RMarkdown Assignment - Exercise 9

Pushkar Chougule

Oct 4th 2020

R Markdown

** Reading the csv into data frame **

```
student_df <- read.csv("student-survey.csv", header = TRUE)
```

a. Covariance Calculations

```
library(pander)
pander(cov(student_df), caption = "Covariance Calculations - Raw data")
```

Table 1: Covariance Calculations - Raw data

	TimeReading	TimeTV	Happiness	Gender
TimeReading	3.055	-20.36	-10.35	-0.08182
TimeTV	-20.36	174.1	114.4	0.04545
Happiness	-10.35	114.4	185.5	1.117
Gender	-0.08182	0.04545	1.117	0.2727

Answer a.

> As we learn from the (Field, Miles, and Field 2012) book, Covariance is the simplest way to look at whether two variables are associated with each other. It is calculated by using deviation of the one variable from the mean and corresponding deviation of second variable from it's own mean as well. If both of the individual deviations are Positive or both are Negative, the result is positive covariance (meaning both variables change in the same directions). Whereas, if one deviation is positive and other is negative, then Covariance is Negative, indicating the two variables in question change in oppositive directions. e.g. From above results, we can see that covariance value for TimeReading / TimeTV combination is approx. -20.36, which means the two variables change in opposite directions.Whereas TimeTV and Happiness combination have covariance of approx. 114.37 , which means they both change in the same direction.

b. Measurements being used for the variables and effect of change in measurements

```
## converting Time Reading to minutes, assuming it is mentioned in hours

student_df$TimeReading <- student_df$TimeReading * 60
```

```
# covariance calculation for all the variables

pander(cov(student_df),
       caption = "Covariance Calculations - Formatted data")
```

Table 2: Covariance Calculations - Formatted data

	TimeReading	TimeTV	Happiness	Gender
TimeReading	10996	-1222	-621	-4.909
TimeTV	-1222	174.1	114.4	0.04545
Happiness	-621	114.4	185.5	1.117
Gender	-4.909	0.04545	1.117	0.2727

Answer b.

In the provided data, the measurement scale have not been mentioned. But, looking at the numbers, assuming that TimeReading is in hours, TimeTV is minutes, Happiness is percentages (pure number on scale of 100) and Gender is categorical variable (0 or 1). So, looking at above calculations, we can see the formatting of data (conversion to same scale), has big effect on Covariance values. e.g. TimeReading / TimeTV covariance value increased to -1222, from previously being -20.36. Thus, Covariance may not be very reliable indicator of the size of the effect / relation between the two variables, as it can reasily get affected by measurement scale used. Better alternative for this would be using Correlation values - using cor() function and methods associated. Correlation coeficients are standardizations of Covariance and they use respective Standard deviations as well for calculations. The values vary between +1 (positive correlation - both variables change in same direction) and -1 (negative correlation - two variables in review change in opposite directions). Correlation coeeficient value of 0 indicates there is no relation between the two variables.

C. Type of Correlation test to use and reasoning

Answer c.

Looking at the Covariance results, we can see that the covariance values for possible combinations of variables are positively / negatively related to each other, as explained earlier. So, I would predict that the Correlation test would yield similar results for all the variables, except that the range of all the respective covariances will remain between +1 and -1. e.g. TimeReading and TimeTV - covariance value is -1222, with formatted data and we expect the direction to remain the same (negative correlation) Below first calculating the correlation coefficents using all three available methods viz. pearson, spearman and kendall.

```
pander(cor(student_df, method = "pearson"),
       caption = "Student Survey Pearson Correlation")
```

Table 3: Student Survey Pearson Correlation

	TimeReading	TimeTV	Happiness	Gender
TimeReading	1	-0.8831	-0.4349	-0.08964
TimeTV	-0.8831	1	0.6366	0.006597
Happiness	-0.4349	0.6366	1	0.157
Gender	-0.08964	0.006597	0.157	1

```
pander(cor(student_df, method = "spearman"),
       caption = "Student Survey Spearman Correlation")
```

Table 4: Student Survey Spearman Correlation

	TimeReading	TimeTV	Happiness	Gender
TimeReading	1	-0.9073	-0.4065	-0.08801
TimeTV	-0.9073	1	0.5662	-0.029
Happiness	-0.4065	0.5662	1	0.1155
Gender	-0.08801	-0.029	0.1155	1

```
pander(cor(student_df, method = "kendall"),
       caption = "Student Survey Kendall Correlation")
```

Table 5: Student Survey Kendall Correlation

	TimeReading	TimeTV	Happiness	Gender
TimeReading	1	-0.8045	-0.2889	-0.07825
TimeTV	-0.8045	1	0.463	-0.02508
Happiness	-0.2889	0.463	1	0.09847
Gender	-0.07825	-0.02508	0.09847	1

Answer c.

Looking at above mentioned results, we can confirm that the directions of relations between the possible combinations of the two variables are remaining the same and values vary between +1 and -1. e.g. TimeReading an TimeTV correlation value for "Pearson" method shows approx. -0.88 value. So, we can confirm that they are still negatively correlated to each other. And with TimeTV and Happiness variables - correlation coefficient value for "Pearson" method is approx. +0.64, which confirms the Positive relation between these two variables. I would prefer to use "Pearson method" since our data is interval data and later on we have to calculate correlation matrix for the variables with rcorr() function, for which Kendall method is not supported. And, Spearman method needs data variables being compared to be ordinal data, which can be categorized into finite number of categories. But our data being compared e.g. Time Reading / Time TV can take on many different values and there can not be categories.

d. Correlation Analysis

`1. All Variables`

Used pander for correlation matrix display in table format It can also be done using ggpairs() - GGally package (Lander 2014).

Table 6: Correlation table using Pearson method

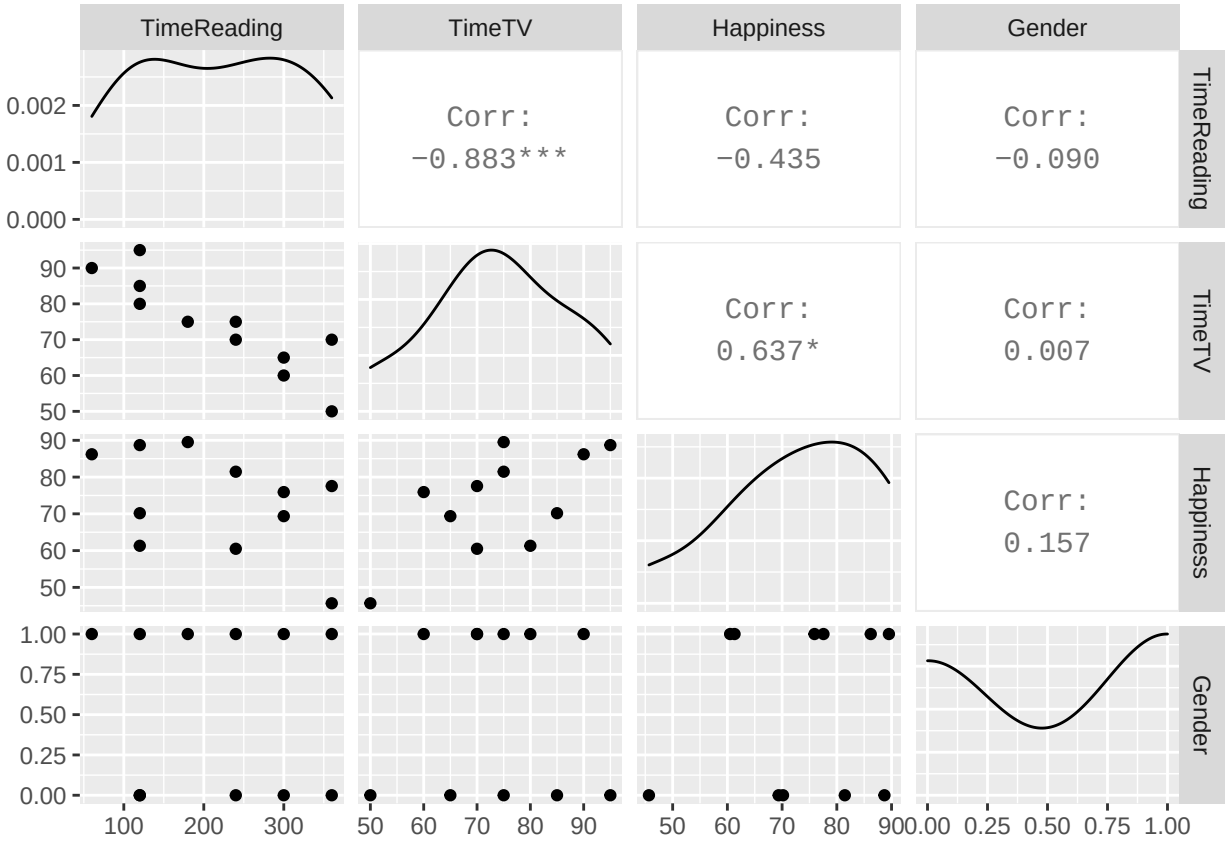
	TimeReading	TimeTV	Happiness	Gender
TimeReading	1	-0.8831	-0.4349	-0.08964
TimeTV	-0.8831	1	0.6366	0.006597
Happiness	-0.4349	0.6366	1	0.157
Gender	-0.08964	0.006597	0.157	1

```
## Loading required package: ggplot2

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

##
## Attaching package: 'GGally'

## The following object is masked from 'package:pander':
##
##     wrap
```



2. Single correlation between two pairs of variables

```r
cor.test(student_df$TimeReading, student_df$TimeTV, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  student_df$TimeReading and student_df$TimeTV
## 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
```

```r
cor.test(student_df$TimeTV, student_df$Happiness, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  student_df$TimeTV and student_df$Happiness
## t = 2.4761, df = 9, p-value = 0.03521
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.05934031 0.89476238
## sample estimates:
##      cor
## 0.636556
```

3. Single correlation between two pairs of variables

```r
cor.test(student_df$TimeReading, student_df$TimeTV, method = "pearson", conf.level = 0.99)
```

```
##
##  Pearson's product-moment correlation
##
## data:  student_df$TimeReading and student_df$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
```

```r
cor.test(student_df$TimeTV, student_df$Happiness, method = "pearson", conf.level = 0.99)
```

```
##
##  Pearson's product-moment correlation
##
## data:  student_df$TimeTV and student_df$Happiness
```

```
## t = 2.4761, df = 9, p-value = 0.03521
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.1570212  0.9306275
## sample estimates:
##      cor
## 0.636556
```

```
4. Explanation of the two correlation tests above
```

correlation tests in Step 2 and 3 for **TimeReading / TimeTV** : In Step both 2 and 3, the p value of this relation is 0.0003153, which much lower than 0.05, indicating the relation between these two variables is significant i.e. chances of coefficient value being -0.88 when there is no relationship between these two variables in the given dataset are 0.0003153 or 0.03%, which indicates Null Hypothesis is NOT true. For the confidence interval (CI) for 95% case is between (-0.9694145 to -0.6021920) and CI for 99% case is between (-0.9801052 to -0.4453124). This confirms that the Negative correlation between TimeReading / TimeTV holds good almost all of the times.

correlation tests in Step 2 and 3 for **TimeTV / Happiness** : In Step both 2 and 3, the p value of this relation is 0.03521, which is lower than 0.05, indicating the relation between these two variables is significant i.e. chances of coefficient value being approx +0.64 when there is no relationship between these two variables in the given dataset are 0.03521 or 3.5%, which indicates Null Hypothesis is NOT true. For the confidence interval (CI) for 95% case is between (0.05934031 0.89476238) and CI for 99% case is between (-0.1570212 0.9306275). This indicates that the Positive correlation between TimeTV / Happiness holds good most of the times, but in some cases, the correlation may turn negative i.e. TimeTV and Happiness may change in opposite directions as we move away from the mean value of the observations.

e. correlation coefficient and coefficient of determination

```
pander(cor(student_df),
       caption = "Correlation Coefficients table")
```

Table 7: Correlation Coefficients table

	TimeReading	TimeTV	Happiness	Gender
TimeReading	1	-0.8831	-0.4349	-0.08964
TimeTV	-0.8831	1	0.6366	0.006597
Happiness	-0.4349	0.6366	1	0.157
Gender	-0.08964	0.006597	0.157	1

```
pander(cor(student_df)^2,
       caption = "Coefficient of Determination R^2 table")
```

Table 8: Coefficient of Determination R^2 table

	TimeReading	TimeTV	Happiness	Gender
TimeReading	1	0.7798	0.1891	0.008036
TimeTV	0.7798	1	0.4052	4.352e-05

	TimeReading	TimeTV	Happiness	Gender
Happiness	0.1891	0.4052	1	0.02465
Gender	0.008036	4.352e-05	0.02465	1

Answer e. > We have earlier reviewed the Correlation Coefficients values previously. So, we will review the Coefficient of Determination values. Coefficient of Determination value between TimeReading and TimeTV is approx 0.78 or 78%, which is very significant. It means TimeTV may share 78% of the TimeReading variations, but we may not necessarily be able to say that it is the cause behind the changing TimeReading, because the relation can also be interpreted the other way round i.e. TimeReading can cause the changes in TimeTV. Coefficient of Determination value between Happiness and TimeTV is approx 0.405 or 40.5%, which is very significant. It means TimeTV may share 40.5% of the Happiness variations, but it may not necessarily mean that it is the cause behind this. One can also suggest that being more Happy may make someone watch more TV.

f. watching more TV causing students to read less

Answer f. > As we discussed in the explanation of Answer e, below is the highlight Coefficient of Determination value between TimeReading and TimeTV is approx 0.78 or 78%, which is very significant. It means TimeTV may share 78% of the TimeReading variations, but we may not necessarily be able to say that it is the cause behind the changing TimeReading, because the relation can also be interpreted the other way round i.e. TimeReading can cause the changes in TimeTV. Same explanation can be provided using correlation value of -0.88 that we can see negative correlation between TimeReading and TimeTV. But we may not be able to definitely say that whether TimeReading causes changes in TimeTV or vice-a-versa. That is we may not be able to clearly conclude which variable is independent variable and which one is the dependent variable, just by looking at the values.

g. Partial correlation

```
library(ggm)

pcor(c("TimeReading", "TimeTV", "Happiness"), var(student_df))
```

```
## [1] -0.872945
```

g. Answer > The above value of correlation coefficient : -0.872945, indicates that even if we keep the effects of Happiness variable constant, is fairly close match to the pairwise correlation coefficient value of -0.88. So, the relationship between TimeReading and TimeTV, which was determined earlier, still holds true.

References

Field, A., J. Miles, and Z. Field. 2012. *Discovering Statistics Using R*. SAGE Publications. https://books.google.com/books?id=wd2K2zC3swIC.

Lander, J. P. 2014. *R for Everyone: Advanced Analytics and Graphics*. Addison-Wesley Data and Analytics Series. Addison-Wesley. https://books.google.com/books?id=3eBVAgAAQBAJ.