assignment_07_02_PothineniKalyan

Pothineni, Kalyan

2023-04-29

Student Survey

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

The covariance is a measure of the linear association between two variables. A positive covariance indicates that when one variable increases, the other tends to increase as well, while a negative covariance indicates that when one variable increases, the other tends to decrease. The magnitude of the covariance depends on the strength of the linear association between the two variables.

calculate covariance between columns from student survey file

```
student_df <- read.csv("data/student-survey.csv")

coveriance_matrix <- cov(student_df[, c("TimeReading","TimeTV","Happiness","Gender")])
coveriance_matrix
```

```
##              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
```

This will give a 4x4 matrix with the covariances between the four variables (Timereading, timetv, happiness, and gender). The diagonal elements of the matrix represent the variances of each variable, while the off-diagonal elements represent the covariances between pairs of variables.

In this case, the covariance matrix allows us to examine the relationships between Timereading, timetv, happiness, and gender.

The covariance between Timereading and gender provides information on the association between the amount of time spent reading and gender. Similarly, the covariance between timetv and gender would provide information on the association between the amount of time spent watching TV and gender.

Overall, the covariance matrix provides a useful summary of the relationships between multiple variables in a dataset, and can help us identify potential patterns or associations that might be of interest. Additionally, the interpretation of the covariance depends on the scale and units of the variables being measured, so it is important to consider these factors when interpreting the results.

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

The "Timereading" and "timetv" variables are measured using an interval or ratio scale, since they are integer variables representing quantities of time. The "happiness" variable is measured using a numeric scale, which could be either ordinal or interval/ratio, depending on how the scale is defined and what kind of assumptions are being made about the distances between different levels of happiness. Finally, the "gender" variable is measured using a nominal scale, where the values represent categories without any inherent order or numerical value.

Changing the measurement scale for a variable can have a significant effect on the covariance calculation. In particular, if the variables are not measured on the same scale, it can be difficult to compare or combine their covariances meaningfully. For example, if one variable is measured in seconds and another is measured in minutes, the resulting covariance will depend on the units used, and it may not be immediately clear how to interpret or compare the resulting values.

To address this issue, one potential solution is to standardize the variables by converting them to z-scores. This involves subtracting the mean value of each variable from its individual values, and dividing by the standard deviation. This will result in variables that are expressed in terms of standard deviations from the mean, which allows them to be more easily compared and combined. Once the variables have been standardized, we can then calculate the covariance matrix using the standardized values.

One alternative to using covariance is to use correlation, which is a standardized measure of the linear relationship between two variables. Correlation is less affected by differences in measurement scale, as it is based on the standardized values of the variables rather than the raw scores. However, correlation still assumes a linear relationship between the variables, and may not capture more complex associations. Additionally, it is important to use appropriate measurement scales for the variables when calculating correlation.

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?

Lets examine the relationship between Timereading and timetv variables in the student survey file, a suitable correlation test would be the Pearson correlation coefficient.

The Pearson correlation coefficient is appropriate for this analysis because both Timereading and timetv are continuous variables and the relationship between them can be assumed to be linear.

Based on the nature of the variables, it is difficult to make a prediction about whether the Pearson correlation coefficient will yield a positive or negative correlation. It is possible that students who spend more time reading also spend less time watching TV (negative correlation), or that there is no significant relationship between these variables (zero correlation). However, it is also possible that students who spend more time reading also spend more time watching TV (positive correlation), although this may seem less likely based on common assumptions about how students allocate their time.

Therefore, it is important to perform the correlation analysis to determine the nature and strength of the relationship between Timereading and timetv variables in the student survey file.

Calculate the Pearson correlation coefficient

```
correlation <- cor(student_df$TimeReading, student_df$TimeTV, method = "pearson")
print(correlation)
```

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

4. To perform a correlation analysis of all variables in the student survey file, we can use the cor() function in R as follows:

Calculate the correlation matrix for all variables

```
correlationMatrix <- cor(student_df, method = "pearson")
print(correlationMatrix)
```

```
##             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
```

This will output a matrix showing the Pearson correlation coefficients between all pairs of variables in the data set.

To perform a correlation analysis between a pair of variables, we can modify the above code as follows:

Calculate the Pearson correlation coefficient between Happiness and TimeReading

```
correlation <- cor(student_df$Happiness, student_df$TimeReading, method = "pearson")
print(correlation)
```

```
## [1] -0.4348663
```

This will output a single value showing the Pearson correlation coefficient between the Happiness and TimeReading variables.

To repeat the correlation test in step 2 with a 99% confidence interval, we can modify the code as follows:

Calculate the 99% confidence interval for the Pearson correlation coefficient between Timereading and timetv

```
correlation <- cor.test(student_df$Happiness,
                        student_df$TimeReading, method = "pearson",
                        conf.level = 0.99)
print(correlation)
```

```
##
##  Pearson's product-moment correlation
##
## data:  student_df$Happiness and student_df$TimeReading
## t = -1.4488, df = 9, p-value = 0.1813
## alternative hypothesis: true correlation is not equal to 0
## 99 percent confidence interval:
##  -0.8801821  0.4176242
## sample estimates:
##        cor
## -0.4348663
```

This will output a result showing the Pearson correlation coefficient between the Happiness and TimeReading variables, along with the 99% confidence interval for this coefficient.

The correlation matrix for all variables in the student survey file suggests that there is a weak negative correlation between Timereading and Happiness (correlation coefficient of -0.43), and a moderate positive correlation between Timetv and Happiness (correlation coefficient of 0.63). There is also a weak negative correlation between Timereading and timetv (correlation coefficient of -0.88), indicating that students who spend more time reading tend to spend less time watching TV.

When considering a single correlation coefficient between two variables, the Pearson correlation coefficient between Happiness and TimeReading is -0.43, which indicates a weak negative correlation. When this correlation test is repeated with a 99% confidence interval, the results suggest that the true population correlation coefficient is likely to lie between -0.88 and 0.41. This wider confidence interval indicates that the estimate of the correlation coefficient is less precise, and that the strength of the relationship between Happiness and TimeReading may be less clear.

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

Calculate the Pearson correlation coefficient

```
correlation <- cor(student_df$Happiness, student_df$TimeReading, method = "pearson")
```

```r
# Calculate the coefficient of determination
model <- lm(Happiness ~ TimeReading, data = student_df)
r_squared <- summary(model)$r.squared

# Print the results
print(paste("Correlation coefficient:", correlation))
```

```
## [1] "Correlation coefficient: -0.434866331376903"
```

```r
print(paste("Coefficient of determination:", r_squared))
```

```
## [1] "Coefficient of determination: 0.189108726165207"
```

The Pearson correlation coefficient between Happiness and TimeReading variables is -0.43, indicating a weak negative correlation. The coefficient of determination, also known as R-squared, is 0.189, which means that 18.9% of the variation in happiness can be explained by the variation in TimeReading

Overall, this suggests that there is a negative relationship between the amount of time spent reading and happiness, although other factors may also play a role in determining happiness. It is important to note that correlation does not necessarily imply causation, so we cannot conclude that reading more directly causes greater happiness..

6. Based on your analysis can you say that watching more TV caused students to read less? Explain

No, based on the analysis performed, we cannot conclude that watching more TV caused students to read less. The correlation coefficient between TimeReading and TimeTV variables is negative (-0.88), indicating a weak negative correlation.

Furthermore, the correlation analysis only provides information about the strength and direction of the relationship between the variables, but it does not provide information about causality or the direction of causality. To establish causality, we would need to conduct further research, such as a randomized controlled trial or a natural experiment, in which the amount of TV watched or time spent reading is manipulated while controlling for other factors that could influence the outcome

7.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.

To perform a partial correlation analysis, we will control for the effect of a third variable on the relationship between two variables. In this case, we will use the pcor.test() function from the ggm package to calculate the partial correlation coefficients between three variables in the student survey file.

Let's say we are interested in examining the relationship between TimeReading and Happiness, while controlling for the effect of gender. We can perform a partial correlation analysis as follows:

Calculate the partial correlation coefficient between TimeReading and happiness, controlling for gender

```
library(ppcor)
```

```
## Warning: package 'ppcor' was built under R version 4.2.3
```

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

```
partial_correlation <- pcor.test(student_df$TimeReading,
                                 student_df$Happiness,student_df$Gender)$estimate

# Print the result
print(paste("Partial correlation coefficient:", partial_correlation))
```

```
## [1] "Partial correlation coefficient: -0.42779849556702"
```

This code will calculate the partial correlation coefficient between Timereading and happiness, controlling for gender.

If the partial correlation coefficient is positive and statistically significant, it means that the relationship between Timereading and happiness remains even after accounting for the influence of gender. If the partial correlation coefficient is negative and statistically significant, it means that the relationship between Timereading and happiness is weakened or even disappears after controlling for gender.

By performing a partial correlation analysis, we can get a more nuanced understanding of the relationship between two variables and the potential impact of a third variable on that relationship. In this example, the partial correlation is weaker or negative, it indicates that gender could be a confounding variable that needs to be taken into account when interpreting the relationship between Timereading and happiness.