  $\beta = 0.986$ .  $\beta$  is large,  $\alpha$  is small.

$\beta$  is very high, so it is likely that my failure to reject the null hypothesis (two variables - length of sleep and room temperature - are independent) is false. This means that it is likely that the null is false - variables are not independent.

The two error tests conclude that:

- (Linear regression null hypothesis: length of sleep and room temperature are independent) is very likely to be false
- (Fisher's exact test null hypothesis: length of sleep and room temperature are independent) is very likely to be false

## Conclusion

I can conclude that it is very likely that length of sleep and room temperature are not independent of each other.

```
max(mydata$`Room temperature`)
```

```
## [1] 22
```

```
min(mydata$`Room temperature`)
```

```
## [1] 18
```

This seems like a reasonable conclusion to draw. My dataset had a relatively small range of values, of standard room temperature (18 - 22 degrees celsius). Lower standard room temperatures will be less comfortable, hence disturbing my sleep.

From the previous linear regression, the R-squared value is somewhat low -  $R\text{-squared} = 0.24$ . This means that the linear regression model does not fit the datapoints for this dataset very well. Based on my previous conclusion it is possible that this linear model (i.e,  $\text{length of sleep} = 0.49 \times (\text{room temperature}) - 2.94$ ) was suitable to fit the data I provided, but may not be suitable to model the dataset for more extreme values - very cold AND hot temperatures will be uncomfortable and disturb my sleep - hence the low p-value and R-squared value.

I did not find a direct linear relationship between length of sleep and water consumption or stress level. However, since it is possible that the linear regression model I plotted for the relationship between room temperature and length of sleep is unsuitable, a multivariate regression model that includes these variables could improve the accuracy of the linear regression model.

## Summary

I analysed a dataset that includes data about my length of sleep, room temperature upon waking up, daily water intake, and stress levels.

Linear regression ( $p=0.01$ ) and Fisher's exact test ( $p=0.5$ ) provided two opposing conclusions, but doing error testing for the two analyses ( $\alpha = 0.045$ ,  $\beta = 0.986$ , respectively) allowed me to find that length of sleep and room temperature are not independent of each other.

The room temperature impacted my length of sleep, probably because colder temperatures disturb my sleep whilst warmer (more comfortable) temperatures did not. I did not find a direct relationship between length of sleep and water intake or stress level, based on scattergraphs.