How to Easily Perform Bivariate Analysis in R

Authored by

stats writer

December 2, 2025

Bivariate analysis is a fundamental statistical method essential for data exploration, specifically designed to investigate the nature and strength of the relationship between exactly two variables. By examining how changes in one variable correspond to changes in the other, analysts can gain crucial insights into causality, association, and dependency. This process is distinct from univariate analysis (analyzing one variable) and multivariate analysis (analyzing three or more variables).

The statistical programming language R provides a robust environment for conducting bivariate tests efficiently, utilizing powerful built-in functions. Key functions like `cor()`, `cov()`, and `t.test()` allow users to identify associations, calculate covariance, and perform formal hypothesis testing, such as a t-test, to assess the statistical significance of observed differences or relationships. Understanding the proper application of these tools is paramount for accurate data interpretation.

In the context of quantitative research and data science, the primary objective of bivariate analysis remains constant: to rigorously understand the interplay between two specific measures. We will explore three of the most common and powerful techniques employed to visualize and quantify this relationship in R.

The term **bivariate analysis** is derived directly from its function, where the prefix "bi" signifies two, confirming that the scope of the analysis is limited to two interdependent or independent variables. The following methods offer complementary perspectives—visualization, quantification, and modeling—to fully characterize the relationship:

Scatterplots: Visual exploration of data distribution and relationship direction.
Correlation Coefficients: Numerical quantification of the linear association strength.
Simple Linear Regression: Statistical modeling to predict outcomes and define the mathematical relationship.

Setting Up the Data Environment in R

To demonstrate these three forms of bivariate analysis, we will utilize a simulated dataset representative of a typical educational study. This dataset comprises 20 observations, recording two key variables for each student: the total **Hours spent studying** (our potential independent variable) and the final **Exam score received** (our dependent variable). This scenario is ideal for exploring a potential linear relationship where increased effort (hours) is hypothesized to lead to improved performance (score).

We begin by creating this data structure within R using the `data.frame()` function. A data frame is the standard organizational structure in R for storing tabular data, where variables are columns and observations are rows. The following code initializes the dataset named `df` and displays the first few entries to confirm successful data loading.

#create data frame in R
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 the data frame
head(df)

hours score
1 1 75
2 1 66
3 1 68
4 2 74
5 2 78
6 2 72

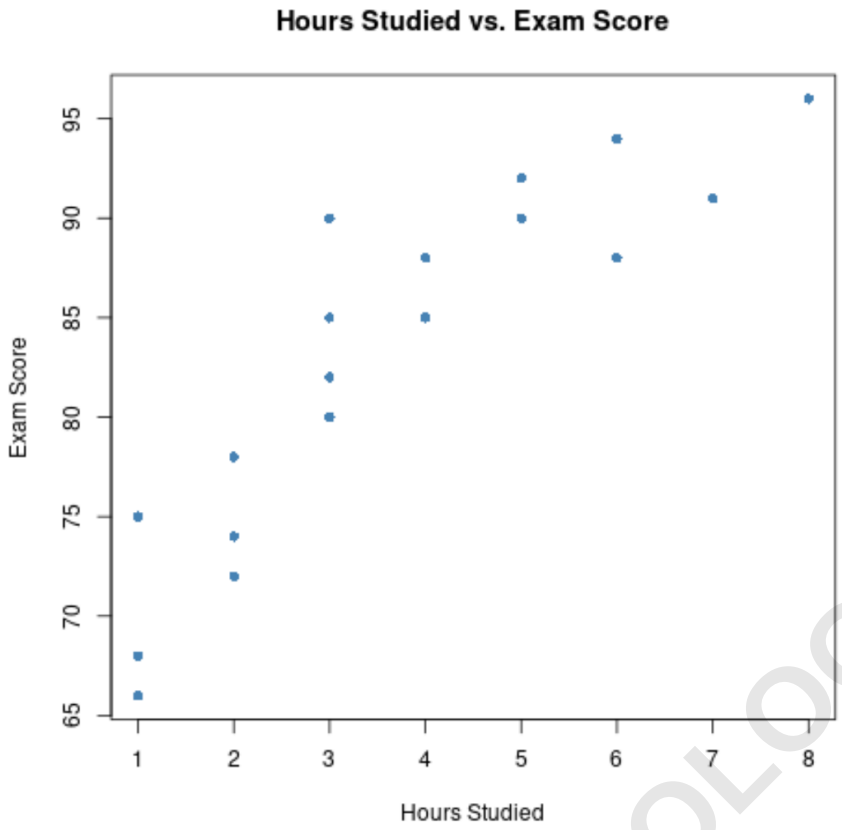
Visualizing Relationships with Scatterplots

The initial step in any bivariate analysis should involve visualization. Scatterplots are the most effective graphical tool for this purpose, providing a direct visual representation of the joint distribution of two continuous variables. By plotting the independent variable (Hours Studied) on the x-axis and the dependent variable (Exam Score) on the y-axis, we can immediately assess the form, direction, and strength of their relationship, as well as identify potential outliers.

In R, the base plotting function `plot()` is sufficient for generating a clear scatterplot. We specify the variables using the standard data frame indexing notation (`df$hours` and `df$score`). We also apply aesthetic enhancements, such as setting the point character (`pch=16`), assigning a color (`col='steelblue'`), and clearly labeling the axes and providing a descriptive title to ensure the plot is self-explanatory.

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

The resulting graph visually confirms the nature of the relationship:



Upon inspection of the scatterplot, we observe a clear pattern. The points generally ascend from the lower left corner to the upper right corner of the plot. This visual trend is indicative of a **positive relationship**: as the value of the independent variable (Hours Studied on the x-axis) increases, the value of the dependent variable (Exam Score on the y-axis) tends to increase concurrently. Furthermore, the points appear to cluster relatively closely around what could be a straight line, suggesting that the relationship is not only positive but also highly linear.

Quantifying Linear Association with Correlation Coefficients

While a scatterplot provides qualitative confirmation of a relationship, a Pearson Correlation Coefficient (r) offers a precise quantitative measure of the strength and direction of the **linear** association between two continuous variables. The correlation coefficient ranges from -1.0 to +1.0. A value near +1.0 signifies a strong positive linear relationship, a value near -1.0 indicates a strong negative linear relationship, and a value near 0 suggests a weak or non-existent linear relationship.

In R, calculating this essential metric is straightforward using the built-in `cor()` function, which defaults to computing the Pearson product-moment correlation coefficient. By supplying the two variable vectors, `df$hours` and `df$score`, we instruct R to measure how closely these variables move together across all observations.

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

0.891306

The calculated correlation coefficient is approximately **0.891**. Interpreting this result confirms the visual evidence from the scatterplot. Because 0.891 is very close to the maximum positive value of 1.0, it denotes an extremely strong, positive linear correlation between the hours a student spends studying and the score they receive on the exam. This high value suggests that studying time is an excellent linear predictor of exam performance in this sample.

Modeling Relationships using Simple Linear Regression

The most sophisticated form of bivariate analysis for continuous variables is Simple Linear Regression (SLR). SLR aims to model the relationship between the two variables by fitting a straight line (the regression line) that minimizes the sum of squared errors (or residuals) between the observed data points and the line. This method provides an algebraic equation that can be used for explicit relationship interpretation and for making predictions.

The general form of the simple linear regression model is $Y = beta_0 + beta_1 X + epsilon$, where Y is the dependent variable (Score), X is the independent variable (Hours), $beta_0$ is the intercept, $beta_1$ is the slope, and $epsilon$ represents the error term. In R, the `lm()` function (for linear model) is used to compute the coefficients ($beta_0$ and $beta_1$) that define the line of best fit. The syntax uses the formula structure `Y ~ X`, indicating that we are modeling `score` as a function of `hours`.

#fit simple linear regression model using the lm() function
fit <- lm(score ~ hours, data=df)

#view comprehensive summary of the model results
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 output of the `summary()` function provides all necessary statistical details for model evaluation. Focusing on the `Coefficients` table, we extract the estimated values for the intercept and the slope, which define our specific regression equation:

Fitted Regression Equation:
$$ text{Exam Score} = 69.0734 + 3.8471 times (text{Hours Studied}) $$

Interpreting the Regression Model Output

The coefficients extracted from the linear model are critical for understanding the quantitative relationship between studying hours and exam scores. The **Intercept** ($beta_0 = 69.0734$) represents the predicted exam score when the hours studied is zero. While this interpretation may not be strictly meaningful in all real-world scenarios, it serves as the baseline value for the prediction model.

More importantly, the **Hours** coefficient (the slope, $beta_1 = 3.8471$) dictates the marginal effect of the independent variable. This value signifies that, holding all other factors constant, for every additional hour a student spends studying, their predicted exam score increases by an average of **3.8471 points**. This slope is statistically significant, as indicated by the high t-value and the extremely low p-value, suggesting that the relationship observed is highly unlikely to be due to random chance.

Beyond the coefficients, the model summary provides crucial goodness-of-fit statistics. The **Multiple R-squared** value, which is 0.7944, indicates that approximately 79.44% of the variability in the Exam Score can be explained by the variation in Hours Studied, a strong explanatory power for a simple model. The remaining variability is accounted for by the residuals, which show the difference between the observed scores and the scores predicted by the fitted line.

Using the Model for Prediction

One of the primary advantages of fitting a Simple Linear Regression model is its predictive capability. Once the equation is established, we can input any value for the independent variable

(Hours Studied) within the range of our observed data to estimate the expected outcome (Exam Score). This predictive utility allows educators or students to gauge expected performance based on preparation effort.

For instance, to predict the exam score for a student who studies exactly 3 hours, we substitute $X=3$ into the derived regression equation. The prediction calculation proceeds as follows:

Exam Score = 69.0734 + 3.8471*(hours studied)
Exam Score = 69.0734 + 3.8471*(3)
Exam Score = **81.6147**

Therefore, a student who dedicates three hours to studying is predicted to achieve a score of **81.6147** based on the established linear model. It is important to remember that regression models only describe association and are subject to model assumptions and limitations, such as the independence of errors and the linearity of the relationship.

Conclusion and Further Study

Mastering bivariate analysis in R is essential for moving from simple data summarization to insightful statistical modeling. By employing scatterplots for initial inspection, correlation coefficients for quantification, and Simple Linear Regression for prediction, analysts can thoroughly characterize the relationship between any pair of variables. These foundational techniques pave the way for more complex multivariate analyses and sophisticated statistical inference.

For those seeking to deepen their knowledge, further exploration into the assumptions underlying linear regression (such as normality and homoscedasticity) and alternative methods, such as non-parametric correlation tests or generalized linear models, is highly recommended.