Assignment5

Mahaldar Arti

10/11/2021

## Load the `data/student-survey.csv` to  
studentsurvey <- read.csv('C:/Users/Sandeep Raina/Documents/dsc520/data/student-survey.csv')  
str(studentsurvey)

## 'data.frame': 11 obs. of 4 variables:  
## $ TimeReading: int 1 2 2 2 3 4 4 5 5 6 ...  
## $ TimeTV : int 90 95 85 80 75 70 75 60 65 50 ...  
## $ Happiness : num 86.2 88.7 70.2 61.3 89.5 ...  
## $ Gender : int 1 0 0 1 1 1 0 1 0 0 ...

#2. Examine the Survey data variables. What measurement is being used for the variables? Explain what effect changing the measurement being used for the variables would have on the covariance calculation. Would this be a problem? Explain and provide a better alternative if needed.  
#install.packages("psych")  
#library(psych)   
  
  
str(studentsurvey)

## 'data.frame': 11 obs. of 4 variables:  
## $ TimeReading: int 1 2 2 2 3 4 4 5 5 6 ...  
## $ TimeTV : int 90 95 85 80 75 70 75 60 65 50 ...  
## $ Happiness : num 86.2 88.7 70.2 61.3 89.5 ...  
## $ Gender : int 1 0 0 1 1 1 0 1 0 0 ...

summary(studentsurvey)

## TimeReading TimeTV Happiness Gender   
## Min. :1.000 Min. :50.00 Min. :45.67 Min. :0.0000   
## 1st Qu.:2.000 1st Qu.:67.50 1st Qu.:65.34 1st Qu.:0.0000   
## Median :4.000 Median :75.00 Median :75.92 Median :1.0000   
## Mean :3.636 Mean :74.09 Mean :73.31 Mean :0.5455   
## 3rd Qu.:5.000 3rd Qu.:82.50 3rd Qu.:83.83 3rd Qu.:1.0000   
## Max. :6.000 Max. :95.00 Max. :89.52 Max. :1.0000

# In order to use describe on package  
  
#describe(studentsurvey)  
  
#correlation  
cor(studentsurvey)

## TimeReading TimeTV Happiness Gender  
## TimeReading 1.00000000 -0.883067681 -0.4348663 -0.089642146  
## TimeTV -0.88306768 1.000000000 0.6365560 0.006596673  
## Happiness -0.43486633 0.636555986 1.0000000 0.157011838  
## Gender -0.08964215 0.006596673 0.1570118 1.000000000

class(studentsurvey$TimeReading)

## [1] "integer"

class(studentsurvey$TimeTV)

## [1] "integer"

sapply(studentsurvey,class)

## TimeReading TimeTV Happiness Gender   
## "integer" "integer" "numeric" "integer"

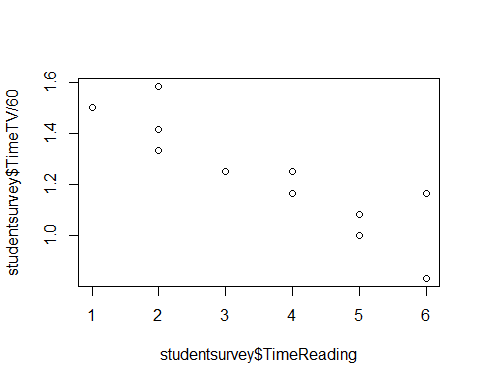
#Converting the class   
  
studentsurvey$TimeReading.f = factor(studentsurvey$TimeReading/60,ordered = TRUE)  
#3. Choose the type of correlation test to perform, explain why you chose this test, and make a prediction if the test yields a positive or negative correlation?  
  
#Correlation method can be pearson, spearman or kendall  
  
#The pearson correlation coefficient measure the linear dependence between two variables.  
  
#If method is âkendallâ or âspearmanâ, Kendallâs tau or Spearmanâs rho statistic is used to estimate a rank-based measure of association. These are more robust and have been recommended if the data do not come from a bivariate normal distribution.  
  
pearson\_test<-cor.test(studentsurvey$TimeReading, studentsurvey$TimeTV/60, method="pearson")  
pearson\_test

##   
## Pearson's product-moment correlation  
##   
## data: studentsurvey$TimeReading and studentsurvey$TimeTV/60  
## t = -5.6457, df = 9, p-value = 0.0003153  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.9694145 -0.6021920  
## sample estimates:  
## cor   
## -0.8830677

spearman\_test<-cor.test(studentsurvey$TimeReading, studentsurvey$TimeTV/60, method="spearman",exact=FALSE)  
  
spearman\_test

##   
## Spearman's rank correlation rho  
##   
## data: studentsurvey$TimeReading and studentsurvey$TimeTV/60  
## S = 419.6, p-value = 0.0001152  
## alternative hypothesis: true rho is not equal to 0  
## sample estimates:  
## rho   
## -0.9072536

plot(studentsurvey$TimeReading, studentsurvey$TimeTV/60)



#Looking at the plot Pearson\_test method sounds the best for this  
  
#4. Perform a correlation analysis of:  
  
#All Variables  
  
studentsurvey <- read.csv('C:/Users/Sandeep Raina/Documents/dsc520/data/student-survey.csv')  
cor(studentsurvey)

## TimeReading TimeTV Happiness Gender  
## TimeReading 1.00000000 -0.883067681 -0.4348663 -0.089642146  
## TimeTV -0.88306768 1.000000000 0.6365560 0.006596673  
## Happiness -0.43486633 0.636555986 1.0000000 0.157011838  
## Gender -0.08964215 0.006596673 0.1570118 1.000000000

## TimeReading vs. TimeTV  
cor(studentsurvey$TimeReading, studentsurvey$TimeTV/60, method = c("pearson", "kendall", "spearman"))

## [1] -0.8830677

### TimeReading vs. Happiness  
cor(studentsurvey$TimeReading, studentsurvey$Happiness, method = c("pearson", "kendall", "spearman"))

## [1] -0.4348663

### TimeReading vs. Gender  
cor(studentsurvey$TimeReading, studentsurvey$Gender, method = c("pearson", "kendall", "spearman"))

## [1] -0.08964215

#Repeat your correlation test in step 2 but set the confidence interval at 99%  
timereadingvstimetv<-cor.test(studentsurvey$TimeReading, studentsurvey$TimeTV/60,method = c("pearson", "kendall", "spearman"),exact = NULL, conf.level = 0.95, continuity = FALSE)  
  
#Describe what the calculations in the correlation matrix suggest about the relationship between the variables. Be specific with your explanation.  
names(timereadingvstimetv)

## [1] "statistic" "parameter" "p.value" "estimate" "null.value"   
## [6] "alternative" "method" "data.name" "conf.int"

timereadingvstimetv$conf.int

## [1] -0.9694145 -0.6021920  
## attr(,"conf.level")  
## [1] 0.95

timereadingvstimetv$p.value

## [1] 0.0003153378

timereadingvstimetv$method

## [1] "Pearson's product-moment correlation"

timereadingvstimetv$null.value

## correlation   
## 0

timereadingvstimetv$statistic

## t   
## -5.645664

#Most variables show very little relationship and negative correlationship  
  
#Calculate the correlation coefficient and the coefficient of determination, describe what you conclude about the results.  
cofofdet <- lm( studentsurvey$Gender ~ studentsurvey$TimeReading + TimeTV + Happiness, data=studentsurvey)  
summary(cofofdet)

##   
## Call:  
## lm(formula = studentsurvey$Gender ~ studentsurvey$TimeReading +   
## TimeTV + Happiness, data = studentsurvey)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.5728 -0.4535 0.1275 0.3960 0.6930   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 2.38109 3.04494 0.782 0.460  
## studentsurvey$TimeReading -0.18042 0.24260 -0.744 0.481  
## TimeTV -0.03057 0.03752 -0.815 0.442  
## Happiness 0.01481 0.01894 0.782 0.460  
##   
## Residual standard error: 0.589 on 7 degrees of freedom  
## Multiple R-squared: 0.1097, Adjusted R-squared: -0.2719   
## F-statistic: 0.2874 on 3 and 7 DF, p-value: 0.8333

#Based on your analysis can you say that watching more TV caused students to read less? Explain.  
  
#Yes, That what it looks like  
  
#Pick three variables and perform a partial correlation, documenting which variable you are controlling. Explain how this changes your interpretation and explanation of the results.  
#install.packages('ppcor')  
##install.packages('MASS')  
library(ppcor)

## Loading required package: MASS

pcor(studentsurvey)

## $estimate  
## TimeReading TimeTV Happiness Gender  
## TimeReading 1.0000000 -0.8827973 0.4013124 -0.2706036  
## TimeTV -0.8827973 1.0000000 0.6311611 -0.2943135  
## Happiness 0.4013124 0.6311611 1.0000000 0.2833152  
## Gender -0.2706036 -0.2943135 0.2833152 1.0000000  
##   
## $p.value  
## TimeReading TimeTV Happiness Gender  
## TimeReading 0.000000000 0.001615344 0.28437887 0.4812716  
## TimeTV 0.001615344 0.000000000 0.06832112 0.4420392  
## Happiness 0.284378868 0.068321119 0.00000000 0.4600603  
## Gender 0.481271572 0.442039185 0.46006033 0.0000000  
##   
## $statistic  
## TimeReading TimeTV Happiness Gender  
## TimeReading 0.0000000 -4.9720962 1.1592148 -0.7436966  
## TimeTV -4.9720962 0.0000000 2.1528933 -0.8147673  
## Happiness 1.1592148 2.1528933 0.0000000 0.7816064  
## Gender -0.7436966 -0.8147673 0.7816064 0.0000000  
##   
## $n  
## [1] 11  
##   
## $gp  
## [1] 2  
##   
## $method  
## [1] "pearson"

Note that the echo = FALSE parameter was added to the code chunk to prevent printing of the R code that generated the plot.