Untitled

2022-06-16

## Loading libraries required

library(data.table)  
library(tidyverse)  
library(ggplot2)  
library(dplyr)  
library(RColorBrewer)  
library(readxl)

# loading dataset from excel

SAGE <- read\_excel("SAGE.xlsx")

# Getting overview of the data

format(glimpse(SAGE))

## Rows: 2,356  
## Columns: 10  
## $ ID <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17,…  
## $ centre <chr> "China", "China", "China", "China", "China", "China", "Chi…  
## $ gender <dbl> 1, 2, 1, 2, 1, 2, 2, 1, 2, 2, 1, 1, 1, 2, 1, 2, 1, 2, 1, 1…  
## $ age <dbl> 67, 53, 60, 50, 57, 55, 65, 73, 61, 77, 57, 77, 57, 58, 60…  
## $ ideation <dbl> 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0…  
## $ attempt <dbl> 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0…  
## $ health <dbl> 1, 2, 2, 1, 3, 2, 3, 1, 1, 1, 1, 1, 3, 2, 2, 1, 3, 3, 2, 1…  
## $ sleeplength <dbl> 22.0, 8.0, 7.0, 9.0, 10.0, 5.0, 8.0, 7.0, 7.5, 7.0, 8.0, 9…  
## $ sleepqual <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0…  
## $ married <dbl> 2, 2, 1, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 1, 1, 1, 2, 2…

## [1] "# A tibble: 2,356 × 10"   
## [2] " ID centre gender age ideation attempt health sleeplength sleepqual"  
## [3] " <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>"  
## [4] " 1 1 China 1 67 0 0 1 22 0"  
## [5] " 2 2 China 2 53 0 0 2 8 0"  
## [6] " 3 3 China 1 60 1 1 2 7 0"  
## [7] " 4 4 China 2 50 0 0 1 9 0"  
## [8] " 5 5 China 1 57 0 0 3 10 0"  
## [9] " 6 6 China 2 55 1 0 2 5 1"  
## [10] " 7 7 China 2 65 0 0 3 8 0"  
## [11] " 8 8 China 1 73 0 0 1 7 0"  
## [12] " 9 9 China 2 61 1 0 1 7.5 1"  
## [13] "10 10 China 2 77 0 0 1 7 0"  
## [14] "# … with 2,346 more rows, and 1 more variable: married <dbl>"

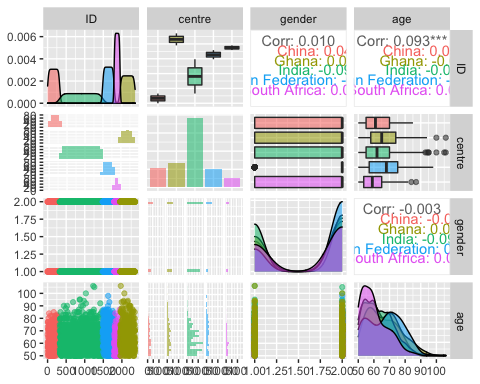
format(head(SAGE))

## [1] "# A tibble: 6 × 10"   
## [2] " ID centre gender age ideation attempt health sleeplength sleepqual"  
## [3] " <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>"  
## [4] "1 1 China 1 67 0 0 1 22 0"  
## [5] "2 2 China 2 53 0 0 2 8 0"  
## [6] "3 3 China 1 60 1 1 2 7 0"  
## [7] "4 4 China 2 50 0 0 1 9 0"  
## [8] "5 5 China 1 57 0 0 3 10 0"  
## [9] "6 6 China 2 55 1 0 2 5 1"  
## [10] "# … with 1 more variable: married <dbl>"

There is no missing data in the dataset. The data is clean and ready to be used. The data is also in a tidy format. # Corelation Analysis

Now We can start with the some basic analysis. First of all we will make pairplot to analyse the corelations between the variables.

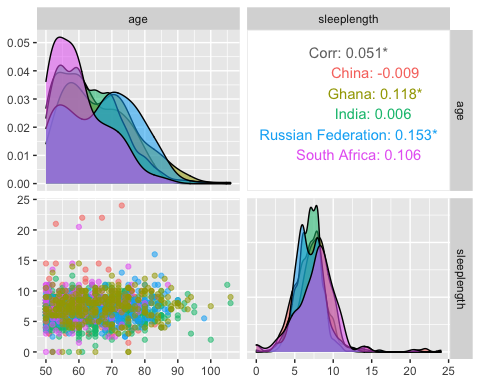
# install.packages("GGally")  
library(GGally)  
  
ggpairs(SAGE, #   
 columns = 1:4,   
 aes(color = centre,   
 alpha = 0.5))



The pairplot is a great way of analysing the datasets and finding the corelations between the variables. In the above pairplot we can observe that for 5 different countries the corelation coefficent values are vey low or negative between numeric variables. The boxplot between age and centre shows that age is average age is highest for South Africa and lowest for Russian Federation. There are some outliers in the age variable as shown by boxplot. The density plots show that the age is left skewed. Simialrly the corelation coefficent values with aesterics also show that signficant corelation between the variables for different centres. Data from Indian and Russian Federation are highliy signficant from each other. It can also be read by barplot with high values for Russian centr as compared to Indian centre.

According to the requirements of the question we can choose age and sleep length as variables to analyse the corelation between them.

ggpairs(SAGE, #   
 columns = c("age","sleeplength"),   
 aes(color = centre,   
 alpha = 0.5))



The above graph shows that age and sleep length are not highly corelated with each other for different centres. The scatter plots also shows a flat trend which doesn’t show any positive corelation. Moreover as shown above the age is right skewed. Density plot for sleep length shows that it is almost normally distrbuted for 4 countries exceot for Russian Federation. Moreover the overall corelation coefficent is 0.051 which is very small and it indicates that that sleep length shorttage cannot be related to age of the persons in 5 centres. ’Normally distrubution can be check by following test.

Before conducting the regression analysis we can check the normality of the our columns by Shapiro Wilk test

shapiro.test(SAGE$age)

##   
## Shapiro-Wilk normality test  
##   
## data: SAGE$age  
## W = 0.95448, p-value < 2.2e-16

For the age column the p-value is less than 0.05 which is chosen signficance level. Since for the normality test we should have value >0.05 to consider the data to be normal so can conclude that for age data is is not normal.Now we can check for sleep length variable

shapiro.test(SAGE$sleeplength)

##   
## Shapiro-Wilk normality test  
##   
## data: SAGE$sleeplength  
## W = 0.93587, p-value < 2.2e-16

At 5% significance level and 95% confidence interval the result of shapiro wilk normality test shows that data for sleep length is normally distrbuted since p < .

We can further analyse our two variables by a Linear regression model.

model <- lm(SAGE$age~SAGE$sleeplength)  
  
summary(model)

##   
## Call:  
## lm(formula = SAGE$age ~ SAGE$sleeplength)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.711 -8.332 -1.580 6.546 41.919   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 62.0756 0.7538 82.350 <2e-16 \*\*\*  
## SAGE$sleeplength 0.2507 0.1022 2.453 0.0142 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 9.793 on 2271 degrees of freedom  
## (83 observations deleted due to missingness)  
## Multiple R-squared: 0.002643, Adjusted R-squared: 0.002204   
## F-statistic: 6.018 on 1 and 2271 DF, p-value: 0.01424

Result of the linear regression model indicates many important features which are described as

* p-value: at 5% significance level and 95% confidence interval our p-value is <0.05 which indicates that our model is statistically significant.
* F-statistics: Higher the value of F-stats better it is. our F-stats value is 6.018 with degree of freedom = 2271. THis value indicates if one of variable has significant effect.
* Multiple R-squared = .002. This tells us that only 0.02% of time our independent variable is able to predict the dependent variables. This value is low and it also indicates why our corelation coefficent values are very low in the pair plots above.
* Intercept: It is the intercept in the linear regression model of y=mx+c where m is slope of linear regression line. The 3 aesterics next to sleep length indicates that our p-value is in range 0.01 and 0,05. Higher the number of aesteric higher the significant the value is.
* Linear regression equation for above model results can be written as

where 0.2507 is the coefficient estimate and 62.07 is the intercept. It shows that for very 1 year increase in age we should expect an increase of 62 in sleep length.

We can extract this equation by formulas as well

library(equatiomatic)  
extract\_eq(model,use\_coefs=T)

$$
\operatorname{\widehat{SAGE\$age}} = 62.08 + 0.25(\operatorname{SAGE\$sleeplength})
$$

* Residual Error: Smaller residual values are considere better.
* Coefficent of determination R2: 0.002 which is very low hence our model is not good to predict the sleep length based on age value.
* Std Error: It is the measure of uncertainty in coefficents.
* Residual Standard error: Value of 9.793 indicate that actual values are this much far away from regression line.
* P|t| value: It is actually the t-value calculated by dividing coefficient by standard deviation. We can further analyse results by

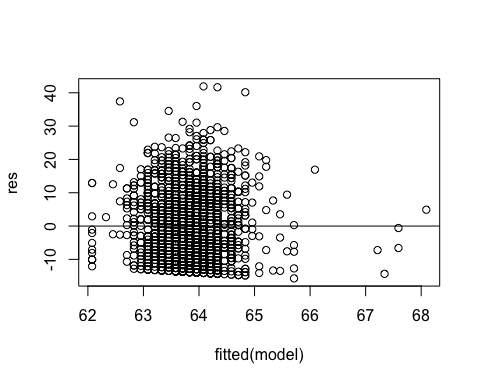
confint(model)

## 2.5 % 97.5 %  
## (Intercept) 60.59740503 63.553828  
## SAGE$sleeplength 0.05029108 0.451088

It varifies that our model results are at 95% confidence interval.

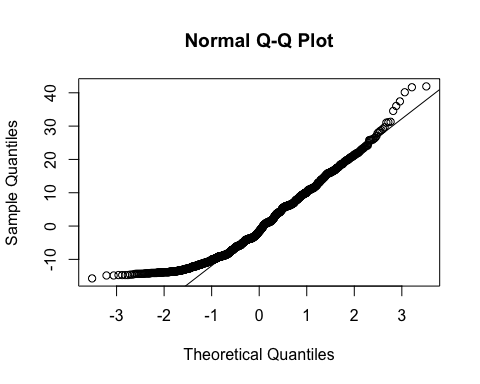
Now we will plot the residual and other plots for our model.

#define residuals  
res <- resid(model)  
  
#produce residual vs. fitted plot  
plot(fitted(model), res)  
  
#add a horizontal line at 0   
abline(0,0)



It shows that residuals are not normally distributed around 0 hence homoscedasticity is not violated.

#create Q-Q plot for residuals  
qqnorm(res)  
  
#add a straight diagonal line to the plot  
qqline(res)

 The qqplot shows that data is not normally distributed. since at the edges the scatter plot doesn’t follow the straight line.