Week7 Assignment #02

Mcouto

2023-07-23

As a data science intern with newly learned knowledge in skills in statistical correlation and R programming, you will analyze the results of a survey recently given to college students. You learn that the research question being investigated is: "Is there a significant relationship between the amount of time spent reading and the time spent watching television?"

You are also interested if there are other significant relationships that can be discovered?

The survey data is located in this StudentSurvey.csv file. 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.

```
studentsurvey_df <- read.csv("student-survey.csv")
surv_cov <-cov(studentsurvey_df[, c("TimeReading",
                                    "TimeTV", "Happiness", "Gender")])
library(knitr)
kable(surv_cov, caption = "Survey Covariates")
```

Table 1: Survey Covariates

	TimeReading	TimeTV	Happiness	Gender
TimeReading	3.0545455	-20.3636364	-10.350091	-0.0818182
TimeTV	-20.3636364	174.0909091	114.377273	0.0454545
Happiness	-10.3500909	114.3772727	185.451422	1.1166364
Gender	-0.0818182	0.0454545	1.116636	0.2727273

Covariance measures the extent of the relationship between two variables.A positive result indicates that as one variable increases, the other increases with it showing a positive linear relationship. Whereas a negative result shows that as one increases, the other decreases.

Negative Covariance from student survey

TimeReading and TimeTV **-20.36**
TimeReading and Happiness **-10.35**
TimeReading and Gender **-0.081**

Positive Covariance from Student Survey

TimeTV and Happiness **114.38**
TimeTV and Gender **0.045**
Happiness and Gender **1.12**

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.

Variables	Measure	Type
TimeReading	Hours	• Discrete
TimeTV	Minutes	• Discrete
Happiness	Scale	• Continuous
Gender	1/0	• Categorical

Considering the difference between the number on TimeReading vs TimeTv, Even though it doesn't indicate on the table, I think the TimeReading variable is measured in hours. The outcome on the covariate test would be more meaningful if the variables we're comparing, i.e. time spent reading vs tv have the same units of measurement. However, since these measures are relative, it may not make much of a difference on the positive/negative correlation since minutes and hours are proportional to each other

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 allows us to dive deeper into the outcomes of our covariance test. Correlation gives us an indication of the direction and strength of the relationship within the variables we're testing. The Pearson's product-moment correlation coefficient is the default test in R. My prediction is that the Pearson's correlation will mirror the correlation test done previously However, its important to note that Pearson assumes normality for all the variables, to which, the variable gender is not. So, if we want to measure the correlation between gender and the other variables, we can use the Point-Biserial Correlation. In statistics, I find, that there is no one-size fits all approach to the tests we perform. Often, we need to understand the data we're using and then choose the appropriate test to get insights from the information we have

Perform a correlation analysis of: 1. All variables

```
surv_cor <-cor(studentsurvey_df[, c("TimeReading",
                                    "TimeTV", "Happiness", "Gender")])
library(knitr)
kable(surv_cor, caption = "Survey Correlation All Variables")
```

Table 3: Survey Correlation All Variables

	TimeReading	TimeTV	Happiness	Gender
TimeReading	1.0000000	-0.8830677	-0.4348663	-0.0896421
TimeTV	-0.8830677	1.0000000	0.6365560	0.0065967
Happiness	-0.4348663	0.6365560	1.0000000	0.1570118
Gender	-0.0896421	0.0065967	0.1570118	1.0000000

2. A single correlation between two a pair of the variables - Point-Biserial Correlation

```
gender_happiness <- cor.test(studentsurvey_df$Happiness,studentsurvey_df$Gender)

gender_happiness
```

```
##
##  Pearson's product-moment correlation
##
## data:  studentsurvey_df$Happiness and studentsurvey_df$Gender
## t = 0.47695, df = 9, p-value = 0.6448
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4889126  0.6917342
## sample estimates:
##       cor
## 0.1570118
```

2. A single correlation between two a pair of the variables - Spearman Correlation

```
Read_Happy_Cor <- cor.test(studentsurvey_df$TimeReading, studentsurvey_df$Happiness, method = "spearman

Read_Happy_Cor
```

```
##
##  Spearman's rank correlation rho
##
## data:  studentsurvey_df$TimeReading and studentsurvey_df$Happiness
## S = 309.43, p-value = 0.2147
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.4065196
```

2. A single correlation between two a pair of the variables - Spearman Correlation at 99% confidence interval

```
Read_Happy_Cor <- cor.test(studentsurvey_df$TimeReading, studentsurvey_df$Happiness, method = "spearman

Read_Happy_Cor
```

```
##
##  Spearman's rank correlation rho
##
## data:  studentsurvey_df$TimeReading and studentsurvey_df$Happiness
## S = 309.43, p-value = 0.2147
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.4065196
```

Describe what the calculations in the correlation matrix suggest about the relationship between the variables. Be specific with your explanation.

```
cor_matrix <-cor(studentsurvey_df[,c("TimeReading","TimeTV","Happiness")])

cor_matrix
```

```
##             TimeReading     TimeTV  Happiness
## TimeReading   1.0000000 -0.8830677 -0.4348663
## TimeTV       -0.8830677  1.0000000  0.6365560
## Happiness    -0.4348663  0.6365560  1.0000000
```

The values shown on the calculation for the correlation between reading and happiness at both tests with or without the 99% confidence interval produced the same results.The value of S at 309.43 shows us the strength of the significance of the relationship between the two variables, which is then used to calculate the p-value. The conventional guidelines tells us that the p-value of .21 (rounded) is greater than 0.05 meaning that our test is not statistically significant. Finally, the samples estimate: rho value of -0.41 shows a negative correlation. Meaning, a weak inverse relationship between the two. Weak because of its distance from the value of -1 and inverse meaning that as more time spent reading, happiness decreases.

The relationship between TimeReading and TimeTV is a strong, negative relationship

The relationship between TimeReading and Happiness is a negative relationship, but is not very strong.

The relationship between TimeReading and Gender is a very weak, negative relationship

The relationship between TimeTV and Happiness is a positive and moderately strong one

The relationship between TimeTV and Gender is a very weak, negative one

The relationship between Happiness and Gender is a positive, weak correlation

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

```
surv_cor <-cor(studentsurvey_df[, c("TimeReading",
                                    "TimeTV", "Happiness", "Gender")])
surv_cor
```

```
##             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
```

The first step to calculate the coefficient of determination is to use the formula for linear regression to calculate R-squared

```
step1 <- lm(TimeReading ~ TimeTV ,data = studentsurvey_df)
```

Once the R-squared is obtained, we use the summary and coefficients function

```
summary(step1)
```

```
##
## Call:
## lm(formula = TimeReading ~ TimeTV, data = studentsurvey_df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9452 -0.4922 -0.2846  0.3851  1.8851
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30287    1.55704   7.901 2.44e-05 ***
## TimeTV      -0.11697    0.02072  -5.646 0.000315 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8645 on 9 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7553
## F-statistic: 31.87 on 1 and 9 DF,  p-value: 0.0003153
```

```
summary(step1)$r.squared
```

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

```
coefficients(step1)
```

```
## (Intercept)      TimeTV
##  12.3028721  -0.1169713
```

Based on your analysis can you say that watching more TV caused students to read less? Explain. *Using the dataset to explain the relationship between watching tv and reading, we could make the determination that R-squared value of 78 tells us that TimeTV is a good predictor for TimeReading. However, since the correlation is inversely related, it tells us that as time spent watching TV increases, the time spent reading decreases.*

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.

```
library(ppcor)
par_cor_gender <- pcor.test(studentsurvey_df$TimeReading,studentsurvey_df$TimeTV,studentsurvey_df$Gende

par_cor_gender
```

```
##     estimate      p.value statistic  n gp  Method
## 1 -0.8860628 0.0006411949 -5.406281 11  1 pearson
```

```
par_cor_happiness <- pcor.test(studentsurvey_df$TimeReading,studentsurvey_df$TimeTV,studentsurvey_df$Ha

par_cor_happiness
```

```
##    estimate      p.value statistic  n gp  Method
## 1 -0.872945 0.0009753126 -5.061434 11  1 pearson
```

After looking at the outcome for TimeTV and TimeReading, I thought it would be interesting to see how the other variables affect this outcome. Partial correlation is the correlation of two variables while controlling for gender and happiness. The outcome table indicates that the partial correlation after controlling for the effects of GENDER = -0.89 and the partial correlation after controlling for the effects of HAPPINESS = -0.87