Sleep Study

Michelle Rice

2/25/2021

# Project Summary

I found this sleep dataset interesting for a number of reasons. As someone who has dealth with insomnia in the past and now sometimes does not get enough sleep due to overloaded schedules, family commitments, etc., I often feel tired. It also seems as though many others in the workplace, social circles, etc, often mention being tired. We hear that the goal for sleep is 8 hours, but some people seem to function just fine on much less while others can get 8 hours of sleep and still complain of being tired. As the old adage goes, "everyone is different." So, I wouldn't expect everyone to have the same needs or experience, but I was curious to see if there was a strong correlation that could be found to other factors causing us to still feel tired.  
  
The problem statement addressed is, how much sleep is enough (and what does "enough" really mean) and what factors might contribute to feeling tired or not feeling you got enough sleep.  
  
  
Methodologies I used include creating scatterplots of data to identify any visual trends and correlatons, using linear regression for my continuous variables and logistic regression for categorical variables.   
  
One interesting insight that I saw right away in the data was that the correlation between hours of sleep and feeling tired was not as linear as I would have expected.   
  
A challenge that I faced was using different variable types and knowing if I am modeling correclty based on the type of data that I have. I was surprised by my findings that breakfast had a stronger correlation to feeling of tiredness than hours of sleep, but I'm also not completely sure that I did this correctly due to it being a nominal variable.   
  
I feel that I could continue with this analysis and look more closely at the "enough" variable to see if there are trends there that might help with the analysis. One problem with this dataset is that the two variables that I was trying to predict, enough and tiredness rating, are based on personal feeling, so two people could have all other variables equal but one person still rates themself as more tired or not getting enough sleep. Or a level 3 tiredness rating could mean different things to different people. But I still think it might be valuable to dig into it more and see if there is a common expectation of "enough."

## Import Dataset

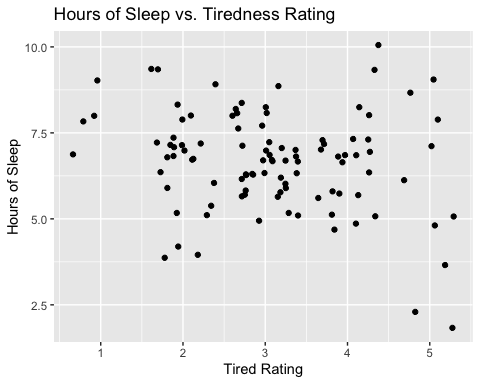
## Enough Hours PhoneReach PhoneTime Tired Breakfast  
## 1 No 2 Yes Yes 5 No  
## 2 No 2 No No 5 No  
## 3 No 4 No Yes 2 Yes  
## 4 No 4 Yes Yes 2 Yes  
## 5 No 4 No No 5 No  
## 6 No 4 Yes Yes 2 Yes

## Enough Hours PhoneReach PhoneTime   
## Length:102 Min. : 2.000 Length:102 Length:102   
## Class :character 1st Qu.: 6.000 Class :character Class :character   
## Mode :character Median : 7.000 Mode :character Mode :character   
## Mean : 6.657   
## 3rd Qu.: 7.000   
## Max. :10.000   
## Tired Breakfast   
## Min. :1.000 Length:102   
## 1st Qu.:2.000 Class :character   
## Median :3.000 Mode :character   
## Mean :3.088   
## 3rd Qu.:4.000   
## Max. :5.000

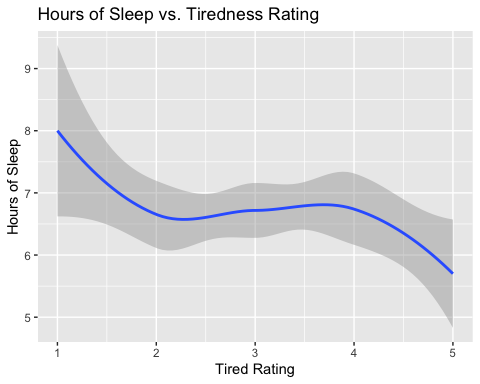
In order to work with the categorical variables, I converted yes/no responses to 0/1. I then created a linear regression model with tired rating as the dependent variable and hours as the predictor The p-value is .053 which is generally considered a significant predictor:

##   
## Call:  
## lm(formula = sleepStudy$Tired ~ sleepStudy$Hours, data = sleepStudy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.04105 -0.90355 -0.04105 0.82144 2.23395   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.00357 0.47849 8.367 3.66e-13 \*\*\*  
## sleepStudy$Hours -0.13750 0.07032 -1.955 0.0533 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.002 on 100 degrees of freedom  
## Multiple R-squared: 0.03683, Adjusted R-squared: 0.0272   
## F-statistic: 3.824 on 1 and 100 DF, p-value: 0.05332

For a visual of the relationship, I plotted hours against tiredness rating. Since the scatter plot was difficult to see any trends, I used the jitter function and also smooth to better show the linear relationship between the variables. This was not quite as linear as I expected.

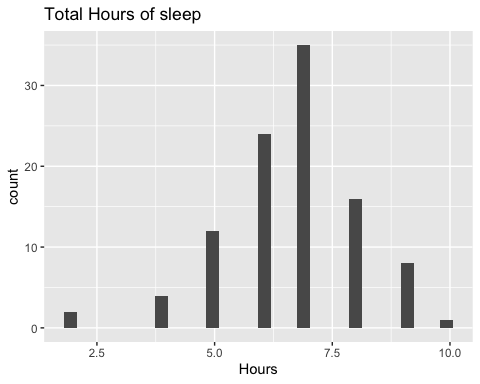


## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



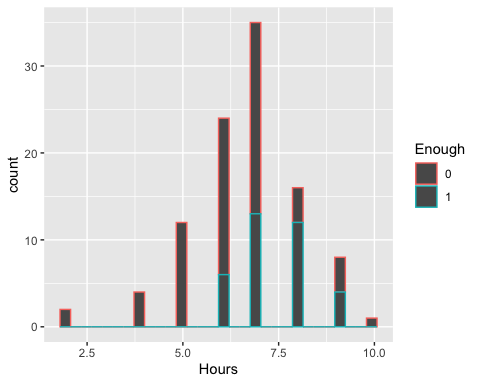
I then plotted a histogram of the total hours of sleep reported just to see what the distribution looked like. Overall the distribution looks normal but we can see a clear outlier on the lower end with less than 2.5 hours of sleep and an outlier at the other end with 10 hours, although this does not appear to be as far outside the distribution. The data appears to be slightly left skewed with more data at about 7 hours and less.

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



I also plotted a histogram of the responses for “enough” sleep, color coding the yes and no responses to see them comparitively.

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



Next I created a logistic regression model with tired as the dependent variable and hours, breakfast, phone in reach and phone time before bed as predictor variables. As expected, hours of sleep was a very strong predictor of enough sleep, as was tiredness rating, both having p-values near 0. The other three variables all had p-values of .6 or higher, indicating that while there might be some relationship, these are not significant predictors of someone reporting enough sleep.

##   
## Call:  
## glm(formula = as.factor(Enough) ~ Hours + Tired + Breakfast +   
## PhoneReach + PhoneTime, family = binomial(), data = sleepStudy)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5673 -0.7131 -0.4750 0.6938 2.4045   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -2.43410 1.87873 -1.296 0.195111   
## Hours 0.83938 0.23888 3.514 0.000442 \*\*\*  
## Tired -1.19710 0.33052 -3.622 0.000292 \*\*\*  
## Breakfast1 -0.25674 0.56049 -0.458 0.646905   
## PhoneReach1 -0.30345 0.60481 -0.502 0.615859   
## PhoneTime1 -0.08718 0.72809 -0.120 0.904686   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 131.191 on 101 degrees of freedom  
## Residual deviance: 95.588 on 96 degrees of freedom  
## AIC: 107.59  
##   
## Number of Fisher Scoring iterations: 5

I will now split the data to create a train model and run it on the test set to see how accurate it is. Create train and test data by splitting the datasest:

## [1] TRUE FALSE TRUE FALSE TRUE TRUE

## Enough Hours PhoneReach PhoneTime   
## Length:68 Min. : 2.000 Length:68 Length:68   
## Class :character 1st Qu.: 6.000 Class :character Class :character   
## Mode :character Median : 7.000 Mode :character Mode :character   
## Mean : 6.676   
## 3rd Qu.: 7.250   
## Max. :10.000   
## Tired Breakfast   
## Min. :1 Length:68   
## 1st Qu.:2 Class :character   
## Median :3 Mode :character   
## Mean :3   
## 3rd Qu.:4   
## Max. :5

## Enough Hours PhoneReach PhoneTime   
## Length:34 Min. :2.000 Length:34 Length:34   
## Class :character 1st Qu.:6.000 Class :character Class :character   
## Mode :character Median :7.000 Mode :character Mode :character   
## Mean :6.618   
## 3rd Qu.:7.000   
## Max. :9.000   
## Tired Breakfast   
## Min. :2.000 Length:34   
## 1st Qu.:2.250 Class :character   
## Median :3.000 Mode :character   
## Mean :3.265   
## 3rd Qu.:4.000   
## Max. :5.000

# Create the train model:

##   
## Call:  
## glm(formula = as.factor(Enough) ~ Hours + Tired + Breakfast +   
## PhoneReach + PhoneTime, family = binomial(), data = trainData)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.5136 -0.5381 -0.3486 0.3091 2.6550   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.5937 2.8188 -0.565 0.57183   
## Hours 0.9863 0.3394 2.906 0.00366 \*\*  
## Tired -1.6355 0.5362 -3.050 0.00229 \*\*  
## Breakfast1 -0.4194 0.8121 -0.516 0.60552   
## PhoneReach1 -0.6350 0.9495 -0.669 0.50362   
## PhoneTime1 -0.6417 0.9678 -0.663 0.50728   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 87.021 on 67 degrees of freedom  
## Residual deviance: 51.310 on 62 degrees of freedom  
## AIC: 63.31  
##   
## Number of Fisher Scoring iterations: 6

# Summary of test model and train model

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0004102 0.0582206 0.1781129 0.2644041 0.4076535 0.8347840

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.0001144 0.0752720 0.2154580 0.3382353 0.5849575 0.9811317

# Validate the Model - Confusion Matrix

## Predicted\_value  
## Actual\_Value FALSE TRUE  
## 0 40 5  
## 1 8 15

# Calculate Accuracy

## [1] 0.8088235

The test model is 81% accurate in predicting if a person reports feeling rested enough.

Since the dataset includes a response of rested enough and a tiredness level, which give us similar information, I wanted to look at both. I created a logistic regression model to predict tiredness level by using hours, breakfast, phone in reach, and phone time:

##   
## Call:  
## glm(formula = Tired ~ Hours + +Breakfast + PhoneReach + PhoneTime,   
## data = sleepStudy)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.20511 -0.74752 0.05144 0.73633 2.10923   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.30055 0.51317 8.380 4.15e-13 \*\*\*  
## Hours -0.10580 0.07156 -1.478 0.1425   
## Breakfast1 -0.51615 0.20921 -2.467 0.0154 \*   
## PhoneReach1 -0.35485 0.22147 -1.602 0.1124   
## PhoneTime1 0.05857 0.25716 0.228 0.8203   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.9631535)  
##   
## Null deviance: 104.206 on 101 degrees of freedom  
## Residual deviance: 93.426 on 97 degrees of freedom  
## AIC: 292.51  
##   
## Number of Fisher Scoring iterations: 2

In my initial linear model with hours and tiredness rating, the p-value for hours was .05, very significant. Now, when we have added other variables, the p-value for hours is only .143 and breakfast appears to have the most significant value with a p-value of .015. Since the phone variables are less significant, I removed those and ran the model again with only hours and breakfast.

##   
## Call:  
## lm(formula = Tired ~ Breakfast, data = sleepStudy)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.3902 -0.8852 0.1148 0.6098 2.1147   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.3902 0.1546 21.934 <2e-16 \*\*\*  
## Breakfast1 -0.5050 0.1999 -2.527 0.0131 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.9897 on 100 degrees of freedom  
## Multiple R-squared: 0.06001, Adjusted R-squared: 0.05061   
## F-statistic: 6.384 on 1 and 100 DF, p-value: 0.01309

The p-value for breakfast predicting tiredness level was .013, which would make eating breakfast a stronger predictor of tiredness leven than hours of sleep.

# Conclusions

Based on this analysis, it appears that both hours of sleep and eating breakfast have a strong correlation to tiredness level, which has a strong correlation to feeling of getting enough sleep. The variables regarding phone use did not appear to have a significant correlation.