DSC_520_week7_Assignment01

Reenie Christudass

2022-07-24

Contents

Load Libraries

```
if(!require('pander')) {
  install.packages('pander')
  library('pander')
}
```

```
## Loading required package: pander
```

```
## Warning: package ’pander’ was built under R version 4.2.1
```

```
if(!require('ggplot2')) {
  install.packages('ggplot2')
  library('ggplot2')
}
```

```
## Loading required package: ggplot2
```

```
## Warning: package ’ggplot2’ was built under R version 4.2.1
```

```
if(!require('knitr')) {
  install.packages('knitr')
  library('knitr')
}
```

```
## Loading required package: knitr
```

```
if(!require('tinytex')) {
  install.packages('tinytex')
  library('tinytex')
}
```

```
## Loading required package: tinytex
```

```
if(!require('ppcor')) {
  install.packages('ppcor')
  library('ppcor')
}
```

```
## Loading required package: ppcor
```

```
## Warning: package ’ppcor’ was built under R version 4.2.1
```

```
## Loading required package: MASS
```

```
if(!require('formatR')) {
  install.packages('formatR')
  library('formatR')
}
```

```
## Loading required package: formatR
```

```
## Warning: package 'formatR' was built under R version 4.2.1
```

```
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)
```

Read the Student Survey

```
## Set the working directory to the root of your DSC 520 directory
setwd("C:/Users/chris/dsc520/data")
```

```
## Load the `data/student-survey` to
survey_df <- read.csv("C:/Users/chris/dsc520/data/student-survey.csv")
head(survey_df)
```

```
##   TimeReading TimeTV Happiness Gender
## 1           1     90     86.20      1
## 2           2     95     88.70      0
## 3           2     85     70.17      0
## 4           2     80     61.31      1
## 5           3     75     89.52      1
## 6           4     70     60.50      1
```

Question 1

```
## 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
## indicate.

## Covariance between two variable (Happiness and TimeTV)
Mood = survey_df$Happiness
TVTime = survey_df$TimeTV
cov(Mood, TVTime)
```

```
## [1] 114.3773
```

```
## Covariance between two variable (TimeReading and TimeTV)
Mood = survey_df$TimeReading
TVTime = survey_df$TimeTV
cov(Mood, TVTime)
```

```
## [1] -20.36364
```

```
print(paste(" Happiness vs TVTime - As one variable changes, the other variable
will change in the same direction with a magnitude of 114.3773. Humans are happy
when they watch TV."))
```

```
## [1] " Happiness vs TVTime - As one variable changes, the other variable \nwill change in the same di
```

```
print(paste(" TimeReading vs TVTime - The calculated covariance of TimeReading
and TVTime is negative -20.3636, indicating the two variables are negatively
related. As one variable changes, the other variable will change in the opposite
direction with a magnitude of -20.3636. "))
```

```
## [1] " TimeReading vs TVTime - The calculated covariance of TimeReading \nand TVTime is negative -20.3
```

```
print(paste("Covariance is used to measure variables that have different units of
measurement. The unit of measurement is not standarized "))
```

```
## [1] "Covariance is used to measure variables that have different units of\nmeasurement. The unit of m
```

Question 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.

# Changing the scale for column Happiness and TimeTV. Multiple it by constant
# value 20

survey_df_change_measure <- survey_df
survey_df_change_measure$Happiness <- survey_df$Happiness * 20
survey_df_change_measure$TimeTV <- survey_df$TimeTV * 20
head(survey_df_change_measure)
```

```
##   TimeReading TimeTV Happiness Gender
## 1           1   1800    1724.0      1
## 2           2   1900    1774.0      0
## 3           2   1700    1403.4      0
## 4           2   1600    1226.2      1
## 5           3   1500    1790.4      1
## 6           4   1400    1210.0      1
```

```
## Covariance between two variable
Mood = survey_df_change_measure$Happiness
TVTime = survey_df_change_measure$TimeTV
cov(Mood, TVTime)
```

```
## [1] 45750.91
```

```
print(paste("Change of scale affects covariance. For example, if the value of two variables is multipli
```

```
## [1] "Change of scale affects covariance. For example, if the value of two variables is multiplied by
```

Question 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?


## Covariance between two variable
Mood = survey_df$TimeReading
TVTime = survey_df$TimeTV
cov(Mood, TVTime)
```

```
## [1] -20.36364
```

```
Mood = survey_df$Happiness
TVTime = survey_df$TimeTV
cov(Mood, TVTime)
```

```
## [1] 114.3773
```

```
print(paste("CONCLUSION:Pearson Correlation between TimeReading and TimeTV found above
is -0.88( Negative correlation), meaning that the 2 variables vary in opposite
direction. More the amount of time you spend in reading book and less time is
spend in watching TV.The correlation coefficient (the closer to -1 or 1) shows
the stronger the relationship"))
```

```
## [1] "CONCLUSION:Pearson Correlation between TimeReading and TimeTV found above \nis -0.88( Negative c
```

```
print(paste("CONCLUSION:Pearson Correlation between Happiness and TimeTV found above
is 0.64( Positive correlation), meaning that the 2 variables vary in same direction."))
```

```
## [1] "CONCLUSION:Pearson Correlation between Happiness and TimeTV found above \nis 0.64( Positive corr
```

```
print(paste("CONCLUSION:A correlation close to 0 indicates that the two variables
 are independent - TimeTV and Gender"))
```

```
## [1] "CONCLUSION:A correlation close to 0 indicates that the two variables \n are independent - TimeTV
```

Question 4

```
## 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.

round(cor(survey_df), digits = 2  # rounded to 2 decimals
)
```

```
##             TimeReading TimeTV Happiness Gender
## TimeReading        1.00  -0.88     -0.43  -0.09
## TimeTV            -0.88   1.00      0.64   0.01
## Happiness         -0.43   0.64      1.00   0.16
## Gender            -0.09   0.01      0.16   1.00
```

```
## correlation between two variable
Mood = survey_df$TimeReading
TVTime = survey_df$TimeTV
cor(Mood, TVTime)
```

```
## [1] -0.8830677
```

```
## Pearson correlation test
res <- cor.test(survey_df$TimeReading, survey_df$TimeTV, method = "pearson", conf.level = 0.99)
res
```

```
##
##  Pearson's product-moment correlation
##
## data:  survey_df$TimeReading and survey_df$TimeTV
## t = -5.6457, df = 9, p-value = 0.0003153
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.9801052 -0.4453124
## sample estimates:
##        cor
## -0.8830677
```

```
# Extract 99 percent confidence interval
res$conf.int
```

```
## [1] -0.9801052 -0.4453124
## attr(,"conf.level")
## [1] 0.99
```

```
# Extract the p.value
res$p.value
```

```
## [1] 0.0003153378
```

```
# Extract the correlation coefficient
res$estimate
```

```
##        cor
## -0.8830677
```

Question 5

```
## Calculate the correlation coefficient and the coefficient of determination,
## describe what you conclude about the results.

## correlation between two variable
Mood = survey_df$TimeReading
TVTime = survey_df$TimeTV
cor(Mood, TVTime)
```

```
## [1] -0.8830677
```

```
## coefficient of determination between two variable
eruption.lm = lm(TimeReading ~ TimeTV, data = survey_df)
summary(eruption.lm)$r.squared
```

```
## [1] 0.7798085
```

```
mod1 <- lm(TimeReading ~ TimeTV, data = survey_df)
mod1_summ <- summary(mod1)
mod1_summ$coefficients
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 12.3028721 1.55704477  7.901425 2.443632e-05
## TimeTV      -0.1169713 0.02071878 -5.645664 3.153378e-04
```

```
mod1_summ$r.squared
```

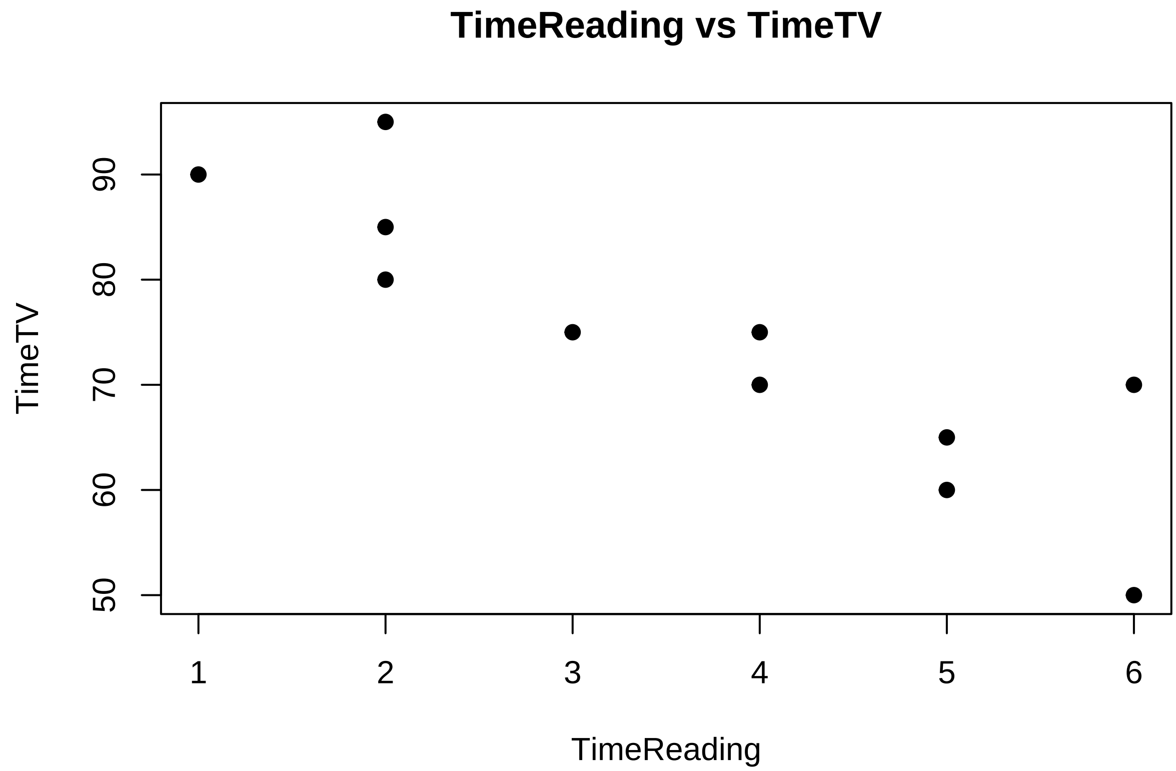
```
## [1] 0.7798085
```

Question 6

```
## Based on your analysis can you say that watching more TV caused students to
## read less? Explain. A student who watches more TV has decreased the reading
## time.Negative Correlation

x <- survey_df$TimeReading
y <- survey_df$TimeTV

plot(x, y, main = "TimeReading vs TimeTV", xlab = "TimeReading", ylab = "TimeTV",
    pch = 19)
```



```r
print(paste("A student who watches more TV has decreased the reading time.Negative Correlation"))
```

```
## [1] "A student who watches more TV has decreased the reading time.Negative Correlation"
```

Question 7

```r
## 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.

pcor(survey_df)
```

```
## $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"
```

```
print(paste("
Suppose we use a set of data  which lists three (TimeReading, TimeTV, Happiness). Each child was
tested for TimeReading (Y) and TimeTV (X2), and their Happiness was also noted. A correlation
statistic was desired which predicts Y (TimeReading) from X1 and X2 (TimeTV and Happiness).

Normally, in a situation where X1 and X2 were independent random variables, we'd find out how
important each variable was by computing a squared coefficient of correlation between X1 and X2
and the dependent variable Y. We would know that these squared coefficients of correlation were
equal to the square multiple coefficient of correlation. But in a case like ours, X1 and X2 are
anything but independent. TimeReading is highly dependent on Happiness, and so using the squared
coefficient will count the contributions of each variable several times over.
"))
```

```
## [1] "\nSuppose we use a set of data  which lists three (TimeReading, TimeTV, Happiness). Each child w
```