Modeling Sleep Related Physiological Variables Using Linear Regression Techniques

Reference Paper and Data: *SaYoPillow: Blockchain Integrated Privacy Assured IoMT Framework for Stress Management Considering Sleeping Habits*, Laavanya Rachakonda et al (2020)

Abstract

SaYoPillow: Blockchain Integrated Privacy Assured IoMT Framework for Stress Management Considering Sleeping Habits presents a dataset that has nine variables. The reference paper seeks to generate sleep and stress-related machine learning models with the experimental data generated. In this paper, the sleep dataset will be explored. After exploring the properties of the dataset, a linear regression model will be generated to explore the relationship between physiological variables. Criterion-based methods such as AIC and BIC will be utilized to determine the model's accuracy.

STAT 4010: Statistical Methods and Applications 2

Author: Samuel Kwon

Department of Applied Mathematics
University of Colorado Boulder

04/29/22

Introduction

Sleep quality plays a role in mental and physical wellbeing. Important functions such as organ repair, cell communication, restoration of energy, and control of hormonal levels happen during sleep. Because of the significance of sleep, it is important to get an adequate amount and good quality of sleep. *SaYoPillow: Blockchain Integrated Privacy Assured IoMT Framework for Stress Management Considering Sleeping Habits* proposes a smart healthcare system that measures physiological changes during sleep. The authors propose an e-textile pillow that measures different sleep variables, which is then stored in a blockchain for consumers to view. The pillow measures variables such as hours of sleep, snoring range (dB), respiration rate (bpm), heart rate (bpm), blood oxygen levels, eye movement rate (REM), limb movement rate, body temperature, and stress state. An algorithm will consider the eight physiological variables to generate a stress prediction level. It is important to note that the stress prediction level in the dataset is a discrete variable. On the other hand, the eight physiological factors are continuous variables. The proposed dataset was collected through an experimental study by the authors, then utilized to generate a machine learning model for stress prediction. The dataset utilized for machine learning in the reference paper will be explored in this paper. Then, a linear regression model will be generated to predict and relate different physiological variables. This exploration process aims to determine how physiological variables impact or relate to other variables during sleep. Criterion based fitting techniques such as AIC and BIC will be utilized to determine a good fit for the model.

Statistical Methods

Linear Regression Assumptions

(1) Linearity

Linearity can be defined as $\mathbf{Y} = X\beta + \epsilon$ where the estimated $\mathbf{E}(\epsilon) = 0$ giving the equation $\mathbf{y_i} = \beta_0 + \beta_1 x_{i,1} + ... + \beta_p x_{i,p}$. There are three characteristics of a linear function. (1) The estimated $\mathbf{E}(\mathbf{y_i})$ is a linear function of $\mathbf{x_{i,j}}$. (2) The slope of the line that relates $\mathbf{E}(\mathbf{y_i})$ to $\mathbf{x_{i,j}}$ doesn't depend on any other values. (3) The $\mathbf{x_{i,j}}$ variable is additive to $\mathbf{E}(\mathbf{y_i})$.

(2) Independence

Independence is when the dataset variables are uncorrelated and the errors are independent. The independence assumption can be examined by plotting the residuals and time/index plot or the Durbin Watson test. The Durbin Watson method tests the null and alternative hypothesis. The null hypothesis is that there is no correlation among the residuals. The alternative hypothesis is that the residuals are correlated.

(3) Homoskedasticity (Constant Variance)

Constant variance can be examined by plotting the residuals and fitted plots. If the dataset does not follow this assumption, transformations or weights can be placed. Square root or log transformations may allow the set to meet the constant variance assumption.

(4) Normality

The normality assumption assumes that the dataset follows a normal distribution. Deviations from this assumption can be observed by plotting a QQ plot, residuals and fitted plot, or the Shapiro Wilk test. The Shapiro Wilk method tests a null and alternative hypothesis. The null hypothesis is that the residuals come from a normal distribution. The alternative hypothesis is that the residual does not come from a normal distribution. If the Shapiro Wilk test reports a p-value less than 0.05, the null hypothesis is rejected, indicating that the set does not come from a normal distribution.

Criterion Based Methods

(1) Akaike Information Criterion (AIC)

$$AIC = 2(p+1) + nlog(\frac{RSS}{n}) \tag{1}$$

Where **p** is the number of predictors and **n** is the number of units

(2) Bayes Information Criterion (BIC)

$$BIC = (p+1)log(n) - 2logL(\hat{\beta}) \tag{2}$$

Where **p** represents the number of predictors and **n** represents the number of units. The function $\mathbf{2logL}(\hat{\beta})$ measures how well the least squared estimate fits the set.

Multicollinearity
Multicollinearity occurs if one of the predictors is a linear combination of others. This will lead to decreased accuracy of the model.
(1) Variance Inflation Factors (VIF)

$$var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_{i,j} - \bar{x_j})^2} \times \frac{1}{1 - R_j^2}$$

Note: The multiple $\frac{\mathbf{1}}{\mathbf{1-R_j^2}}$ represents the inflation factor.

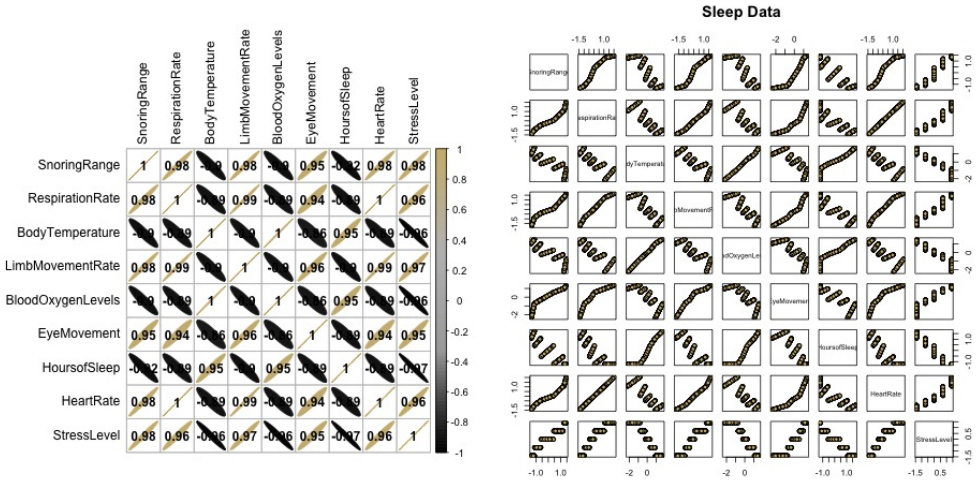
(3)

(2) Condition Number (κ)

$$\kappa = \sqrt{\frac{\lambda_{largest}}{\lambda_{smallest}}} \tag{4}$$

Results

The dataset contains nine variables. They are hours of sleep, snoring range (dB), respiration rate (bpm), heart rate (bpm), blood oxygen levels, eye movement rate (REM), limb movement rate, body temperature, and stress state. As mentioned earlier, it is important to note that the stress state is a discrete variable while the other variables are continuous. When initially exploring the dataset, an attempt was made to generate a stress model. Since the stress variable is discrete, an attempt was made to convert the values into a binomial representation. The dataset was transformed so that a stress value less than two corresponded to zero and stress value greater than 3 corresponded to one. When running the binomial regression, multiple issues were reported because of the noncontinuous properties of some variables. In addition to binomial regression, a Poisson regression was also generated; however, the same issues were present. As a result, physiological variables were explored instead of stress state. Before generating linear regression models for physiological variables,the dataset was normalized using the scale function in R. Correlation and pair plots were generated to make initial observations. In figure one, it should be noted that some of the variables have negative correlations. Furthermore, the pairs plot indicates that some variables have nonlinear relationships with other measured variables. For example, the plot of blood oxygen level and hours of sleep is non linear. Initially, there is a constant increase and then a linear increase. Since the stress level data is discrete, the plots of stress levels with other variables will not be continuous. It is comparable to a step function in that it only increases and stays constant at some value **x** and has a repeating box type graph.



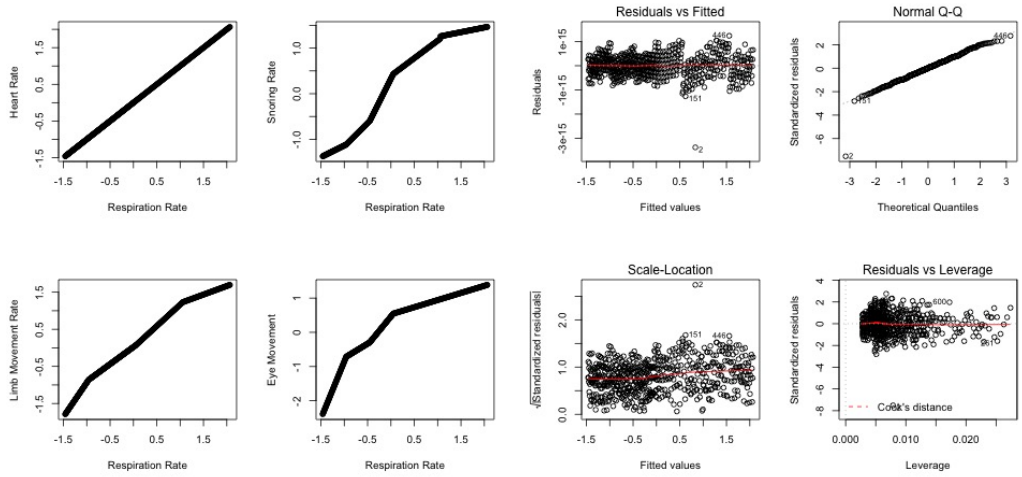
(a) Correlations (b) Normalized Pairs

Figure 1: Data Observations

The model explored will be a linear regression of respiration rate as a response. Prediction variables will be chosen based on the pairs plot and linearity of each variable.

Respiration Rate Model

Four predictors were utilized in the respiration rate model. Heart rate, snoring range, limb movement rate, and eye movement variables have some continuous relationships with respiration rate. An attempt was made to generate a full model with all the variables included. However, because of nonlinear and discrete properties in some variables, the model violated all of the linear regression assumptions. Figure two shows the predictor and response variables. Respiration rate and heart rate have a near-perfect linear relationship. The respiration rate and snoring rate have some type of sigmoidal activity; however, they were considered in the general model because limited variables were present. Respiration rate and limb movement rate have a linear relationship. Although it is not perfectly linear like respiration rate and heart rate, it does possess some linear properties. Finally, respiration rate and eye movement have some linear properties in the beginning and end sections of the plot.



(a) Predictors and Response Plots (b) Linear Regression Assumptions

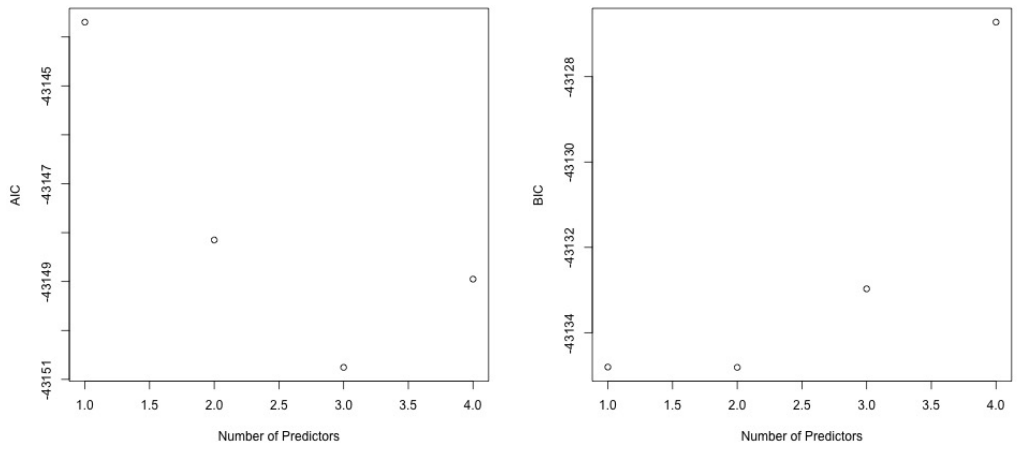
Figure 2: Data Observations

The linear regression model for respiration follows the linear regression assumptions. The residuals and fitted values plot should show a random scatter of data points at the zero horizontal axis. Figure two presents the linear regression plot where the assumptions are examined. The model's residuals and fitted values plot has scattered points at the zero horizontal axis. However, there is a sudden shift in data points when the fitted value is 0.5. This observation should be noted for future assessment. The model follows the linearity and constant variance assumption based on the fitted and residuals plot. In examining the normality assumption, the QQ plot should have a linear line with the data points roughly following a linear pattern. In this case, the model follows the normality assumption. For the independence assumption, a Durbin Watson test was performed. The p-value for this test was 0.234, which is greater than 0.05, indicating that the null hypothesis cannot be rejected. Based on this statistic, it can be assumed that the model does follow the independence assumption. In the section below, criterion based method will be performed to determine how many predictors can accurately model respiration rate.

Criterion Based Methods

The criterion-based method has two tests: AIC and BIC. AIC is mainly utilized if prediction is the goal of the model. BIC is utilized if explanation is the goal of the model. In this paper, both AIC and BIC will be performed and compared. The model has a total of four predictors. They are heart rate, limb movement rate, eye movement, and snoring range. The code for criterion based methods will generate a true and false table to indicate the optimized number and type of predictors. In figure three, the AIC value for three predictors is the lowest indicating that three predictors are suitable for the model. The corresponding true and false table indicates that heat rate, limb movement rate, and eye movement rate should be utilized as the predictors in the model.

The BIC method produces different results than the AIC method. Figure three shows the BIC value and number of predictors. A low BIC value is favorable. As a result, the BIC test shows that two predictors are favorable for explaining the model. The two predictors should be heart rate and limb movement rate. Since both the AIC and BIC methods produced different results, both optimizations will be taken into consideration when producing the final respiration model. The following section will explore the linear regression model for both AIC and BIC. Analysis of the model will be made to determine a good model for respiration rate.



(a) Predictors and Response Plots (b) Linear Regression Assumptions

Figure 3: Data Observations

AIC Adjusted Linear Regression Model

The AIC criteria indicate that three predictors should be used in the model. The three variables are heart rate, limb movement rate, and eye movement rate. The linear regression for these variables was generated. When plotting the adjusted linear regression model, the same trends were observed from the original model with four variables. The fitted and residuals graph has some type of sigmoidal curve. The p-value for the limb movement rate is greater than 0.05, indicating that it is not statistically significant. However, heart rate and eye movement have a p-value of less than 0.05, indicating some significance in the model.

BIC Adjusted Linear Regression Model

The BIC criteria indicate that two predictors should be used in the model. The two variables are heart rate and limb movement rate. The BIC adjusted model also has similar trends to the AIC adjusted and original model plots in that the fitted and residuals plot has a sigmoidal type curve. Similar to the AIC model, the heart rate p-value is less than 0.05, while the p-value for limb movement rate is greater than 0.05. In some ways, the results produced from the AIC and BIC model are similar. Both consider the limb movement rate which does not have significance in the model. Taking the results into consideration, the next section will explore multicollinearity and what variables to use to produce the best model.

Multicollinearity

Multicollinearity can be examined through the Variance Inflation Factor (VIF) and Condition Number (κ). The variance inflation factor was measured for the four variables utilized in the general model. The VIF for heart rate is 105.92, the snoring range is 28.44, the limb movement rate is 183.05, and the eye movement rate is 24.86. The general rule is that if the VIF value is greater than 5, it is evidence of some multicollinearity. If the VIF value is greater than 10, it is strong evidence of multicollinearity. Based on the VIF data obtained, there is evidence of multicollinearity. On the other hand, the constant kappa is 25.60. Generally, when the constant kappa is greater than 30, it is evidence of multicollinearity. The two metrics presented indicate some level of multicollinearity in the general respiration model. Since the VIF is greater than 10 for all variables, it will be hard to remove multicollinearity from the model. As a result, the model lacks some accuracy. The same trend of multicollinearity is present in the AIC adjusted and BIC adjusted model.

Final Model

Considering the AIC, BIC, and p-values, the best model seems to be with two predictors. The two predictors are heart rate and eye movement. Although BIC suggests that the two predictors should be heart rate and limb movement rate, the p-value for limb movement rate suggests it is not useful in the model. Since the p-value for eye movement is less than 0.05, it replaced the limb movement rate variable. The respiration model that utilizes heart rate and eye movement as the predictors also decreases the multicollinearity observed with the AIC and BIC suggested models. The VIF for heart rate is 8.01 and for eye movement is 8.01. The constant kappa is 5.52. Both VIF and constant kappa decreased significantly compared to the original, AIC, and BIC models. Although the multicollinearity values decreased significantly, the VIF values still indicate some multicollienarity. As a result, it can be concluded that the variables heart rate and eye movement have some relations.

Discussion and Conclusion

The final proposed model for respiration rate is to use heart rate and eye movement rate as predictors. This conclusion was made because of the criterion-based methods and p values associated with each model. However, it should be noted that the problem of multicollinearity is still present with the current model. In addition, when running the linear regression, a warning was issued. The warning states that the model is a perfect fit, and the summary may be unreliable. Based on this observation, the question of independence and the relationship between the variables is explored. The multicollinearity present in the current model indicates that the variables may be linear combinations of each other. Thus, the variables heart rate and eye movement are related. When comparing the p values of the four original variables, the snoring range and limb movement rate had p values greater than 0.05, indicating that the data values are nearly similar. It can be stated that all of the physiological variables utilized in this model are related. Because heart rate is a significant physiological variable that controls the body's response during sleep, it would be safe to say that the heart rate impacts eye movement, limb movement, and snoring range. Since one variable impacts the outcome of the other, this relationship would influence the regression model. This project emphasizes exploring relationships between predictor variables before generating a linear regression model. It is important to consider how the variables may influence each other before exploring the dataset.

For future applications, a regression model of stress can be explored. It would be interesting to model the dynamics of stress and how that can impact sleep quality. In this dataset, stress is measured as a discrete variable. It would be interesting to generate continuous stress data and explore those values for future applications. Another application would be to explore the physiological models through non-parametric modeling. Although non-parametric modeling was introduced near the end of the course, the trends observed in the dataset would be modeled better through a nonlinear additive regression technique.

References

(1) L. Rachakonda, A. K. Bapatla, S. P. Mohanty, and E. Kougianos, "SaYoPillow: Blockchain-Integrated Privacy-Assured IoMT Framework for Stress Management Considering Sleeping Habits", IEEE Transactions on Consumer Electronics (TCE), Vol. 67, No. 1, Feb 2021, pp. 20-29.

(2) L. Rachakonda, S. P. Mohanty, E. Kougianos, K. Karunakaran, and M. Ganapathiraju, "Smart-Pillow: An IoT based Device for Stress Detection Considering Sleeping Habits", in Proceedings of the 4th IEEE International Symposium on Smart Electronic Systems (iSES), 2018, pp. 161–166.

Appendix

R Code

```
library(ggplot2)
Data=read.table('SaYoPillow.csv',header=TRUE, sep=",")
names(Data)
Sleep=data.frame(SnoringRange=Data$sr, RespirationRate=Data$rr, BodyTemperature=Data$t,
                 LimbMovementRate=Data$lm, BloodOxygenLevels=Data$bo, EyeMovement=Data$rem,
                 HoursofSleep=Data$m, HeartRate=Data$hr, StressLevel=Data$sl)
summary(Sleep)

par(mfrow = c(1,3))
boxplot(Sleep$SnoringRange, main='Snoring Range',col="#CFB87C")
boxplot(Sleep$RespirationRate, main='Respiration Rate',col="#CFB87C")
boxplot(Sleep$BodyTemperature, main='Body Temperature',col="#CFB87C")

par(mfrow = c(1,3))
boxplot(Sleep$LimbMovementRate, main='Limb Movement Rate',col="#CFB87C")
boxplot(Sleep$BloodOxygenLevels, main='Blood Oxygen Levels',col="#CFB87C")
boxplot(Sleep$EyeMovement, main='Eye Movement',col="#CFB87C")

par(mfrow = c(1,3))
boxplot(Sleep$HoursofSleep, main='Hours of Sleep',col="#CFB87C")
boxplot(Sleep$HeartRate, main='Heart Rate',col="#CFB87C")
boxplot(Sleep$StressLevel, main='Stress Level',col="#CFB87C")

# Correlation and Pairs Plot
library(corrplot)
col4 = colorRampPalette(c("black", "darkgrey", "grey","#CFB87C"))
corrplot(cor(Sleep), method = "ellipse", col = col4(100),
         addCoef.col = "black", tl.col = "black")
pairs(Sleep, main = "Sleep Data",  pch = 21,bg = c("#CFB87C"))

#GLM Using Poisson
glmod2_stress=glm(StressLevel~SnoringRange+RespirationRate+BodyTemperature+
                  LimbMovementRate+BloodOxygenLevels+EyeMovement+HoursofSleep,
                  data=Sleep,family=poisson)
summary(glmod2_stress)
par(mfrow=c(2,2))
plot(glmod2_stress)

# Code to Modify Stress Into Binomial
library(readr)
library(dplyr)
library(tidyr)
SleepB=mutate(Sleep, StressLevel=ifelse(StressLevel>=2,1,0))
summary(SleepB)

# Binomial Regression Attempt
glmodB_stress=glm(StressLevel~SnoringRange+RespirationRate+BodyTemperature+
                  LimbMovementRate+BloodOxygenLevels+EyeMovement+HoursofSleep,
                  data=SleepB,family=binomial)
summary(glmodB_stress)

```

```
# Full Linear Regression Without Any Data Transformation
lmod_REM2=lm(EyeMovement~SnoringRange+RespirationRate+BodyTemperature+LimbMovementRate+
               BloodOxygenLevels+HoursofSleep,data=Sleep)
summary(lmod_REM2)
par(mfrow=c(2,2))
plot(lmod_REM2)

#Modified Sqrt Transformation of Original Data
Sleep$EyeMovementTransform=sqrt(Sleep$EyeMovement)
Sleep$SnoringRangeTransform=sqrt(Sleep$SnoringRange)
Sleep$RespirationRateTransform=sqrt(Sleep$RespirationRate)
Sleep$BodyTemperatureTransform=sqrt(Sleep$BodyTemperature)
Sleep$LimbMovementRateTransform=sqrt(Sleep$LimbMovementRate)
Sleep$BloodOxygenLevelsTransform=sqrt(Sleep$BloodOxygenLevels)
Sleep$HoursofSleepTransform=sqrt(Sleep$HoursofSleep)

lmod_REM3=lm(EyeMovementTransform~SnoringRangeTransform+RespirationRateTransform+
               BodyTemperatureTransform+LimbMovementRateTransform+BloodOxygenLevelsTransform+
               HoursofSleepTransform,data=Sleep)
summary(lmod_REM3)

par(mfrow=c(2,2))
plot(lmod_REM2)

#Modified Scale Transformation (Normalization )
Sleep3=as.data.frame(scale(Sleep[1:9]))

#Only Linear Section RespirationRate
library(car)
lmod_Resp=lm(RespirationRate~HeartRate+SnoringRange+LimbMovementRate+EyeMovement,data=Sleep3)
summary(lmod_Resp)
par(mfrow=c(2,2))
plot(lmod_Resp)
durbinWatsonTest(lmod_Resp)

# Adjusted Model AIC
lmod_RespAdj=lm(RespirationRate~HeartRate+EyeMovement,data=Sleep3)
summary(lmod_RespAdj)
par(mfrow=c(2,2))
plot(lmod_RespAdj)

# Adjusted Model BIC
lmod_RespAdj2=lm(RespirationRate~HeartRate+LimbMovementRate,data=Sleep3)
summary(lmod_RespAdj2)
par(mfrow=c(2,2))
plot(lmod_RespAdj2)

# AIC
library(leaps)
library(MASS)
n = dim(Sleep3)[1];
reg1 = regsubsets(RespirationRate~HeartRate+SnoringRange+LimbMovementRate+EyeMovement,
  data = Sleep3)
rs = summary(reg1)
rs$which
AIC=2*(2:5)+n*log(rs$rss/n)
plot(AIC~I(1:4),xlab="Number of Predictors", ylab="AIC")

#BIC
BIC = log(n)*(2:5) + n*log(rs$rss/n)
#BIC
plot(BIC ~ I(1:4), xlab = "Number of Predictors", ylab = "BIC")
vif(lmod_RespAdj)
kappa(lmod_RespAdj)
```