

# How do I perform bivariate analysis in R, and what are some examples?

Authored by  
**stats writer**

July 1, 2024

## RECOMMENDED CITATION

stats writer (2024). *How do I perform bivariate analysis in R, and what are some examples?*. PSYCHOLOGICAL SCALES. Retrieved from <https://scales.arabpsychology.com/?p=165413>

Bivariate analysis is a statistical technique used to determine the relationship between two variables. In order to perform bivariate analysis in R, one can use various built-in functions and packages such as "cor" and "lm". These functions allow for the calculation of correlation coefficients and the creation of scatter plots to visualize the relationship between the two variables. Other techniques such as linear regression and ANOVA can also be used for bivariate analysis in R. Some examples of when bivariate analysis may be useful include studying the relationship between age and income, or examining the correlation between education level and job satisfaction. By performing bivariate analysis in R, researchers and analysts can gain valuable insights and make informed decisions based on the relationship between two variables.

## **Perform Bivariate Analysis in R (With Examples)**

**The term bivariate analysis refers to the analysis of two variables. You can remember this because the prefix "bi" means "two."**

**The purpose of bivariate analysis is to understand the relationship between two variables**

**There are three common ways to perform bivariate analysis:**

**1. Scatterplots**

**2. Correlation Coefficients**

**3. Simple Linear Regression**

**The following example shows how to perform each of**

these types of bivariate analysis using the following dataset that contains information about two variables: (1) Hours spent studying and (2) Exam score received by 20 different students:

```
#create data frame
```

```
df <- data.frame(hours=c(1, 1, 1, 2, 2, 2, 3, 3, 3, 3,  
3, 4, 4, 5, 5, 6, 6, 6, 7, 8),  
score=c(75, 66, 68, 74, 78, 72, 85, 82, 90, 82,  
80, 88, 85, 90, 92, 94, 94, 88, 91, 96))
```

```
#view first six rows of data frame
```

```
head(df)
```

```
hours score
```

```
1 1 75
```

```
2 1 66
```

```
3 1 68
```

```
4 2 74
```

```
5 2 78
```

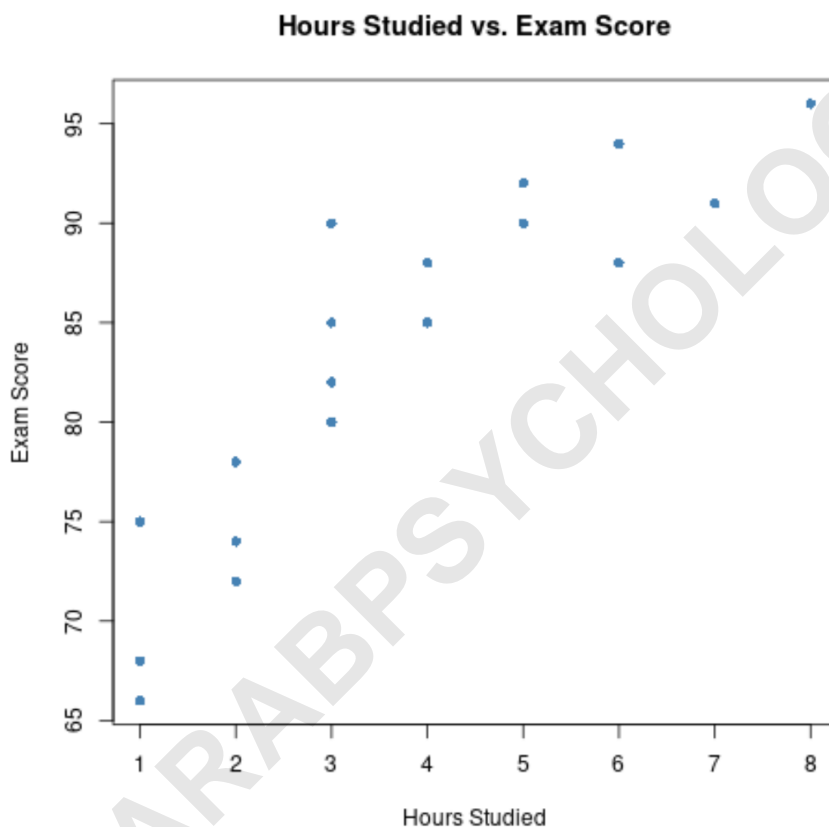
```
6 2 72
```

## 1. Scatterplots

We can use the following syntax to create a scatterplot

## of hours studied vs. exam score in R:

```
#create scatterplot of hours studied vs. exam score  
plot(df$hours, df$score, pch=16, col='steelblue',  
main='Hours Studied vs. Exam Score',  
xlab='Hours Studied', ylab='Exam Score')
```



The x-axis shows the hours studied and the y-axis shows the exam score received.

From the plot we can see that there is a positive

**relationship between the two variables: As hours studied increases, exam score tends to increase as well.**

## **2. Correlation Coefficients**

**A Pearson Correlation Coefficient is a way to quantify the linear relationship between two variables.**

**We can use the `cor()` function in R to calculate the Pearson Correlation Coefficient between two variables:**

```
#calculate correlation between hours studied and exam score received  
cor(df$hours, df$score)
```

**0.891306**

**The correlation coefficient turns out to be 0.891.**

**This value is close to 1, which indicates a strong positive correlation between hours studied and exam score received.**

## **3. Simple Linear Regression**

**Simple linear regression is a statistical method we can**

use to find the equation of the line that best "fits" a dataset, which we can then use to understand the exact relationship between two variables.

We can use the `lm()` function in R to fit a for hours studied and exam score received:

```
#fit simple linear regression model
```

```
fit <- lm(score ~ hours, data=df)
```

```
#view summary of model
```

```
summary(fit)
```

**Call:**

```
lm(formula = score ~ hours, data = df)
```

**Residuals:**

```
Min 1Q Median 3Q Max
```

```
-6.920 -3.927 1.309 1.903 9.385
```

**Coefficients:**

```
Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 69.0734 1.9651 35.15 < 2e-16 ***
```

```
hours 3.8471 0.4613 8.34 1.35e-07 ***
```

```
---
```

**Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1**

**Residual standard error: 4.171 on 18 degrees of freedom**

**Multiple R-squared: 0.7944, Adjusted R-squared: 0.783**

**F-statistic: 69.56 on 1 and 18 DF, p-value: 1.347e-07**

**The fitted regression equation turns out to be:**

**Exam Score = 69.0734 + 3.8471\*(hours studied)**

**This tells us that each additional hour studied is associated with an average increase of 3.8471 in exam score.**

**We can also use the fitted regression equation to predict the score that a student will receive based on their total hours studied.**

**For example, a student who studies for 3 hours is predicted to receive a score of 81.6147:**

**Exam Score = 69.0734 + 3.8471\*(hours studied)**

**Exam Score = 69.0734 + 3.8471\*(3)**  
**Exam Score = 81.6147**

## Additional Resources

**The following tutorials provide additional information about bivariate analysis:**

ARABPSYCHOLOGY.COM