assignment_07_MukherjeeChitramoy-01

Mukherjee Chitramoy

`r Sys.Date()`

{r setup, include=FALSE} knitr::opts_chunk$set(echo = TRUE) ## Install and Load required packages : "'{r, echo=FALSE} # Package names # Package names # packages <- c("ggplot2","dplyr","tidyr","magrittr","tidyvers
packages <- c("ggplot2","dplyr","magrittr","tidyverse","purrr","GGally","scales","reshape","moments","ggpubr")

Install packages not yet installed

installed_packages <- packages %in% rownames(installed.packages()) if (any(installed_packages == FALSE)) { install.packages(packages[!installed_packages]) }

Packages loading

invisible(lapply(packages, library, character.only = TRUE))

```
#### Set the working directory to the root of your DSC 520 directory
setwd("C:/Users/chitro/Desktop/dsc520-fork-chitro")

#### Load the `data/student-survey.csv` to
ssurvey_df <- read.csv("data/student-survey.csv")


## Using `cor()` compute correlation coefficients for
```

```{r , echo=TRUE}

setwd("C:/Users/chitro/Desktop/dsc520-fork-chitro")
ssurvey_df <- read.csv("data/student-survey.csv")

ssurvey_df
ssurvey_df[,c(2,2:4)]

#--------------------------------------#
#       **** Assignment-I *****        #
#--------------------------------------#
# Assignment-I : Use R to calculate the covariance of the Survey variables
#                and provide an explanation of why you would use this calculation and what the results i

#-- Explanation :
```

```
#   Cor/Cov/Var function will compute variance of x or covariance or corelation
#   of x and y. Applying cor() function on survey variables, will produce
#   corelations matrix values betweem 1 amd -1, higeher positive number means
#   closer relationshi between the variables, and negative number mean inverse.
#   Give out for survey results indicate +ve corelation between TimeTV vs
#   Happiness (0.63) and -ve corelation between TimeTV and TimeReading (-0.88).
#   Results can be visualize using GGally::ggpairs.

library(GGally)
cor(ssurvey_df)
cov(ssurvey_df)
var(ssurvey_df)
cor(ssurvey_df, method = c("pearson", "kendall", "spearman"))
GGally::ggpairs(ssurvey_df)
#help -- ?cor()

#---------------------------------------#
#       **** Assignment-II *****       #
#---------------------------------------#
# 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.


#-- Explanation :

cor(ssurvey_df)
cov(ssurvey_df)

#cov(ssurvey_df)
#             TimeReading       TimeTV  Happiness      Gender
#TimeReading   3.05454545 -20.36363636 -10.350091 -0.08181818
#TimeTV      -20.36363636 174.09090909 114.377273  0.04545455
#Happiness   -10.35009091 114.37727273 185.451422  1.11663636
#Gender       -0.08181818   0.04545455   1.116636  0.27272727

# The diagonal elements 3,174,185 and 0.2 indicate the variance in data sets
#(lowest variance: 0.27 and Highest variance:185.451422), variance positive 174
#co-variance between TimeTV and Happiness indicates, happiness increases and
#TVtime goes up, however negative -20 variance indicates oppsite with TimeTv
#and TimeReading variance. positive 0.04 variance has minimal impact with Gender
#and TimeTV. Changing measures in Covariance unit will change the result/outcome.

# Problem is covariance -
# The main problem with covariance interpretation is that the wide range of
# results,it hard to interpret sometime. ( 0.2 to 185 in survey data frame.)

# Alternative : Correlation Coefficient method do have several advantages over
#               covariance for determining    strengths of relationships:
# Covariance can take on practically any number while a correlation is limited: -1 to +1.
# Because of it's numerical limitations, correlation is more useful for - determining how strong the rel
# Correlation does not have units. Covariance always has units
# Correlation isn't affected by changes in the center (i.e. mean) or scale of the variables
```

```
#-- References :
#-- https://www.cuemath.com/algebra/covariance-matrix/
#-- https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/covariance/
#-- https://www.mygreatlearning.com/blog/covariance-vs-correlation/#variance

# Variance - Variance is the expectation of the squared deviation of a random
# variable from its mean
# Standard Deviation
# Standard deviation is a measure of the amount of variation or dispersion of a
# set of values. A low standard deviation indicates that the values tend to be
# close to the mean of the set, while a high standard deviation indicates that the
# values are spread out over a wider range. It essentially measures the absolute
# variability of   a random variable.
# Covariance and correlation are related to each other, in the sense that
# covariance determines the type of interaction between two variables, while
# correlation determines the direction as well as the strength of the
# relationship between two variables.

# To find coorelation,columns/df variables needs to be integer.
# str(ssurvey_df)
# summary(ssurvey_df)
# cor(ssurvey_df, use = "complete.obs")
# cov(ssurvey_df)
# cov(ssurvey_df$TimeTV,ssurvey_df$TimeReading)
# cov(ssurvey_df$TimeTV,ssurvey_df$Happiness)
# cov(ssurvey_df$TimeReading,ssurvey_df$Happiness)

#---------------------------------------#
#       **** Assignment-III *****       #
#---------------------------------------#
# 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?

#-- Explanation :

#  Considering student survey dataset with no missing and NULL values and
#  skewness ratio, and positive and negative relationship between with variables
#  TimeTV/Happiness and TimeTV/TimeReading, I prefer to go with "Pearson" method.

#install.packages("moments")
library(moments)
skewness(ssurvey_df)

cor(ssurvey_df, use = "complete.obs", method = c("pearson"))
# cor(ssurvey_df, use = "complete.obs", method = c("pearson", "kendall", "spearman"))
#cor(heights_df$ed,heights_df$earn, method = 'kendall')
#cor(heights_df$ed,heights_df$earn, method = 'pearson')

#-  Visual inspection of the data normality using Q-Q plots (quantile-quantile
#   plots). Q-Q plot draws the correlation between a given sample and the normal
#   distribution.
#   http://www.sthda.com/english/wiki/correlation-test-between-two-variables-in-r
```

```
#installed_packages("ggpubr")
library("ggpubr")
# Happiness
ggqqplot(ssurvey_df$Happiness, ylab = "Happiness")
# TimeTV
ggqqplot(ssurvey_df$TimeTV, ylab = "TimeTV")


# References -
#   https://ademos.people.uic.edu/Chapter22.html

# The Pearson product-moment correlation is one of the most commonly used
# correlations in statistics. It's a measure of the strength and the direction
# of a linear relationship between two variables.
  # Your data is interval or ratio
  # Pearson only works with linear data. That means that your two correlated
    # factors have to approximate a line,and     not a curved or parabolic shape
  # Outliers in your data can really throw off a Pearson correlation

# Skewness interpretation :
#  As a general rule of thumb: If skewness is less than -1 or greater than 1,
# the distribution is highly skewed.If skewness is between -1 and -0.5 or
# between 0.5 and 1, the distribution is moderately skewed. If skewness is
# between -0.5 and 0.5, the distribution is approximately symmetric.
# he data you are analyzing needs to be normally distributed. This can be done
# in a couple of ways (Skewness, Kurtosis) but it can also be done in a
# quick and dirty manner through histograms

#---------------------------------------#
#       **** Assignment-IV *****        #
#---------------------------------------#
# Assignment-IV : Perform a correlation analysis of:
  # - All variables
  # - A single correlation between two a pair of the variables
  # - Repeat your correlation test in step 2 but set the confidence interval at 99%
  # - Describe what the calculations in the correlation matrix suggest about
  #   the relationship between the variables. Be specific with your explanation.

#-- Explanation :

# Cor() function define corelation between all the variables with values between -1 to 1.
cor(ssurvey_df)
# - Corelation between thevariables ssurvey_df$TimeTV and ssurvey_df$Happiness,
#   using default method pearson.
cor.test(ssurvey_df$TimeTV,ssurvey_df$Happiness,method="pearson")
# Corelation with confidence level 0.99
cor.test(ssurvey_df$TimeTV,ssurvey_df$Happiness,method="pearson",conf.level = 0.99 )
#- Matix values interpretation :
  # - Values 0.63 represent positive relationship between variables
  #   TimeTV and Happiness.
  # - Values -0.88 represent negative relationship between variables
  #   TimeTV and TimeReading.

#---------------------------------------#
```

```
#       **** Assignment-V  *****        #
#---------------------------------------#

# Assignment-V : Calculate the correlation coefficient and the coefficient of
# determination, describe what you conclude about the results.

#-- Explanation :

# Objective is to find the co-relation between predictor variables TimeTV and
# TimeReading, Positive corelation coefficients (0.63) positive colinear
# relationships between them, however coefficient of detemination prediction
# (0.47), means a 47% variation in the Happiness can be explained by the time
# spend on watching TV and reading time.

# correlation coefficient
cor.test(ssurvey_df$TimeTV,ssurvey_df$Happiness,method="pearson")

# coefficient of determination
ss_model <- lm(ssurvey_df$Happiness ~ ssurvey_df$TimeTV + ssurvey_df$TimeReading, data=ssurvey_df)

#view model summary
summary(ss_model)
summary(ss_model)$r.squared

# References -
# https://www.statology.org/good-r-squared-value/

#---------------------------------------#
#       **** Assignment-VI  *****       #
#---------------------------------------#

# Assignment-VI : Based on your analysis can you say that watching more TV caused
# students to read less? Explain.

#-- Explanation :
# - Negative co-relation between the variables TimeTV and TimeReading,
#   and correlation coefficient (-0.88)
# - indicates student who spend more time watching TV will spend less
#   hours on reading and vise versa.

#?ggplot()
ggplot(ssurvey_df, aes(x=ssurvey_df$TimeTV, y=ssurvey_df$TimeReading)) +
  geom_point() +
  xlab("TimeTV") +
  ylab("TimeReading")

GGally::ggpairs(ssurvey_df)

#---------------------------------------#
#       **** Assignment-VII  *****      #
#---------------------------------------#

# Assignment-VII : 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.

#-- Explanation :

#  With vector V1, Partial correlation value between variables TimeTV and Happiness
#  is 0.63, which signifies that both variables highly consistent and they increase
#  with each other.

#  With vector V2, partial correlation value between variables TimeTV and
#  Happiness changed, TimeTV and Happiness vector is still the same because the
#  vector TimeReading affecting them. So now the correlation value dropped to
#  0.63 to 0.59 because TimeTV and TimeReading are inconsist with the
#  value of -0.8729450.


# install.packages("ppcor")
# install.packages("dplyr")
library(ppcor)
library(dplyr)
library(purrr)

V1 <- ssurvey_df %>% dplyr::select(TimeTV,Happiness)
V2 <- ssurvey_df %>% dplyr::select(TimeTV,Happiness,TimeReading)
ppcor::pcor(V1)
ppcor::pcor(V2)
#pcor(ssurvey_df)

#- Reference - https://www.statology.org/partial-correlation-r/
#              https://www.geeksforgeeks.org/how-to-calculate-partial-correlation-in-r/
```