

Bivariate Relationships Featuring Two Continuous Variables

8

Abstract

In this chapter, we discuss bivariate relationships between two continuous variables. In research, these are the relationships that occur the most often. We can express bivariate relationships between continuous variables in three ways: (1) through a graphical representation in the form of a scatterplot, (2) through a correlation analysis, and (3) through a bivariate regression analysis. We describe and explain each method and show how to implement it in SPSS and Stata.

8.1 What Is a Bivariate Relationship Between Two Continuous Variables?

A bivariate relationship involving two continuous variables can be displayed graphically and through a correlation or regression analysis. Such a relationship can exist if there is a *general* tendency for these two variables to be related, even if it is not a completely determined rule. In statistical terms, we say that these two variables “vary together”; this means that values of the variable x (the independent variable) tend to occur more often with some values of the variable y (the dependent variable) than with other values of the variable y .

8.1.1 Positive and Negative Relationships

When we describe relationships between variables, we normally distinguish between positive and negative relationships.

Positive relationship High, or above average, values of x tend to occur with high, or above average, values of y . Also, low values of x tend to occur with low values of y .

Examples:

- Income and education
- National wealth and degree of democracy
- Height and weight

Negative relationship High, or above average, values of x tend to occur with *low*, or *below* average, values of y . Also, low values of x tend to occur with high values of y .

Examples:

- State social spending and income inequality
- Exposure to Fox News and support for Democrats
- Smoking and life expectancy

8.2 Scatterplots

A scatterplot graphically describes a quantitative relationship between two continuous variables: Each dot (point) is one individual observation's value on x and y . The values of the independent variable (X) appear in sequence on the horizontal or x -axis. The values of the dependent variable (Y) appear on the vertical or y -axis. For a positive association, the points tend to move diagonally from lower left to upper right. For a negative association, the points tend to move from upper left to lower right. For NO association, points are scattered with no discernable *diagonal* line.

8.2.1 Positive Relationships Displayed in a Scatterplot

Figure 8.1 displays a positive association, or a positive relationship, between countries' per capita GDP and the amount of energy they consume. We see that even if the data do not exactly follow a line, there is nevertheless a tendency that countries with higher GDP per capita values are associated with more energy usage. In other words, higher values of the x -axis (the independent variable) correspond to higher values on the y -axis (the dependent variable).

8.2.2 Negative Relationships Displayed in a Scatterplot

Figure 8.2 displays a negative relationship between per capita GDP, and the share agriculture makes up of a country's GDP; that is, our results indicate that the richer a country becomes, the more agriculture loses its importance for the economy. In statistical terms, we find that low values of the x -axis correspond to high values on the y -axis and high values on the x -axis correspond to low values on the y -axis.

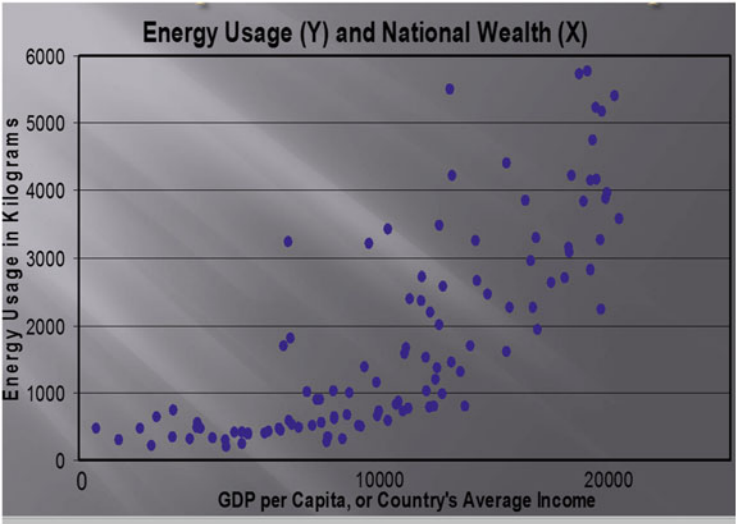


Fig. 8.1 Bivariate relationship between the GDP per capita and energy spending

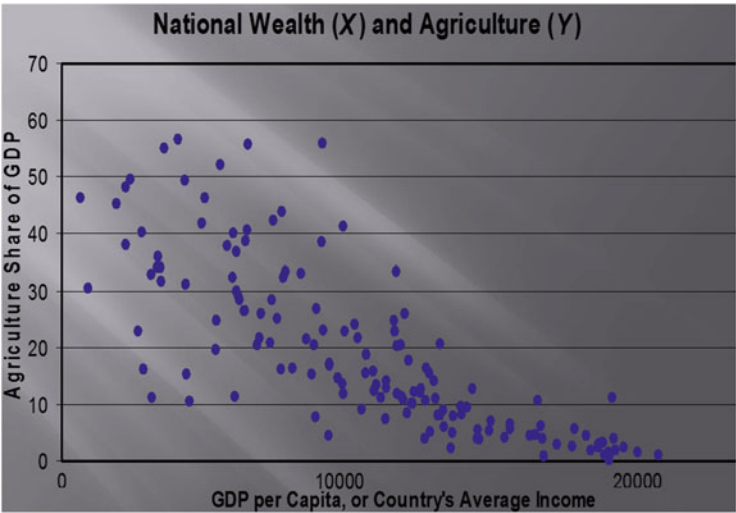


Fig. 8.2 Bivariate relationship between national wealth and agriculture

8.2.3 No Relationship Displayed in a Scatterplot

Figure 8.3 displays an instance in which the independent and dependent variable are unrelated to one another. In more detail, the graph highlights that the affluence of a country is unrelated to its size. In other words, there is no discernable direction to the points in the scatterplot.

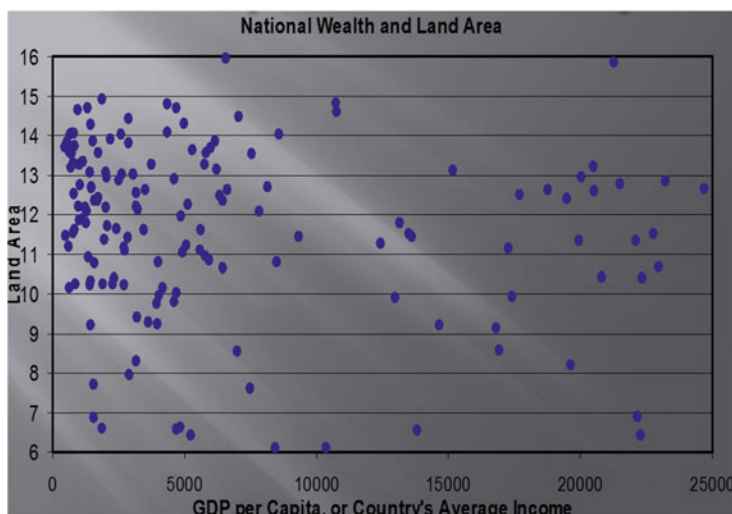


Fig. 8.3 Relationship between a country's GDP per capita and its size

8.3 Drawing the Line in a Scatterplot

The line we draw in a scatterplot is called the ordinary least square line (OLS line). In theory, we can draw a multitude of lines, but in practice we want to find the best fitting line to the data. The best fitting line is the line where the summed up distance of the points from below the line is equal to the summed up distance of the points from above the line. Figure 8.4 clearly shows a line that does not fit the data properly. The distance of all the points toward the line is much larger for the points below the line in comparison to the points above the line. In contrast, the distance of the points toward the line is the same for the points above the line as for the points below the line in Fig. 8.4. In contrast, the line in Fig. 8.5 is the best fitting line (i.e. the sum of the distance of all the points from the line is zero).

8.4 Doing Scatterplots in SPSS

For our scatterplot, we will use money spent partying as the dependent variable. For the independent variable, we will use quality of extra-curricular activities, hypothesizing that students who enjoy the university-sponsored activities will spend less money partying. Rather than going out and party, they will be in sports and social or political university clubs and partake in their activities. A scatterplot can help us confirm or disconfirm this hypothesis.

Step 1: Go to Graphs—Legacy Dialogs—Scatter/Dot (see Fig. 8.6).

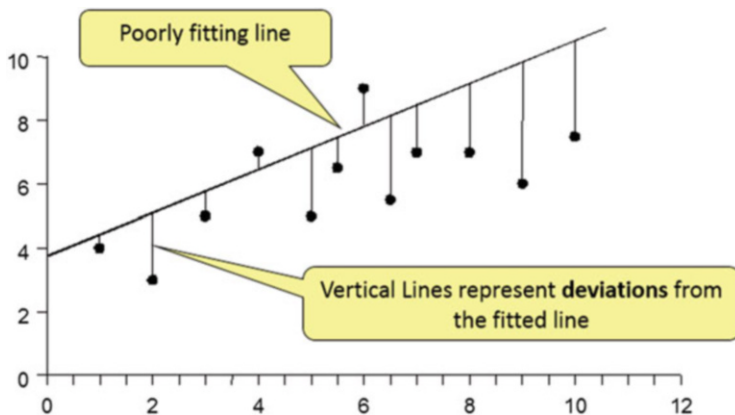


Fig. 8.4 An example of a line that fits the data poorly

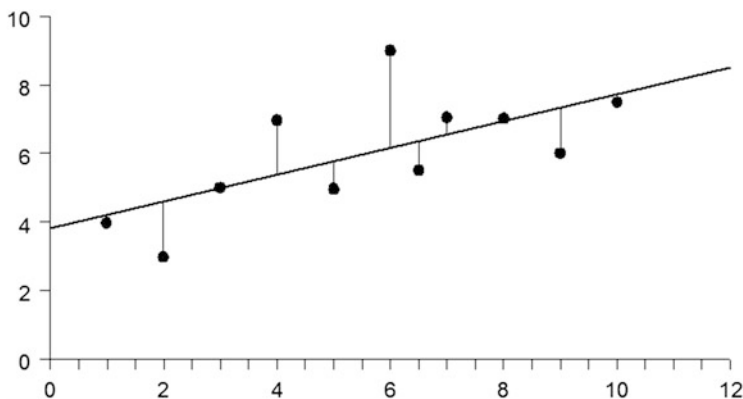


Fig. 8.5 An example of the best fitted line

Step 2: Then you see the box below—click on Simple Scatter and Define (see Fig. 8.7).

Step 3: Put the dependent variable (i.e., money spent partying) on the y-axis and the independent variable (quality of extra-curricular activities) on the x-axis. Click okay (see Fig. 8.8).

Step 4: After the completion of the steps explained in step 3, the scatterplot will show up. However, it would be nice to add a line, which can give us a more robust estimate of the relationship between the two variables. In order to include the line, double-click on the scatterplot in the output window. The chart builder below will appear. Then, just click on the line icon on the second row (the middle icon), and the scatterplot with the line will appear (see Fig. 8.9).

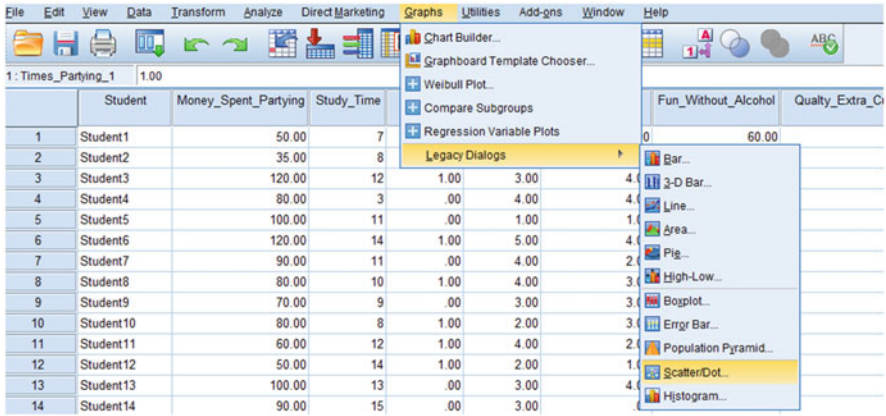


Fig. 8.6 Doing a scatterplot in SPSS (first step)

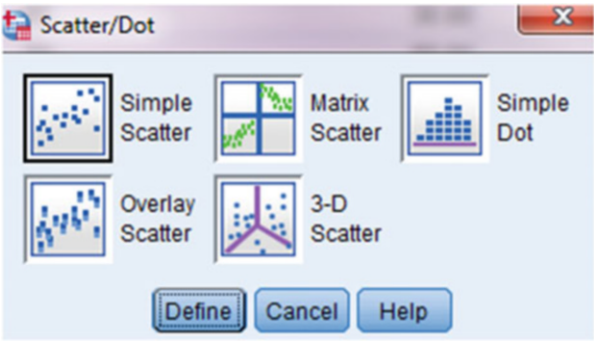


Fig. 8.7 Doing a scatterplot in SPSS (second step)

As expected, Fig. 8.9 displays a negative relationship between the quality of extra-curricular activities and the money students spent partying. The graph displays that students who think that the extra-curricular activities offered by the university are poor do in fact spend more money per week partying. In contrast, students who like the sports and social and political clubs at their university are less likely to spend a lot of money partying. Because we can see that the line is relatively steep, we can already detect that the relationship is relatively strong; a bivariate regression analysis (see below) will give us some information about the strength of this relationship.

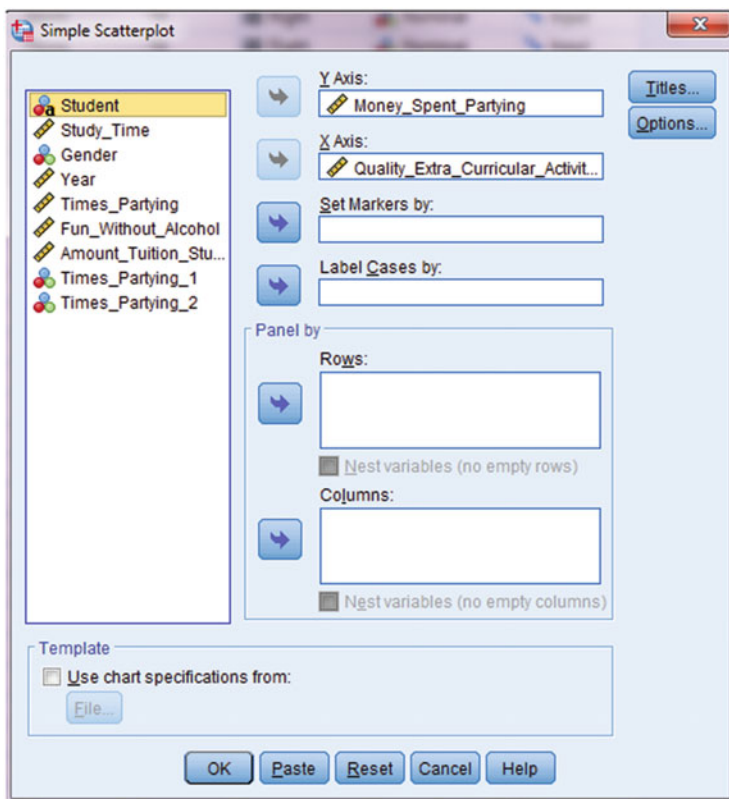


Fig. 8.8 Doing a scatterplot in SPSS (third step)

8.5 Doing Scatterplots in Stata

For our scatterplot, we will use money spent partying as the dependent variable. For the independent variable, we will use quality of extra-curricular activities, hypothesizing that students who enjoy the university-sponsored free time activities will spend less money partying. Rather than going out and party, they will be involved in sports or social and political university clubs and partake in their activities. A scatterplot can help us confirm or disconfirm this hypothesis.

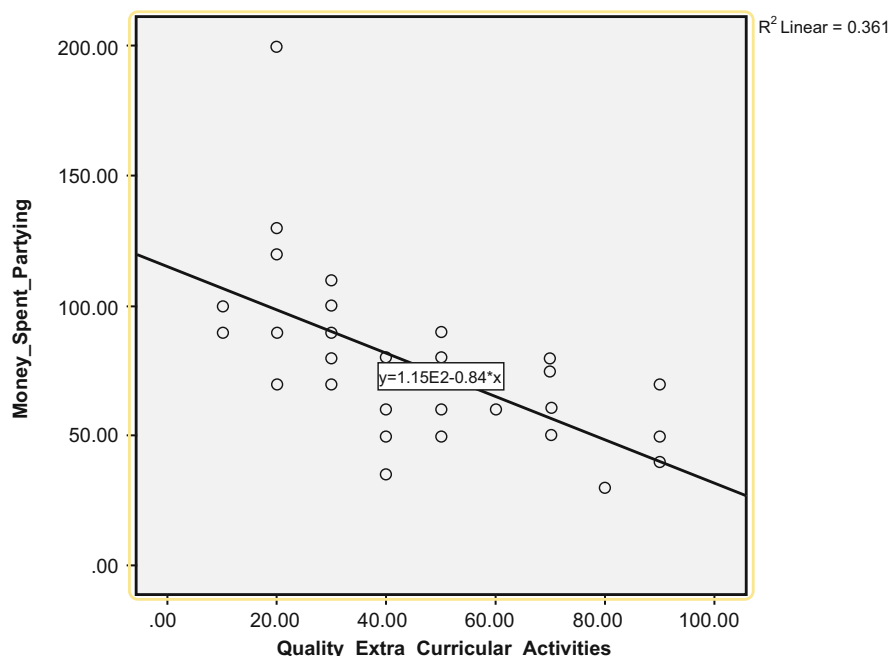


Fig. 8.9 SPSS scatterplot between the quality of extra-curricular activities and money spent partying per week

Step 1: Type in the command editor:

graph twoway (scatter Money_Spent_Partying Quality_Extra_Curricular_Activ)
(lfit Money_Spent_Partying Quality_Extra_Curricular_Activ)¹ (see Fig. 8.10).

Figure 8.11 displays a negative slope, that is, the graph displays that students who think that the extra-curricular activities offered by the university are poor do in fact spend more money partying. In contrast, students who like the sports, social, and political clubs at their university are less likely to spend a lot of money partying. Because we can see that the line is relatively steep, we can already detect that the relationship is relatively strong; a bivariate regression analysis (see below) will give us some information about the strength of this relationship.

```
Command
graph twoway (scatter Money_Spent_Partying Quality_Extra_Curricular_Activities) (lfit Money_Spent_Partying Quality_Extra_Curricular_Activities)
```

Fig. 8.10 Doing a scatterplot in Stata

¹In the dataset, you might want to write Aktiv because writing out activities makes the word too long to fit into the data field in Stata.

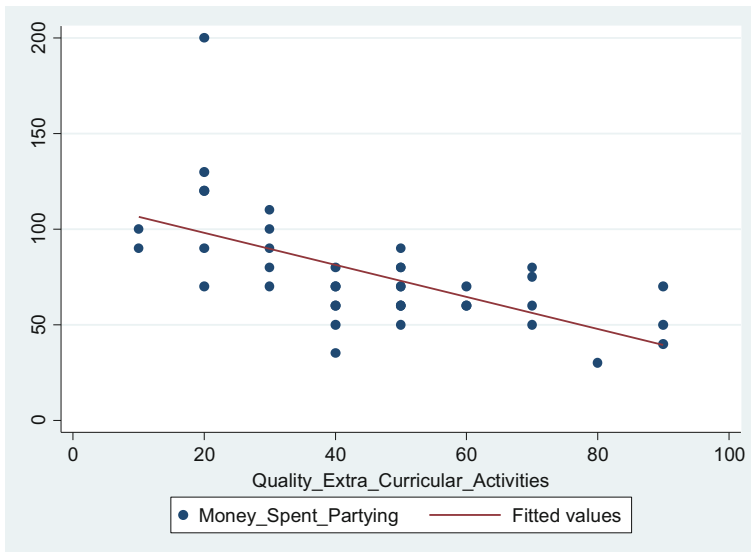


Fig. 8.11 Stata scatterplot between the quality of extra-curricular activities and the money spent partying

Step 2: Stata allows us to include the confidence interval around the fitted line:

`graph twoway (scatter Money_Spent_Partying Quality_Extra_Curricular_Activ)`
`(lfitci Money_Spent_Partying Quality_Extra_Curricular_Activ)` (see Fig. 8.12).

Assuming that our sample was randomly picked from a population of university students, the confidence interval in Fig. 8.13 depicts the range around the fitted line in which the real relationship falls. In other words, the population line displaying the relationship between the independent and dependent variable should be anywhere in the gray shaded area.

```
Command
graph twoway (scatter Money_Spent_Partying Quality_Extra_Curricular_Activities) (lfitci Money_Spent_Partying Quality_Extra_Curricular_Activities)
```

Fig. 8.12 Doing a scatterplot with confidence interval in Stata

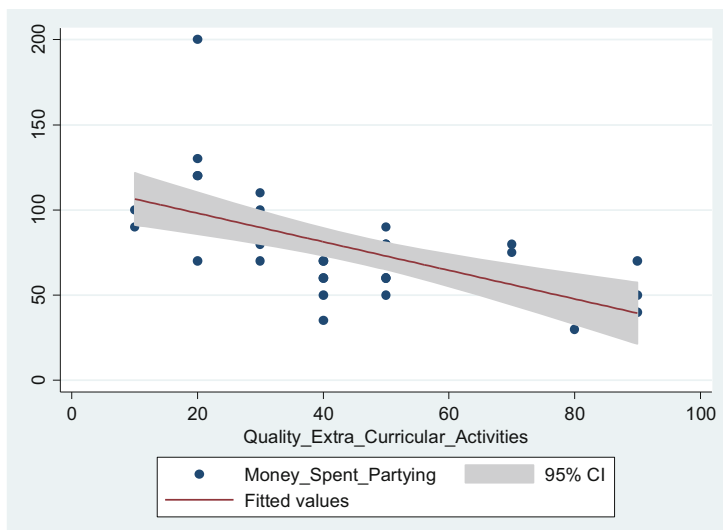


Fig. 8.13 Stata scatterplot between the quality of extra-curricular activities and the money spent partying with the confidence interval

8.6 Correlation Analysis

A correlation analysis is closely linked to the analysis of scatterplots. In order to do a correlation analysis, the relationship between independent and dependent variable must be linear. In other words, we must be able to use a line to express the relationship between the independent variable x and the dependent variable y . To interpret a scatterplot, two things are important: (1) the direction of the line (i.e., a relationship can only exist if the fitted line is either positive or negative) and (2) the closeness of the points toward the line (i.e., the closer the points are clustered around the line, the stronger the correlation is). In fact, in a correlation analysis, it is solely the second point, the closeness of the points to the line that helps us determine the strength of the relationship. To highlight, if we move from graph 1 to graph 4 in Fig. 8.14, we can see that the points get closer to the line for each of the four graphs. Hence, the correlation between the two variables becomes stronger.

In statistical terms, the correlation coefficient, also called Pearson correlation coefficient, is denoted by r . It expresses both the *strength* and *direction* of a relationship between two variables in a single number. If the dots line up to exactly one line, we have a perfect correlation or a correlation coefficient of 1. In contrast, the more all over the place the dots are, the more the correlation coefficient approaches 0 (see Fig. 8.3). In terms of direction, a correlation coefficient with a positive sign depicts a positive correlation, whereas a correlation coefficient with a negative sign depicts a negative correlation.

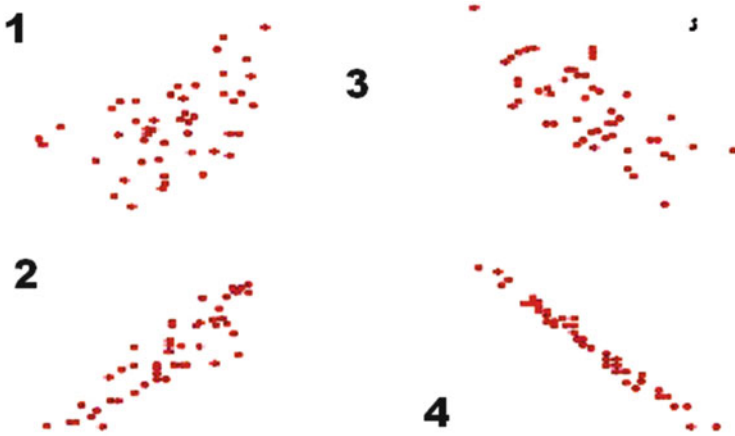


Fig. 8.14 Assessing the strength of relationships in correlation analysis

Formula for r :

$$r = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{S_x} \right) \left(\frac{y_i - \bar{y}}{S_y} \right)$$

Properties of r :

- $-1 < r < 1$.
- $r > 0$ means a positive relationship—the stronger the relationship, the closer r is to 1.
- $r < 0$ means a negative relationship—the stronger the relationship, the closer r is to -1 .
- $r = 0$ means no relationship.

Benchmarks for establishing the level of correlation:

- (-) $0.3 < r < (-) 0.45$ = weak correlation
- (-) $0.45 < r < (-) 0.6$ = medium strong correlation
- $r < (-) 0.6$ = strong correlation

R , like the mean and the standard deviation, is sensitive to outliers. For example, Fig. 8.15 displays nearly identical scatterplots, with the sole difference being that we add an outlier to graph 2. Adding this outlier decreases the correlation coefficient by 0.2 points. Given that the line is drawn so that the sum of the points below the line and the sum of the points above the line equal 0, an outlier pushes the line in one direction (in our case downward), thus increasing the distance of each point toward the line, which, in turn, decreases the strength of the correlation.

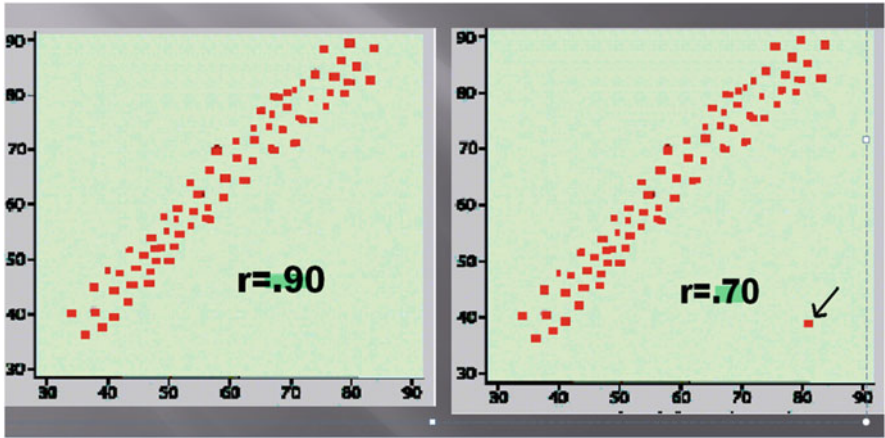


Fig. 8.15 R with and without an outlier

There are two important caveats with correlation analysis. First, since a correlation analysis defines strength at how close the points in a scatterplot are toward a line, it does not provide us with any indication of the strength in impact, in the substantial sense, of a relationship of an independent on a dependent variable. Second, a correlation analysis depicts only whether two variables are related and how closely they follow a positive or a negative direction. It does not give us any indication which variable is the cause and which is the effect.

8.6.1 Doing a Correlation Analysis in SPSS

Step 1: Go to Analyze—Correlate—Bivariate (see Fig. 8.16).

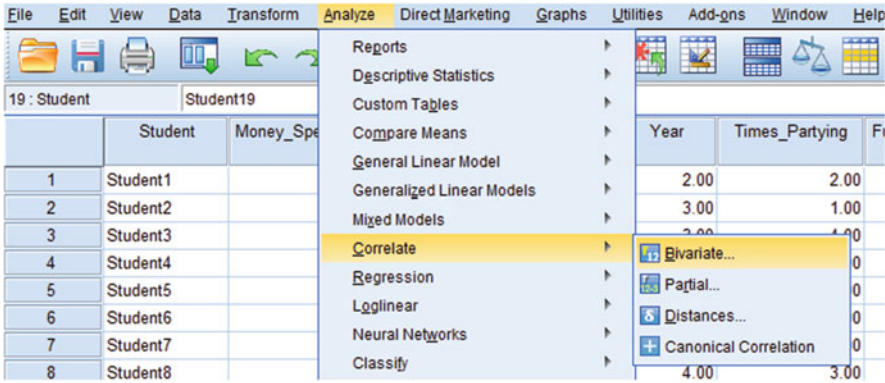


Fig. 8.16 Doing a correlation in SPSS (first step)

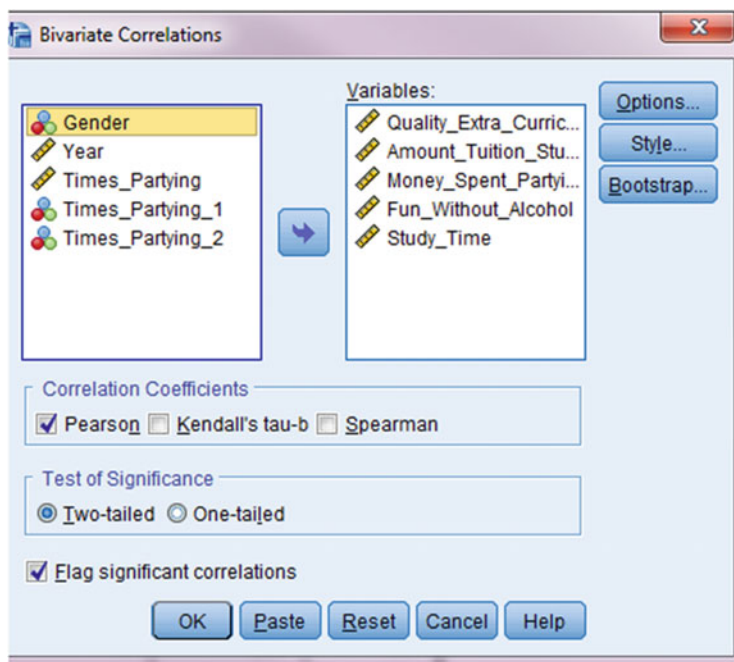


Fig. 8.17 Doing a correlation in SPSS (second step)

Step 2: Choose the continuous variables you want to correlate and click okay (see Fig. 8.17).

It is important to note the correlation analysis only functions with continuous variables. We have five continuous variables in our data and will include them all in the correlation matrix. These five variables are (1) money spent partying, (2) quality of extra-curricular activities, (3) the amount of tuition students pay themselves, (4) whether or not students can have fun without alcohol when they go out, and (5) their study time per week.

8.6.2 Interpreting an SPSS Correlation Output

The correlation output in Table 8.1 displays the bivariate correlations between the five continuous variables in our dataset. To determine whether two variables are correlated, we first look at the significance or alpha level for each correlation. If we find that $\text{sig} > 0.05$, we can conclude that the two variables are not correlated. In this case, we do not interpret the correlation coefficient, which should be small anyways. If we find a significant p -value ($\text{sig}, \leq 0.05$), we can go on and interpret the strength of the correlation using the benchmarks provided in Sect. 8.6. To highlight, let us

Table 8.1 SPSS correlation output

Correlations						
		Quality_Extra_Curricular_Activities	Amount_Tuition_Student_Pays	Money_Spent_Partying	Fun_Without_Alcohol	Study_Time
Quality_Extra_Curricular_Activities	Pearson Correlation	1	-.090	-.600**	.062	-.133
	Sig. (2-tailed)		.580	.000	.706	.414
	N	40	40	40	40	40
Amount_Tuition_Student_Pays	Pearson Correlation	-.090	1	.247	.010	-.195
	Sig. (2-tailed)	.580		.125	.953	.228
	N	40	40	40	40	40
Money_Spent_Partying	Pearson Correlation	-.600**	.247	1	-.206	.171
	Sig. (2-tailed)	.000	.125		.201	.291
	N	40	40	40	40	40
Fun_Without_Alcohol	Pearson Correlation	.062	.010	-.206	1	-.777**
	Sig. (2-tailed)	.706	.953	.201		.000
	N	40	40	40	40	40
Study_Time	Pearson Correlation	-.133	-.195	.171	-.777**	1
	Sig. (2-tailed)	.414	.228	.291	.000	
	N	40	40	40	40	40

** . Correlation is significant at the 0.01 level (2-tailed).

look at the correlation between the amount of tuition students pay and the amount of time students generally dedicate to their studies per week; we find that there is no correlation between these two variables. The significance or alpha level (denoted p in statistical language) is 0.228, which is much higher than the benchmark of 0.05. Hence, we would conclude from there that there is no relationship between these two variables. In other words, we would stop our interpretation here, and we would not interpret the correlation coefficient, because it is not statistically significant. In contrast, if we look at the relationship between the variable money spent partying and the variable quality of extra-curricular activities, we find that the significance level is 0.000. This means that we can be nearly 100% sure that there is a correlation between the two variables. Having established that a correlation exists, we can now look at the sign and the magnitude of the coefficient. We find that the sign is negative, indicating that the more money students spend while going out, the less they participate in extra-curricular activities in their university. Knowing that there is a negative correlation, we can now interpret the magnitude of the correlation coefficient, which is -0.600 , indicating that we have medium strong to strong negative correlation. In addition to the correlation between the quality of extra-curricular activities and money spent partying, there is one more statistically significant and substantively strong correlation, namely, between students' study time per week and whether or not they can have fun partying without alcohol. This correlation is significant at the 0.000 level and substantively strong (i.e., $r = -0.777$). The negative sign further highlights that it is negative indicating that high values for one variable trigger low values for the other variables. In this case, this implies that students that study a lot also think that they can have a lot of fun without alcohol when they party.

8.6.3 Doing a Correlation Analysis in Stata

It is important to note that correlation analyses only function with continuous variables. We have five continuous variables in our data and include them in the Stata correlation matrix: (1) money spent partying, (2) quality of extra-curricular activities, (3) the amount of tuition students pay themselves, (4) whether or not students can have fun without alcohol when they go out, and (5) their study time per week.

Step 1: Write in the command editor
pwcorr Money_Spent_Partying Study_Time Fun_Without_Alcohol
Quality_Extra_Curricular_Activ Amount_Tuition_Student_Pays, sig (see Fig. 8.18)

The correlation output in Table 8.2 displays the bivariate correlations between the five continuous variables in our dataset. For each bivariate correlation, Stata provides the Pearson correlation coefficient and the significance level (*p*). For example, 0.1711 is the correlation coefficient between study time and money spent partying. The second number (0.29) is the corresponding significance level. To determine whether two variables are correlated, we first look at the significance or alpha level for each correlation. If we find that sig >0.05, we can conclude that the two variables are not correlated. In this case, we do not interpret the correlation coefficient, which should be small anyways. In cases where we find a significant *p*-value (sig <0.05), we can go on and interpret the strength of the correlation using the benchmarks provided in Sect. 8.6. To highlight, let us look at the correlation between

```
Command  
pwcorr Money_Spent_Partying Study_Time Fun_Without_Alcohol Quality_Extra_Curricular_Activ Amount_Tuition_Student_Pays, sig
```

Fig. 8.18 Doing a correlation in Stata

Table 8.2 Stata correlation output

	Money_~g	Study_~e	Fun_Wi~l	Qualit~v	Amount~s
Money_Spen~g	1.0000				
Study_Time	0.1711 0.2912	1.0000			
Fun_Withou~l	-0.2064 0.2012	-0.7774 0.0000	1.0000		
Quality_Ex~v	-0.6005 0.0000	-0.1329 0.4136	0.0616 0.7057	1.0000	
Amount_Tui~s	0.2468 0.1247	-0.1950 0.2279	0.0096 0.9533	-0.0901 0.5802	1.0000

the amount of tuition students pay and the amount of time students generally dedicate to their studies per week; we find that there is no correlation between these two variables. The significance or alpha level (denoted p in statistical language) is 0.228, which is much higher than the benchmark of 0.05. Hence, we would conclude from there that there is no correlation between these two variables. In other words, we would stop our interpretation here, and we would not interpret the correlation coefficient, because it is not statistically different from zero. In contrast, if we look at the relationship between the variable money spent partying and the variable quality of extra-curricular activities, we find that the significance level is 0.000. This means that we can be nearly 100% sure that there is a correlation between the two variables. Having established that a correlation exists, we can now look at the sign and the magnitude of the coefficient. We find that the sign is negative, indicating that the more students spend money while going out, the less they participate in extra-curricular activities in their university. Knowing that there is a negative correlation, we can now interpret the magnitude of the correlation coefficient, which is -0.601 , indicating that we have a medium strong to strong negative correlation. In addition to the correlation between the quality of extra-curricular activities and money spent partying, there is one more statistically significant and substantively strong correlation, namely, between students' study time per week and whether or not they can have fun partying without alcohol. This correlation is significant at the 0.000 level and substantively strong (i.e., $r = -0.762$). The negative sign further highlights that it is a negative relationship, indicating that high values for one variable trigger low values for the other variables. In this case, this implies that students that study a lot also think that they can have a lot of fun without alcohol when they party.

8.7 Bivariate Regression Analysis

In correlation analyses, we look at the direction of the line (positive or negative) and at how closely the points of the scatterplot follow that line. This allows us to detect the degree to which two variables covary, but it does not allow us to determine how strongly an independent variable influences a dependent variable. In regression analysis, we are interested in the magnitude of the influence of independent on dependent variable, as measured by the steepness of the slope. To determine the influence of an independent variable on a dependent variable, two things are important: (1) the steeper the slope, the more strongly the independent variable impacts the dependent variable. (2) The closer the points are to the line, the more certain we can be that this relationship actually exists.

8.7.1 Gauging the Steepness of a Regression Line

To explain the notion that a steeper slope indicates a stronger relationship, let us compare the two graphs in Fig. 8.19. Both graphs depict a perfect relationship, meaning that all the points are on a straight line. The correlation for both would be 1. However, we can see that the first line is much steeper than the second line.

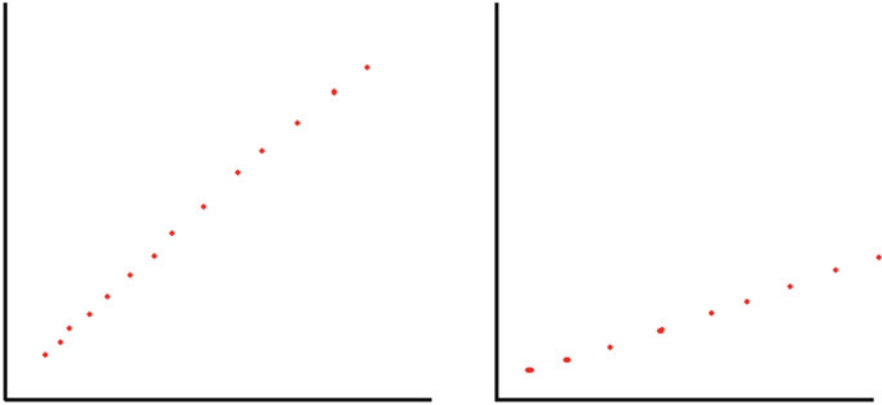


Fig. 8.19 Two regression lines featuring a strong and weak relationship, respectively

In other words, the y value grows stronger with higher x values. In contrast, the second line only moves slightly upward. Consequently, the regression coefficient is higher in the first compared to the second graph.

To determine the relationship between x and y , we use a point slope:

$$y = a + bx$$

y is the dependent variable.

x is the independent variable.

b is the slope of the line.

a is the intercept (y when $x = 0$).

The formula for the regression line:

$$b = r \frac{S_y}{S_x} \quad a = \bar{y} - b\bar{x}$$

When we draw the regression line, we could in theory draw an infinite number of lines. The line that explains the data best is the line that has the smallest sum of squared errors. In statistical terms, this line is called the least square line (OLS line).

Figure 8.20 displays the least square line between the independent variable average GDP per capita per country and the dependent variable energy usage in kilograms of oil per person. The equation denoting this relationship is as follows: energy usage = 318 + 0.25 average country GDP per capita.

This means 318 kg is the amount of energy that an average citizen uses when GDP is at 0. The equation further predicts that for every 1 unit (dollar) increase on the y -axis, y values increase by 0.25 kg. So in this example, this means that for each extra dollar the average citizen in a country becomes richer, she uses 0.25 kg more of energy (oil) per year. To render the interpretation more tangible, the equation would

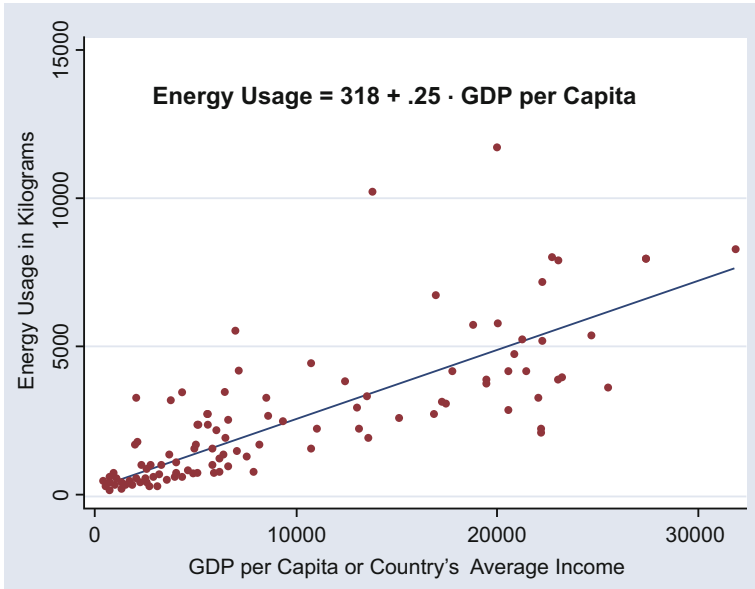


Fig. 8.20 The equation between per capita GDP and energy usage

predict that a resident in a country, where the average citizens earn 10,000 dollars, is predicted to consume 2818 kg of energy ($318 + 0.25 \times 10000 = 2818$).

8.7.2 Gauging the Error Term

The metric we use to determine the magnitude of a relationship between an independent and a dependent variable is the steepness of the slope. As a rule, we can say that the steeper the slope, the more certain we can be that a relationship exists. However, the steepness of the slope is not the only criterion we use to determine whether or not an independent variable relates to a dependent variable. Rather, we also have to look how close the data points are to the line. The closer the points are to the line, the less error there is in the data. Figure 8.21 displays two identical lines measuring the relationship between age in months and height in centimeters for babies and toddlers. What we can see is that the relationship is equally strong for the two lines. However, the data fits the first line much better than the second line. There is more “noise” in the data in the second line, thus rendering the estimation of each of the points or observations less exact in the second line compared to the first line.

Figure 8.22 graphically explains what we mean by error term or residual. The error term or residual in a regression analysis measures the distance from each data point to the regression line. The farther any single observation or data point is away from the line, the less this observation fits the general relationship. The larger the average distance is from the line, the less well does the average data point fits the linear prediction. In other words, the greater the distance between the average data

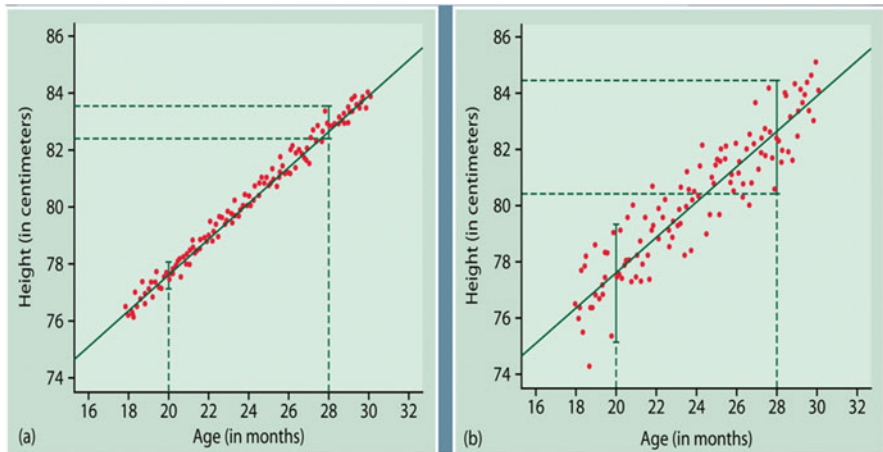


Fig. 8.21 Better and worse fitting data

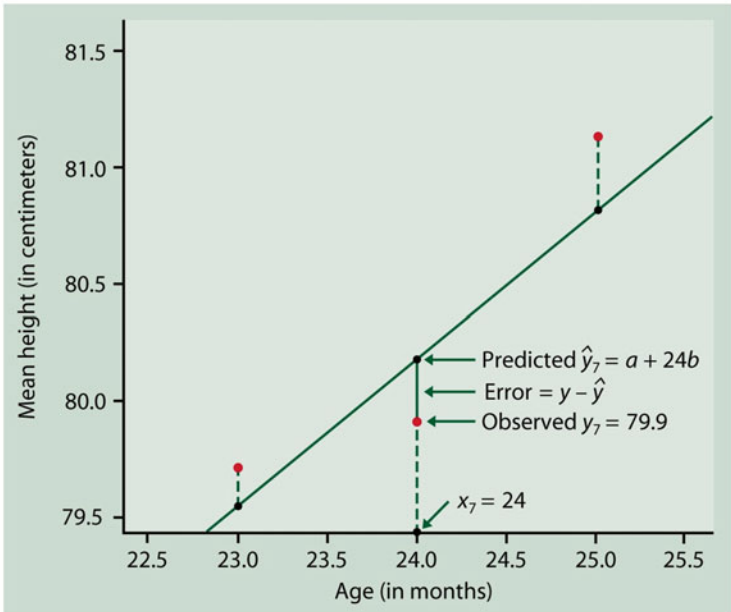


Fig. 8.22 The error term in a regression analysis

point and the line, the less we can be assured that the relationship portrayed by the regression line actually exists.

8.8 Doing a Bivariate Regression Analysis in SPSS

To conduct the bivariate regression analysis, we take the same variables we used for the scatterplots, that is, our dependent variable money spent partying and the independent variable quality of extra-curricular activities

- Step 1:** Go to—Analyze—Regression—Linear (see Fig. 8.23).
- Step 2:** Choose the dependent variable—choose the independent variable—click okay (Fig. 8.24).

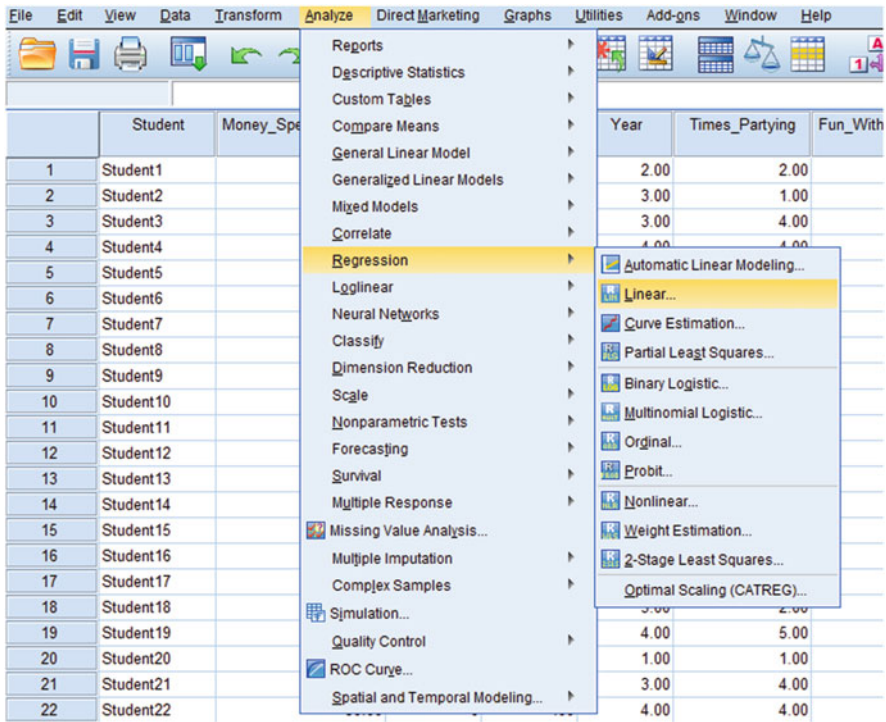


Fig. 8.23 Doing a regression analysis in SPSS (first step)

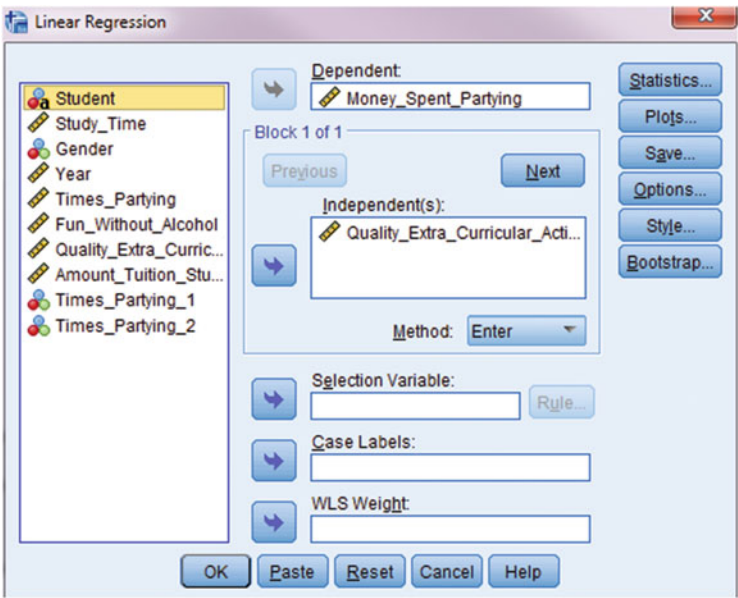


Fig. 8.24 Doing a regression analysis in SPSS (second step)

8.9 Interpreting an SPSS (Bivariate) Regression Output

The SPSS regression output consists of three tables: (1) a model summary table, (2) an ANOVA output, and (3) a coefficients' table. We interpret each table separately

8.9.1 The Model Summary Table

The model summary table (see Table 8.3) indicates how well the model fits the data. It consists of four parameters: (1) the Pearson correlation coefficient (r), (2) the R -squared value, (3) the adjusted R -squared value, and (4) the standard error of the estimate.

Table 8.3 SPSS model summary table

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.600 ^a	.361	.344	24.42734
a. Predictors: (Constant), Quality_Extra_Curricular_Activities				

R is the correlation coefficient between the real data points and the values predicted by the model. For the example at hand, the correlation coefficient is high indicating that real values and predicated values correlate at the level of 0.600.

R -squared is the most important parameter in the model summary statistic's table and is a measure of model fit (i.e., it is the squared correlation between the model's predicted values and the real values). It explains how much of the variance in the dependent variable the independent variable(s) in the model explain. In theory, the R -squared values can range from 0 to 1. An R -squared value of 0 means that the independent variable(s) do not explain any of the variance of the dependent variable, and a value of 1 signifies that the independent variable(s) explain all the variance in the dependent variable. In our example, the R -squared value of 0.361 implies that the independent variable—the quality of extra-curricular activities—explains 36.1% of the variance in the dependent variable, the money students spent partying per week.

The adjusted R -squared is a second statistic of model fit. It helps us compare different models. In real research, it might be helpful to compare models with a different number of independent variables to determine which of the alternative models is superior in a statistical sense. To highlight, the R -squared will always increase or remain constant if I add variables. Yet, a new variable might not add anything substantial to the model. Rather, some of the increase in R -squared could be simply due to coincidental variation in a specific sample. Therefore, the adjusted R -squared will be smaller than the R -squared since it controls for some of the idiosyncratic variance in the original estimate. As such, the adjusted R -squared is a measure of model fit adjusted for the number of independent variables in the model. It helps us compare different models; the best fitting model is always the model with the highest adjusted R -squared (not the model with the highest R -squared).² In our sample, the adjusted R -squared is 0.344 (We do not interpret this estimator in bivariate regression analysis).

The standard error of the estimate is the standard deviation of the error term and the square root of the mean square residual (or error). (Normally, we do not interpret this estimator when we conduct regression models.) In our sample, the standard error of the estimate is 24.43.

8.9.2 The Regression ANOVA Table

The f -test in a regression model works like an f -test or ANOVA analysis (see Table 8.4). It is an indication of the significance of the model (does the regression equation fit the observed data adequately?). If the f -ratio is significant, the regression equation has predictive power, which means that we have at least one statistically significant variable in the model. In contrast, if the f -test is not significant, then none of the variables in the model is statistically significant, and the model has no predictive power. In our sample, the f -value is 21.43, and the corresponding significance level is 0.000. Hence, we can already conclude that our independent

²Please note that we can only compare adjusted R -squared values of different models, if these models have the same number of observations.

Table 8.4 SPSS ANOVA table of the regression output

ANOVA ^a						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	12785.588	1	12785.588	21.427	.000 ^b
	Residual	22674.412	38	596.695		
	Total	35460.000	39			

a. Dependent Variable: Money_Spent_Partying

b. Predictors: (Constant), Quality_Extra_Curricular_Activities

variable—the quality of extra-curricular activities—influences our dependent variable, the money students spend per week partying.

8.9.3 The Regression Coefficient Table

The coefficients’ table (see Table 8.5) is the most important part of a regression output. It tells us how the independent variable relates to the dependent variable. A coefficients’ table has the following statistics:

Unstandardized Regression Weights The first column portrays the unstandardized regression weights, which give us an indication of how the independent variable (s) relates to the dependent variable. In our example, the unstandardized regression weight or coefficient depicting the influence of extra-curricular activities on students’ party spending is –0.839. The constant is 114.87. In other words, the –0.839 is the slope coefficient; it indicates the change in the DV associated with a 1-unit change in the IV. In our example, this implies that for each point on the 0–100 scale, students like the extra-curricular activities more, their party spending decreases by 0.839 dollars per week. The 114.87 is the predicted spending pattern if the independent

Table 8.5 The SPSS regression coefficients’.

Coefficients ^a					
Model		Unstandardized Coefficients		Standardized Coefficients	Sig.
		B	Std. Error	Beta	
1	(Constant)	114.869	9.145		.000
	Quality_Extra_Curricular_Activities	-.839	.181	-.600	.000
a. Dependent Variable: Money_Spent_Partying					

variable has the value 0 (i.e., if students think that the extra-curricular activities at their university are very bad).

We can also express this relationship in a regression equation: $y = 114.87 - 0.839x$.

The Standard Error This statistic gives us an indication of how much variation there is around the predicted coefficient. As a rule, we can say that the smaller the standard error in relation to the regression weight is, the more certainty we can have in the interpretation of our relationship. In our case, the standard error behind the relationship between extra-curricular activities and money spent partying is 0.181.

Standardized Regression Coefficients (Beta) In multiple regression models, it is impossible to compare the magnitude of the coefficients of the independent variables. To highlight, the variable gender is a dummy variable coded 0 for guys and 1 for girls. In contrast, the variable fun without alcohol is a variable coded on a 100-point scale (i.e., 0–100). Unstandardized weights are standardized coefficients that allow us to compare the effect size of several independent variables. To do so, SPSS converts the unstandardized coefficients to *z*-scores (i.e., weights with mean of zero and standard deviation of 1.0). The size of the standardized regression weights gives us some indication of the importance of the variable in the regression equation. In a multiple regression analysis, it allows us to determine the relative strength of each independent variable in comparison to other variables on the dependent variable. The standardized beta is -0.600 in our bivariate model.

***T*-value** The *t*-test compares the unstandardized regression against a predicted value of zero (i.e., no contribution to regression equation). It does so by dividing the unstandardized regression coefficient (i.e., our measure of effect size) by the standard error (i.e., our measure of variability in the data). As a rule, we can say that the higher the *T*-value, the higher the chance that we have a statistically significant relationship (see significance value). In our example, the *T*-value is -4.629 ($-0.839/0.181$).

Significance Value (Alpha Level) The significance level (*sig*) indicates the probability that a specific independent variable impacts the dependent variable. In our example, the significance level is 0.000, which implies that we can be nearly 100% sure that the quality of extra-curricular activities influences students' spending patterns while partying.

8.10 Doing a (Bivariate) Regression Analysis in Stata

To conduct the regression analysis, we will use the same variables that we used for the scatterplots, that is, our dependent variable is money spent partying and our independent variable is the quality of extra-curricular activities (see Table 7.11).


```
Command
reg Money_Spent_Partying Quality_Extra_Curricular_Activ
```

Fig. 8.25 Doing a bivariate regression analysis in Stata

Step 1: Write into the command editor: `reg Money_Spent_Partying Quality_Extra_Curricular_Activ` (see Fig. 8.25).

8.10.1 Interpreting a Stata (Bivariate) Regression Output

On the following pages, we explain how to interpret a Stata bivariate regression output (see Table 8.6):

Number of obs indicates how many observations the model is based upon

The *f*-test provides the *f*-test value for the regression analysis. It works like an *f*-test or ANOVA analysis and gives an indication of the significance of the model (i.e., it measures whether the regression equation fits the observed data adequately). If the *f*-ratio is significant, the regression equation has predictive power, which means that we have at least one statistically significant variable in the model. In contrast, if the *f*-test is not significant, then none of the variables in the model is statistically significant, and the model has no predictive power. In our sample, the *f*-value is 21.43 and the corresponding significance level is 0.000 (Prob > *F*). Hence, we can already conclude that our independent variable—the quality of extra-curricular activities—influences our dependent variable, the money students spend per week partying. (The Table you see to the left of the summary statistics (i.e., Number of obs, *F*, Prob > *F*, *R*-squared, Adj *R*-squared, Root MSE) is the summary statistics of the *f*-test (see section on *t*-test or ANOVA for more details).)

Table 8.6 The Stata bivariate regression

Source	SS	df	MS	Number of obs	=	40
Model	12785.588	1	12785.588	F(1, 38)	=	21.43
Residual	22674.412	38	596.695054	Prob > F	=	0.0000
				R-squared	=	0.3606
				Adj R-squared	=	0.3437
Total	35460	39	909.230769	Root MSE	=	24.427

Money_Spent_Partying	Coef.	Std. Err.	t	P> t	[95% Conf. Interval	
Quality_Extra_Curricular_Activ	-.8386742	.1811795	-4.63	0.000	-1.205453	-.471895
_cons	114.8693	9.144632	12.56	0.000	96.357	133.381

***R*-squared** is an important parameter when we interpret a regression output. It is a measure of model fit (i.e., it is the squared correlation between the model's predicted values and the real values). It explains how much of the variation in the dependent variable is explained by the independent variables in the model. In theory, the *R*-squared values can range from 0 to 1. An *R*-squared value of 0 means that the independent variable(s) do not explain anything in the variance of the dependent variable. An *R*-squared value of 1 signifies that the independent variable(s) explain all the variance in the dependent variable. In our example, the *R*-squared value of 0.361 implies that the independent variable—the quality of extra-curricular activities—explains 36.1% of the variance in the dependent variable, the money students spent partying per week.

The adjusted *R*-squared is a second statistic of model fit that helps us compare different models. In real research, it might be helpful to compare models with a different number of independent variables to determine which of the alternative models is superior in a statistical sense. To highlight, the *R*-squared will always increase or remain constant if I add variables even though a new variable might not add anything substantial to the model. Rather, some of the increase in the *R*-squared could be simply due to coincidental variation in a specific sample. Therefore, the adjusted *R*-squared will be smaller than the *R*-squared since it controls for some of the idiosyncratic variance in the original estimate. Therefore, the adjusted *R* is a measure of model fit adjusted for the number of independent variables in the model. It helps us to compare different models; the best fitting model is always the model with the highest adjusted *R*-squared (not the model with the highest *R*-squared).³ In our sample, the adjusted *R*-squared is 0.344. (We do not interpret this estimator in bivariate regression analysis.)

The root MSE (root mean squared error) is the standard deviation of the error term and the square root of the mean square residual (or error). (Normally, we do not interpret this estimator when we conduct regression models.) In our sample, the standard error of the estimate is 24.43.

Coef. (coefficient) The first column portrays the unstandardized regression weights, which give us an indication of how the independent variable(s) relate to the dependent variable. In our example, the unstandardized regression weight or coefficient depicting the influence of extra-curricular activities on students' party spending is -0.839 . The constant is 114.87. In other words, the -0.839 is the slope coefficient; it indicates the change in the dependent variable associated with a 1-unit change in the independent variable. In our example, this implies that for each additional point on the 0–100 scale that students like the extra-curricular activities, their party spending decreases by 0.839 dollars per week. The constant (cons) coefficient of 114.87 is the predicted spending pattern if the independent variable

³Ibid.

has the value 0 (i.e., this implies that if students do not like the extra-curricular activities at their school at all, they are predicted to spend 115 dollars per week partying).

We can also express this relationship in a regression equation: $y = 114.87 - 0.839x$

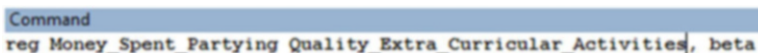
The standard error gives us an indication of how much variation there is around the predicted coefficient. As a rule, we can say that the smaller the standard error relative to the regression weight, the more certainty we can have in the interpretation of our relationship between independent and dependent variable. In our case, the standard error behind the relationship between extra-curricular activities and money spent partying is 0.181.

Standardized regression coefficients (beta) In multiple regression models, it is impossible to compare the magnitude of the coefficients of the independent variables. To highlight, the variable gender is a dummy variable coded 0 for guys and 1 for girls. In contrast, the variable fun without alcohol is a variable coded on a 100-point scale (i.e., 0–100). Standardized weights are standardized coefficients that allow us to compare the effect size of several independent variables. To do so, Stata converts the unstandardized coefficients to z-scores (i.e., weights with mean of zero and standard deviation of 1). The size of the standardized regression weights gives us some indication of the importance of the variable in the regression equation. In a multiple regression analysis, it allows us to determine the relative strength of each independent variable in comparison to other variables on the dependent variable. The standardized beta is not one of the default options in the Stata package. However, it can be easily calculated. To do so, add the option `beta` at the end of the regression analysis. Put in the command field:

`reg Money_Spent_Partying Quality_Extra_Curricular_Activities, beta` (see Fig. 8.26)

In our bivariate example, the standardized beta coefficient is -0.600 (see Table 8.7).

T-value The *t*-test compares the unstandardized regression weight against a predicted value of zero (i.e., no contribution to regression equation). It does so by dividing the unstandardized regression coefficient (i.e., our measure of effect size) by the standard error (i.e., our measure of variability in the data). As a rule, we can say that the higher the *t*-value, the higher the chance that we have a statistically significant relationship (see significance value). In our example, the *t*-value is -4.629 ($-0.839/0.181$).



```
Command
reg Money_Spent_Partying Quality_Extra_Curricular_Activities, beta
```

Fig. 8.26 Doing a bivariate regression analysis with standardized beta coefficients

Table 8.7 The Stata bivariate regression table with standardized coefficients

Source	SS	df	MS	Number of obs	=	40
				F(1, 38)	=	21.43
Model	12785.588	1	12785.588	Prob > F	=	0.0000
Residual	22674.412	38	596.695054	R-squared	=	0.3606
				Adj R-squared	=	0.3437
Total	35460	39	909.230769	Root MSE	=	24.427

Money_Spent_Partying	Coef.	Std. Err.	t	P> t	Beta
Quality_Extra_Curricular_Activ	-.8386742	.1811795	-4.63	0.000	-.6004695
_cons	114.8693	9.144632	12.56	0.000	.

$P > t$ is the significance value or alpha level The significance level determines the probability with which we can determine that a specific independent variable impacts the dependent variable. In our example, the significance level is 0.000, which implies that we can be nearly 100% sure that the quality of extra-curricular activities influences students’ spending patterns while partying.

(95% conf. interval) Assuming that our sample was randomly picked from a population of university students, the confidence interval depicts the interval in which the real relationship between the perceived quality of extra-curricular activities and students spending patterns lies. Our coefficient for the quality of extra-curricular activities has a confidence interval that ranges from -1.21 to -0.472 . This means that the relationship in the population could be anywhere in this range. Similarly, the confidence interval around the constant implies that somebody who rates the quality of extra-curricular activities at 0 is expected to spend between 96.36 dollars and 133.38 for partying in a week.

8.10.2 Reporting and Interpreting the Results of a Bivariate Regression Model

When we report the results of bivariate regression model, we report the coefficient, standard error, and significance level, as well as the R -squared and the number of observations in the model. In an article or report, we would report the results as follows: Table 8.8 reports the results of a bivariate regression model measuring the influence of the quality of extra-curricular activities on students’ weekly spending patterns when they party based on a survey conducted with 40 undergraduate students at a Canadian university. The model portrays a negative and statistically significant relationship between the two variables. In substantive terms, the model predicts that for every point students’ evaluation of the extra-curricular activities at their university increases, they spent 84 cents less per week partying. This influence is substantial. For example, somebody, who thinks that the extra-curricular activities at her university are poor (and rates them at 20), is predicted to spend approximately 98 dollars for party activities. In contrast, somebody, who likes the extra-curricular

Table 8.8 Reporting a bivariate regression outcome

	Coefficient	Std. error	Sig
Quality of extra-curricular activities	−0.839	0.181	0.000
Constant	114.87	9.14	0.000
R-squared	0.34		
N	40		

activities (and rates them at 80), is only expected to 48 dollars for her weekly partying. The *R*-squared of the model further highlights that the independent variable, the quality of extra-curricular activities, explains 34% of the variance in the dependent variable partying spending per week.

Further Reading

Gravetter, F. J., & Forzano, L. A. B. (2018). *Research methods for the behavioral sciences*. Boston: Cengage Learning (chapter 12). Nice introduction into correlational research; covers the data and methods for correlational analysis, applications of the correlational strategy, and strength and weakness of the correlational research strategy.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (Vol. 821). San Francisco: John Wiley & Sons (chapters 1 and 2). Chapters 1 and 2 provide a hands on and practical introduction into linear regression modelling, first in the bivariate realm and then in the multivariate realm.

Ott, R. L., & Longnecker, M. T. (2015). *An introduction to statistical methods and data analysis*. Toronto, ON: Nelson Education (chapter 11). Provides a good introduction into correlation and regression analysis clearly highlighting the differences between the two techniques.

Roberts, L. W., Wilkinson, L., Peter, T., & Edgerton, J. (2015). *Understanding social statistics*. Don Mills: Oxford University Press (chapters 10–14). These chapters offer students the basic tools to examine the form and strength of bivariate relationships.